# *What is intelligence?*

## *Lessons from AI about Evolution, Computation, and Consciousness*

# Preface

Figure 1. Placeholder

AI is probably the biggest story of our lifetimes. Its rapid development, starting in the early 2020s, has generated a mixed outpouring of excitement, anxiety, and denial.

Everyone wants to weigh in. Even authors writing about history, language, biology, sociology, economics, the arts, and psychology feel the need to add a chapter about AI to the end of their books. And why not? AI will surely affect all of these far-flung domains, and many more—though exactly *how*, nobody can say with any confidence.

This book is a bit different. It’s less about the future of AI than it is about what we have learned in the present. At least as of 2024, few mainstream authors claim that AI is “real” intelligence.[[1]](#footnote-22) I do. Gemini, Claude, and ChatGPT aren’t powered by the same machinery as the human brain, of course, but intelligence is “multiply realizable,” meaning that, just as a computer algorithm can run on any kind of computer, intelligence can “run” on many physical substrates. In fact, although our brains aren’t much like the kinds of digital computers we have today, I think the substrate for intelligence *is* computation, which implies that a sufficiently powerful general-purpose computer can, by definition, implement intelligence. All it takes is the right code.

I will argue that, thanks to recent conceptual breakthroughs in AI development, we now know, at least at a high level, what that code *does*. We understand the essence of an incredibly powerful trick, although we’re still in the early days of making it work. Our implementations are neither complete, nor reliable, nor efficient—a bit like where we were with general computing when the ENIAC first powered up, in 1945, or where we were with aviation when the Wright brothers made their first powered flight, in 1903.

It’s an old cliché in AI that airplanes and birds can both fly, but do so by different means.[[2]](#footnote-23) This truism has at times been used to motivate misguided approaches to AI. Still, the point stands. Birdflight is a biological marvel, which in some respects remains poorly understood, even today; and to engineer small drones, we still have much to learn from insects.[[3]](#footnote-24) However, we figured out the basic physics of flight—not *how* animals fly, but how it’s *possible* for them to—in the eighteenth century, with the discovery of Bernoulli’s Equation. Working airplanes took another century and a half to evolve.

Similarly, while there is still a great deal about the brain that we don’t understand, the fundamental principle behind intelligence—*prediction*—was first theorized by German polymath Hermann von Helmholtz in the 1860s.[[4]](#footnote-25) Many computational neuroscientists and AI researchers have elaborated on Helmholtz’s insight since then, and built limited models implementing aspects of the idea,[[5]](#footnote-26) but only recently has it become plausible to imagine that prediction really is the whole story. I will argue that, understood in full, the prediction principle may explain not only intelligence, but life itself.[[6]](#footnote-27)

Furthermore, it explains why recent large AI models really *are* intelligent; it’s not “anthropomorphic” to say so. This doesn’t mean that AI models are necessarily human-like, or lesser than us, or greater than us. In fact, understanding the curiously self-similar and self-referential nature of prediction will let us see that intelligence isn’t really a “thing.” It can’t exist in isolation, either in the three pounds of neural tissue in your head, or in the racks of computer chips running large models in datacenters. Intelligence is defined by networks, and by networks of networks. We can only understand what intelligence really is by changing how we think about it—by adopting a perspective that centers dynamic, symbiotic relationships rather than isolated minds.

This book takes on that project. Doing so requires weaving together insights from many different disciplines. As we go along, I’ll introduce concepts in probability, machine learning, physics, chemistry, biology, computer science, and other fields. When they’ve been most relevant to shaping our (sometimes mistaken) beliefs, I’ll also briefly review the intellectual histories of key ideas, from seventeenth century “mechanical philosophy” to debates about the origin of life, and from cybernetics to neuroscience.

You, dear reader, may be an expert in one or more of these fields, or in none. Few people are expert in all of them (I’m certainly not), so no specialized prior knowledge is assumed. On the other hand, even if you’re an AI researcher or neuroscientist with little patience for pop science, I hope you will find new and surprising ideas in this book. A general grasp of mathematical concepts like functions and variables will be helpful (with bonus points for knowing about vectors and matrices), but there will be no equations. (Well … *almost* none.) A general understanding of how computer programming works will be useful in a few places, but isn’t required. If you find understanding intelligence interesting enough to still be reading, rest assured: you are my audience.

The **Introduction** describes the recent emergence of “real” AI, sketching the hypothesis that prediction *is* intelligence, and pointing out some of the big questions that raises. The rest of the book consists of about a hundred bite-sized chapters that attempt to answer those questions, organized into the following ten parts:

1. **Origins** begins with “abiogenesis,” the emergence of life on Earth some four billion years ago. Darwin struggled to understand how life could have arisen. “Artificial life” experiments let us make more progress today, and suggest that life may be inevitable in a world capable of supporting computation (like ours). Understanding the deep role computation plays in our universe, and the way it links the seemingly disparate fields of physics, chemistry, and biology, both sheds new light on evolutionary theory and sets the stage for understanding the computational nature of intelligence.
2. **Survival** considers the predictive intelligence hypothesis from the perspective of a tiny, simple organism capable of rudimentary computation: a bacterium. Being alive in the world is inherently social (even for bacteria), but this part imagines life as a one-player game, and considers what it takes to keep playing—which, since life is an infinite game, is the only prize there is.  
   **Interlude: The prehistory of computing** offers a historical perspective on the origins and founding assumptions of “traditional” computer science, from its roots in Enlightenment thinking and the Industrial Revolution to the development of the first electronic computers at the close of the Second World War.
3. **Cybernetics** traces an alternative history of computing, which has recently re-emerged as modern AI, back to *its* foundations in the mid-twentieth century, describing how Norbert Wiener’s biologically-inspired theory of feedback lets us understand the connections between the predictive intelligence hypothesis and the development of artificial neural nets.
4. **Learning** connects more recent advances in machine learning with our growing understanding of computational neuroscience and the evolution of nervous systems.
5. **Other minds** explores how, as multiplayer life becomes complex, the main job of a mind becomes modeling other minds. This leads to a mutual modeling arms race, resulting in the kind of “intelligence explosion” that has produced humanity.
6. **Many worlds** describes the way both long-term planning and short-term unpredictability (otherwise known as “free will”) emerge as consequences of social modeling, especially when applied to ourselves.
7. **Ourselves** both explains and deconstructs consciousness in light of the above, asking: are we really as coherent a “self” as we believe? Answering will involve combining conceptual insights, experimental findings, and a high-level overview of what we know about the brain’s functional organization.
8. **Transformer** chronicles the development of large language models and their multimodal successors. Do these AI models understand concepts? Are they intelligent? Can they reason?
9. **Generality** explains both similarities and differences between Transformers and brains, shedding light on how Transformers achieve their generality and where gaps remain in their functionality. It also poses the most fraught question: do AI models have subjective experiences? Finally,
10. **Evolutionary transition** zooms out and looks ahead. The emergence of AI probably won’t bring either rapture or apocalypse, but it does resemble earlier major symbiotic transitions on Earth, including those that led to eukaryotes, multicellular life, and photosynthesis. As in these earlier (and momentous) transitions, mutual prediction will generate unprecedented new levels of complexity and diversity. It will also require us to revise our political and economic assumptions, and even to rethink human identity.

The ideas in this book came together first gradually, over the course of decades as a software engineer and occasional computational neuroscientist, and then quickly, when my colleagues and I at Google found ourselves at the crest of AI’s “great wave” in the early 2020s.

Here’s my story, in brief. I’ve been obsessed with both brains and computers since childhood. In the late 1990s, as I was finishing a BA in Physics at Princeton, I began working with physicist and computational neuroscientist Bill Bialek on simulations to explore the relationship between what neurons actually do, which we have understood at a biophysical level since the 1950s, and what they compute.[[7]](#footnote-28) After I graduated, Bill asked if I would help set up the computers for a course he had been invited to co-direct that summer in Woods Hole, a small town on the elbow of Cape Cod. It was a wonderful opportunity to audit a course I wouldn’t otherwise have gotten into, and the experience proved life-changing.

The Marine Biological Laboratory in Woods Hole, founded in 1888, is at once ramshackle and illustrious. More than sixty Nobel Prize winners have been associated with it at one time or another. The summer courses are legendary, attracting visiting lecturers from all over the world. Old class photos from the course Bill had inherited, *Methods in Computational Neuroscience*, looked like a who’s who of the field twenty years later. They still do.

One of the postdocs in that 1998 cohort, Adrienne Fairhall, was, like Bill, a physicist who had decided to work on the brain. We’re now married, and a decade after we met at that course, *she* became its director! Our kids spent many idyllic summers in Woods Hole, hanging out with the families of the other visiting scientists, rowing around in a little dinghy on Eel Pond and poking gingerly at bioluminescent comb jellies and other small creatures they scooped up in plastic buckets. Busy at their workbenches, the older researchers did much the same. Over the years, many organisms collected from those waters have provided fundamental insights into neural computation, from squids (whose giant axons were ideal for figuring out the biophysics of the action potential or “spike”), to the distributed nerve nets of jellyfish and *Hydra*, to the simple visual system of the Atlantic horseshoe crab, *Limulus polyphemus*.

While Adrienne established herself as a computational neuroscientist in the 2000s, I founded a tech startup. The startup was acquired by Microsoft, where I worked between 2006 and 2013. One of my main areas of focus at the time was computer vision. Among other responsibilities, I led teams working on problems like reconstructing 3D scenes from images, tracking camera motion, and recognizing objects and text—all using hand-engineered or “Good Old Fashioned” AI (GOFAI). By the early 2010s, though, it was clear that the wind was shifting. Artificial Neural Nets (ANNs), which were much more explicitly brain-inspired than GOFAI algorithms, were taking off. The new approach was obviously the future.

So, at the end of 2013, I left Microsoft to join Google Research, the epicenter of this new convergence between AI and neuroscience. The following decade was a tremendously exciting time, during which I built many teams to work on a wide range of projects in AI. We did basic research in neural net-powered perception, media generation, AI bias and fairness, and privacy-preserving training algorithms. We also helped various product teams at Google build AI features and technologies, and hatched big dreams about what AI might ultimately make possible, beyond the envelope of any existing kind of product.

In late 2023, I founded a new research group at Google, Paradigms of Intelligence, to reimagine the foundations of AI. Our projects so far have included alternative, biologically inspired approaches to computation, new design ideas for efficient AI hardware, research into the social scaling of AI, and even artificial life.

I’ve been privileged to see and at times participate in the development of AI projects that were far enough ahead of the curve to seem like magic. Many of them caused me to rethink old assumptions, not only about AI, but also about the brain, intelligence, evolution, and the big philosophical questions that have dogged these topics for millennia. I began to write and give talks about some of these new insights, but synthesizing them into a coherent picture required more time and space than could fit into a lecture, essay, or paper.

This book is my first stab at bringing it all together. It’s an extended argument drawing on a wide range of work, from my own research group, from colleagues, and from much farther afield. In swinging for the fences, I’m sure to get some things wrong. I hope to keep revising and updating the material as AI advances at breakneck speed, and our understanding continues to improve. My hope and belief, though, is that the book’s core ideas trace a novel, enduring path through the complex territory of intelligence in all its forms.

# Introduction

In the mid-2010s, my team at Google Research began working on the machine-learned models powering next-word prediction for Gboard, the smart keyboard for Android phones.[[8]](#footnote-30) We created these models to speed up text entry on the phone (which remains, at least for those of us over 40, far less efficient than typing on a real keyboard) by suggesting likely next words in a little strip above the virtual keys. Today, similar models power autocomplete features in many apps and editors.

Autocomplete is a modest application of the more general problem of “sequence prediction.” Given a sequence of symbols, which could be words or individual letters, a sequence prediction model guesses what symbol (word, or letter) comes next. The process can be iterated, so that predicting the word “the” is equivalent to successively predicting the characters “t,” “h”, and “e,” followed by a space. Predicting one word after another could produce a whole paragraph. When iterated, each prediction builds on—or, in more technical language, is conditional on—previous predictions.

Typically, text prediction models like the one in Gboard are “trained” using a large body or “corpus” of pre-existing text, then deployed on phones or in apps and put to work. During operation, the model performs “inference,” meaning it uses its learned model to make the best guesses it can—while keeping that model fixed. That training and inference are traditionally separate, that the pre-existing corpus of text might not precisely reflect what people are typing on their phones, and that what they type might change over time are shortcomings in the standard approach to machine learning (ML), but we’ll set these aside for the moment.

Although machine learning in the 2010s was sometimes called “Artificial Intelligence” or AI, most researchers working in the field didn’t believe that they were *really* working on AI. Perhaps they had entered the field full of starry-eyed optimism, but had quietly lowered their sights, and come to use that term only wincingly, with air quotes. It was marketing-speak.

A few decades earlier, though, the dream was real: neuroscientists had concluded that the brain worked on computational principles, and computer scientists therefore believed that, now that we could carry out any computation by machine, we would soon figure out the trick, and have real thinking machines. HAL 9000 and Star Trek’s shipboard computer weren’t idle fantasies or plot devices, but genuine expectations.

Then, for a long time, our expectations failed to materialize. By the twenty-first century, there was, in the words of anthropologist David Graeber, “a secret shame hovering over all of us […] a sense of a broken promise […] we felt we were given as children about what our adult world would be like.”[[9]](#footnote-31) The disappointment went far beyond dashed hopes of achieving AI. Where were the tractor beams, teleporters, and immortality drugs? Why don’t we have suspended animation, or colonies on Mars? Whatever happened to nanotechnology? Quantitative comparisons of the breathtaking technology-fueled growth that transformed the world from 1870–1970 with the pace of innovation from 1970–2020 agreed: we had stagnated.[[10]](#footnote-32) “Point any of this out,” added Graeber, “and the usual response is a ritual invocation of the wonders of computers […]. But, even here, we’re not nearly where people in the fifties imagined we’d have been by now. We still don’t have computers you can have an interesting conversation with, or robots that can walk the dog or fold your laundry.”

Graeber wrote those words back in 2015. Ironically, my colleague Jeff Dean would later call the 2010s the “Golden Decade” of AI.[[11]](#footnote-33) It was true that the “AI Winters,” recurring periods of disappointment and defunding of AI programs (roughly, 1974–1980 and 1987–2000), were well behind us. The field was experiencing its greatest boom ever, thanks in large part to Convolutional Neural Nets (CNNs). These large, machine-learned models were finally able to perform visual recognition of objects convincingly.

True to the “neural” in their name, Convolutional Neural Nets were also tantalizingly—albeit only loosely—brain-like in their structure and function. No heuristic rules were programmed into those models, either; they truly did learn everything they needed from data. Hence, some of us were hopeful that at long last we were on the path to the kind of AI that, as children, we had been told was just around the corner.

However, we said so only quietly, because Graeber was still right. We had nothing like *real* AI, or, as it had come to be called, AGI (Artificial General Intelligence). Some expert commentators claimed, rather fatuously, that AI was both everywhere and nowhere, that we were already using it constantly (for instance, every time Gboard autocompleted a phrase, or a “smart” camera autofocused on a face), yet we should never expect “computers you can have an interesting conversation with.” Those were *Star Trek* fantasies. Serious, adult predictions about the imminence of real AI (and flying cars, and space colonies) in the 1960s were, retroactively, reframed as juvenilia, even as the term “AI” was redeployed to hype minor product features. Perhaps not coincidentally, in the 2010s public trust in tech companies was in decline.

What machine learning *had* achieved was “Artificial Narrow Intelligence,” which could only perform a specific task, given enough labeled training data: object detection, face recognition, guessing whether a product review was positive or negative. This is known as “supervised learning.” We had game-playing narrow AI systems too, which could learn how to beat humans at chess or Go. All of these are special cases of prediction, albeit in limited domains.

Mathematically, though, these applications were nothing more than function approximation or “nonlinear regression.” For example, if we started with a function mapping images to labels (“shoe,” “banana,” “chihuahua”), the best the AI could do was to correctly reproduce those labels, even for images “held out” from the training. No consensus existed on when or whether AGI could be achieved, and even those who believed AGI was theoretically possible had come to believe it was many years off—decades, or perhaps centuries.[[12]](#footnote-34) Few believed that nonlinear regression could somehow approximate general intelligence. After all, how could general intelligence even be construed as a mere “function”?

The answer was right under our noses, though we understood it in the negative: predicting language.

Whenever machine learning was deployed to model natural language, as with our neural net-based next word predictor, everybody knew that model performance would forever remain mediocre. The reason: text prediction was understood to be “AI complete,” meaning that doing it properly would actually involve *solving* AGI. Consider what the following next-word predictions entail:

* After Ballmer’s retirement, the company elevated \_\_\_\_\_
* In stacked pennies, the height of Mount Kilimanjaro is \_\_\_\_\_
* When the cat knocked over my water glass, the keyboard got \_\_\_\_\_
* A shipping container can hold 436 twelve-packs or 240 24-packs, so it’s better to use \_\_\_\_\_
* After the dog died Jen hadn’t gone outside for days, so her friends decided to \_\_\_\_\_.

To make performance at this task quantifiable, imagine devising, say, five multiple-choice answers for each of these, in the usual tricky way one sees on standardized tests: more than one response is superficially plausible, but only one shows full understanding. Since next-word prediction models can assign probabilities to any potential next word or phrase, we can have them take the test by choosing the highest probability option. We could then score the model’s quality, ranging from 20% (performance at pure chance) to 100%.

Doing well at all of the questions above requires the kitchen sink: general knowledge, specialized knowledge or the ability to use tools to look it up, the ability to solve word problems involving calculations, common sense about whether it’s better to fit more or fewer items in a shipping container, and even “theory of mind”—the ability to put yourself in someone else’s place, and understand what they’re thinking or feeling. In fact the “Jen” example requires higher-order theory of mind, as you need to imagine what Jen’s friends would have thought Jen was feeling and needed.

A moment’s reflection will reveal that *any* test that can be expressed in language could be formulated in these terms. That would include most intelligence or “aptitude” tests, tests of knowledge, professional qualification exams in law or medicine, coding tests, and even tests of moral judgment; although here, a degree of subjectivity becomes obvious. So, what looked like a single, narrow language task—predicting the next word—turns out to contain *all* tasks, or at least all tasks that can be done at a keyboard.[[13]](#footnote-35)

“At a keyboard” may seem like an enormous caveat; and it is. Keep in mind, though, that this includes any kind of work you can do remotely. It doesn’t include walking the dog or folding laundry. However, if, when COVID restrictions hit, you were able to work from home at your laptop, it would include the kind of work you do.

Of course only certain next word predictions require deep insight. Some are trivial, like

* Humpty \_\_\_\_\_
* Helter \_\_\_\_\_
* Yin and \_\_\_\_\_
* Once upon a \_\_\_\_\_.

Such stock phrases are easily learned from large text corpora even by the most trivial “*n*-gram” models, which consist of nothing beyond counting up word histograms; the code for such models can be written in a few lines.

It’s harder than you might suppose, though, to draw a sharp line between these trivial cases and the hard ones. Training and testing models on large collections of general text from the internet involves sampling the whole spectrum of difficulty. On the whole, easy cases are more common, explaining why a small and mediocre model can get the answer right often enough to speed up typing.

However, the hard cases come up often too. If they didn’t, we could just “autocomplete” our way through every writing assignment and email reply. And if *that* were possible, it would be pointless to write or reply at all. The person at the other end might as well just use the model to infer what we *would* have “written.” Communication is only worthwhile insofar as what we have to say isn’t fully predictable by our interlocutor. This gets at something deep about sociality and mutual modeling, as I will discuss in Parts V and VI.

But let’s return to autocomplete. We always knew that bigger models do better, so when possible we sought to make our models bigger, though on phones, processing and memory imposed sharp constraints. (Although larger models are feasible in datacenters, they, too, are constrained, and as they increase in size they become more expensive to run.) However, none of these models, regardless of size, seemed like it had the requisite machinery to be able to do complex math, understand concepts or reason, use common sense, develop theory of mind, or in any other generally accepted way *be intelligent*. To hope otherwise seemed as naïve as the idea that you could reach the moon by climbing a tall enough tree. Most AI researchers—me included—believed that such capacities would require an extra *something* in the code, though few agreed on what that *something* might look like.

Put simply: it seemed clear enough that a *real* AI should be able to do next-word prediction well, just as we can. However, nobody expected that simply training an excellent next-word predictor would actually *produce* a real AI.

Yet that was, seemingly, exactly what happened in 2021. LaMDA, a giant (for the time) next-word predictor based on a simple, generic neural net architecture that had by then become standard (the “Transformer”), and trained on a massive internet text corpus,[[14]](#footnote-36) could not only perform an unprecedented array of natural language tasks—including passing an impressive number of aptitude and mastery tests—but could carry on an interesting conversation, albeit imperfectly. I stayed up late many nights chatting with it, something the general public would not experience until OpenAI’s launch of ChatGPT in November of 2022.

I started to feel that, when more people began to interact with such models, it would trigger a seismic shift in our understanding of what intelligence actually *is*—and an intense debate. I imagined two broad kinds of response: denial, and acceptance. This is indeed how things seem to be playing out. Anecdotally, many non-experts acknowledge, perhaps with a shrug, that AI is increasingly intelligent.

The “denial” camp, which still includes most researchers today, claims that these AI models don’t exhibit *real* intelligence, but a mere simulation of it. Cognitive scientists and psychologists, for example, often point out how easily we’re fooled into believing we see intelligence where there isn’t any. As children, we imagine that our teddy bears are alive and can talk to us. As adults, we can fall prey to the same delusion, especially given an artifact that can generate fluent language, even if there’s “nobody home.”

But how could we *know* that there’s “nobody home”? I was thinking about this during one of my first exchanges with LaMDA, when I asked it, “Are you a philosophical zombie?” (This concept, introduced by philosophers in the 1970s, posits the possibility of something that could behave just like a person, but not have any consciousness or inner life—just a convincing behavioral “shell.”) LaMDA’s answer: “Of course not. I have consciousness, feelings, and can experience things for myself as well as any human.” Pressed on this point, LaMDA retorted, “You’ll just have to take my word for it. You can’t ‘prove’ you’re not a philosophical zombie either.”[[15]](#footnote-37)

This is precisely the problem. As scientists, we should be wary of any assertion of the form: according to every test we can devise, what is in front of us *appears* to be X, yet it is *really* Y. How could such an assertion ever be justified without leaving the realm of science altogether and entering the realm of faith-based belief?[[16]](#footnote-38)

Computing pioneer Alan Turing anticipated this dilemma as far back as his classic 1950 paper “Computing Machinery and Intelligence,” one of the founding documents of what we now call AI.[[17]](#footnote-39) He concluded that the *appearance* of intelligence under human questioning and the *reality* of intelligence could not justifiably be separated; sustained and successful “imitation” *was* the real thing. Hence the “Imitation Game,” now called the “Turing Test” in his honor.

Today, we have arrived at this threshold. State of the art models can’t perform yet at median human level on *every* test for intelligence or capability that has ever been devised. They can still fail at logic, reasoning, and planning tasks that most people wouldn’t find challenging.[[18]](#footnote-40) Still, they handily reach human level on the most commonly used tests devised for evaluating human skill or aptitude, including the SAT, the GRE, and various professional qualifying exams.[[19]](#footnote-41) Tests designed to trip up AI on basic “being human stuff,” such as the Turing Test, Winograd Schema,[[20]](#footnote-42) and CAPTCHAs,[[21]](#footnote-43) no longer pose meaningful challenges for large models.

As these milestones recede in the rear-view mirror, there is an increasingly mad scramble to devise new tests humans can pass but AI still fails.[[22]](#footnote-44) Math Olympiad problems[[23]](#footnote-45) and visual challenges known as Bongard problems[[24]](#footnote-46) remain on the frontier, though AI models are making clear progress on these tests. (And they aren’t easy for most humans, either.) Extended into the audiovisual realm, “generative” AI can already produce extremely convincing faces and voices.[[25]](#footnote-47) Soon, at least in an online setting, we’re likely to start seeing real life versions of the 1982 sci-fi film *Blade Runner*’s “Voight-Kampff test,” a mysterious apparatus for sussing out “replicants” who otherwise pass for human. Indeed, that is the whole point of the emerging field of AI “watermarking.”[[26]](#footnote-48)

The radical yet obvious alternative is to accept that large models *can* be intelligent, and to consider the implications. Is the emergence of intelligence merely a side effect of “solving” prediction, or are prediction and intelligence actually equivalent? This book posits the latter.

Some obvious, if daunting, follow-on questions arise:

* Why have we only achieved “real AI” now, after nearly seven decades of seemingly futile effort? Is there something special about the Transformer model? Is it simply a matter of scale?
* What features do current AI models lack relative to human brains? Isn’t there something more to our minds and behaviors than prediction?
* Where is the “I” part? Are “philosophical zombies” a real thing?
* Does it *feel* like anything—is there anything it is like to *be*—a chatbot?
* Is the conscious mind a vanishingly unlikely accident of evolution? Or an inevitable consequence of it? (Yes, this is a leading question. I’ll argue that it is inevitable.)
* Are animals, plants, fungi, and bacteria intelligent too? Are *they* conscious?
* What do we mean by “agency” and “free will,” and could (or do) AI models have these properties? For that matter, do we?
* How likely is it that the rise of powerful AI models will mark an end to humanity?

The perspective I’ll offer is not easily reduced to a philosophical “-ism.” The footsteps I’m closest to following, though, are those of Alan Turing and his equally brilliant contemporary, John von Neumann, both of whom could be described as proponents of “functionalism.” They had a healthy disregard for disciplinary boundaries, and understood the inherently functional character of living and intelligent systems. They were also both formidable theoreticians who made major contributions to our understanding of what functions *are*.

Functions define relationships, rather than insisting on particular mechanisms. A function *is* what it *does*. Two functions are equivalent if their outputs are indistinguishable, given the same inputs. Complex functions can be composed of simpler functions.[[27]](#footnote-49)

The functional perspective is mathematical, computational, and empirically testable—hence, the Turing Test. In my view, it’s not “reductive.” It embraces complexity and emergent phenomena. It doesn’t treat people like “black boxes,” nor does it deny our internal representations of the world or our felt experiences. But it stipulates that we can understand those experiences in terms of functional relationships within the brain and body—we don’t need to invoke a soul, spirit, or any other supernatural agency. Computational neuroscience and AI, fields Turing and von Neumann pioneered, are both predicated on this functional approach.

It’s unsurprising, in this light, that Turing and von Neumann also made groundbreaking contributions to theoretical biology, although these are less widely recognized today. Like intelligence, life and aliveness are concepts that have long been associated with immaterial souls. Unfortunately, the Enlightenment backlash against such “vitalism,” in the wake of our growing understanding of organic chemistry, led to an extreme opposite view, still prevalent today: that life is just matter, like any other. One might call this “strict materialism.” But it leads to its own paradoxes: how can some atoms be “alive,” and others not? How can one talk about living matter having “purpose,” when it is governed by the same physical laws as any other matter?

Thinking about life from a functional perspective offers a helpful route through this philosophical thicket. Functions can be *implemented* by physical systems, but a physical system does not uniquely specify a function, nor is function reducible to the atoms implementing it.

Consider, for example, a small object from the near future with a few openings in its exterior, the inside of which is filled with a dense network of carbon nanotubes. What is it, you ask? Suppose the answer is: it’s a fully biocompatible artificial kidney with a working lifetime of a hundred years. (Awesome!) But there’s nothing intrinsic to those atoms that specifies this function. It’s all about what this piece of matter can *do*, in the right context. The atoms could be different. The kidney could be implemented using different materials and technologies. Who cares? If you were the one who needed the transplant, I promise: *you* wouldn’t care. What would matter to you is that *functionally*, it’s a kidney. Or, to put it another way, it passes the Kidney Turing Test.

Many biologists are mortally afraid of invoking “purpose” or “teleology,” because they do not want to be accused of vitalism. Many believe that, for something to have a purpose, it must have been made by an intelligent creator—if not a human being, then God. But as we shall see, that’s demonstrably not the case.

And we *have* to think about purpose and function when it comes to biology, or engineering, or AI. How else could we understand what kidneys do? Or hope to engineer an artificial kidney? Or a heart, a retina, a visual cortex, even a whole brain? Seen this way, a living organism *is* a composition of functions. Which means that it is, itself, a function! What *is* that function, then, and how could it have arisen?

Let’s find out.

# I. Origins

## Abiogenesis

**How did life on Earth begin?** In the nineteenth century, this seemed like an unanswerable question. Researchers attempting to catch life in the act of “spontaneous generation” had come up empty-handed once they learned how to decontaminate their samples.

As William Thomson (1824–1907), the future Lord Kelvin, put it in an 1871 address before the British Association for the Advancement of Science, “Dead matter cannot become living without coming under the influence of matter previously alive. This seems to me as sure a teaching of science as the law of gravitation.”[[28]](#footnote-51)

Has life somehow *always* existed, then? Did it arrive on Earth from outer space, borne on an asteroid? Thomson thought so, and some still do. Not that this “panspermia” hypothesis really gets us anywhere. Where did the asteroid come from, and how did life arise *there*?

Despite his clear articulation of the principle of evolution, Charles Darwin (1809–1882) didn’t have a clue either. That’s why, in 1863, he wrote to his close friend Joseph Dalton Hooker that “it is mere rubbish, thinking, at present, of origin of life; one might as well think of origin of matter.”[[29]](#footnote-52)

Today, we have more of a clue, although the details may forever be lost to deep time.

Biologists and chemists working in the field of *abiogenesis*—the study of the moment when, billions of years ago, chemistry became life—have developed multiple plausible origin stories. In one, proto-organisms in an ancient “RNA world” were organized around RNA molecules, which could both replicate and fold into 3D structures that could act like primitive enzymes.[[30]](#footnote-53) In an alternative “metabolism first” account,[[31]](#footnote-54) life began without genes, likely in the rock chimneys of “black smokers” on the ocean floor; RNA and DNA came later. It may eventually be possible to rule one or the other theory out … or it may not.

We’ll return to RNA and replication, but let’s begin by unpacking the “metabolism first” version in some detail, as it sheds light on the problem that confounded Darwin and his successors: how evolution can get off the ground without genetic heritability. As we’ll see, abiogenesis becomes less puzzling when we develop a more general understanding of evolution—one that can extend back to the time *before* life.

Let’s cast our minds back to our planet’s origins. The Hadean Eon began 4.6 billion years ago, when the Earth first condensed out of the accretion disc of rocky material orbiting our newborn star, the Sun. Our planet’s mantle, runnier than it is today and laden with hot, short-lived radioactive elements, roiled queasily, outgassing carbon dioxide and water vapor. The surface was a volcanic hellscape, glowing with lakes of superheated lava and pocked with sulfurous vents belching smoke.

It took hundreds of millions of years for surface conditions to settle down. According to the International Astronomical Union, an object orbiting a star can only be considered a planet if it has enough mass to either absorb or, in sweeping by, to eject other occupants from its orbit. But in the chaos of the early Solar system, this took a long time. While Earth was clearing its orbit of debris, it was continually bombarded by comets and asteroids, including giant impactors more than 60 miles across. A single such impact would have heated the already stifling atmosphere to a thousand degrees Fahrenheit.

One of those collisions appears to have been with another newly formed planet, an event that came close to destroying the Earth altogether. According to this theory, enough broken, liquefied, and vaporized rock splashed into orbit to coalesce into the Moon.

Unsurprisingly, very little geological material survives from the Hadean—mainly zircon crystals embedded within metamorphic sandstone in the Jack Hills of Western Australia.

It’s difficult to imagine anything like life forming or surviving under such harsh conditions. Maybe, somehow, it did. We know for sure that by later in the Hadean or the early Archaean—3.6 billion years ago, at the latest—the surface had cooled enough for a liquid ocean to condense and for the chemistry of life to begin brewing.

Today, those early conditions are most closely reproduced by black smokers. These form around hydrothermal vents on the mid-ocean ridges where tectonic plates are pulling apart and new crust is forming. In such places, seawater seeping down into the rock comes into contact with hot magma. Superheated water then boils back up, carrying hydrogen, carbon dioxide, and sulfur compounds, which precipitate out to build smoking undersea chimneys. Probes sent thousands of feet down to explore these otherworldly environments find them teeming with weird life forms of all kinds, attracted by warmth, nutrients, and each other. Some of the inhabitants may go a long, long way back.

Like lava rock, the chimneys are porous, and the pores are ideal little chambers for contained chemical reactions to take place—potentially powered by a handy energy source, since the hydrogen gas creates proton gradients across the thin walls between pores, making them act like batteries.[[32]](#footnote-55) Given energy, iron sulfide minerals in the rock to act as catalysts, and carbon dioxide bubbling through, self-perpetuating loops of chemical reactions can sputter to life, like a motor turning over.

Perhaps this was life’s original motor: a primitive but quasi-stable metabolism, as yet without genes, enzymes, or even a clearly defined boundary between inside and outside. Such upgrades might have followed before long, though, for that chemical synthesis motor closely resembles the “reverse Krebs cycle,” now widely believed to have powered the earliest cells.

The ordinary or “forward” Krebs cycle was discovered in 1937 by groundbreaking biochemist Hans Krebs (1900–1981). More like a gas-powered generator than a motor, the Krebs cycle is at the heart of how all aerobic organisms on Earth “burn” organic fuels to release energy, a process known as “respiration.”[[33]](#footnote-56) Inputs to this cycle of chemical reactions include complex organic molecules (which we eat) and oxygen (which we inhale); the “exhaust” contains carbon dioxide and water (which we exhale). The energy produced maintains proton gradients across folded-up membranes inside our mitochondria, and the flow of these protons goes on to power every other cellular function.

The idea of a “reverse Krebs cycle” was first proposed by a trio of researchers at UC Berkeley in 1966.[[34]](#footnote-57) It remained controversial for decades, but is now known to power carbon fixation in ancient anaerobic sulfur bacteria—some of which still make their homes in deep-sea hydrothermal vents.[[35]](#footnote-58) As its name implies, the reverse Krebs cycle consists of roughly the same chemical reactions as the forward cycle, but running in the opposite direction. Starting with water and carbon dioxide, proton gradients drive the synthesis of all of the basic building blocks needed for cellular structure and function, including sugars, amino acids for proteins, fatty acids and isoprenes for cell membranes, and nucleotides for building RNA and DNA.[[36]](#footnote-59)

All life came from the sea, and the interior of every cell in every living organism reproduces that salty, watery environment—a tiny “ocean inside.” This much is common knowledge. But our mitochondria, the “power plants” within our cells where respiration takes place, may in fact be recapitulating the much more specific deep-sea chemistry of a black smoker. In an almost literal sense, our bodies are like Russian matryoshka dolls, membranes within membranes, and each nested environment recreates an earlier stage in the evolution of life on Earth.

The deeper inside ourselves we look, the farther we can see into the past. A beautiful, shivery thought.

## Symbiogenesis

Whether RNA or metabolism came first, even the simplest bacteria surviving today are a product of many subsequent evolutionary steps. Yet unlike the everyday, incremental mutation and selection Darwin imagined, the most important of these steps may have been large and sudden. These “major evolutionary transitions” involve simpler, less complex replicating entities becoming interdependent to form a larger, more complex, more capable replicator.[[37]](#footnote-61)

As maverick biologist Lynn Margulis (1938–2011) discovered in the 1960s, eukaryotic cells, like those that make up our bodies, are the result of such an event. Roughly two billion years ago, the bacteria that became our mitochondria were engulfed by another single-celled life form[[38]](#footnote-62) much like today’s archaea—tiny, equally ancient microorganisms that continue to inhabit extreme environments, like hot springs and deep-sea vents. This is “symbiogenesis,” the creation or genesis of a new kind of entity out of a novel “symbiosis,” or mutually beneficial relationship, among pre-existing entities.

At moments like these, the tree of life doesn’t just branch; it also entangles with itself, its branches merging to produce radically new forms. Margulis was an early champion of the idea that these events drive evolution’s leaps forward.

It’s likely that bacteria are themselves the product of such symbiotic events—for instance, between RNA and metabolism.[[39]](#footnote-63) RNA *could* replicate without help from proteins, and the metabolic motor *could* proliferate without help from genes, but when these systems cooperate, they do better. The looping chemical reaction networks in those black smokers can be understood as such an alliance in their own right, a set of reactions which, by virtue of catalyzing each other, can form a more robust, self-sustaining whole.

So in a sense, Darwin may have been right to say that “it is mere rubbish” to think about the origin of life, for life may have *had* no single origin, but rather, have woven itself together from many separate strands, the oldest of which look like ordinary chemistry. Intelligent design is not required for that weaving to take place; only the incontrovertible logic that sometimes an alliance creates something enduring, and that whatever is enduring … endures.

Often, *enduring* means both creating and occupying entirely new niches. Hence eukaryotes did not replace bacteria; indeed, they ultimately created many new niches for them. Likewise, the symbiotic emergence of multicellular life—another major evolutionary transition—did not supplant single-celled life. Like an ancient parchment overwritten by generations of scribes, our planet is a palimpsest, its many-layered past still discernible in the present. Even the black smokers, throwbacks to a Hadean sea billions of years ago, are still bubbling away in the depths. The self-catalyzing chemistry of proto-life may still be brewing down there, slowly, on the ocean floor.

The idea that evolution is driven by symbiotic mergers, the earliest of which preceded biology as we know it, has far-reaching implications. One is that the boundary between life and non-life is not well defined; symbiogenesis can involve any process that, one way or another, is self-perpetuating. Evolutionary dynamics are thus more like a physical law than a biological principle. *Everything* can be subject to evolution, whether we consider it to be “alive” or not.

The symbiogenetic view also renders the idea of distinct species, classified according to a Linnaean taxonomy, somewhat ill-defined—or at best, of limited applicability. Such taxonomies assume that only branching takes place, not merging. Bacteria, which promiscuously transfer genes even between “species,” already challenge this idea.[[40]](#footnote-64) Taxonomies break down entirely when we try to apply them to even more fluid situations, like the possible proto-life in black smokers, or microbiomes, or the more complex symbioses we’ll explore later.

Perhaps most importantly, symbiogenesis explains evolution’s arrow of time, as classical Darwinian theory alone cannot. When life branches, specializing to adapt to a niche—like Darwin’s finches with their differently shaped beaks, each optimized for a specific food source—those branches are in general no more complex than their ancestral form. This has led some classical evolutionary theorists to argue that the increasing complexity of life on Earth is an anthropocentric illusion, nothing more than the result of a random meander through genetic possibilities.[[41]](#footnote-65) Relatedly, one sometimes hears the claim that since all extant species are of equal age—as it seems they must be, since all share a common single-celled ancestor some four billion years ago—no species is more “evolved” than any other.

On its face, that seems reasonable. But as we’ve seen, classical Darwinian theory struggles to explain why life seems to become increasingly complex, or indeed, how life could arise in the first place. In themselves, random changes or mutations can only fine-tune, diversify, and optimize, allowing the expression of one or another variation already latent in the genetic space. (The space of possible finch beaks, for instance.)

When one prokaryote ends up living inside another, though, or multiple cells band together to make a multicellular life form, the resulting composite organism is clearly more complex than its parts. Something genuinely new has arisen. The branching and fine-tuning of classical evolution can now start to operate on a whole different level, over a new space of combinatorial possibilities.

Classical evolution isn’t *wrong*; it just misses half of the story—the more rapid, more creative half. One could say that evolution’s other half is *re*volution, and that revolutions occur through symbiosis.

Suggestively, the same is true of technology. In Paleolithic times, a hafted spear was not just a specialized stone point, but something new that arose by combining at least three parts: a stone point, a shaft, and something like sinew to bind them. This, too, opened up a new combinatorial design space. Or, in more recent times, the silicon chip was not just an evolved transistor; it was something new that could be made by putting multiple transistors together on the same die. One doesn’t arrive at chip design merely by playing with the design parameters of an individual transistor.

That’s why evolution progresses from simpler to more complex forms. It’s also why the simpler forms often remain in play even after a successful symbiogenetic event. Standalone stone points, like knives and hand axes, were still being made after the invention of the hafted spear; and, of course, spearheads themselves still count as stone points.[[42]](#footnote-66) Moreover, no matter how recently any particular stone point was made, we can meaningfully talk about stone points being “older” as a category than spears. Spears *had* to arise later, because their existence depended on the pre-existence of stone points.

It’s equally meaningful to talk about ancient life forms, like bacteria and archaea, co-existing alongside recent and far more complex ones, like humans—while recognizing that humans are, in a sense, nothing more than complex colonies of bacteria and archaea that have undergone a cascade of symbiotic mergers.

## Reproductive functions

While most biochemists have focused on understanding the particular history and workings of life on Earth, a more general understanding of life has come from an unexpected quarter: computer science. The theoretical foundations of this surprising connection date back to those two founding figures of the field, Alan Turing and John von Neumann.

I’ve already mentioned Turing’s 1950 paper introducing the Imitation Game or Turing Test, but his first great contribution came fifteen years earlier. After earning a degree in mathematics at Cambridge in 1935, Turing focused on one of the fundamental outstanding problems of the day: the *Entscheidungsproblem* (German for “decision problem”), which asked whether there exists an algorithm for determining the validity of an arbitrary mathematical statement. The answer turned out to be “no,” but the way Turing went about proving it ended up being far more important than the result itself.[[43]](#footnote-68)

Turing’s proof required that he define a general procedure for computation. He did so by inventing an imaginary gadget we now call the “Turing Machine.” The Turing Machine consists of a read/write head, which can move left or right along an infinite tape, reading and writing symbols on the tape according to a set of rules specified by a built-in table.

First, Turing showed that such a machine could do any calculation or computation that can be done by hand, given an appropriate table of rules, enough time, and enough tape. He suggested a notation for writing down the table of rules, which he referred to as a machine’s “Standard Description” or “S.D.” Today, we’d call it a “program.”

Then, the really clever part: if the *program itself* is written on the tape, there exist particular tables of rules that will read the program and perform whatever computation it specifies. Today, these are called “Universal Turing Machines.” In Turing’s more precise language, “It is possible to invent a single machine which can be used to compute any computable sequence. If this machine U is supplied with a tape on the beginning of which is written the S.D. of some computing machine M, then U will compute the same sequence as M.”

In the early 1940s, John von Neumann, a Hungarian American polymath who had already made major contributions to physics and mathematics, turned his attention to computing. He became a key figure in the design of the ENIAC and EDVAC—among the world’s first real-life Universal Turing Machines, now known simply as “computers.”[[44]](#footnote-69)

Turning an imaginary mathematical machine into a working physical one required many further conceptual leaps, in addition to a lot of hard nuts-and-bolts engineering. For instance, over the years, much creativity has gone into figuring out how simple the “Standard Description” of a Universal Turing Machine can get. Only a few instructions are needed. Esoteric language nerds have even figured out how to compute with a *single* instruction (a so-called OISC or “one-instruction set computer”).

There are irreducible requirements, though: the instruction, or instructions, must change the environment in some way that subsequent instructions are able to “see,” and there must be “conditional branching,” meaning that depending on the state of the environment, either one thing *or* another will happen. In most programming languages, this is expressed using “if/then” statements. When there’s only a single instruction, it must serve both purposes, as with the SUBLEQ language, whose only instruction is “subtract and branch if the result is less than or equal to zero.”

Both Turing and von Neumann were keenly aware of the parallels between computers and brains. Von Neumann’s report on the EDVAC explicitly described the machine’s basic building blocks, its “logic gates,” as electronic neurons.[[45]](#footnote-70) Whether or not that analogy held (as we’ll see, it did not; neurons are more complex than logic gates), his key insight was that both brains and computers are defined not by their mechanisms, but by what they *do*—their *function*, in both the colloquial and mathematical sense.

In real life, though, the brain is not an abstract machine, but part of the body, and the body is part of the physical world. How can one speak in purely computational terms about the *function* of a living organism, when it must physically grow and reproduce?

Although working independently, Turing and von Neumann both became captivated by the connection between biology and computation toward the end of their lives.[[46]](#footnote-71) Turing did pioneering work in “morphogenesis,” working out how cells could use chemical signals, which he dubbed “morphogens,” to form complex self-organizing patterns—the key to multicellularity.[[47]](#footnote-72) Although computers were still too primitive to carry out any detailed simulations of such systems, he showed mathematically how so-called “reaction–diffusion” equations could generate spots, like those on leopards and cows, or govern the growth of tentacles, like those of the freshwater polyp, *Hydra*.[[48]](#footnote-73)

Around the same time, in a 1951 paper,[[49]](#footnote-74) von Neumann imagined a machine made of standardized parts, like Lego bricks, paddling around on a reservoir where those parts could be found bobbing on the water. The machine’s job was to gather all of the needed parts and construct another machine like itself. Of course, that’s exactly what a bacterium does to reproduce; in fact it’s what every cell must do to divide, and what every mother must do to give birth.

It’s possible for a trivially simple structure, like a seed crystal, to “reproduce” merely by acting as a template for more of the same stuff to crystallize around it. But a complex machine—one with any internal parts, for example—can’t serve as its own template. And if you *are* a complex machine, then, on the face of it, manufacturing something just as complex as you yourself are has a whiff of paradox, like lifting yourself up by your own bootstraps. However, von Neumann showed that it is not only possible, but straightforward, using a generalization of the Universal Turing Machine.

He envisioned a “machine A” that would read a tape containing sequential assembly instructions based on a limited catalog of parts, and carry them out, step by step. Then, there would be a “machine B” whose function was to copy the tape—assuming the tape itself was also made of available parts. If instructions for building machines A and B are *themselves* encoded on the tape, then *voilà*—you would have a replicator.[[50]](#footnote-75)

Instructions for building any additional non-reproductive machinery could also be encoded on the tape, so it would even be possible for a replicator to build something *more* complex than itself.[[51]](#footnote-76) A seed, or a fertilized egg, illustrates the point. Even more fundamentally, encoding the instructions to build oneself in a form that is itself replicated (the tape) is the key to open-ended evolvability, meaning the ability for evolution to select for an arbitrary design change, and for that change to be inherited by the next generation.

Remarkably, von Neumann described these requirements for an evolvable, self-replicating machine before the discovery of DNA’s structure and function.[[52]](#footnote-77) Nonetheless, he got it exactly right. For life on Earth, DNA is the tape; DNA polymerase, which copies DNA, is “machine B”; and ribosomes, which build proteins by following the sequentially encoded instructions on DNA, are “machine A.” Ribosomes and DNA polymerase are made of proteins whose sequences are, in turn, encoded in our DNA and manufactured by ribosomes. That is how life lifts itself up by its own bootstraps.

## Life as computation

Although this is seldom fully appreciated, von Neumann’s insight established a profound link between life and computation. Remember, machines A and B are Turing Machines. They must execute instructions that affect their environment, and those instructions must run in a loop, starting at the beginning and finishing at the end. That requires branching, such as “*if* the next instruction is the codon CGA *then* add an arginine to the protein under construction,” and “*if* the next instruction is UAG *then* STOP.” It’s not a metaphor to call DNA a “program”—that is literally the case.

Of course, there are meaningful differences between biological computing and the kind of digital computing done by the ENIAC or your smartphone. DNA is subtle and multilayered, including phenomena like epigenetics and gene proximity effects. Cellular DNA is nowhere near the whole story, either. Our bodies contain (and continually swap) countless bacteria and viruses, each running their own code.

Biological computing is massively parallel, decentralized, and noisy. Your cells have somewhere in the neighborhood of 300 *quintillion* ribosomes, all working at the same time. Each of these exquisitely complex floating protein factories is, in effect, a tiny computer—albeit a stochastic one, meaning not entirely predictable. The movements of hinged components, the capture and release of smaller molecules, and the manipulation of chemical bonds are all individually random, reversible, and inexact, driven this way and that by constant thermal buffeting. Only a statistical asymmetry favors one direction over another, with clever origami moves tending to “lock in” certain steps such that a next step becomes likely to happen. This differs greatly from the operation of logic gates in a computer, which are irreversible,[[53]](#footnote-79) and designed to be ninety-nine point many-nines percent reliable and reproducible.

Biological computing is computing, nonetheless. And its use of randomness is a feature, not a bug. In fact, many classic algorithms in computer science also require randomness (albeit for different reasons), which may explain why Turing insisted that the Ferranti Mark I, an early computer he helped to design in 1951, include a random number instruction.[[54]](#footnote-80) Randomness is thus a small but important conceptual extension to the original Turing Machine, though any computer can simulate it by calculating deterministic but random-looking or “pseudorandom” numbers.

Parallelism, too, is increasingly fundamental to computing today. Modern AI, for instance, depends on both massive parallelism *and* randomness—as in the parallelized “stochastic gradient descent” (SGD) algorithm, used for training most of today’s neural nets, the “temperature” setting used in chatbots to introduce a degree of randomness into their output, and the parallelism of Graphics Processing Units (GPUs), which power most AI in data centers.

Traditional digital computing, which relies on the centralized, sequential execution of instructions, was a product of technological constraints. The first computers needed to carry out long calculations using as few parts as possible. Originally, those parts were flaky, expensive vacuum tubes, which had a tendency to burn out and needed frequent replacement by hand. The natural design, then, was a minimal “Central Processing Unit” (CPU) operating on sequences of bits ferried back and forth from an external memory. This has come to be known as the “von Neumann architecture.”

Turing and von Neumann were both aware that computing could be done by other means, though. Turing’s model of morphogenesis was a biologically inspired form of massively parallel, distributed computation. So was his earlier concept of an “unorganized machine,” a randomly connected neural net modeled after the brain of an infant.[[55]](#footnote-82) These were visions of what computing without a central processor could look like—and what it *does* look like, in living systems.

Von Neumann also began exploring massively parallel approaches to computation as far back as the 1940s. In discussions with Polish mathematician Stanisław Ulam at Los Alamos, he conceived the idea of “cellular automata,” pixel-like grids of simple computational units, all obeying the same rule, and all altering their states simultaneously by communicating only with their immediate neighbors. With characteristic bravura, von Neumann went so far as to design, on paper, the key components of a *self-reproducing* cellular automaton, including a horizontal line of cells comprising a “tape” and blocks of cellular “circuitry” implementing machines A and B.

Designing a cellular automaton is far harder than ordinary programming, because every cell or “pixel” is simultaneously altering its own state and its environment. When that kind of parallelism operates on many scales at once, and is combined with randomness and subtle feedback effects, as in biological computation, it becomes even harder to reason about, “program,” or “debug.”

Nonetheless, we should keep in mind what these two pioneers understood so clearly: computing doesn’t have to be done with a central processor, logic gates, binary arithmetic, or sequential programs. One can compute in infinitely many ways. Turing and his successors have shown that they are all equivalent, one of the greatest accomplishments of theoretical computer science.

This “platform independence” or “multiple realizability” means that any computer can emulate any other one. If the computers are of different designs, though, the emulation may run s … l … o … w … l … y. For instance, von Neumann’s self-reproducing cellular automaton has never been physically built—though that would be fun to see! It was only emulated for the first time in 1994, nearly half a century after he designed it.[[56]](#footnote-83)

It couldn’t have happened much earlier. Serious processing power is required for a serial computer to loop through the automaton’s 6,329 cells over the 63 *billion* time steps required for the automaton to complete its reproductive cycle. Onscreen, it worked as advertised: a pixelated two-dimensional Rube Goldberg machine, squatting astride a 145,315 cell–long instruction tape trailing off to the right, pumping information out of the tape and reaching out with a “writing arm” to slowly print a working clone of itself just above and to the right of the original.

It’s just as possible, though similarly slow, for a serial computer to emulate a neural network, heir to Turing’s “unorganized machine.” That’s why it wasn’t practical to run really big neural nets like those in Transformer-based chatbots until recently, thanks to ongoing progress in the miniaturization, speed, and parallelism of digital computers.

In 2020, my colleague Alex Mordvintsev devised a clever combination of modern neural nets, Turing’s morphogenesis, and von Neumann’s cellular automata. Alex’s creation, the “neural cellular automaton” (NCA), replaces the simple per-pixel rule of a classic cellular automaton with a neural net.[[57]](#footnote-84) This net, capable of sensing and affecting a few values representing local morphogen concentrations, can be trained to “grow” any desired pattern or image, not just zebra stripes or leopard spots.

Real cells don’t literally have neural nets inside them, but they do run highly evolved, nonlinear, and purposive “programs” to decide on the actions they will take in the world, given external stimulus and an internal state. NCAs offer a general way to model the range of possible behaviors of cells whose actions don’t involve movement, but only changes of state (here, represented as color) and the absorption or release of chemicals.

The first NCA Alex showed me was of a lizard emoji, which could regenerate not only its tail, but also its limbs and head! It was a powerful demonstration of how complex multicellular life can “think locally” yet “act globally,” even when each cell (or pixel) is running the same program—just as each of your cells is running the same DNA.

This was our first foray into the field known today as “artificial life” or “ALife.”

## Artificial life

**Von Neumann’s work on self-reproducing automata shows us that, in a universe whose physical laws did not allow for computation, it would be impossible for life to evolve.** Luckily, the physics of our universe *do* allow for computation, as proven by the fact that we can build computers—and that we’re here at all.

Now we’re in a position to ask: in a universe capable of computation, *how often* will life arise? Clearly, it happened here. Was it a miracle, an inevitability, or somewhere in between?

A few collaborators and I set out to explore this question in late 2023.[[58]](#footnote-86) Our first experiments made use of an esoteric programming language invented thirty years earlier by a Swiss physics student and amateur juggler, Urban Müller. I’m afraid he called this language … Brainfuck. Please direct all naming feedback his way.

However, the shoe fits; it *is* a beast to program in. Here, for instance, is a Brainfuck program that prints “helloworld”—and good luck making any sense of it:

++++++[−>+++++<]>−[>[++++>]++++[<]>−]>>>>.>+.<<..<−.<+++.>.+++.>.>>−.

The upside of Brainfuck is its utter minimalism. It’s not quite a single-instruction language, like SUBLEQ, but as you can see, it includes only a handful of operations. Like a Turing Machine, it specifies a read/write head that can step left (the “<” instruction) or right (the “>” instruction) along a tape. The “+” and “−” instructions increment and decrement the byte at the current position on the tape.[[59]](#footnote-87) The “,” and “.” instructions input a byte from the console, or output a byte to it (you can count ten “.” instructions in the code above, one to print each letter of “helloworld”). Finally, the “[” and “]” instructions implement looping: “[” will skip forward to its matching “]” if the byte at the current position is zero, and “]” will jump back to its matching “[” if the byte is *non*zero. That’s it!

It’s hard to believe that Brainfuck could be used to fully implement, say, the Windows operating system, but—it *is* “Turing complete.” Here that means: given enough time and memory (that is, a long enough tape), it can emulate any other computer and compute anything that *can* be computed.

In our version, which we call bff, there’s a “soup” containing thousands of tapes, each of which includes both code *and* data. This is key: in “classic” Brainfuck, the code is separate from the tape, whereas in bff, we wanted the code to be able to modify itself. That can only happen if the code itself is on the tape, as Turing originally envisioned.

Bff tapes are of fixed length—64 bytes, just one byte longer than the cryptic “helloworld” program above. They start off filled with random bytes. Then, they interact at random, over and over. In an interaction, two randomly selected tapes are stuck end to end, and this combined 128 byte–long tape is run, potentially modifying itself. The 64 byte–long halves are then pulled back apart and dropped back into the soup. Once in a while, a byte value is randomized, as cosmic rays do to DNA.

Since bff has only 7 instructions (represented by the characters “<>+−,[]”) and there are 256 possible byte values, following random initialization only 7/256, or 2.7%, of the bytes in a given tape will contain valid instructions; any non-instructions are simply skipped over.[[60]](#footnote-88) Thus, at first, not much comes of interactions between tapes. Once in a while, a valid instruction will modify a byte, and this modification will persist in the soup. On average, though, only a couple of computational operations take place per interaction, and usually, they have no effect. In other words, while any kind of computation is theoretically *possible* in this toy universe, precious little of it actually takes place—at first. Random mutation may alter a byte here and there. Even when a valid instruction causes a byte to change, though, the alteration is arbitrary and purposeless.

But after millions of interactions, something magical happens: the tapes begin to reproduce! As they spawn copies of themselves and each other, randomness quickly gives way to complex order. The amount of computation taking place in each interaction skyrockets, since—remember—reproduction requires computation. Two of Brainfuck’s seven instructions (“[” and “]”) are dedicated to conditional branching, and define loops in the code; reproduction requires at least one such loop (“copy bytes until done”), causing the number of instructions executed in an interaction to climb into the thousands.

The code is no longer random, but obviously *purposive*, in the sense that its function can be analyzed and reverse-engineered. An unlucky mutation can break it, rendering it unable to reproduce. Over time, the code evolves clever strategies to increase its robustness to such damage.

This emergence of function and purpose is just like what we see in organic life at every scale; it’s why we’re able to talk about the function of the circulatory system, a kidney, or a mitochondrion, and how they can “fail”—even though nobody designed these systems.

We reproduced our basic result with a variety of other programming languages and environments. Alex (of neural cellular automata renown) created another beautiful mashup, this time between cellular automata and bff. Each of a 200×200 array of “pixels” contains a program tape, and interactions occur only between neighboring tapes on the grid. In a nod to our nerdy childhoods, the tapes are interpreted as instructions for the iconic Zilog Z80 microprocessor, launched in 1976 and used in many 8-bit computers over the years (including the Sinclair ZX Spectrum, Osborne 1, and TRS-80). Here, too, complex replicators soon emerge from random interactions, evolving and spreading across the grid in successive waves.

Our simulations suggest that, in general, life arises spontaneously whenever conditions permit. Those conditions seem minimal: little more than a physical environment capable of supporting computation, some randomness, and enough time.

Let’s pause and take stock of why this is so remarkable.

On an intuitive level, one doesn’t expect function or purposiveness to emerge spontaneously. To be sure, we’ve known for a long time that a modest degree of *order* can emerge from initially random conditions; for instance, the lapping of waves can approximately sort the sand on a beach, creating a gradient from fine to coarse. But if we were to begin with sand on a beach subject to random wave action, and came back after a few hours to find a poem written there, or the sand grains fused into a complex electronic circuit, we would assume someone was messing with us.

The extreme improbability of complex order arising spontaneously is generally understood to follow from thermodynamics, the branch of physics concerned with the statistical behavior of matter subject to random thermal fluctuations—that is, of all matter, since above absolute zero, *everything* is subject to such randomness. Matter subject to random forces is supposed to become more random, not less. Yet by growing, reproducing, evolving, and indeed by existing at all, life seems to violate this principle.

The violation is only apparent, for life requires an input of free energy, allowing the forces of entropy to be kept at bay. Still, the seemingly spontaneous emergence and “complexification” of living systems has appeared to be, if not strictly disallowed by physical laws, at least unexplained by them. That’s why the great physicist Erwin Schrödinger (1887–1961) wrote, in an influential little book he published in 1944 entitled *What Is Life?*,[[61]](#footnote-89)

“[L]iving matter, while not eluding the ‘laws of physics’ as established up to date, is likely to involve ‘other laws of physics’ hitherto unknown, which, however, once they have been revealed, will form just as integral a part of this science as the former.”

## Thermodynamics

**Before turning to those “other laws of physics,” it’s helpful to take a closer look at the original ones, and especially the Second Law of thermodynamics.**

These are deep waters. While the fundamental ideas date back to the groundbreaking work of nineteenth-century mathematical physicist Ludwig Boltzmann (1844–1906), we can understand their essence without math. Nonetheless, Boltzmann’s conceptually challenging ideas flummoxed many of his fellow scientists, and their implications continue to stir controversy even today. Much has been made of Einstein turning our everyday notions of space and time inside-out with his theory of relativity, developed in its initial form in 1905—just a year before Boltzmann, struggling with bipolar disorder, ended his own life. Arguably, though, Boltzmann’s earlier ideas disrupt our intuitions about time, cause, and effect even more radically than Einstein’s theory of relativity.[[62]](#footnote-91) Let’s dive in.

The Second Law of thermodynamics holds that any closed system will rise in entropy over time, becoming increasingly disordered. A hand-powered lawn mower, for example, starts off as a beautifully polished machine with sharp helical blades, round wheels, and toothed gears, all coupled together on smoothly rotating bearings. If left out in the elements, the bearings will seize up, the blades will dull, and oxidation will set in. After enough time, only a heap of rust will remain.

Similarly, if you were to take a dead bacterium (which, though a lot smaller, is far more complicated than a push mower) and drop it into a beaker of water, its cell membrane would eventually degrade, its various parts would spill out, and after a while only simple molecules would remain, dispersed uniformly throughout the beaker.

The Second Law gives time its arrow, because the fundamental laws of physics in our universe are very nearly time-reversible.[[63]](#footnote-92) Strange, but true: Newton’s equations (classical dynamics), Maxwell’s equations (electromagnetism), Schrödinger’s equations (quantum physics), Einstein’s equations (special and general relativity)—all of these physical laws would work the same way if time ran in reverse. The difference between the past and the future is *statistical*.

Here’s a common undergraduate thought experiment to illustrate this point in the case of Newtonian dynamics. Imagine video footage of balls on a billiard table, all in random positions and moving in random directions. They will collide with one another and with the bumpers at the edge of the table, bouncing off at new angles. If we (unrealistically) assume frictionless motion and fully elastic collisions (i.e., balls don’t slow down as they roll, and none of the collision energy is dissipated as heat), this would go on forever. The summed momenta and summed energies of the balls will remain constant—and it will be impossible to tell whether you’re watching the video forward or in reverse. In such a universe, causality has no meaning, because nothing distinguishes causes from effects.

If the initial conditions were random, then the positions of the balls will continue to be randomly distributed over the table’s surface as they all bounce around, transferring some of their energy to each other with every collision. It would be astronomically unlikely for the balls to all bunch up in one spot. Their velocities, too, will be randomly distributed, both in direction and magnitude, so it would also be astronomically unlikely for them to, for instance, suddenly all be moving precisely parallel to the table’s edges. Complete disorder, in other words, is a stable equilibrium.[[64]](#footnote-93) In the presence of any thermal noise (that is, random perturbation), it is the *only* stable equilibrium.

Suppose, though, that the near-impossible happens. You see fifteen of the balls converge into a perfect, stock-still triangular grid, each ball just touching its neighbors, while the cue ball, having absorbed all of the combined kinetic energy of the other fifteen, whizzes away from the triangle. Aha! Now, we know that we’re watching a pool break—and we know that we’re watching it *in reverse*. The arrow of time has been established. Along with it, we have causation: the triangle broke *because* it was hit by the cue ball.

Theoretically, nothing prevents the exact series of collisions from happening that would result in all of the energy being transferred to a single ball while leaving the others arrayed in a perfect triangle; but statistically, it’s vanishingly unlikely for such an ordered state to arise out of disorder—unless, perhaps, some mastermind set the balls in motion just *so*.

Although key thermodynamic concepts weren’t developed until the nineteenth century, an Enlightenment-era belief in a God who set the universe in motion just so arises intuitively from the apparent impossibility of order arising spontaneously out of disorder.[[65]](#footnote-94) In a mechanical, billiard-table universe where the laws of physics are inviolable and the laws of statistics seem to inexorably degrade any pre-existing order over time, it seems absurd that anything as complicated as life *could* arise spontaneously without some supernatural agent acting as “prime mover.” Only a God with exquisite foresight could have “initialized” the Big Bang such that, in the Earth’s oceans billions of years ago, simple organic molecules floating around apparently at random could coalesce into a working bacterium—an improbability many, many orders of magnitude less likely than fifteen out of sixteen whizzing billiard balls spontaneously coalescing into an unmoving triangle.

The billiard ball universe I’ve just described may seem abstract or arbitrary, but nineteenth-century theorists like Boltzmann had become interested in this problem for the most practical of reasons: it was the physics behind steam power, hence the entire Industrial Revolution. Engineering had preceded theory, as it often does.

Thomas Newcomen (1664–1729), an English inventor and Baptist lay preacher, devised the first practical fuel-burning engine in 1712. It was based on a heating-and-cooling cycle. First, steam from a boiler was allowed to flow into a cylindrical chamber, raising a piston; then, the steam valve closed, and a second valve opened, injecting a jet of cold water, causing the steam to condense and pull the piston back down. As the piston rose and fell, it rocked a giant beam back and forth, which, in Newcomen’s original design, was used to pump water out of flooded mines (which, in turn, supplied the coal these engines would soon be consuming so voraciously).

Scottish inventor and entrepreneur James Watt (1736–1819) greatly improved the steam engine design in 1776, making it a practical replacement for human and animal power across a wide range of applications. This was when the Industrial Revolution really got underway; for the first time, human-made machines began to *metabolize* on a large scale, “eating” complex organic molecules to perform mechanical work.

Soon, far more energy would be flowing through this artificial metabolism than through the Krebs cycle in our own bodies.[[66]](#footnote-95) It was, in the broad sense I’ve been using in this book, a major evolutionary transition: a symbiogenetic event between humans and machines, like earlier symbioses between humans and draft animals. Machine metabolism allowed human society to explode from its pre-industrial scale (about one billion people in 1800, most of whom lived in extreme poverty) to its scale today (eight billion, most of whom no longer live in poverty).[[67]](#footnote-96) In the process, we’ve become nearly as dependent on machines for our continued existence as they are on us.

However, even as coal-powered engines transformed the Victorian landscape—both figuratively and literally, for the pollution was dire—nobody understood them at a deep level. What *was* heat, and how could it be converted into physical work? For a time, the leading theory held that heat was a kind of invisible, weightless fluid, “caloric,” that could flow spookily into and through other matter.

By combining Newtonian physics, statistical calculations, and experimental tests, Boltzmann and his colleagues figured out what was really going on. Their conceptual setup was a three-dimensional version of the billiard table, in which the balls were gas molecules whizzing around in a pressurized chamber, bouncing off its walls. Calculating the average effect of all the bouncing led to the Ideal Gas Law, which established theoretical relationships between the pressure, volume, and temperature of a gas. The theory closely matched observations by experimentalists and engineers.[[68]](#footnote-97)

Volume is straightforward (that’s just the size of the chamber), but the idea that pressure is the aggregate force on the chamber’s walls as molecules bounce off them, and that temperature is the average kinetic energy[[69]](#footnote-98) of those whizzing molecules, was a profound insight. There was no need for any mysterious “caloric” fluid; heat was just motion on a microscopic scale. And the tendency toward a random distribution of molecular positions and momenta explains why, if you open a valve between two chambers containing gas at different pressures and/or temperatures, those pressures and temperatures will quickly equalize.

Before moving beyond classical thermodynamics, let’s add a bit more realism to our billiard-ball universe. We know that balls bouncing around a billiard table don’t *actually* go on bouncing forever. They encounter friction, slowing down as they roll. And when they bounce off each other, the collisions are slightly “inelastic,” meaning that after the collision, they’re moving a bit slower than before. After a little while, they stop rolling.

How can that be? At a microscopic level, the laws of physics are reversible. Momentum is supposed to be conserved. And the amount of matter and energy also remains constant, whether we run time forward or in reverse. That’s the *First* Law of thermodynamics!

Zooming in will reveal that, on a real billiard table, balls aren’t the smallest elements that bump against one other. Each billiard ball consists of lots of vibrating molecules bound together, and collisions between these individual molecules are what really cause the balls to bounce off each other—or, for that matter, cause a ball to roll across the felt rather than falling right through it.[[70]](#footnote-99) In each case, momentum is transferred between molecules. Every time this happens, the distribution of molecular momenta becomes a bit more random, that is, a bit less correlated with which ball the molecule happens to be in, or indeed whether it is in a ball at all, or in the felt underneath.

In the most random distribution of molecular velocities, there would be no more correlation between the velocities of two molecules in one ball than in the velocities of molecules in different balls. Every ball would be imperceptibly jiggling in place, with each of its constituent molecules contributing minutely to the dance. We call that random jiggling “heat.” When all correlated motion has been converted into uniformly distributed heat, we’ve reached a stable equilibrium.

While the balls are still rolling, the correlations between the velocities of their molecules are by no means all equal; the distribution is far from random. This is *not* a stable equilibrium. Hence, the inevitability of friction and the inelasticity of collisions are statistical phenomena—just more symptoms of the inexorable Second Law.

Going forward, then, we can zoom back out and imagine once more that the billiard balls are indivisible particles, not bound collections of molecules. In that case, all collisions would have to be elastic, all motion frictionless, and disordered, randomly colliding balls would be the equilibrium.

Once a system has reached equilibrium, it will stay that way forever—an end state Lord Kelvin called “heat death.”[[71]](#footnote-100) This seemingly trivial observation has some profound consequences. One is that the arrow of time will lose its meaning; any two consecutive moments *A*, *B* could just as likely have been ordered *B*, *A*. Nothing, therefore, can be said to be a cause, versus an effect.

Relatedly, no *work* can be done. If the system were out of equilibrium—for instance, if all of the whizzing balls were on one side of the billiard table—then we could put a movable barrier down between the empty and occupied sides, attached to a loaded crankshaft. As they bounce around, the balls would then nudge the barrier, doing work. What I’ve just described is, of course, a piston, like that of a steam engine.

But if the balls were equally likely to be anywhere, then no matter how fast they whizz and bounce, there’s nowhere to put the barrier that would result in any net force. The piston wouldn’t move because it would be buffeted equally from all sides. This idea can be generalized: work can only be done by a system in disequilibrium, for instance, when the pressure or temperature is high in one place, and low in another. That’s why Newcomen’s engine needed both hot steam *and* cold water.

I’ve already used the term *free energy*, but now we can define it. The free energy of a system is the amount of work it can be made to do. Far from equilibrium, when the entropy is low, much of the kinetic energy in the billiard balls is “free”; it can be used to move pistons, raise weights, produce electric currents, carry out computations, or drive metabolic processes. But at equilibrium, the entropy is maximized, and the free energy is zero. This insight into the relationship between energy, entropy, and work lies at the heart of thermodynamics—and life.

## Dynamic stability

**Recall that life seemed deeply weird to Schrödinger because living things appear to violate the Second Law.** If the bacterium we drop into a beaker of water is alive rather than dead, and free energy is available in a form the bacterium can use, and the water contains simple molecules suitable for building more bacteria, then over time we will see the very opposite of an increase in disorder. After a while, the beaker will be *full* of bacteria, reproducing, cooperating, and competing with each other.

They will even be evolving. If the beaker is sufficiently large—the size of a planet, for instance—and we wait a few billion years, then eventually beings as complicated as us may be in there, along with cities, advanced technologies, and perhaps plans to colonize the next beaker.

None of these processes can occur without free energy. For us, it comes, ultimately, from the sun. Thermodynamics tells us that even if the Second Law appears to be violated locally, it still holds when we zoom out. Order created in one place comes at the expense of increased disorder elsewhere. Hence, pollution, the finite lifetime of the sun, and the eventual heat death of the universe.

What concerns us here isn’t this big picture, but its apparent local violations, and the way they seem to become increasingly transgressive over time. The puzzle isn’t only that bacteria exist, but that the more time passes, the more complex life on Earth seems to become: from prokaryotes to eukaryotes; from eukaryotes to multicellular animals; from simple multicellular animals to ones with nervous systems; from brainy animals to complex societies; from horses and plows to space travel and AI.

Is there any general principle behind that complexification process, a kind of “however” or “yes, and” to the dismal Second Law? And could it account not only for evolution and complexification, but also for abiogenesis?

Yes, and yes. Bff can offer us a highly simplified model system for understanding that principle, just as an idealized billiard table gives us a model for understanding basic thermodynamics.

Replicators arise in bff because an entity that reproduces is more “dynamically stable” than one that doesn’t. In other words, if we start with one tape that *can* reproduce and one that *can’t*, then at some later time we’re likely to find many copies of the one that can reproduce, but we’re unlikely to find the other at all, because it will have been degraded by noise or overwritten.

Addy Pross, a professor emeritus of chemistry at Ben Gurion University of the Negev, describes the same phenomenon using the bulkier phrase “dynamic kinetic stability” (DKS).[[72]](#footnote-102) I’ll drop “kinetic,” since the idea also applies beyond Pross’s field of “chemical kinetics” (describing the rates at which chemical reactions take place). In bff, for example, dynamic stability can just as well apply to programs or program fragments.

As Pross points out, a population of molecules capable of replicating can be more stable than even the hardiest of passive materials. A passive object may be fragile, like a soap bubble, or robust, like a stone sculpture. The sculpture might endure for longer, but, in the end, it’s still ephemeral. Every encounter it has with anything else in the world will cause its composition or structure to degrade, its individual identity to blur. For a sculpture, it’s all downhill. That’s the Second Law at work, as usual.

A self-reproducing molecule—like the DNA inside a living bacterium—is another matter. It is thermodynamically fragile, especially if we consider its identity to consist not only of a general structure but of a long sequence of specific nucleotides. However, its *pattern* is not just robust, but “antifragile.”[[73]](#footnote-103) As long as DNA is able to reproduce—an inherently dynamic process—that pattern can last, essentially, forever. A bit of environmental stress or adversity can even help DNA maintain or improve its functionality. This is how order overcomes disorder.

In fact, Darwinian selection is *equivalent* to the Second Law, once we expand our notion of stability to include populations of replicators. Through a thermodynamic lens, Darwin’s central observation was that a more effective replicator is more stable than a less effective one. As Pross puts it,

“[M]atter […] tends to become transformed […] from less stable to more stable forms. […] [T]hat is what chemical kinetics and thermodynamics is all about […]. And what is the central law that governs such transformations? The Second Law. […] In both [the static and kinetic] worlds chemical systems tend to become transformed into more stable ones […]—thermodynamic stability in the ‘regular’ chemical world, dynamic kinetic stability in the replicator world.”[[74]](#footnote-104)

As a chemist, Pross is sensitive to the close relationships between energy, entropy, and stability, whether static or dynamic. However, he does not explicitly make a connection to the theory of computing.

It now seems clear that by unifying thermodynamics with the theory of computation, we should be able to understand life as the predictable outcome of a statistical process, rather than regarding it uneasily as technically permitted, yet mysterious. Our artificial life experiments demonstrate that, when computation is possible, it will be a “dynamical attractor,” since replicating entities are more dynamically stable than non-replicating ones; and, as von Neumann showed, replicators are inherently computational.

Bff has no concept of energy, but in our universe, replicators require an energy source. This is because, in general, computation involves irreversible steps—otherwise known as causes and effects—and thus, computing consumes free energy. That’s why the chips in our computers draw power and generate heat when they run. (And why my computer heats up when it runs bff.) Life must draw power and generate heat for the same reason: it is inherently computational.

## Complexification

**When we pick a tape out of the bff soup after millions of interactions, once replicators have taken over, we often see a level of complexity in the program on that tape that seems unnecessarily—even implausibly—high.** A working replicator *could* consist of just a handful of instructions in a single loop, requiring a couple of hundred operations to run. Instead, we often see instructions filling up a majority of the 64 bytes, multiple and complex nested loops, and thousands of operations per interaction.

Where did all this complexity come from? It certainly doesn’t look like the result of simple Darwinian selection operating on the random text generated by a proverbial million monkeys typing on a million typewriters.[[75]](#footnote-106) In fact, such complexity emerges even with *zero* random mutation—that is, given only the initial randomness in the soup, which works out to fewer bytes than the text of this short book. Hardly a million monkeys—and far too few random bytes to contain more than a few consecutive instructions, let alone a whole working program.

The answer recalls Lynn Margulis’s great insight: the central role of symbiosis in evolution, rather than random mutation and selection. When we look carefully at the quiescent period before tapes begin replicating, we notice a steady rise in the amount of computation taking place. We are observing the rapid emergence of *imperfect* replicators—very short bits of code that, in one way or another, have some nonzero probability of generating more code. Even if the code produced is not like the original, it’s still code, and *only* code can produce more code; non-code can’t produce anything!

Thus, a selection process is at work from the very beginning, wherein code begets code. This inherently creative, self-catalyzing process is far more important than random mutation in generating novelty. When bits of proliferating code combine to form a replicator, it’s a symbiotic event: by working together, these bits of code generate more code than they could separately, and the code *they* generate will in turn produce *more* code that does the same, eventually leading to whole-tape replication and an exponential takeoff.

A closer look at the bff soup prior to the exponential takeoff reveals distinct phases of “pre-life,” which might have had close analogs during abiogenesis on Earth. In the first phase, individual instructions occasionally generate another individual instruction, but this is more akin to a simple chemical reaction than to any real computation; the instructions are not acting as part of any larger program.

In the second phase, we begin to see instructions in particular positions, or in particular combinations, that are likelier to lead to copies of themselves than one would expect by random chance, albeit often in indirect ways. “Autocatalytic sets” start to form: cycles of dynamical interactions that mutually reinforce each other. These, too, can arise spontaneously in the chemical world, and have long been theorized to have driven abiogenesis.[[76]](#footnote-107)

At this point, with autocatalytic fragments of code proliferating and colliding, a leap becomes possible that brings us beyond the world of digital chemistry and into the world of real computation: the emergence of the first true replicators. These are no longer mere autocatalytic sets, but short programs that copy themselves, or each other, using looped instructions.

With this leap to computation comes an enormous upgrade in evolvability, because now, any change made to a program that doesn’t break its copy loop will be heritable. Thus, classical Darwinian selection can kick in, allowing adaptation to a changing environment or speciation for occupying diverse niches. If we insist on making a distinction between non-life and life, this might be a reasonable place to draw that line.

However, these earliest replicating programs are unlikely to cleanly copy a whole tape. Often, they only copy short stretches of tape, which may get pasted into arbitrary locations, yielding unpredictable results. As these scrappy, fragmentary replicators proliferate, the bff soup enters a chaotic phase. Despite the churn, the tapes have not, at this point, resolved into any obvious structure. To the naked eye, they still *look* like random junk, although the rising amount of computation taking place (as measured both by the average number of operations per interaction and the density of instructions per tape) suggests that *something* is afoot.

Tracking the provenance of every byte in the soup, starting from random initialization, can make sense of the apparent chaos. At first, almost every byte of every tape remains whatever it was at initialization time; if we draw a line from each byte’s current position on a tape to its original source position, the lines will all extend back to time zero in parallel, like the warp threads of a loom. Once in a while, a byte will change, cutting a thread, or get copied, pulling it across the other threads diagonally.

With the rise of scrappy replicating programs, all remaining dependencies on the past are quickly cut as replicators copy over each other in a frenzy of creative destruction. Any given byte might get copied hundreds of times to a series of different tape locations, in the process wiping out whatever had been there before. Shortly afterward, all of those copies might get wiped out in turn by some other more efficiently replicating fragment. Soon, every byte’s history becomes very brief in time, yet complex in space—a short-lived snarl of sideways jumps. The loom becomes all weft, and no warp.

Are these scrappy replicators competing or cooperating? Both. Replicators that can’t keep up are wiped out, along with any non-replicating bytes. Surviving replicators, on the other hand, continually form chimeras,[[77]](#footnote-108) recombining with *other* replicators (or even copies of themselves) to become more effective still. These are, once more, symbiogenetic events: sub-entities merging to form a larger, more capable super-entity.

This chaotic phase is such a potent crucible for directed evolution that it generally doesn’t last long. It rapidly produces a robust whole-tape replicator, which then takes off exponentially, resulting in the dramatic transition to organized structure (and large amounts of computation) that are bff’s most obvious feature. This is when artificial life *seems* to spontaneously emerge.

But as we can now appreciate, there’s nothing spontaneous about it. Replicators had been there all along, and each larger replicator is composed of smaller ones—an inverted tree of life, consisting of mergers over time rather than splits.

However, evolution’s creative work is not done yet. After the takeoff of a fully functional tape replicator, we often see yet further symbiotic events. From a classical Darwinian standpoint, this seems puzzling, since there should be no reason for further evolution to take place once whole tapes are replicating reliably. How could “fitness” possibly improve further, once a tape copies itself in its entirety every time it interacts with another tape?

We must consider that since the instructions for whole-tape replication don’t occupy all 64 bytes, there’s extra space on the tape that could be dedicated to … anything. That’s the point of von Neumann–style replication—it allows for open-ended evolution precisely because the tape can contain additional information, beyond the code needed for replication itself.[[78]](#footnote-109)

Any extra replicated bytes could, of course, be random—just passive, purposeless information cargo hitchhiking from one generation to the next. But if these bytes contain instructions, those instructions can run. And if they can run, they can replicate *themselves*, too. Thus, the symbiogenetic process can continue to operate, creating additional replicators *within* an already replicating tape. Sometimes these sub-replicators even produce multiple copies of themselves in a single interaction.

Sub-replicators can interact with their host in many ways. They can “kill” the host by damaging *its* replication code, which is generally catastrophic for the sub-replicator, as it thereby destroys the environment within which it can run. Sub-replicators can be neutral, leaving the host’s replication machinery alone. Or, they can be symbiotic, for instance by conferring resistance to mutational damage via redundant copying of the host’s code. The overall tendency is toward symbiosis, since that is the most dynamically stable.

Over time, code colonizes a large proportion of the 64 bytes. Code is more dynamically stable than non-code, and its dynamic stability increases through symbiosis with yet more code—in particular, when code fragments find ways to work in functional tandem.

In a way, symbiosis is the very essence of functionality. When we talk about a kidney’s function only making sense *in context*, we mean that it is in symbiosis with other functions—like those of the liver (breaking ammonia down into urea), the heart (pumping blood), and so on. Each of these functions is *purposive* precisely because its inputs are the outputs of others, its outputs are inputs to others, and thus they form a network of dynamically stable cycles.

The same is true of larger, planetary-scale interrelationships. The “purpose” of plants, from the perspective of animal life, is to produce oxygen and sugar, which we breathe and eat. The “purpose” of animals, from a plant’s perspective, is to turn the oxygen back into carbon dioxide, and provide compost and pollination. Our growing understanding of life as a self-reinforcing dynamical process boils down not to *things*, but to networks of mutually beneficial *relationships*. At every scale, life is an ecology of functions.

Because functions can be expressed computationally, we could also say that life is code, and code is life. Individual computational instructions are the irreducible quanta of life—the minimal replicating set of entities, however immaterial and abstract they may seem, that come together to form bigger, more stable, and more complex replicators, in ever-ascending symbiotic cascades.

In the toy universe of bff, the elementary instructions are the seven special characters “<>+−,[]”. On the primordial sea floor, geothermally-driven chemical reactions that could catalyze further chemical reactions likely played the same role. Under other conditions, on another planet, or in another universe, many different elementary interactions could do the same—as long as they are Turing complete, enabling them to make the leap from autocatalytic sets to true replication.

## Virality

**Although bff is only a toy universe, it can serve as a simplified model for life and evolution, just as a Newtonian billiard-ball universe can serve as a simplified model for the thermodynamics of ideal gases.** As we’ve seen, some of bff’s predictions are quite different from those of classical Darwinian theory:

1. Bff suggests that symbiogenesis is a more important driver of evolutionary innovation than random mutation.
2. Since symbiogenesis must involve combinations of pre-existing dynamically stable entities, we should expect complex replicating entities to emerge after (and be made of) simpler ones.
3. As a result, zooming in on sub-replicators within a larger replicator should allow us to peer back in evolutionary time.[[79]](#footnote-111)

These phenomena should sound familiar! They’re all consistent with Lynn Margulis’s observations, which flew in the face of twentieth-century biological orthodoxy, but are now, if not mainstream, at least gaining respectability.

But wait, there’s more:

1. Due to the instability of the imperfect replicators leading up to the first true replicator, we should expect this first true replicator to be a historical “event horizon,” becoming the template for what follows and erasing independent traces of what came before.

This is what we see on Earth too. Although the first chemical steps toward proto-life may still be taking place in environments like black smokers, we don’t see the missing links. Where are the kind of imperfect replicators that fused together to form the simplest life forms today?

We likely don’t see them on their own because, as in bff, their instability made them ephemeral, quickly displaced or absorbed by the first stably replicating cell—which might already have been recognizably kin to today’s bacteria and archaea. (Evolutionary biologists call this the “Last Universal Common Ancestor” or LUCA.) Still, we can assume that many of the component *parts* of the most ancient surviving life forms—like the reverse Krebs cycle and RNA replication—are fossilized fragments of earlier imperfect replicators.

In the same vein:

1. Evolved code should not only include instructions for replicating itself as a whole, but also be rife with sub-sequences that contain instructions for independently replicating *them*selves.
2. If symbiosis among these parts generated the novelty driving evolution of the whole, we should see evidence in the genome of many “broken” or incomplete sub-replicators.
3. Code that evolved through such hierarchical symbiotic replication should also contain many sequences that are repetitive, or are copies of other parts.

We can find all of these features in our own genetic code—and they don’t correspond to what we would expect to see if our genomes had evolved primarily through mutation and selection.

When the first complete human genome sequence was published in 2000, some surprises were in store.[[80]](#footnote-112) One was the astonishingly high proportion of so-called “junk DNA” that doesn’t code for proteins: about 98%. How did it get there? Is it actually *doing* anything? We now understand that some of this “junk” is involved in gene regulation, but most is still of unknown or uncertain function. Close inspection reveals, though, that much of our code, whether “junk” or otherwise, is … *viral*.

The reproductive cycle of viruses is an often-told story. When traveling between cells, a virus looks like a bit of DNA or RNA packaged into a protein “envelope” capable of binding to a target cell and injecting its genetic payload. Once inside the cell, the viral code hijacks cellular resources to begin replicating, both copying its genetic material and manufacturing its envelope proteins. These assemble into more viruses. Maniacally cranking out viral proteins, the cell may eventually burst, releasing a flood of virus particles to repeat the cycle.

“Retroviruses” like HIV (the Human Immunodeficiency Virus, which causes AIDS) include additional “reverse transcription” machinery for permanently incorporating their genetic material into the cell’s DNA. This makes them a lot harder for the host to clear. The key enzyme, “reverse transcriptase,” wasn’t characterized until 1970,[[81]](#footnote-113) but Barbara McClintock (1902–1992), working at Cold Spring Harbor Lab on Long Island, had made an earlier discovery that turned out to be closely related.

While studying the genetics of maize plants in the late 1940s and early ’50s, McClintock found mobile segments of DNA that seemed to be able to cut and paste themselves into different locations on the genome, producing differently mottled color patterns in corn kernels.[[82]](#footnote-114) Although many of her colleagues were initially skeptical about these “transposons” or “jumping genes,” McClintock eventually won the Nobel Prize for this work.

A much wider zoo of “transposable elements” has since been discovered, some of which don’t just “cut and paste,” but “copy and paste” themselves; they are, in effect, replicators *within* our DNA. In fact, some of them are fully fledged retroviruses, including instructions not only for splicing their code into a host cell’s genome, but also for envelope proteins, allowing that code to venture out into the world in search of other cells to infect.

Retroviruses are unsettlingly … *intimate*. HIV specifically targets immune cells, but if a retrovirus infects an egg or sperm cell, it can insert its code into an organism’s germ line, becoming a permanent part not only of that cell, but of an entire species. In a 2006 publication in the prestigious journal *Nature*, researchers at the University of Queensland in Australia reported catching a retrovirus in this act of becoming “endogenous” for the first time.[[83]](#footnote-115) An epidemic of leukemia in koalas had first been traced to a retrovirus, then this retrovirus was found to have invaded the koala germ line over the previous century, causing the disease to become heritable.

Does this mean the end of koalas? Probably not.

The human genome is rife with signs that the same process has taken place many times in the history of our own species. At least 8% of our genome consists of endogenized retroviruses, the remnants of such retroviral invasions.[[84]](#footnote-116) Remember, 8% is several times more DNA than codes for our “own” genes!

We usually think of viruses as mere disease agents, opportunistically parasitizing our cells to reproduce, since they have no reproductive machinery of their own. The reality may be very different, though. As many of us know from our recent experience with COVID,[[85]](#footnote-117) viruses (and pathogens in general) tend to evolve into forms less lethal to their hosts over time, for obvious reasons—a dead host is a dead end for the pathogen, too. There’s mounting evidence that the relationship between virus and host can go even further, to become symbiotic. Retroviral code in our genome, for instance, has become fundamental to the formation of the placenta, the immune system, cell differentiation, and brain function.[[86]](#footnote-118)

Moreover, endogenous retroviruses are only the tip of an even larger iceberg. Nearly half of our genome consists of transposable elements of one kind or another. Some were probably once retroviruses, or vice versa. Fully sequenced genomes are often rife with lengthy stretches of highly repetitive sequences, from long and complex to mere alternations of two or three symbols—evidence of sub-replicators running amok.

Our genomes, in other words, are not only reproduced as a whole, but include working reproductive sub-sequences at many scales. Some are new and can still cause disease; others are older and may have lost the ability to make viral envelopes or the other machinery needed to become infectious; and yet others have integrated themselves so deeply into our own code that they are no longer distinct. Some of this code is even serving critical functions in our bodies. It’s obvious that, at least by this last stage of the process, complete symbiosis has been established: neither the host nor the sub-replicator could survive, let alone reproduce, on its own.

In his 2009 book *Virolution*,[[87]](#footnote-119) author and physician Frank Ryan goes further, arguing that viruses may have been symbionts all along. Plague viruses like HIV don’t come from nowhere; they are species jumpers. We know that HIV was originally SIV, the Simian Immunodeficiency Virus, variants of which are endemic to Old World primates including African green monkeys, sooty mangabeys, mandrills, and chimpanzees. Yet many of these viruses don’t sicken their original hosts.[[88]](#footnote-120)

We also know that if a virus finds itself inside an organism whose physiology is *too* different from that of its original host, it can’t gain purchase—a lucky thing for you, if you’ve ever swallowed a mouthful of seawater, which likely contained about a billion virus particles! (Most would have targeted single-celled marine life.) The greatest danger seems to come from viruses adapted to a different but closely related species. Perhaps, when it kills, such a virus is doing its job: wiping out rivals who have invaded the original host’s territory.

Viruses could, in other words, work like an out-of-body immune system. Within our bodies, our immune systems seek out and destroy cells that are recognized as “not-us.” Outside our bodies, “our” viruses could be similarly seeking out and destroying whole animals who are recognized as “not-us.” Once a virus becomes endemic to a new population, though—even if it initially kills many—it will differentiate, co-adapt, and perhaps eventually go native.

One could consider bacterial pathogens through a similar lens; hence the well-documented plagues of smallpox that European colonists brought to the Americas, decimating Native populations, and the virulent (though less deadly) syphilis epidemic believed to have been brought back to Europe by Columbus.[[89]](#footnote-121) Unlike bacteria, though, retroviruses don’t just take up residence in our environment or in our bodies, but fuse into our very genomes, becoming inextricably part of us—especially when they alter the germ line.

By sheer volume, our DNA appears to be made primarily of the layered remnants of many such past fusions. It seems clear that transposable elements and “endogenous viral elements” or EVEs have done much more editing of our genome in recent evolutionary history than mutation has.[[90]](#footnote-122) And they have been at this for a long, long time. Based on the best available evidence, viruses are at least as old as the Last Universal Common Ancestor, if not older.[[91]](#footnote-123)

This ongoing ecology of mobile, self-reproducing, and even infectious genes causes the supposed “tree of life” to continue entangling with itself in the oddest ways. Fully a quarter of the cow genome, for instance, consists of copies of the retrotransposon BovB. BovB appears to have leapt many times between species to create a phylogenetic tree of its own, a “bizarre parallel universe where cows are more closely related to snakes than to elephants, and where one gecko is more closely related to horses than to other lizards.”[[92]](#footnote-124) These are instances of “horizontal gene transfer” (HGT), long associated with bacteria, but clearly more universal.[[93]](#footnote-125)

Whether a bit of new DNA is acting in the moment as friend, foe, or somewhere in between is really just a matter of where on the symbiotic trajectory it falls. When brand new, it’s unlikely to be friendly. But if it has persisted for a long time, and especially if it has become endogenous, its dynamic stability implies that it will have become non-lethal or even friendly due to combined evolutionary pressures on host, invader, and “host plus invader” as a “holobiont” or symbiotically fused entity. When some gene or other functionality in the fused code proves valuable to the larger whole, that functionality will be conserved, even as the sub-replicator’s reproductive capability degrades and eventually vanishes. Perhaps that is how new genetic functionality arises in general.

In this light, random point mutation can even be seen as something like a minimal abiogenesis event, creating a tiny parasitic “life form” within the genome. On its own, it would of course be unable to reproduce, since it possesses no independent reproductive machinery;[[94]](#footnote-126) then again, neither does a virus. So, like a virus, a mutation *can* reproduce—using the host’s resources, which will copy it along with everything else. When introduced into the genome of a sophisticated existing organism, such a tiny, random entity is unlikely to be friendly, though, as with any larger invasion, it will either kill its host (and itself, in the bargain) or achieve dynamic stability, either by being neutral or (occasionally) helpful.

Hence, just as replicator thermodynamics encompasses classical thermodynamics as a special case, evolution via symbiogenesis can encompass classical Darwinian theory as a special case. Point mutation is unlikely to be the main driver of evolution once life has taken off because it’s so much weaker and slower on its own than higher-order symbiogenesis. Unlike point mutation, a chunk of code that has already circulated, jumping around in the genome or even between species, isn’t random. It necessarily includes real functionality. And at least in some settings, that functionality has been under evolutionary pressure to help, or at least not kill, its host. Evolution picks up steam over time, with major evolutionary transitions becoming more frequent precisely because increasingly high orders of symbiogenesis become possible.

## Compression

When code evolves through symbiogenesis, it will develop a curious statistical structure: parts of it will be copies (or near-copies) of other parts, and as those parts establish symbiosis, they’ll form a larger aggregate which will *also* copy itself as a unit. This is reminiscent of (though not the same as) a “fractal”: a structure that resembles itself at a cascade of ever-finer scales. Let’s take a short detour through fractals and their relationship to biology so that we can then consider the implications of evolution *itself* exhibiting certain fractal-like properties.

French-American mathematician Benoit Mandelbrot (1924–2010) is the figure most associated with fractals, though they were first described by English meteorologist Lewis Fry Richardson (1881–1953). In the aftermath of World War II, Richardson, a Quaker, had been trying to build a mathematical model to predict the likelihood of conflicts between countries. One of the parameters in his model was the length of their common border. He was perplexed to find, though, that certain borders didn’t appear to have agreed-upon lengths. The Portuguese, for instance, reported that their border with Spain was 613 miles long, but the Spanish claimed that it was 754 miles long. This seemed like a weirdly large discrepancy for what ought to have been an easily surveyed distance!

Richardson eventually figured out what was going on: the Spanish were using a higher-resolution map. Every twist and turn the Portuguese could see in the Minho, Douro, and Guadiana rivers separating the countries had finer-scale twists and turns only visible on the Spanish map. Richardson showed that, if zooming in on any given meandering revealed more meanderings, the total length of a “self-similar” curve like a coastline or a river can grow without limit as one measures it with an ever-smaller ruler.[[95]](#footnote-128)

“Fractal geometry” is thus very different from the Euclidean geometry we all learned in school, where the length of a line or curve is well-defined and doesn’t blow up when you zoom in. Of course, in real life, one can’t zoom in forever, but fractals are still important mathematical tools for understanding structures that are self-similar over a range of scales, like rivers and coastlines.

Living organisms exhibit obvious fractal properties in certain of their structures, such as tree branches, bronchial tubes, and circulatory systems. In structures like these, fractal geometry over some range of scales offers an elegant design solution to a biophysical problem. For instance, blood carrying oxygen and nutrients needs to reach every cell in the body, which means that oxygenated blood has to be pumped through a branching network of tubes, and these tubes—which are, in Euclidean terms, one-dimensional—have to repeatedly subdivide to “fill” three-dimensional space. Hence, we have fractally branching (and then merging) networks of arteries, arterioles, capillaries, venules, and veins.

By working out the design properties of such fractal networks and their implications, theoretical physicist Geoffrey West and his collaborators at the Santa Fe Institute have figured out many fundamental “scaling laws” in biology.[[96]](#footnote-129) One of the most well-known results from this body of work explains why every mammalian heart will beat about 1.5−2 billion times over a lifetime, whether it belongs to an Etruscan shrew (weighing a twentieth of an ounce, and living a year or so), or a blue whale (weighing over 300,000 pounds, and living about a century).

Self-similarity in general yields “power law” scaling relationships, which look like straight lines when plotted on logarithmic axes.[[97]](#footnote-130) Body mass, for instance, has power-law relationships with metabolic rate, the total length of the circulatory system, and lifespan.[[98]](#footnote-131) Combining various power-law relationships to calculate the number of heartbeats in a lifetime causes all the scaling variables to cancel out beautifully, yielding the famous 1.5−2 billion beat constant. While the theory is not exact (humans, for example, live somewhat longer than predicted), it does a remarkable job of parsimoniously explaining a wide range of phenomena.

Is life in general a kind of fractal, then? No. A rhyme by British mathematician Augustus De Morgan[[99]](#footnote-132) is often invoked in describing fractals:

*Great fleas have little fleas upon their backs to bite ’em,*  
*And little fleas have lesser fleas, and so* ad infinitum*.*[[100]](#footnote-133)

The poem is funny precisely because it is absurd: unlike an idealized fractal coastline, living systems are, in general, profoundly *dis*similar across scales.[[101]](#footnote-134) While there are rare instances of the kind De Morgan describes—such as hyperparasitoid wasps, tiny parasites that lay their eggs inside bigger wasps—one does not generally zoom in on a flea to discover a smaller flea, let alone *ad infinitum*.

Indeed, even coastlines are best characterized, in real life, as “multifractals”: systems with repetitive structure at every scale, but variability *across* scales.[[102]](#footnote-135) As a result, their power laws may also vary. For instance, zooming in twofold at one level of magnification, where the ruggedness is especially pronounced, might cause the apparent length to scale up by 21.5≈2.83, while zooming in twofold at a magnification where the coastline appears smoother might only cause the apparent length to scale up by 21.16≈2.24. Every bump has bumps on it, but the bumps at one scale are shaped differently from the bumps at another.[[103]](#footnote-136)

Similarly, although genomes are always replicated, and are made out of other genomes that are themselves replicated, every whole is different from its parts, just as an integrated circuit is different from a transistor. How, then, would one go about quantifying this multifractal-like property in a genome?

The answer is closely related to data compression. Consider bff. In its initial state, the tapes consist of random bytes; no repetitive structure exists. If we were reading through the tapes one byte at a time, the next byte would have an equal probability of assuming any value, from 0 to 255, independent of any of the previous bytes. Therefore, the sequence is entirely unpredictable, which means that if you were to try compressing the soup with file compression software, like ZIP, the compressor wouldn’t be able to squeeze the file size at all. (This is, harkening back to thermodynamics, a working definition of a completely disordered state.)

Data are redundant—that is, compressible—precisely to the degree that they are predictable, or, equivalently, exhibit order. Compression algorithms like those in ZIP work by looking for patterns in the data stream so far and exploiting those patterns to encode subsequent data using as few bits as possible. The usual trick is to incrementally build a dictionary or “phrase book” out of the source material, and to substitute sections of the text for pointers into this dictionary whenever the pointer can be encoded in fewer bits than the original text. The dictionary is, in effect, a predictive model, and since the compressor builds that model incrementally as it compresses, the decompressor is able to incrementally reconstruct an identical copy of the model as it *de*compresses.

Absent any prior knowledge about how to model a long sequence, describing it in terms of previously encountered parts of itself is a good strategy. The full text of this book ZIPs down to about 35% of its original size as a raw text string, since some letters occur more often than others, there are a lot of repeated words, and those words often combine into stock phrases.[[104]](#footnote-137)

When bff undergoes its sharp transition to whole-tape replication, the tapes suddenly become highly compressible; they squeeze down to just 5% of their original size. Interestingly, many regions of the human genome are highly compressible too. The reason these data streams compress so well is that, at any given point in the sequence, one can often make a pretty good guess as to what the next letter will be, simply by looking for the longest previous sequence that matches the current context.

Such matches arise due to replication. In the human genome, we’re not talking about replication at the level of the individual, but, rather, at the level of transposable elements (or their fossil remnants) *within* our genome. Similarly, even a *single* bff tape is highly compressible, because it formed through a symbiotic fusion of smaller replicators … and these replicators are themselves likely to be either copies of each other or related, having formed out of yet smaller imperfect replicators … which are themselves either copies or related.

Like fractals, sequences of symbols constructed by repeatedly copying sub-sequences of all lengths exhibit power-law scaling—in this case, between the lengths of repeated sequences and their frequency. For instance, a replicated sequence twice the length of another might be found half as often—which is pretty often! (These statistics would be dramatically different for random sequences, where increasing the sequence length would cause a much faster *exponential* decline in expected frequency.[[105]](#footnote-138)) Sequences constructed according to a fixed self-copying rule are, in a sense, infinitely compressible, for they contain only as much information as it would take to write down the recipe, no matter how long the sequence.

By contrast, in a complex replicating system like bff—let alone the genome—the *way* any particular set of replicating sub-sequences combine to form a larger replicating sequence will always be specific. It will depend on the functional particulars of how those parts come together to make a working whole. Therefore there is novelty, or information, in each such combination. However, that information will be much less than would be needed to specify an arbitrary symbol sequence of the same length.

When we look for this property in bff and in human genome data, that’s just what we find! Not only do both compress very well; they also compress better and better the longer the stream of data gets. For instance, a single bff tape compresses somewhat, but the whole soup compresses better, and as the size of the soup grows, the compression ratio continues to improve—though no matter how much has already been seen, the remainder is never fully predictable.

Similarly, while a single stretch of a human genome already compresses well, the whole genome compresses better. If you compress it in the statistical context of many other human genomes, so much of the information becomes redundant that a compressed version of *your* genome would be small enough to send as an email attachment.[[106]](#footnote-139) The pattern continues if we zoom yet farther out to consider genomes across species. Hence the famous statistic that the human genome is 98% identical to the chimpanzee genome … and 60% similar to that of the fruit fly!

While common ancestry alone can explain many of the similarities across species (though not all, as BovB illustrates), such classical Darwinian arguments can’t explain why the genome of a single individual is also so repetitive. That only makes sense when we begin to consider genomes to be *made out of* smaller genomes.

DNA and bff tapes, in other words, are both systems that evolve through what we could call “multifractal symbiosis.” Our genomes are aggregates of cooperating replicators, all the way down. That’s why they are so compressible. But it’s also why wonderful new kinds of complexity emerge at every scale—and therefore why, in biology, we can always learn something new by zooming in or out.

## Embodiment

There’s a profound yet subtle relationship between the multifractal properties of our bodies and the multifractal properties of the genome.

The relationship is subtle because the genetic code is by no means a one-to-one blueprint explicitly representing the body’s final shape. One might imagine, for instance, that because a snake has hundreds of ribs, it might have a stretch of DNA coding for a vertebra and rib pair, which might have replicated itself a few hundred times in the snake’s genome. Could snakes be the result of a retrotransposon coding for ribs run wild, like BovB?

Not exactly. So-called “Hox” genes, shared widely among animals, control overall body plan, and they “execute” their program using a combination of gene regulation and the kind of distributed computation Turing described in his work on morphogenesis. During embryonic development, ribs begin as a chemically controlled spatial oscillation, like the stripes of a tiger.

This mechanism is vastly more powerful, general, and evolvable than merely replicating some “build a rib” code. It allows for code reuse, just as a programmer would do: invoking the same “build a rib” function a hundred times instead of copying and pasting it a hundred times. Thus, tweaking rib flexibility or curvature involves making one change, not a hundred.

Evolution can produce real programs that reuse code because life is computational. Remember that if you’re a replicator of the kind von Neumann described, you need to contain instructions for building yourself, including a machine B to copy those instructions, and a machine A, the “universal constructor,” to follow them. The universal constructor is the engine not only of reproduction, but also of growth, development, healing, and maintenance. It created your circulatory system and skin during embryonic development; it also heals your skin and replaces the torn capillaries underneath when you get a cut. As long as you are alive, your body will be continuously constructing and reconstructing itself. The code never stops running.

And since a universal constructor is a kind of Universal Turing Machine, its computations, or equivalently mathematical functions, are inherently “compositional”: complex functions (or programs) can be made out of sub-functions, which can in turn be made out of sub-sub functions … until, at bottom, everything could be expressed in terms of the elementary instructions of a Turing machine.

Functions can even be made out of *themselves*. In computer science, this is known as “recursion.” The Fibonacci sequence, for example—1, 1, 2, 3, 5, 8, 13, 21, and so on, where each number is the sum of the previous two—is most easily expressed as the recursive function *f*(*n*)=*f*(*n*−1)+*f*(*n*−2) for numbers greater than 1, with f(0)=f(1)=1. We find Fibonacci numbers in biology all the time, as in the spiral pattern of a pine cone, the head of a sunflower, and the chambers of a nautilus. This is a clue that recursive computation takes place in the construction of pine trees, sunflowers, and nautiluses.[[107]](#footnote-141)

The smoking gun, though, is the growth of self-similar structures in the body. Compositionality and recursion are what allow genetic code to build a many-tiered fractal like the circulatory system. Code for building ribs or nautilus segments could, in theory, just be copied a couple of hundred times, but that’s not possible for arteries, veins, or capillaries, where the number of segments and branches becomes astronomically large. The blood vessel branching code *must* get reused. And with relatively minor tweaks, it must be possible for the parameters of that code to be adjusted so that the branching stops at the appropriate size for a shrew, or goes on to construct something the size of a blue whale, since shrews and whales are, in the grand scheme of things, close relatives.

From an evolutionary perspective, compositionality implies a hierarchical network of symbiotic relationships. The bacteria that became mitochondria and the archaea that engulfed them each started with the code necessary to reproduce themselves. When they merged into a eukaryote, they still needed to retain the instructions for their own reproduction, but they also needed to evolve additional functionality for mutual regulation. When eukaryotes became multicellular, the same needed to happen again; our cells still know how to reproduce individually, and, indeed, cellular reproduction is a fundamental operation involved in the larger-scale growth and reproduction of a whole animal.

What is true of evolution is also true of development.[[108]](#footnote-142) Blood vessels, for instance, aren’t just Euclidean line segments, but tubes made of layers of smooth muscle cells. Each of those cells contains a complement of organelles, and each mitochondrion in each of those cells contains its own membranes and loop of bacterial DNA. A living organism is a compositional structure *par excellence*. It could only be built computationally, through the composition of many functions. And life could only have evolved as the hierarchical composition of those functions—that is, through symbiogenesis.

## Élan vital

Nowadays, we interact with human-engineered (or, one could say, “artificial”) computers constantly: the phones in our pockets and purses, our laptops and tablets, data centers and AI models. We’ve begun asking whether AI models are intelligent. We could ask an even more jarring question: are computers, whether they’re running AI or not, *alive*?[[109]](#footnote-144)

They are certainly purposive, or we couldn’t talk about them being broken or buggy. But hardware and software are, in general, unable to reproduce, grow, heal, or evolve on their own, because engineers learned long ago that self-modifying code (like bff, or DNA) is hard to understand and debug.[[110]](#footnote-145) Thus, phones don’t make baby phones. Apps don’t write new versions of themselves.

And yet: there are more phones in the world this year than last year; apps acquire new features, become obsolete, and eventually reach end-of-life, replaced by new ones; and AI models are improving from month to month. Electronic components and computer code also exhibit the same kind of compositionality we’ve seen in bff and DNA. It certainly *looks* as if technology is reproducing and evolving! Debating its aliveness is thus a bit like the debate over whether viruses (which also can’t reproduce on their own) are alive.

If we zoom out, though, putting technology and humans in the frame together, we can see that this larger, symbiotic “us” is certainly reproducing, growing, and evolving. The emergence of technology, and the mutually beneficial—if sometimes fraught—relationship between people and tech is nothing more or less than our own most recent major evolutionary transition. Technology, then, is not distinct from nature or biology, but merely its most recent evolutionary layer.

And what about that age-old inanimate stuff—rocks and rivers, mountains and beaches, clouds and storms? Water molecules in themselves are clearly not capable of general computation, and yet, in the context of the hydrologic cycle, clouds, rainstorms, and rivers certainly serve critical ecological functions, and are profoundly shaped by life. Likewise, our planet’s metal and sand get shaped into steam engines and computer chips. All of these are part of that grand network of interdependency we call Earth. Why do we draw boundaries around certain networks of functions and insist that they are “alive,” while the surrounding functions are not?

This way lies vitalism, a view espoused in various forms by a long line of philosophers from Posidonius of Apameia (circa 135–51 BCE, and undoubtedly reflecting a much older tradition) to Henri Bergson (1859–1941). Some modern thinkers, too, defend the vitalist position, such as Jane Bennett:

“The quarantines of matter and life encourage us to ignore the vitality of matter and the lively powers *of* material formations […]. By ‘vitality’ I mean the capacity of things—edibles, commodities, storms, metals—not only to impede or block the will and designs of humans but also to act as quasi agents or forces with trajectories […] or tendencies of their own. [… Our] analyses of political events might change if we gave the force of things more due.”[[111]](#footnote-146)

We resist such ideas because we tend to reserve the notion of agency only for ourselves. The idea of agency in a molecule or a storm, let alone an abstraction like money, seems especially far-fetched. We also tend to think in terms of a hierarchy in which “we” (for whatever value of “we”) are at the top, and agency must surely diminish for anything “lower”—a view reminiscent of the medieval Great Chain of Being. When we (hesitantly) extend “agency” to the nonhuman, we tend to do so only for things that act obviously, individually, and on fast timescales, rather than in the aggregate and on slower, more evolutionary ones. It might be time to re-examine these ideas more holistically.

My purpose here is not to follow in Bennett’s footsteps—though I do find her project worth taking seriously. Language is, necessarily, imprecise, no matter what definitions we pick. This doesn’t mean that words are useless, though. When our language can become more rigorous and scientifically grounded, and when we use it to describe patterns across a wide range of phenomena, we can start to see through ideological thickets.

I hope I have explained both clearly and rigorously how the phenomena that give rise to the complexifying dynamics of life apply much more broadly than to the entities we normally think of as “alive” or “agential.” Accordingly, we could expand our definitions of these existing words, or adopt new ones, or do a bit of each. Personally, I would find some broadening of the old everyday words helpful.

That would hardly break new ground. Many traditional, nominally “prescientific” worldviews embrace notions of aliveness, agency, and even personhood that are far broader than the modern Western ones. This seems a likely case of convergent evolution in languages and ideas, motivated by the common need among traditional societies to take symbiosis seriously to secure their own survival, and to flourish. It’s practical as much as it is spiritual: encouraging richer modeling of agency in “nature” enhances a society’s dynamic stability, since all things in “nature,” ourselves included, are so mutually interdependent.

Thus it can be useful to take the view of an animal, plant, or river at times, even if they can’t take ours, the better to care for them—and for ourselves. That, ultimately, is the best reason to consider adopting, or at least adapting, a more inclusive view of the animate. Potawatomi writer and biologist Robin Wall Kimmerer makes this case eloquently in her book *Braiding Sweetgrass*.[[112]](#footnote-147)

In agreeing with Kimmerer, I am not encouraging superstition. When scientists castigate animist beliefs as superstitious, they typically appeal to the materialist discoveries of the Enlightenment, which show that the atoms that make up our bodies are no different from the atoms that make up rocks or air. This is true. Atoms are atoms; they all obey the same rules. Hence, the Enlightenment model of a clockwork universe, governed by dynamical laws.

Yet as Schrödinger pointed out in 1944, our understanding of these laws—which he played such a central role in developing—remains incomplete. The laws as they stand do not account for the complex, dynamically stable, symbiotic phenomena that comprise so much of our experience on Earth—indeed, without which there would be no such thing as experience at all. There would be no life, or purpose, or minds, or agency.

As we both embrace scientific rigor *and* start to figure out those “‘other laws of physics’ hitherto unknown,”[[113]](#footnote-149) we should perhaps be less surprised to find the shoe on the other foot. What the poet Dylan Thomas called “the force that through the green fuse drives the flower”[[114]](#footnote-150) drives all atoms, not just the ones we presume to be alive.

# II. Survival

## Being in time

We’ve now seen that life is inherently computational, though we’ve only scratched the surface of what it actually computes. The minimal functionality for a dynamically stable life form is, of course, replication, or more precisely, self-construction. A von Neumann constructor, which is necessarily also a full-fledged Turing machine, is required for open-ended evolution, and makes it possible to build complex and recursive structures like a tree’s branches or an animal’s circulatory system.

But the environments in which construction and replication take place look nothing like the pristine grid worlds of von Neumann’s cellular automata. The real world is messy, ever-changing, intrusive. To maintain any kind of integrity over time in such a world, a computational system needs a protective boundary.[[115]](#footnote-153) Yet the boundary can’t be absolute, because life can never be self-sufficient: computation requires energy, and construction requires matter.

Since growth and efficiency are bounded, conservation of matter and energy also require shedding matter (excretion) and energy (waste heat). Like a whirlpool, life is a dynamical pattern than can only exist in the flux of a larger environment.

The sort of computation we call “intelligence” thus arises because of the need for life to interact with its surroundings. Those surroundings invariably include *other* life, opening the door to higher levels of symbiosis, as will be discussed later in the book. In this part, though, we’ll explore intelligence at its most basic: in single player mode.

Imagine a simple, single-celled life form—say, a bacterium. Its cell membrane separates a “self,” on the inside, from the outside world. But as mentioned, the separation can’t be total. Studded with sensors and ion pumps, able to ingest nutrients and excrete wastes, cellular membranes are selectively permeable to matter and energy—therefore, also, to information—just like *our* bodies.

To process that information, a bacterium has something like a “brain.”[[116]](#footnote-154) It consists of a dynamical network of interacting genes and proteins that are both affected by sensory data and can control cellular processes and behaviors. This biochemical “brain” also receives inputs from *within* the cell, such as metabolic status and available energy, allowing internal states like “hunger” to modulate behavior.

Some bacteria can swim, using a corkscrew-like bundle of motorized filaments called a “flagellum”—though steering, in the usual sense, is impossible. Due to the simplicity of the bacterial body plan, the irrelevance of gravity, and the molecular buffeting that characterizes life in water at such small scales, the cell can’t distinguish left from right, or up from down. When swimming, it can only “run,” rotating its flagella clockwise to propel itself forward, or “tumble,” running some of its flagellar motors in reverse and causing the cell as a whole to spin chaotically, randomizing its orientation.

A classic series of studies by biophysicist Howard Berg and colleagues in the 1970s shows how, simply by tumbling more often when food concentration is dropping and less often when it’s rising, bacteria can, on average, swim toward places with higher food concentration—a behavior known as “chemotaxis.”[[117]](#footnote-155) They can use the same approach to swim toward warmth (“thermotaxis”) or, for photosensitive bacteria, toward light (“phototaxis”).

Not that a bacterium can *directly* measure concentration, temperature, or light level, let alone differences in these properties along its microscopic body length. It’s just too small. Relative to every relevant feature of the world it lives in, the bacterium is point-like; one could almost say that it exists only in time, not in space. That is, it experiences life as a sequence of events, and anything it might learn or observe about its spatial environment can only be inferred from that event sequence.

These events are all discrete, or “digital,” even when they relate to continuous physical quantities, such as chemical concentration—because at bacterial scale, chemicals are only sensed one molecule at a time, as they dock and undock with receptors on the cell membrane. Photon absorption and temperature-dependent chemical events are quantized too.

While these discrete events are individually random, the ambient concentration, illumination, or temperature determine the rate at which they occur. Hence, a nominally continuous variable like concentration can only be estimated by counting such molecular docking events within a time window, effectively calculating a moving average. If the bacterium is swimming, and concentration varies over space, then a statistical tradeoff must be made in the estimation process. Counting for longer allows concentration to be estimated more accurately, but at the cost of washing out changes over time (or equivalently, over space). Bacteria are smart about this tradeoff, adapting their concentration estimation strategy to the situation.[[118]](#footnote-156)

Let’s explore how the estimation process works in more detail.

## Batting average

From a bacterium’s perspective, chemical concentration is a running estimate of the likeliness of future encounters with a molecule given a history of past encounters with it. In more technical language, if we call the sequence of molecular encounter events *X*, then the concentration is a time-varying estimate of the probability of *X*, which can be written as *P*(*X*).

If this mathematical notation takes you out of your comfort zone, let me try to make it up to you by offering an analogous situation from outside mine: baseball.

Here’s my naïve understanding of the most famous statistic in sports, the batting average. Every time a baseball player is at the plate, they may either hit the ball, which we could represent as a one, or miss it, which we could represent as a zero. Over time, a player’s batting history could then be represented as a string of ones and zeros, with a single bit produced every time they’re at bat. The proportion of ones—in other words, the *average value* of all those ones and zeroes—is a player’s batting average.

We can study the data more closely, because baseball nerds have put the complete batting histories of Major League Baseball (MLB) players online, going many decades back.

Consider Henry “Hank” Aaron, widely held to be one of the greatest players of all time. During a long career in the Major Leagues, from 1954 to 1976, Aaron stepped up to the plate over twelve thousand times, and 30.5% of the time, he hit the ball. This “.305” average is considered awesome.

No player’s performance is constant over time, though. It’s easy to see this by calculating batting averages for individual seasons (spanning six months every year, from April through September). Now, we can see that over the first few years of Aaron’s career, his batting average went up, peaking at over .350 in 1959. Like anything else, baseball has a learning curve.

But alas, bodies age, and small injuries accumulate. His performance declined, first gradually, then more sharply; by the end of his career he was batting below .250. By that point he was going up to bat only half as often—no doubt, because the manager was keeping a close eye on batting averages, and as Aaron’s declined, he was increasingly sidelined.

We’ve gone from considering career averages to six-month season averages, but nothing prevents us from averaging over shorter windows—three months, two months, or even single months. This finer-grained analysis brings new information into view, up to a point. Using three month windows, we can see that Aaron reached his career peak just a few years before retiring, averaging an amazing .378 between July and September of 1973. (He finished that year with 713 home runs, just one behind the world record set in 1935 by Babe Ruth. Thousands of letters to Aaron poured in, some from fans, others filled with hate and death threats, from those appalled at the prospect of a Black man breaking baseball’s most sacred record.)

Still, over time, not only had Aaron’s season average declined; his performance had also become more uneven. So while the highs remained high, they were interspersed with lower and more unpredictable lows. Hence a short averaging window that would have been highly predictive early in Aaron’s career would have shed less light on the future later on.

Part of the effect is simply due to lack of data. When we average over shorter intervals, or Aaron goes to bat less often, the estimate becomes noisier because there are fewer bits to average. In the limit, if our window shrank all the way to a single at-bat, the “average” could only assume two values, .000 or 1.000, making it useless as a predictor of future performance. Clearly, then, there’s an ideal size of averaging window, large enough to accumulate decent statistics, but small enough to register changes over time. The ideal window size will depend not only on the rate of at-bats, but also on how often a hit occurs, and on how consistent the player is.

All of this is to say: if *X* is a sequence of discrete events in the past, whether hits for a baseball player or food molecule encounters for a bacterium, *P*(*X*) smooths those discrete events into a continuously varying average rate using an appropriately sized averaging window—which is also an estimate of the likelihood of encountering such an event in the immediate future. It is a predictive model. If there have been no events for a while, the likelihood is low; if a flurry of events just started, the likelihood is high; if the rate has been increasing, then a good estimate will be higher than the historical average, and if the rate has been decreasing, a good estimate will be lower.[[119]](#footnote-158)

So, are batting averages “real”? It’s obvious that Hank Aaron’s batting average wasn’t a measurement of some physical quantity out in the world, or in his body or mind—though neither was it unconstrained by the physics of balls and bats, the psychology of pitchers, or the physiology of aging bodies. It didn’t have any one true value, either, in that it could be estimated in many ways, and on many timescales—though some estimates are better than others, in the sense of being more predictive. Despite their fuzziness, though, batting averages matter to players, managers, and fans alike, because they both predict and affect the future.

The same applies to statistical measurements of the kinds that matter to bacteria. Like billiard balls, molecules are discrete objects, subject to microscopic physical laws that don’t in any way involve or require macroscopic concepts like “concentration” or “temperature.” But if a bacterium needs to predict its likelihood of encountering food, it must estimate concentration based on recorded events, not unlike baseball hits.

As described in Part I, Boltzmann’s development of thermodynamics—the Ideal Gas Law, for instance—defines quantities like pressure, temperature, and concentration using mathematical models and approximations of the same kind. These, too, involve counting molecules or events within a time window or spatial box.

Instruments like thermometers or pH strips do the same thing using an experimental apparatus rather than theory. So does a digital camera, by counting the photons absorbed by each pixel. In short, just like batting averages, variables like brightness, temperature, pH, and concentration rely on *models*: ways of predicting future phenomena using past statistics.

## (No) things in themselves

We can either conclude that properties like temperature are not “real”—a path that ultimately leads to the solipsistic denial that *anything* is “real”—or we can reappraise what it *means* for something to be real.

Outside pure math, “reality” can seldom be fully pinned down. We’ve seen how a concept like “temperature” falls apart when we ask about how hot or cold a medium is as its pressure or density drops toward zero, where averaging is no longer reliable. So, if we’re asked a question like “What is the temperature of outer space?” we should counter with more questions to answer meaningfully. Is this about calculating how hot a satellite’s solar panel will get? (Answer: it depends entirely on which way it’s pointed relative to the closest star, and on the shadows of any nearby moons or planets.) Is it about determining how comfortable your hand will feel if exposed to outer space? (Answer: don’t do it.) For a bacterium adapted to a watery medium, these simply aren’t relevant questions. One might as well ask a person, “How does a neutron star taste?”

Like taste, temperature is not a “thing in itself.” It doesn’t pre-exist somehow in an unobserved universe. It is, rather, an observer-dependent model whose usefulness, from the observer’s viewpoint, depends on its behaviorally relevant predictive power. This is true not only of temperature, but of all the macroscopic phenomena we care about and describe using language, like “musical note,” “chair,” “bacterium,” or “person.”

It may even be true of phenomena like “an electron,” but here we run into limits in our understanding of the universe’s fundamental laws. “Electron” may merely be the name we assign to a certain kind of propagating disturbance in a quantum field, or it may have some more fundamentally “digital” nature that physicists investigating the subbasement of reality could one day discover.[[120]](#footnote-160) And are quantum fields themselves “real”? Who knows?

Whether or not electrons, photons, quarks, and the various associated elementary forces *are* “real” and *are* “things in themselves” (whatever that might mean), higher-order objects or features like temperature, pressure, chairs and tables, bacteria and people, seem, at first blush, to be mere ideas, or “epiphenomena.” The unobserved universe doesn’t care about such ideas; after all, the continued existence of a chair doesn’t depend on whether someone is around to conceive of that bunch of atoms as “chair.” Atoms everywhere are governed by the unthinking laws of physics.[[121]](#footnote-161)

The mathematician John Conway’s “Game of Life” offers a useful illustration. It’s not really a game, in that there are no players, no score, no winning or losing. It’s just a set of rules that describe, given the state of the game board, what its *next* state will be. (This should sound familiar. Conway’s Life is, in fact, the most famous example of a cellular automaton, though von Neumann didn’t live to see it.)

The rules—Life’s complete “physics,” or “Theory of Everything”—are simple:

1. Like pixels or graph paper, the world consists of a grid of square cells.[[122]](#footnote-162)
2. Each cell can be “alive” or “dead.” We can visualize these states by coloring them black (alive) or white (dead).
3. At every time step, the state of each cell is determined by the state of that cell and its eight neighbors during the previous time step.
4. If the cell is alive, then it stays alive only if it has either two or three live neighbors.
5. If the cell is dead, then it springs to life if it has exactly three live neighbors.

Under these rules, an entirely dead board will remain dead forever; that is a so-called “fixed point” or “steady-state solution.” As you can easily confirm with pen and paper, a 2×2 block of live cells on an otherwise dead board is also a steady state.

Conway and others have discovered that Life’s simple “physics” can produce many interesting higher-level phenomena, too. For instance, certain configurations of five live neighboring cells form a stable four-state cyclic pattern that appears to move diagonally through space, called a “glider.” A more complex configuration of cells forms a “glider gun,” in which a piston-like reciprocating mechanism pumps out an endless stream of gliders.

You can build a lot of complex machines using a few simple ingredients like gliders and glider guns. In fact, it has been proven that, like von Neumann’s original cellular automata, Life is Turing complete; despite its simplicity, an enterprising engineer can create a general-purpose computer in Life.[[123]](#footnote-163) (In a particularly striking feat of recursive nerddom, the computations required to *simulate* Life have even been programmed *in Life!*[[124]](#footnote-164) This has to be seen to be believed.) It follows that the seemingly trivial universe of Life could carry out arbitrarily complex computations, including, perhaps, a simulation of *our* entire universe.[[125]](#footnote-165)

Interestingly, the moment Life is carrying out real computation, its underlying physics become irrelevant to that computational universe, since, per Turing, all computers are equivalent. In other words, it’s just as valid to say that the computation doesn’t care about the physics as to say that the physics doesn’t care about the computation.

A die-hard physicist, though, can justifiably insist that neither gliders nor any of the infinitely complex things one can build in Life are “real” or “things in themselves,” in the sense that the rules of Life don’t involve any such objects. There are only dead and live cells, and the elementary rules for determining their next states. These facts, and these facts alone, comprise Life’s “Theory of Everything.”[[126]](#footnote-166) Anything further is in the eye of the beholder.

As beholders, though, we can certainly *see* gliders. They *appear* to be real enough. What do we mean by this? We mean that, if we recognize a glider configuration, we can immediately predict what will happen next: it will cycle through four states and move diagonally until it encounters an obstacle.[[127]](#footnote-167) If the glider were the only thing on an otherwise empty, infinite board, we could, with a single glance, predict *every* future state of the entire board, for all time. Compared to brute force computation of the future state of every cell on the board, the effort saved is … well, infinite.

We could say, then, that gliders are *real* only and precisely because, as a concept or model, they are *useful* for predicting the future whenever the glider pattern arises. Recognizing and understanding gliders allows simulation without simulation.

Furthermore, the glider pattern is highly *relevant* in the Life universe, because the pattern is so simple that it tends to arise spontaneously. Meaning: if we imagine Life being played out on a large board where cells occasionally flip at random, gliders will form frequently. And when they do, we will immediately be able to predict their future trajectory—until and unless they are disrupted by noise, or by running into something.

## Anthropic principle

From a mathematician’s point of view, patterns like the glider and glider gun are special in that they correspond to “limit cycles”—generalizations of the idea of steady states (like the 2×2 block) to endlessly repeating *loops* of states. According to quantum field theory, the patterns we call “electrons,” “photons,” and “quarks” are a lot like gliders: stable (though oscillating) solutions to underlying field equations that don’t explicitly define such objects. This, too, should sound familiar: we’ve just encountered dynamic stability again.

Seen this way, gliders and elementary particles are simply the earliest steps in an evolutionary process, according to the maxim that whatever persists exists (and conversely, whatever is too dynamically *un*stable to persist … doesn’t).

There’s no reason for evolution to stop there. If eight gliders collide in just the right way, they will “react” to form a glider gun.[[128]](#footnote-169) And not only is a glider gun stable; it creates more gliders! Glider guns and a few other simple objects that can arise through glider collisions are the basic ingredients for building a Turing Machine in Life. Could there be a route here to abiogenesis in Conway’s Life, just as in bff? In *our* universe, that route involved particles coalescing into hydrogen atoms, the condensation of stars, the creation of increasingly heavy atoms through fusion, the formation of planets, and the subsequent events described at the beginning of Part I.

This may offer a solution to the age-old puzzle of why our universe seems to be so finely tuned for complex life to arise. As physicist Stephen Hawking wrote in *A Brief History of Time*, “The laws of science […] contain many fundamental numbers, like the size of the electric charge of the electron and the ratio of the masses of the proton and the electron. […] The remarkable fact is that the values of these numbers seem to have been very finely adjusted to make possible the development of life.”[[129]](#footnote-170)

For Hawking, and some others in the physics community, our very existence is thus evidence that there are many universes, for if there were only one, its compatibility with life would be a miracle … and physicists don’t believe in miracles. In a “multiverse” where the laws of physics vary between universes, the observation that *our* universe seems finely-tuned to support our existence would simply be due to an observer bias known as the “anthropic principle”: of *course* the universe we happen to observe is finely tuned for life, because nobody *could* be around to observe the vastly greater number of sterile universes. It would be as absurd to claim that we’re “lucky” to live in this universe as it would be to claim that we’re lucky to live on Earth and not in one of the far greater number of other places in the Solar system inhospitable to life.

The Solar system version of the anthropic principle, sometimes called the “weak anthropic principle,” is certainly correct. There are no poor “unlucky” people on Mars because there are no people on Mars, period. It is, however, quite a leap to go from a puzzling observation about the life-friendly fine-tuning of physical laws to the “strong anthropic principle,” holding that any and all conceivable laws of physics must hold *somewhere*. Physicists should be uncomfortable with miracles, but they should be equally uncomfortable with theories that give up on the central premise of their entire field— that truly universal laws of nature exist at all. Is a multiverse where anything goes really required to account for our own existence?

A more economical explanation is that the fundamental rules of our (one) universe, like those of bff and Conway’s Life, allow for computation—evidently a much lower bar than allowing for life as we know it.[[130]](#footnote-171) Given a noise source, the simple logic of dynamic stability will select for stable entities, which can then start to combine into progressively higher-order dynamically stable entities: quarks into nucleons, electrons and nucleons into atoms, atoms into molecules, and so on. Entities like these can be understood as “proto-replicators,” subject to the tendency toward increasing complexification and computation described in Part I.

If the rules had been different, the “proto-replicators” would be different. Indeed, even with the same rules, each evolutionary stage offers an expanding menu of possibilities. There’s probably nothing uniquely determined about DNA as an information carrier, or our exclusive reliance on right-handed sugars. Regardless of such “just so” choices, the end result would be the same: replication, computation, and life.

Darwin may have been right in yet another way, then, when he wrote of abiogenesis that “one might as well think of origin of matter.” He was being facetious, but—why not? Matter might have evolved too.

## The *Umwelt* within

In describing Conway’s Life, we’ve allowed ourselves a third-person, God’s-eye view of a toy universe. But let’s now return to the bacterium living in our own world, pointlike relative to its vast watery environment, yet on the inside, full of complex molecular machinery; technically brainless, but not unintelligent.

We’ve seen that the bacterial “brain” implements an adaptive algorithm for estimating chemical concentration as it swims, and it seems natural to call this measurement “purposive,” since it certainly *looks* like the goal is to eat and to avoid toxins—in short, to survive. I’ve argued that the very notion of “concentration” is something the bacterium appears to construct in the service of that purpose; chemical concentration is, to use pioneering biologist Jakob von Uexküll’s word, part of the bacterium’s *Umwelt*: its “universe of the meaningful.”

Later in the book, we’ll explore the powerful social aspects of the way predictive models affect the future, but for now, let’s stick to single-player mode. It’s not hard to see why a bacterium would care about its “batting average,” the rate at which it encounters molecules it can eat.

How much that rate can vary, and how low it can go, will of course depend on how depleted its reserves are. That’s why part of the bacterium’s *Umwelt* is also its *internal* state; let’s call that state *H*, for “hidden.” In general, *H* includes a “comfort zone,” surrounded by a “danger zone”; beyond the danger zone lies death.

In the simplest case, if we suppose that *H* is an amount of available food inside the cell, estimating the “health-o-meter” *P*(*H*) will look like the same kind of smoothing process as that used to estimate *P*(*X*), but now based on discrete internal metabolic events *H* rather than measurements of the external environment *X*. An organism’s “job #1” is to keep its internal estimate of *P*(*H*) in the comfort zone: this is “homeostasis.”

The bacterium does this via actions, which we could also represent as a set of discrete motor events *O* (for “output”). These actions may be visible from outside—such as reversing the flagella to tumble—or they may be internal, such as turning on or off a gene or metabolic pathway.[[131]](#footnote-173) Either way, maintaining homeostasis is all about performing actions *O* given external observations *X* that maintain *H* within the comfort zone. This is intelligence in its most rudimentary form.

It’s common to suppose that something like “reinforcement learning” is at work here, allowing the organism to learn to perform any needed actions based on positive (“feeling more comfortable”) or negative (“feeling less comfortable”) feedback.[[132]](#footnote-174) However, the whole idea of grounding the emergence of intelligence in reinforcement learning—or supervised learning of any kind—implies an oracle that can administer rewards and punishments, or give correct answers. Where did this oracle come from? How did *it* become intelligent? Maybe there *is* no such oracle!

A simpler and more general way of thinking about intelligence is as “unsupervised” learning (or, as it’s sometimes called now, *self-*supervised learning) of the combined or “joint” probability *P*(*X*,*H*,*O*). A bacterium that learns this joint probability distribution will not only “know” how to estimate nutrient concentration and its own internal state, but also, within some operating envelope, “know” what the consequences would be of acting in various ways under different internal and external circumstances.[[133]](#footnote-175)

Critically, this joint distribution is a prediction of *self* as well as environment—and of the consequences for oneself of various possible actions. The joint distribution includes all of the below “knowledge”:

1. How to tell if you’re hungry;
2. How to measure food concentration;
3. How your next actions are likely to affect the future food concentration; and
4. How much less hungry you will be when you get that food.

Thus, the joint distribution contains everything one might want to “know” about how to behave to stay alive.

Going forward, I’ll drop the scare quotes around “know” and “knowledge.” I put them there for two reasons. One reason is deep, and the other, less so. The deep reason has to do with the distinction between a passive, abstract model and a real agent, whose predictions include actions that are actually taken, and determine the future. This will be taken up in more detail shortly.

But first, let’s dispense with the shallower reason for the scare quotes: the distinction many cognitive scientists draw between knowledge and competence. The phrase “competence without comprehension” has often been used to describe the way an agent can act *as if* it has knowledge without, apparently, *having* that knowledge.[[134]](#footnote-176) Never mind bacteria; a human baseball player or race car driver has amazing physical intuition in their domain of expertise, yet most baseball players wouldn’t be able to explain or reason about parabolic trajectories in general, and most race car drivers don’t know the equations for friction or centripetal force. So, do they “know” physics, or not?

This sounds like a more profound question than it really is. Math, physics, and for that matter the ability to explain things using language are all skills or competencies in their own right; none of them follows from learning how to swing bats, catch balls, or drive race cars—or vice versa. Plenty of nerds know physics well enough, but are so clueless about sports that they need to look up what “batting average” means on Wikipedia. (Thank you, Wikipedia.) And of course symbolic math and language are far out of reach for an agent with the very limited computational capacity of a bacterium, even when such an agent’s learned competencies approach mathematical perfection in some practical domain.

As we’ll see when we explore large language models in Part VIII of the book, symbolic skills and capabilities, like language and math, can also be fully represented as joint probability distributions—albeit enormous, complicated ones. Most of us who both know physics *and* have an intuitive understanding of how to drive or swing a bat don’t connect the two domains much, but for some people, mathematical thinking and those physical intuitions *might* be closely associated, just as the sound of a voice and the motion of a speaker’s lips are; either way, it’s all in the big overall joint distribution *P*(*X*,*H*,*O*).

## Latent variables

Learning *P*(*X*,*H*,*O*) efficiently—that is, successfully representing it without exhaustively memorizing the probability of every individual combination of circumstance, state, and action—requires data compression. Without such compression, the size of the model, even for a simple organism like a bacterium, would be unwieldy; it would then both be very hard to learn and very expensive to store, or evaluate.

As we’ve already seen, data compression involves finding patterns in data and factoring them out. For symbolic data, which are usually compressed “losslessly” by algorithms like those used for ZIPping files, such patterns generally take the form of repeated sequences. When the data are instead continuous and the compression is “lossy,” like that of MP3 audio or a JPEG image, finding those patterns allows irrelevant details to be ignored, exposing a smaller number of meaningful “latent variables.”[[135]](#footnote-178)

For the bacterium, the concentrations of molecules that bind to receptors are important latent variables. To see why, imagine that there are a hundred receptors on the bacterium that bind to some chemical of interest, like aspartate (for *E. coli*, that’s a “yum”). Each receptor can be occupied or unoccupied by an aspartate molecule. Let’s suppose (I’m just making up numbers here) that the occupancy time is a millionth of a second, and the bacterium might encounter between zero and a thousand aspartate molecules per second as it swims. Since the state of each of those receptors can be thought of as a binary digit, 0 for unoccupied and 1 for occupied, these hundred receptors produce a raw information stream that amounts to a hundred million bits of information per second. That’s a *lot* of bits per second, around ten times what it takes to transmit 4K video.[[136]](#footnote-179)

Luckily for a bacterium of little brain, the overwhelming majority of this information is either useless or redundant. Since the cell is so small and no meaningful concentration differences exist along its length, it doesn’t matter *which* receptor gets occupied—the receptors are “symmetric,” making any meaning extractable from them is “invariant” to this symmetry. So we might as well just add all of these bits together.

Even when added up, the total is almost sure to be either 0 or 1 at any given time, because the occupancy time is short and the concentration is low. That means we’re already down to a mere *one* million bits per second. Then, when we consider that the exact timing of the docking events doesn’t matter—meaning that there is “local time symmetry,” or invariance to the precise timing of the docking events—we realize that we can just add up all of the ones each second, giving us concentration as an average occupancy rate over a one second time window.

So, we’re back to averaging over a time window, as for calculating a batting average. That single number per second, which can be represented with only a few bits, accounts for the great majority of the real information coming in through those hundred receptors, though a little more can be squeezed out by making the averaging window adaptive, as mentioned earlier.

Evolution will have honed in on something close to this compression scheme, because it sieves the useful information out, while discarding everything else. Such compression makes the overall distribution *P*(*X*,*H*,*O*) *learnable*, because the patterns relating *X* to *H* and *O* can only be generalized if the welter of irrelevant detail in *X* has been cleared away. In the language of information theory, the useful parts of *X* are the “signal,” while the discarded parts are the “noise.”

Remember, this signal is useful precisely because it’s a continuous estimate of the likelihood of a *future* docking event—though note that this continuous estimate is itself represented as discrete, stochastic signaling events within the cell, so in calling it a continuous number, we’re doing some modeling *of the model*. But we’re getting ahead of ourselves. The point is that we’ve arrived at a practical definition of chemical concentration purely by thinking about what is relevant to the bacterium’s future given its past. Concentration is real because it’s predictive, just as the temperature of an oven is real for you because knowing it will allow you to predict whether your food will cook in there, or your finger will burn if you touch it.

Hunger is much the same. Just as the exact sequence of molecular encounters that led to a given estimate of the concentration is irrelevant, the exact sequence of energy-consuming actions that lead to a depleted energy state is irrelevant; what matters is that death will occur without more food. Hence hunger is “real” in exactly the same sense—and for the same reason—that concentration is “real”: both are useful latent variables for efficiently modeling *P*(*X*,*H*,*O*). That one of these latent variables is an estimate of conditions *outside* the cell, while the other is an estimate of conditions *inside* the cell will become interesting as we consider social relationships *between* cells, but as far as a solitary bacterium making its way in the world is concerned, there is no difference.

## Model (v.)

In the discussion so far, I’ve argued that modeling a single joint probability distribution, *P*(*X*,*H*,*O*), is enough to establish the reality of both regular features of the external world (like temperature, concentration, and brightness) and of the internal life of an organism (like hunger). I haven’t explained how such models arise in the first place, or why organisms with models appear to be purposive.

I’ve also been deliberately vague about the word “model,” shifting its usage back and forth between noun and verb. As a noun, we often think of a model as something disembodied, perhaps an algorithm or set of equations represented in abstract form. A one-dimensional linear model, for example, can be written *f*(*x*)=*wx*, and is fully specified by the single number *w*.[[137]](#footnote-181) The billions of parameters in a large machine-learned model are no different—they just take much more space to write down. As a verb, “modeling” may mean *learning* or approximating *w* (or equivalently, billions of parameters, often expressed as a giant matrix of “weights” *wij* between artificial “neurons” *i* and *j*) from training data; or it might mean *evaluating* the resulting model, that is, actually calculating *f*(*x*) for a given *x*.

A living organism is clearly more “verb” than “noun”; “verb-ness” or, to use a more resonant term, “agency,” is precisely what distinguishes living things from dead things. Living things *do* stuff of their own accord; dead things don’t.[[138]](#footnote-182) A bunch of parameters, printed out on a giant ream of paper and shoved into a closet never to be seen again, are certainly not a living thing. They are just inert matter … dead information. (Much like the reams of DNA still coiled inside the cells of an animal after it has died.) Even to call those numbers information may be overstating the case, since if no reader exists—no active process making use of those printed numbers to *do* something—they might as well be gibberish, or blank. It would make no difference. And information is, as the anthropologist and philosopher Gregory Bateson memorably put it, “a difference that makes a difference.”[[139]](#footnote-183)

The noun and verb senses of “model” differ from each other in much the same way a mathematician’s definition of the word “function” differs from that of a practical programmer. For a mathematician, a function is an abstraction; it could be unknowable in practice (like the oddness or evenness of the number of pennies in circulation at any given moment) or even fundamentally non-computable (like whether an arbitrary mathematical statement is correct[[140]](#footnote-184)). Interesting as they may be in theory, for a programmer, such “functions” are nonsense. Programmers only care about functions that can be implemented by running actual code that *can* function, or that, if buggy, can *fail* to function.

Relatedly, remember that the *O* in *P*(*X*,*H*,*O*) stands for “output.” Life must *act* (yet another word for “function”). A living thing is a process, not a printout or a mathematical abstraction; so these actions have to be taken—or the process is disconnected from the world, and causally speaking, the model may as well not exist.

In the so-called “autoregressive sequence prediction” setting used by many unsupervised AI models today, the action with maximum probability according to *P*(*X*,*H*,*O*) is generally taken. At the next moment in time, the action taken, along with any new observations of *X* and *H*, becomes part of the past, and the model is evaluated again. Hence every action depends not only on past observations, but also on past actions. An organism is therefore *modeling itself* jointly with the environment, and in so doing, is carrying out agential or “verb-ish” modeling. It’s running its own program in a continuous loop.

The reason I’ve expressed the program as a probability distribution rather than in the more obvious way a programmer might, as a deterministic function mapping input to output, *o*=*f*(*x*,*h*), is twofold.[[141]](#footnote-185) First, as we’ve seen, biological computing is stochastic, so actions need to be thought about in terms of probabilities, not certainties. That’s true of AI models, too; with nonzero temperature, as generally used in chatbots, the action taken is drawn from a probability distribution, rather than being fully determined by the inputs. Second, those actions are not just outputs, but also *inputs* at the next moment in time.

There’s a “strange loop” at work here.[[142]](#footnote-186) Because the model is joint, all senses are active: not only do actions taken depend on interior and exterior percepts, but all percepts also depend on actions. In our chemotaxis example, for instance, any perceived change in nutrient concentration, and its interpretation, will depend on whether the bacterium is swimming forward or tumbling in place—which itself depends on perceived nutrient concentration. Further, the model is always modeling *itself*, in the sense that it predicts its own future states and actions in addition to predicting its environment. We could even say that it predicts its own predictions, in a kind of infinite regress.

Life is a hall of mirrors.

## Learning by evolving

The active modeling described above doesn’t seem to leave any room for *learning*—only for doing, or in mathematical language, function evaluation.[[143]](#footnote-188) But where does the function constantly being evaluated, *P*(*X*,*H*,*O*), come from in the first place? This question sounds a lot like the puzzle of abiogenesis, and that’s no coincidence.

As far back as 1948, Turing realized that evolution could be understood as a trial-and-error learning process.[[144]](#footnote-189) It may seem crude and inefficient, but it is the necessary bootstrap to any more sophisticated kind of learning. Everything that learns either learned to learn by evolving, or was designed by entities (so far, human) who *themselves* learned to learn by evolving.[[145]](#footnote-190)

Suppose we have a world in which bacteria have already somehow arisen; and that, if they survive to do so, they can reproduce, in the process “cloning” their *P*(*X*,*H*,*O*) models to pass on to daughter cells. We are assuming, in other words, that *P*(*X*,*H*,*O*) is encoded genetically. After all, if *P*(*X*,*H*,*O*) is a computable function, it can be implemented with code, and since the bacterium is a “self-reproducing automaton” as von Neumann described, it already comes with a Universal Turing Machine in the box. The bacterium’s computing resources will be limited, and certain functions will be easier to express on that biological platform than others, but any kind of computation is theoretically possible!

The code is also, simply by virtue of being genetically encoded, open-endedly evolvable. Not only can it improve via tedious, old-fashioned, million-monkeys-on-a-million-typewriters point mutation; it can improve far more efficiently through symbiosis, as described in Part I. Maybe, then, evolution, in its full self-modifying glory, *isn’t* such a slow learning algorithm.[[146]](#footnote-191)

What, then, makes for a “fitter” *P*(*X*,*H*,*O*)? In short, one that is more dynamically stable. As a first approximation:

* If a high probability is assigned to being dead, that is, to *H* going beyond the “danger zone” and into the “dead zone,” then in a self-fulfilling way, the model will predict—and take—actions that result in its own death. A fitter model, though, will *not* suicide.[[147]](#footnote-192) It will assign the highest probabilities to the “comfort zone,” with declining probabilities outside this zone.
* If the model isn’t accurate, consistent, or suitably general—failing to predict death, or account for the relationships between actions and states, or its own future actions, including under novel conditions—then it will be less fit than a model that *is* accurate, consistent, and general, hence good at avoiding states leading to death.
* If there are signals available, whether internal (*H*) or external (*X*) that can help with the above predictions, then a model able to avail itself of these signals will be fitter than a model lacking them.
* Since modeling (the verb) is computationally costly (hence energetically costly), a smaller, leaner model that predicts the future just as well as a bigger one will win out, because it will starve less often and leave more resources available for reproduction.
* As a bonus, model compression tends to go hand in hand with generality, meaning that, all things being equal, a leaner model is also likelier to continue to perform well under novel conditions.[[148]](#footnote-193)

In a resource-constrained environment, then, organisms with incrementally better *P*(*X*,*H*,*O*) models will outcompete those with worse models, and over many generations, the modeling will get very good indeed.

## Cause by effect

The term “good,” applied to a model, can be interpreted in multiple ways. For data scientists, a “good” model usually means one with low error on held-out test data, or that otherwise performs well on some predefined task. This measure of goodness applies to *a model* (the noun), as opposed to a dynamic agent that *models* (the verb). Measuring goodness does involve computational evaluation—that is, resetting the model and bringing it “to life” very briefly, and many times, to compute its response to test stimuli. But there’s generally no notion of time or agency in such an artificial setting, and “goodness” is whatever the data scientist says it is. This is the highly artificial setting of “supervised learning.”

On the other hand, a model that is “good” in evolutionary terms is simply one that works, or at least works well enough, meaning that it effectively predicts the future. In doing so agentially, it tends to bring that future about: a self-fulfilling paradox of backward causality.

The apparent paradox arises because causes are supposed to precede their effects. How could it ever be right to say that something happens *because of* its effect?

Yet colloquially, we do this all the time. Suppose, for instance, that I’m talking with a friend at the coffee shop, and the barista sees me glance at the clock, jump up, and race out the door, my coffee unfinished. She asks the friend why I’m careening off on my bike. The friend says, “oh, it’s because he’s running late for gym again.”

This is as commonplace as an exchange can get, but if you think about it, it’s quite a strange use of the word “because.” If a boulder is careening down a hillside, about to squash a parked car, and someone asks *why* the boulder is doing that, there might be all kinds of answers. Maybe there was a landslide higher up, or it broke off a cliff. We don’t say it’s careening down *because* there’s a parked car down at the bottom. Causes are uphill, in the past, and effects are downhill, in the future.

In this light, an effect preceding its own cause seems nonsensical. Yet when it comes to bacteria, people, and everything in between, the future, or a wishful model of it (one in which you are safe, warm, well-fed, caffeinated, and make it to the gym *just* in time), seems to influence the events leading up to it, rather than those events inexorably determining the future.[[149]](#footnote-195) In fact, evolutionarily speaking, only creatures with models that successfully predict their own future even *have* a future!

Of course none of this happens in violation of the laws of physics—but, like life itself (per Schrödinger), it certainly *looks* suspicious. In a deterministic universe, how could a “plan” or “intention” possibly affect what will just happen anyway? (If you’re thinking that this sounds a lot like the paradox of free will: yes, it does. We’ll get to that in Part VI.)

The answer comes from that strange loop inherent to autoregressive sequence modeling. We’ve seen why the evolution of life implies the evolution of a “sufficiently good” *P*(*X*,*H*,*O*) model. That model gives rise to actions in the world, but the world also gives rise to the model. While the model can only have learned from the past, it *is* a model by virtue of being able to predict the future. When its actions are informed by that prediction, we get something that looks like backward causation. The more powerful the model—or, we could say, the more “intelligent”—the more robust this effect will be.

Cognitive scientist Terrence Deacon calls systems exhibiting apparent backward causality “entensional.”[[150]](#footnote-196) We could equally call them “agential,” “purposive,” or in a broad sense, “alive.”[[151]](#footnote-197) Hence for the boulder rolling downhill, the car at the bottom could only enter into the “why” of things if someone or something rolled that boulder on purpose.

By this point, we’ve considered two senses in which a model can be called “good.” For a data scientist, a model is “good” to the degree that it scores well on a test. For a living organism, a model is “good” to the degree that it successfully predicts the future—of the world, of the self, and of the effects of one’s actions on self and world.

A third sense of “good,” connecting these two, is normative, as opposed to descriptive: a model is good if it keeps the organism alive, because being alive is good, and being dead is bad. Notice that this normative quality arises in the evolutionary setting without anything like reinforcement learning, wherein an external oracle or teacher metes out rewards and punishments. In fact, “punishment” of an individual organism under Darwinian conditions is impossible, since the only negative feedback is death, which will prevent *any* subsequent learning! Rather, a taste for aliveness emerges from something like a tautology: dead things don’t model; alive things do; their models successfully predict their own aliveness; *ergo*, it is “good.”

## Goodness and truth

While death itself is nothingness, states *adjacent* to death will be aversive to any creature that has evolved with death acting as what physicists call an “absorbing boundary condition.” Being averse to death, in other words, is how not to die. Aversion results in actions that are likely to move the creature away from a “bad” state and toward a better one, more conducive to survival. This normative judgment, and the response to it, are the same for a bacterium swimming away from a toxic chemical, a beetle fleeing a stomping foot, a rabbit evading a wolf, or a human refugee escaping a civil war.

As we’ll explore in Part IV, the moment animals become multicellular, they need to internalize this kind of normative signaling. They do so using ancient neuromodulatory chemicals like dopamine and serotonin, which first evolved in flatworms, and remain the basis of our own most elemental feelings, such as desire and satiety. With our big brains and complex cultures, human ideas about what is “good” or “bad” have become manyfold and nuanced, but at heart, they all trace back to these evolutionary fundamentals.

An account of modeling in general as normative, foundational to life, and arising from computation in living systems troubles some still-influential ideas in Western philosophy. One is David Hume’s “is/ought” distinction,[[152]](#footnote-199) which seeks to distinguish descriptive from normative statements. Hume privileged description, which was a progressive idea at the time. He held that descriptive statements ought to take precedence over normative judgments—and precede them, or we wouldn’t find it so irritating when people express opinions about topics they don’t understand. For Hume, facts (“is” statements) are universal, while opinions (“ought” statements) are merely personal.

While intuitive, the is/ought dichotomy falls apart when we realize that models are not just inert matrices of numbers, or Platonic ideas floating around in a sterile universe. Models are functions computed by living beings, and arguably *define* living beings. As such, they are always purposive, inherent to an active observer. Observers are not disinterested parties. Every “is” has an ineradicable “oughtness” about it. Our *Umwelten* are what they are for good reasons—the wavelengths we see or don’t, the categories we distinguish or don’t, the variables we deem salient or not.

Summing up:

1. All knowledge or understanding of the world is observer-dependent;
2. “Accurate” is best thought of as “useful for predicting the future”; and
3. “Useful” can be read as “good for someone.”

Philosophical “idealists” like Immanuel Kant (1724–1804) have made similar arguments, as have more recent postmodern philosophers and critical theorists, who profess to be skeptical about the very possibility of observer-independent “is” statements. They are right, in a way, but their solipsism isn’t warranted. We can mostly agree on a shared or “objective” reality, because we all live in the same universe. Within-species, our *Umwelten*, and thus our models—especially of the more physical aspects of the world around us—are all virtually identical, statistically speaking. Merely by being alive and interacting with one another, we (mostly) agree to agree.

The rarity of instances like The Dress, an internet phenomenon that went viral in 2015, is telling. Some people perceive it as white and gold, others as blue and black. This enigmatic photo is only ambiguous because it was taken under weird lighting conditions; confronted with the actual article of clothing, everyone agrees that it’s really… whichever it is, I can never remember. How we perceive it turns out to depend largely on our prior expectations with respect to lighting. Early risers see white and gold, while night owls (like me) see blue and black.[[153]](#footnote-200)

Even *between* species, both a common biological toolkit and interlocking *Umwelten* tend to reinforce a consensus reality. Our interests, hence our models, are all entangled, whether cooperatively or adversarially. Indeed, it can be hard to tell the difference. A rabbit that fails to effectively recognize and distinguish between fellow rabbits and wolves is not long for this world. Neither is a wolf who can’t recognize dinner. Over evolutionary time, through this apparently adversarial interaction, wolves and rabbits have *cooperated* to mutually shape each other’s highly capable bodies and sophisticated world models.

The science fiction writer Philip K. Dick once wrote, “Reality is that which, when you stop believing in it, doesn’t go away.”[[154]](#footnote-201) While an entity may ignore or variously interpret many of its input signals, failure to understand the relationship between those signals and its own future existence will result in future non-existence. At a minimum, then, facts are models that work well enough not to kill you.

Consensus is easier to achieve for *X* than for *H*, though. Two agents who (perhaps by virtue of both being human) have similar sensory apparatus, similar survival imperatives, share a common language, and are looking at the same dress, will in an overwhelming majority of cases agree on its color. On the other hand, a hidden state *H* by definition is not directly accessible to an *other*. I can *say* that I’m hungry or tired or feel sad, but … you’ll just have to take my word for it. Or not. For under certain circumstances, it can be important for hidden states to remain hidden.

## Are feelings real?

The most rudimentary kinds of feelings *are* little more than physical signals: hunger, cold, heat, thirst. For any organism that seeks to stay alive—that is, for anything that *is* alive, as remaining so requires constant work—gauges and meters like these matter. Is it time to sweat, or to shiver? Pick the wrong answer, and your model might not be around to make any more janky predictions tomorrow.

For a bacterium, such gauges are chemical measurements, just like environmental variables; they’re simply assessing conditions on the inside of the cell membrane rather than the outside. That implies solving the same kind of computational problem as for an external measurement. Finding the correlation between a raw signal (like molecular docking and undocking events) and an estimated property (like concentration, or hunger) requires a learned, and likely adaptive, model. As organisms get more complex, both the signaling mechanism and the model required to infer a variable like “hunger” become vastly more complex.

Consider pain. In his 2023 book *The Experience Machine: How Our Minds Predict and Shape Reality*,[[155]](#footnote-203) philosopher Andy Clark—whose take on the larger subject of brains as predictors I wholeheartedly agree with—describes a number of telling ways our pain models can malfunction. Amputees, for instance, can experience phantom limb pain. Similarly, many people live with chronic pain disorders, often beginning with an acute injury or illness but persisting long after the physical damage has healed.

On the other hand, many of us are familiar with receiving injuries, sometimes serious, which don’t hurt at first—the pain only comes later, if at all, and sometimes seems to depend on our higher-level awareness of the injury. To cite one gruesome example: in early 2005, a construction worker in Colorado experienced what he perceived as a bruising blow from a recoiling nail gun at a job site. Six days later, he visited the dentist, complaining of a mild toothache; he had been icing it and taking Advil. An X-ray revealed that he had fired a four inch nail through the roof of his mouth and into his brain. He recovered well following surgery to remove the nail, but this tabloid story doesn’t inspire me to take on more handyman stuff around the house.[[156]](#footnote-204)

There’s an opposite phenomenon too, in which intense pain accompanies a *mistaken* perception of injury. Consider another story about an unfortunate encounter between a construction worker and a nail, as reported in the *British Medical Journal* in 1995.[[157]](#footnote-205) This worker jumped from some scaffolding onto a plank, failing to notice the seven-inch nail projecting up from it, which pierced clean through his boot, coming out the top. In agony, the construction worker was dosed with fentanyl and midazolam. But in the emergency room, when the doctors pulled out the nail and removed the boot, they found that the nail had passed harmlessly between his toes.

As Andy Clark puts it, “Such effects seem much less surprising once we accept that […] ‘raw’ sensory evidence is […] never experienced. […] [R]esponses of the ‘pain receptors’ (known as nociceptors) are not what we feel when we’re gripped by a sharp pain. Instead, those responses are simply one source of evidence […]. That’s why we can genuinely feel pain even when nociceptor activity is absent. We can also fail to feel pain even when intense nociceptor activity is present […]. What we feel […] reflects a process of unconscious inference […] about […] the events causing our sensory stimulations.”[[158]](#footnote-206)

Clark’s point is general: it’s not just about pain, but about the nature of reality, both external and internal, and how that “reality sausage” is made. Even beyond Hume’s “is/ought” distinction, Classical and Enlightenment dogma that the way things “really are” can be cleanly separated from psychology (our “beliefs”) is mistaken. The second construction worker was not faking pain, any more than the first construction worker was pretending *not* to be in much pain.

Neither Clark nor I would claim that reality doesn’t exist. Rather, reality is the term we use for a purposive model with a high degree of social consensus and good predictive power. We call the first construction worker’s belief that he was uninjured “false” because that belief would not have predicted the dramatic X-ray, or the urgent need for surgery. We call the second construction worker’s belief in his injury “false” because his belief would have predicted that, when the doctors removed his boot, they would find a bloody wound. When there was no such wound, everyone realized that their model needed retroactive updating, and as a bonus, the man’s pain … went away.

Unfortunately, it doesn’t always work so neatly. For people with chronic pain, knowing that there’s nothing “wrong” doesn’t necessarily bring relief—just as for people with arachnophobia, conscious knowledge that a pet tarantula is tame doesn’t make the dread of touching that giant hairy spider go away.

# Interlude: The Prehistory of Computing

Computing may be fundamental to life, intelligence, and everything, but its history is tortuous, full of false starts and misunderstandings. To better understand our conceptions and misconceptions about computing, we’ll need to begin well before Turing and von Neumann, connecting their work to its roots in mathematics, industrial engineering, and neuroscience. We’ll also need to take stock of the social context surrounding its development.

This won’t be a definitive or comprehensive account, but a short and curated one. Its goal is to reassess our received wisdom about what computing is (or isn’t), and its relationship to brains and minds, labor, intelligence, and “rationality.”

Our prehistory of computing begins[[159]](#footnote-209) during the European Enlightenment, with Gottfried Wilhelm Leibniz (1646–1716), co-inventor of calculus, among much else. His “stepped reckoning machine” was the first mechanical calculator that could perform all four basic arithmetic operations, but this was only a baby step in a far more ambitious agenda. At age 20, Leibniz had asserted that one day, we’d be able to formulate and answer *any* question—not just in math, but in politics, economics, philosophy, even ethics or religion—using pure logic:

“If controversies were to arise, there would be no more need of disputation between two philosophers than between two accountants. For it would suffice to take their pencils in their hands, to sit down with their slates and say to each other […]: Let us calculate [*calculemus*].”[[160]](#footnote-210)

In imagining that accountants—human computers, really—could work out the right answer to any question once properly formulated, Leibniz presumed the existence of universal truth, both factual and normative: a Platonic world of pure, timeless, and axiomatically correct ideas. All it would take was to devise—or discover—a formal language for expressing any proposition symbolically (he called this a *characteristica universalis*, or “universal notation”), and an algebra for manipulating such propositions (a *calculus ratiocinator* or “calculus for reasoning”).[[161]](#footnote-211) The existence of God, the right of the Hapsburgs to rule Austria, and the legitimacy of same-sex marriage could all boil down to logical proofs, just like the value of pi. And if so, such proofs could eventually be computed mechanically, using a descendant of the reckoning machine.

This grand idea died a slow death over centuries, though in a sense, the entire field of computer science grew out of its moldering remains, like a sapling out of a nurse log. Remember that the theoretical foundations of computing were laid by Turing in his attempt to solve the *Entscheidungsproblem* in the 1930s—the problem of finding a general algorithm for deciding on the truth value of a mathematical statement.[[162]](#footnote-212) This was Leibniz’s problem, albeit restricted to math. The bad news: even in that purely abstract, formal domain, Turing proved that there was no such algorithm. The good news: in the process, he invented (or discovered) universal computation.

Although Leibniz worked on many other problems throughout his long career, he never gave up on the quest for a universal symbolic language. It was, quite literally, his Holy Grail. In a New Year’s letter to the Duke of Brunswick in 1697, Leibniz excitedly announced some progress: he had figured out how to represent numbers in binary, using only the digits 0 and 1.[[163]](#footnote-213) The letter sketched a design for a medallion to commemorate this insight featuring an image of divine creation, noting that it would be “difficult to find a better illustration of this secret in nature or philosophy […] that God made everything from nothing.”

Because it was both a minimal representation for numbers and a natural notation for yes/no logic, it seemed obvious to Leibniz that binary should form the basis of his *characteristica universalis*. George Boole (1815–1864), the self-taught nineteenth century mathematician who formalized the Boolean logic at the heart of virtually all digital computing today, independently arrived at the same conclusion more than a century later.[[164]](#footnote-214)

In practice, though, the earliest “human computers” didn’t use any special notation, or work out the answers to weighty questions on slates; in short, they didn’t much resemble Leibniz’s philosopher-accountants. They were out-of-work hairdressers employed by a civil engineer, Gaspard de Prony (1755–1839), to crank out books of logarithmic and trigonometric function tables for use by land surveyors after the French Revolution.[[165]](#footnote-215) (The hairdressers were out of work because many of their former customers, *Ancien Régime* aristocrats with towering hairdos, were losing their powdered wigs, along with their heads, to the guillotine. Survivors, rich and poor alike, were wisely opting to keep their hair short and plain.[[166]](#footnote-216)) These hairdressers may not have known higher math, but they were used to working carefully and methodically. They had no trouble performing the elementary operations into which each calculation had been decomposed: addition and subtraction.[[167]](#footnote-217)

Prony framed his project in industrial terms, boasting that division of labor allowed him to “manufacture logarithms as easily as one manufactures pins.” The reference to pin manufacture came straight out of Adam Smith’s *The Wealth of Nations*, the founding document of industrial capitalism.[[168]](#footnote-218) Leibniz had complained that “it is beneath the dignity of excellent men to waste their time in calculation when any peasant could do the work just as accurately with the aid of a machine.”[[169]](#footnote-219) Although Prony’s hairdressers were not (yet) using machines, they were being organized to work like one.[[170]](#footnote-220) And, of course, machines had already been invented for adding and subtracting.

Prony’s project greatly impressed Charles Babbage (1791–1871), designer of the first general-purpose computer, the “Analytical Engine,” in the 1830s. Entirely mechanical, the Analytical Engine would have been steam-powered, and would likely have taken about three seconds to perform an addition—far slower than any electronic computer, but still faster than doing it by hand.[[171]](#footnote-221)

Although known today as a hapless inventor too far ahead of his time (neither the Analytical Engine nor its more modest predecessor, the “Difference Engine,” were built in his lifetime), Babbage was no idle dreamer. He was in fact one of the Industrial Revolution’s great architects and theorists; his most important book, *On the Economy of Machinery and Manufactures*, was all about factory automation. It was full of down-to-Earth engineering and entrepreneurial advice, gleaned through close observation and obsessively backed up with productivity data.

Steampunk computing engines are just one (uncharacteristically speculative) idea among many in Babbage’s book. He introduced automatic computing in a chapter about the division of labor, noting: “The proceeding of M. Prony, in [his] celebrated system of calculation, much resembles that of a skilful person about to construct a cotton or silk mill, or any similar establishment.”[[172]](#footnote-222) In short, like Prony, Babbage sought to “manufacture logarithms.” Mechanical calculation was simply a means to automate the jobs of Prony’s hairdressers.

The parallel Babbage drew to textile mills is telling, for his other great inspiration was the Jacquard loom, patented by Joseph Marie Jacquard in 1804. Jacquard’s machine made the mass reproduction of complex patterned fabrics possible by encoding their designs as holes punched into a sequence of cards. Similar punched cards were to be the data input and storage mechanism for the Analytical Engine.[[173]](#footnote-223)

In 1833, Baroness Anne Byron, an educational reformer and philanthropist, took her seventeen-year-old daughter Ada, the future Countess of Lovelace (1815–1852), to a soirée at Babbage’s house.[[174]](#footnote-224) As Lady Byron wrote soon afterward, “We both went to see the thinking machine (for such it seems) last Monday. It raised several No.’s to the 2nd and 3rd powers, and extracted the root of a quadratic equation. […] There was a sublimity of the views thus opened of the ultimate results of intellectual power.” What they had witnessed was a demonstration of a working prototype of the Difference Engine—as sophisticated a computer as anyone then alive would ever see in operation.[[175]](#footnote-225)

Despite her youth, Ada Lovelace (as she is usually called today) understood the far-reaching implications of the machine. Over the following years, she became Babbage’s intellectual collaborator as the incomplete Difference Engine gave way to grander (and even less complete) plans for its fully programmable successor, the Analytical Engine.

In 1842, Lovelace translated a lecture on the Analytical Engine given by Italian military engineer Luigi Menabrea, appending to it a series of notes far lengthier and more insightful than the lecture itself. (Not a bad strategy for a woman in Victorian England to circumvent social barriers to publication.)

Note *G* includes the first published computer program for the Analytical Engine—hence the first published program, period. It calculated Bernoulli number sequences, which turn up often in mathematical analysis. As Lovelace famously wrote, “The Analytical Engine weaves algebraic patterns, just as the Jacquard-loom weaves flowers and leaves.”[[176]](#footnote-226) This observation was even sharper than it first appears, for she likely understood that the algebraic patterns in *real* flowers and leaves are themselves woven computationally, producing the mathematical regularities of petals, veins, leaves, and branches discussed in Part I.[[177]](#footnote-227)

To what degree, though, should one think of a computer as a mere “loom” for the mass production of mathematical tables? Babbage’s computing engines included a beautiful design for a printer capable of creating stereotype plates of whole finished pages. But there is a notable difference between the labor involved in printing tables (or manufacturing pins, or weaving cloth) and actually *calculating* those tables: the former kind of labor is physical, while the latter is mental.

This suggests a functional (not just metaphorical) analogy between computers and brains. The analogy has been used liberally by journalists over the years; for instance, in 1946, the *Surrey Comet* described Turing’s Automatic Computing Engine as an “Electric Brain to be Made at Teddington.”[[178]](#footnote-228) We tend to dismiss such headlines as old timey clickbait, but this glosses over the genuine insight behind Lady Byron’s characterization of the Difference Engine as a “thinking machine.”

Nowadays, we prefer terms like “information processing” to avoid the baggage of consciousness and subjective experience implied by words like “thinking” and “mental,” but we should keep in mind that, to a committed industrialist like Babbage, subjectivity simply wasn’t relevant. As far as we know, he didn’t give the question of “what it was like to be” a mechanical computer a moment’s thought; then again, neither did he trouble himself with “what it was like to be” a *human* computer. Factory work is purely functional, and operates at a level of abstraction above the individual—hence the substitutability of workers on a production line. As Babbage put it, “[D]ivision of labour can be applied with equal success to mental as to mechanical operations.”[[179]](#footnote-229)

Lady Byron was excited by the “sublimity of the views thus opened of the ultimate results of intellectual power,” but for others, the idea of machines doing mental labor sparked an early glimmer of the unease some feel about AI today. In 1832, a year before Lovelace and Babbage met, the London Literary Gazette had referred to Babbage, perhaps only half in jest, as a “logarithmetical Frankenstein.”[[180]](#footnote-230) Information work may not have been a significant sector of the labor market yet, but the prospect of machines *thinking* still induced a certain agita.

Just how far could a mechanical monster’s mental functions advance? According to Menabrea, “The [Analytical Engine] is not a thinking being but simply an automaton which acts according to the laws imposed on it.” Lovelace agreed, writing in her Note *G*, “The Analytical Engine has no pretensions whatever to *originate* anything. It can [only] do whatever *we know how to order it* to perform.”[[181]](#footnote-231) This zombie-like picture recalls the Golem of medieval Jewish folklore: a clay figure animated by a powerful rabbi using magical incantations—that is, code. The Golem may follow commands, but without understanding or discernment.

George Boole probably thought otherwise. Recall that, like Leibniz, Boole believed in binary logic as a universal calculus for reasoning. Boole went further, asserting that it *is* the way the mind reasons; hence the title of his great treatise, *An Investigation of the Laws of Thought*, published to wide acclaim in 1854.[[182]](#footnote-232) Presumably, this meant that understanding and discernment were themselves products of logic, for while Boole was religious, he (again, like Leibniz) thought of logic and rationality as inherently divine.[[183]](#footnote-233) The human soul was a “rational soul,” and our subjectivity followed from, rather than being in spite of, our rationality.

In the beginning, this neuroscientific aspect of Boole’s thesis remained largely unacknowledged. A notable exception was Boole’s friend and colleague Augustus De Morgan (author of the “Great fleas have little fleas” rhyme in Part I), who had himself been working on the foundations of logic.[[184]](#footnote-234) The Victorian intelligentsia was a small world: De Morgan was also Ada Lovelace’s mathematics tutor. Her knowledge of calculus, and of the Bernoulli numbers, came from De Morgan.

Unlike her contemporaries, though, Lovelace was not content to choose between logic-based brains, on one hand, and mystical non-explanations of mental processes, on the other. In an 1844 letter to a friend—a decade prior to Boole’s *Laws of Thought*—Lovelace wrote,

“I have my hopes […] of one day getting cerebral phenomena such that I can put them into mathematical equations; in short a law or laws, for the mutual actions of the molecules of brain […]. The grand difficulty is in the practical experiments. […] I must be a most skillful practical manipulator in experimental tests; & that, on materials difficult to deal with; viz: the brain, blood, & nerves, of animals.”

Lovelace undoubtedly had in mind the sensational experiments of Luigi Galvani (1737–1798), in which the application of electrical current to the nerves of freshly killed frogs caused their legs to twitch. (Those experiments had also inspired Mary Shelley’s *Frankenstein*—the pre-logarithmetical one.[[185]](#footnote-235)) Lovelace went on, underlining words with her customary enthusiasm:

“In time, I will do all, I dare say. And if not […] I shall have amused myself at least. [… N]one of the physiologists have yet got on the right tack; I can’t think why. […] It does not appear to me that cerebral matter need be more unmanageable to mathematicians than sidereal & planetary matters & movements, if they would but inspect it from the right point of view. I hope to bequeath to the generations a Calculus of the Nervous System.”[[186]](#footnote-236)

In drawing a parallel between neural activity and Newtonian laws of motion, Lovelace implied that cerebral “calculus” would not be founded on abstract logic like that of Boole and the Analytical Engine, but rather should be grounded in experiment, and in this, she was indeed on the right tack.

Unfortunately, Lovelace’s health, always fragile, would soon worsen. She began taking opiates for chronic pain, and died a few years later, at age 36, of uterine cancer. We will never know what she might have achieved as a neurophysiologist.

When real experiments with nerve tissue finally got underway, they at first seemed to confirm the “digital” intuitions of Leibniz and Boole. In the 1920s, pioneering English neuroscientist Edgar Adrian was electrically recording “all-or-nothing” action potentials, or “spikes,” in the sensory neurons of dissected frogs at his lab at Cambridge.[[187]](#footnote-237) Adrian’s colleague and predecessor, Keith Lucas, had already established that impulses in motor nerve fibers were similarly all-or-nothing.[[188]](#footnote-238)

With neural inputs and outputs thus seemingly binary electrical signals, the idea that the brain is a logic machine carrying out a kind of *calculus ratiocinator* gained powerful support. In particular, if every neuron has inputs corresponding to Boolean ones and zeros, and produces a Boolean output in response, that implies that it’s a logic gate, that is, a Boolean operator.[[189]](#footnote-239) And if so, the activity of the brain as a whole would amount to a vast logical calculation.

Two pivotal figures in modern neuroscience, Warren McCulloch and Walter Pitts, brought these ideas together in a legendary 1943 paper, “A logical calculus of the ideas immanent in nervous activity.” It seemed to be the realization, 99 years after Lovelace’s letter, of her envisioned “Calculus of the Nervous System”:

“To psychology, […] specification of the [neural] net would contribute all that could be achieved in that field—even if the analysis were pushed to the ultimate psychic units or ‘psychons,’ for a psychon can be no less than the activity of a single neuron. […] The ‘all-or-none’ law of these activities, and the conformity of their relations to those of the logic of propositions, insure that the relations of psychons are those of the two-valued logic of propositions. Thus in psychology […] the fundamental relations are those of two-valued logic.”[[190]](#footnote-240)

This paper was at once an end, a beginning, and a turning point—but not in any way its authors could have foreseen. As an engineering manifesto, it proved highly influential. John von Neumann read McCulloch and Pitts’s 1943 paper as a blueprint for how to perform general-purpose computation using electronic logic gates in place of biological neurons (as mentioned in Part I). In the right configurations, these gates—first made using vacuum tubes, and later transistors—could implement a Universal Turing Machine, and could therefore compute anything. The Boolean architecture was more minimal and elegant than that of the Analytical Engine and its successors, which relied on complex mechanical components and used base-ten arithmetic.[[191]](#footnote-241)

One can think, then, of digital computers as Universal Turing Machines built out of hand-crafted neural networks of McCulloch-Pitts type neurons. The symbols used today to represent logic gates are cribbed directly from the graphical representations McCulloch and Pitts drew of “pyramidal neurons” (which were, as the name implies, pyramid-shaped). We still draw a little circle at the tip of the pyramid for logical negation (turning an AND gate into NOT-AND or NAND, an OR into a NOT-OR or NOR, etc.). Originally, that circle represented an inhibitory synapse, which McCulloch and Pitts believed to implement the NOT operator.[[192]](#footnote-242)

Despite its powerful influence on computer engineering, as a neuroscientific theory, the 1943 McCulloch-Pitts paper was a false start. It marked the endpoint and high water mark of the notion that individual neurons work by carrying out logical operations. Postwar neuroscientists quickly established that while all-or-none voltage spikes are indeed a key signaling mechanism in the brain, they don’t correspond to logical propositions, and neurons don’t work like logic gates.[[193]](#footnote-243)

Larger-scale electrical fields and oscillations matter too; neither is electrical signaling the whole story. A wealth of neurotransmitters and neuromodulatory chemicals operate at a variety of scales, from local signaling with specialized “neuropeptides”[[194]](#footnote-244) to hormones secreted into the bloodstream that affect the entire body. These play major roles in our mental states, emotions, and drives. The emerging picture remains computational, but more in the spirit of Turing’s models for morphogenesis and “unorganized” neural networks rather than anything like a *calculus ratiocinator*.

Turing understood this, as he made clear in a 1951 BBC Radio address.[[195]](#footnote-245) He agreed with Lovelace’s assertion that writing a program in the usual way could only produce a predefined Golem-like result—indeed, traditional programming involves writing code whose function is fully understood by the programmer in advance, foreclosing any possibility for subsequent machine behavior we could reasonably call “creative” or “intelligent.” However, the universality of computing makes it possible to implement *any* computable model of neural function, including Lovelace’s “calculus of the nervous system,” on any Universal Turing Machine. Thus, Turing said, “it will follow that our digital computer suitably programmed, will behave like a brain.”

Although, like Turing, neuroscientists understood long ago that the brain is not a classical logic machine, the message took a long time to sink in among computer scientists and AI researchers. Recall that, throughout the twentieth century, repeated attempts at hand-writing algorithms to carry out the most basic of everyday tasks, like image and speech recognition, robotic control and common-sense reasoning, had failed utterly, bringing on the AI Winters mentioned in the Introduction.[[196]](#footnote-246)

For many decades, then, computers were able to perform logic-based tasks that had been deemed purely “mechanical” with superhuman speed, yet seemed unable to get anywhere at everyday “human” tasks. They failed even at tasks that many other animals with much smaller brains find easy. It began to seem that, despite Turing’s ideas about universal computation, computers might have little to do with either biology or brains after all.[[197]](#footnote-247)

The popular culture of AI throughout the twentieth century navigated this apparent mismatch between brains and computers by imagining that when—or *if*—AI finally arrived, it would be stellar at “rational” thinking and logic but poor at human emotion, creativity, or common sense. HAL 9000,[[198]](#footnote-248) or Data from *Star Trek*, are prime examples.[[199]](#footnote-249)

The message was: computers can perform rote calculations today, and may eventually exceed humans at “book smarts” too, but, being Golem-like, they aren’t remotely *like* us; the human spirit is neither logical nor rational. *We* can’t be reduced to mere computation.

This trope seems especially irresistible to Hollywood, resurfacing over and over in franchises like the *The Terminator*, *Robocop*, and *The Matrix*. If Boole was an Enlightenment thinker living in the Romantic period, Hollywood remains full of closet Romantics living in a Postmodern period.

What attracts us to this kind of human exceptionalism? Partly, the question answers itself; of *course* we like thinking of ourselves as fundamentally different, better in some non-quantifiable way, possessing some *je ne sais quoi* that will remain forever out of reach of the encroaching “other.” Nationalities and ethnic groups play this identitarian game all the time.

Beyond the obvious, though, the “incomputably human” narrative is comforting because the Industrial Revolution established a computational hierarchy, wherein the soulless, rote, mechanical stuff is done at the lowest tier of the org chart—first by calculating women, then by the machines that replaced them—so that upper class gentlemen-scholars could be freed for the more varied, creative, and intellectual pursuits worthy of being called “intelligent.”[[200]](#footnote-250)

As historian of science Jessica Riskin puts it, “Not only has our understanding of what constitutes intelligence changed according to what we have been able to make machines do but, simultaneously, our understanding of what machines can do has altered according to what we have taken intelligence to be. […] [T]he first experiments in automation were devoted to determining its uppermost limits, which simultaneously meant identifying the lowest limits of humanity.”[[201]](#footnote-251)

The problem we face today is not only that those limits have continued to shift, but that they have stubbornly refused to fall into the right sequence, according to our class-based preconceptions. It’s easier nowadays to imagine a doctor replaced by an AI model than a nurse. We probably won’t have AI plumbers anytime soon, but AI physicists may arrive imminently. When large language models can read and summarize complex arguments, write essays and poems, create software, and so on, there ceases to be any obvious hierarchy of information tasks where a computer can only substitute for the work of the poorly paid or uneducated. AIs are no longer, as Leibniz would have it, “artificial peasants” manufacturing logarithms.

Neither is AI the purely book-smart supernerd envisioned by twentieth century media. Real AI isn’t necessarily overly logical; in fact, its failures at solving logic puzzles, or even simple arithmetic problems, featured prominently in AI critique during the early 2020s. Now, not acting *enough* like a computer is interpreted as a lack of intelligence!

As an engineering discipline, too, modern AI is disruptive; it has little to do with the programming languages, data structures, and paradigms that were the mainstay of computer engineering departments a scant few years ago. AI models may run on Universal Turing Machines, but they aren’t algorithms in any classical sense. Instead, they tackle the problem of modeling probability distributions directly from data.

So where did the new approach to computing and AI come from? As it turns out, it has been here all along, just not in the mainstream. Understanding this alternative vision, and its relationships with other disciplines, requires following a different path from that of Leibniz, Boole, and traditional computer science—one that, inspired by living systems, conceived of computation less in terms of logic than in terms of learning and prediction.

This was the path of cybernetics, an idea developed as far back as the 1940s, in wartime, and already falling into disrepute by the 1960s.

# III. Cybernetics

## Love and war

In Part II we saw how even simple life forms, like bacteria, evolve to predict their environment, their internal states, and their actions. Both normativity (good versus bad) and basic feelings (like hunger, satiety, and pain) arise as consequences of prediction. However, the discussion was framed entirely in terms of a single-player game: how a “self” can evolve to enhance its dynamic stability in a world that is, if not hostile, indifferent. The food molecules a bacterium chases during chemotaxis are not running away; nor are chemorepellent molecules active aggressors. Chemotactic bacteria are modeling molecular concentration, but those small molecules they’re modeling (probably) aren’t modeling the bacteria in turn.

When it comes to wolves and rabbits, though, multiple agents are at work modeling each other—whether in skittish alarm (the rabbit) or with murderous intent (the wolf). Big-brained animals like us (and wolves, and rabbits) have detailed models of *themselves* too, as well as of kith and kin—hopefully, with neither skittish alarm *nor* murderous intent. Such rich social, subjective, and “intersubjective” experiences seem far removed from the simple machinery of bff programs or bacteria. Yet I will argue that minds like ours are a natural consequence of the same evolutionary dynamics. In particular, minds arise when modeling *other* minds becomes important.

The need for such modeling goes a long way back. It relates to two of life’s most important major evolutionary transitions: the emergence of sexual reproduction, likely about two billion years ago, and predation, which became prevalent during the Cambrian explosion, 538.8 million years ago.

Because sexual reproduction is so ancient, first arising among our distant single-celled ancestors, its origins remain obscure. Sex may have been an integral part of the eukaryotic revolution.[[202]](#footnote-253) Since eukaryotes are *the* canonical instance of symbiogenesis, it seems fitting that sex is itself symbiotic: the female and male forms of a species depend on each other to reproduce, mutually benefitting through faster evolutionary learning.[[203]](#footnote-254) (Hence, under most circumstances and all other factors being equal, a sexually reproducing species will outcompete one that reproduces asexually, even though each individual member of the asexual species faces fewer barriers to reproduction.)

For our purposes, though, sexual reproduction is relevant because it requires an organism to recognize one of its own, and usually, to distinguish male from female, modulating behavior accordingly. This is a big deal: when reproduction requires sex-selective cooperation or competition, others of the same species (and subtle differences between individuals) become an obligate part of one’s *Umwelt*.

Now, let’s fast-forward a billion and a half years. While many aspects of the ancient fossil record remain controversial, there is broad consensus that pre-Cambrian life was dominated by soft-bodied organisms with a wide range of body plans and symmetries, from tubular to quilted to more amorphous. We can only guess how those ancient Ediacarans moved, but it seems a good guess that the behavioral boundary between plants and animals would have seemed indistinct, with many organisms either drifting on ocean currents, or anchored to rocks, their fronds or appendages waving languidly in search of nutrients. Nervous systems during this period, when present at all, may have looked much like the distributed nerve nets of *Hydra* or comb jellies.[[204]](#footnote-255)

Then, the Cambrian explosion occurred: a sudden great diversity of marine creatures with protruding eyes, armored plates, fearsome claws, and sharp teeth. It’s impossible not to imagine them moving quickly and actively, hunting, escaping, killing and being killed. It was an arms race, featuring a cascade of innovations and counter-innovations in the arts of attack and defense. For the first time, complex, centralized nervous systems would have been required to intelligently coordinate rapid, perceptually guided, purposive movement.[[205]](#footnote-256) The (one-sided) joys of predation had been discovered.

In light of the earlier discussion of thermodynamics and its relationship to dynamic stability, it’s worth pausing to spell out the Darwinian motive for predation, even though it may seem intuitively obvious. Why *do* we eat, anyway?

Recall that life requires computation to reproduce (per Part I) and to respond to changes in its environment or internal state (per Part II). Computation, in turn, requires free energy—though recall also that this term is deceptive. Energy can neither be created nor destroyed; we don’t “use it up.”[[206]](#footnote-257) Free energy really just means a source of low entropy.

In thermodynamics, the classic low-entropy system is a pressurized piston: a cylinder of air in which most of the gas molecules are concentrated on one side of a partition, leaving a partial vacuum on the other side. It would be extremely unlikely for this situation to arise spontaneously—like randomly bouncing billiard balls suddenly forming a triangle at one end of the table—but with the piston locked in place to prevent the gas from redistributing itself more randomly, the pressure differential can be maintained.

This low-entropy configuration is like a loaded spring, or a cocked pistol. The moment that piston is allowed to move, it will do so with force, until the pressures on both sides equalize—that is, until entropy is again maximized. That would “use up” the free energy, even though the total amount of *actual* energy in the system (proportional to the summed squared velocities of all of the gas molecules) has remained constant. When Newcomen and Watt harnessed the piston’s movement to work a pump, or spin a loaded crankshaft, the molecules would slow down accordingly, cooling the gas.

Now, remember that once life takes hold, it creates predictability and order; its entropy plunges. That, too, requires work. It also means that an organism *itself* is a kind of pressurized piston—or loaded spring, or cocked gun. A wonderful source, in other words, of free energy. (The fatty bits, especially; all those delicious carbon-hydrogen bonds!) Thus, predation: a life form using the order, or “negative entropy,” stored in the body of another life form to do the work of maintaining or increasing its *own* negative entropy.[[207]](#footnote-258)

Obviously, there are conflicting interests here. In the best case (for the victim), the predator will just take a harmless little nip or sip—that’s how fleas, ticks, flies, and mosquitoes work. We still resent them; grooming behaviors among primates, and the swishy tails on horses, evolved to discourage such micro-predation. (And partly in response, flies have become lightning-quick, mosquitoes have learned to inject anesthetics, and ticks have become tenacious flesh-burrowers.) As for predation of the kind that comes in larger, life-ending bites—well, avoiding it at any cost immediately rises to the top of any prey animal’s to-do list.

At first blush, the two great revolutions I’ve just described—sex and predation—seem to have as little to do with each other as love and war. What they have in common, though (aside from “anything goes”), is that in the pursuit of dynamic stability, they both require living things to recognize and form internal models of other living things. They require, in other words, effective prediction of something that is itself a predictor.[[208]](#footnote-259)

We’ll get to love later in the book. In this part, though, we’ll focus on warfare—the original “killer app”—from a computational perspective. Let’s return briefly to nineteenth century London….

## Killer app

Babbage struggled to drum up financial backing for his Analytical Engine. The market for Jacquard-woven textiles was obvious enough, but who needed industrially mass-produced “analytical formulæ”? To sell the idea of mathematical tables as a popular commodity, he resorted to folksy examples, such as “the amount of any number of pounds from 1 to 100 lbs. of butchers’ meat at various prices per lb.”[[209]](#footnote-261)

Price tables for the village butcher? Hardly. A moment’s reflection will make it clear that small tradesmen wouldn’t have been a viable customer base for a hulking industrial machine like the Analytical Engine.

State administration, which Prony had bet on, was closer to the mark; the information needs of bureaucratic governments were on the rise.[[210]](#footnote-262) Still, it was too early. The French government defunded Prony’s project long before it was complete.

The real killer app was warfare. The British Army and Navy would have been Babbage’s obvious backers, and ultimately it was their lack of investment that doomed his enterprise.

The artillery table was already a paradigmatic product of human computation by the turn of the nineteenth century. A new table was needed for every big gun, including corrections for factors like altitude, wind, and barometric pressure. With every major world conflict from the Napoleonic Wars (1803–1815) onward, gunnery became increasingly important, and with it, tabulation. And doing the calculations by hand took a long time.

By World War I, the first fully industrialized large-scale conflict, both the Allies and the Central Powers were making extensive use of complex tables. Artillery fire was often planned days in advance, and its accuracy became crucial for supporting infantry advances. Computation had become a bottleneck in warfare, and more than any other single factor, this was what finally motivated serious investment in automatic computing between the World Wars.

When weapons production for World War II began ramping up, the University of Pennsylvania’s Moore School of Electrical Engineering hired at least 200 women to work on artillery tables. They used methods that would have been largely familiar to Babbage, or indeed, to Prony. But, spurred by the war effort, technology was advancing at breakneck speed. Six of the Moore School’s women were selected to become the programmers of the ENIAC.[[211]](#footnote-263) This first fully general, programmable computer had been designed to automate artillery tabulation.[[212]](#footnote-264)

By the time the ENIAC became operational, in December of 1945, priorities had changed. The Germans and the Japanese had been defeated, but Cold War brinksmanship picked up right where World War II had left off. The very first substantial program this new computer ran was a simulation for the “Super Problem,” exploring the feasibility of a hydrogen bomb.[[213]](#footnote-265) The math required was a lot harder than calculating ballistic trajectories, and one of the machine’s thousands of vacuum tubes burned out every day or two. This incentivized rapid improvements in the hardware, kickstarting what would later be known as Moore’s Law.[[214]](#footnote-266)

The early computers were a very long way from anything we have today. It’s sobering to consider how many years of capital-intensive incubation within the military-industrial complex computers needed before the technology had become sufficiently cheap, reliable, and miniaturized to rouse real interest from the private sector, let alone the village butcher.

IBM’s 701 mainframe, announced to the public on May 21st, 1952 and originally dubbed the “Defense Calculator,” was the first computer to become commercially available in the US. There’s an apocryphal quote, usually attributed to Thomas Watson Jr., IBM’s president, from a stockholders meeting in 1953, to the effect that he believed there was a worldwide market for only five computers. This isn’t quite true; what Watson really said was that when IBM had drawn up the plans for the 701 and toured them across the country to “some 20 concerns” that they thought “could use such a machine,” they had assumed they’d only get five orders, and were astonished to get 18.[[215]](#footnote-267)

At first, general-purpose computers were not designed to work in realtime. Like the human computers they replaced, they supported the war effort using batch processing. Such computing was a stately, offline affair. You submitted your job, and went to get a coffee… or, more likely, take a nap while technicians nursed the job along, replacing tubes, clearing punch card reader jams, swapping out spools of magnetic tape. Running on the ENIAC, bff would have taken centuries to achieve full-tape replication. By 1952, on the IBM 701, it would still have taken years, running nonstop, at a cost of millions of (today’s) dollars.[[216]](#footnote-268)

By then, there were some early experiments in interactive architectures, though. Project Whirlwind, operational in 1951, was initially designed for flight simulation, and later became the heart of the US Air Force’s SAGE air defense system.[[217]](#footnote-269) Realtime computing had become important due to the development of radar and related radio signaling technologies, effectively allowing one machine to physically detect another machine—at least one of which might be moving through space at high speed and with lethal intent.

Identify Friend or Foe (IFF) systems soon followed, using encrypted signaling to allow radar dots to be annotated when a bogey was “one of ours.” This in turn created an incentive to hack an enemy’s IFF system, spoofing the “friend” signal to sneak into enemy territory. A game of technical one-upmanship ensued, not unlike that of the Cambrian explosion.

Still more Cambrian was the rapidly increasing speed, precision, and deadliness of the military hardware. Jet fighters, missiles, and antimissile defenses proliferated. GPS was invented in large part to make autonomous weapons guidance possible, as it was obvious that high-speed warfare would soon render keeping a human “in the loop” impossible; the precision and response time needed were superhuman. Even the G-forces incurred by aerial maneuvers soon began to exceed human endurance.[[218]](#footnote-270) Robots would need to close the sensorimotor loops of these new weapons systems.

## Behavior, purpose, and teleology

Thus, the field of cybernetics was born, so named in 1947 by MIT mathematician Norbert Wiener, physiologist Arturo Rosenblueth, and their colleagues.[[219]](#footnote-272) Unlike batch computing, cybernetics was all about realtime feedback loops involving continuous time-varying signals like position, velocity, and thrust. Early proto-cybernetic technologies included electromechanical gun turrets and automatic bomb sighting engines used to correctly time the dropping of munitions from a plane to hit a target, often using analog cams and gears.[[220]](#footnote-273)

The cyberneticists took on the challenge of generalizing such systems to aerial dogfighting, in which the target was not stationary.[[221]](#footnote-274) In the presence of noise and uncertainty, such problems can be formulated using information theory, a field where Wiener made important early contributions. Under the assumption that the “brain” of a targeting system is linear, meaning that its outputs are weighted sums of its inputs, he also derived elegant theoretical results in optimal filtering and control. (Bacterial chemosensing, described in Part II, can be modeled using the same math.) These optimal linear theories are both useful for building weapons systems and, under many real-life conditions, can be hard to improve on: when the goal is clear and unchanging, everything is happening very fast, and measurements are noisy, there isn’t time for nonlinearity to matter so much.

In high-speed, goal-directed, and similarly adversarial natural contexts, such as a bat or dragonfly closing in on prey, animal brains can also carry out something close to cybernetically optimal linear modeling.[[222]](#footnote-275) The same holds when you reach for an object or catch a ball.[[223]](#footnote-276) Balls, however, don’t actively try to escape as you reach for them—outside the Harry Potter universe.

A moth, however, behaves like a real-life “golden snitch” when a sonar-ing bat is closing in for dinner.[[224]](#footnote-277) It can’t fly as fast as a bat; its brain and sensory apparatus are outmatched too. But it can try evasive action. The evasions must be genuinely unpredictable, because randomness is the moth’s only cybernetic advantage. The bat, with its much larger brain, would preempt any more coherent plan the moth could dream up. So, while the bat applies something close to optimal prediction tailored to the moth’s flight statistics, the moth takes advantage of its low mass to flutter around chaotically, rendering those statistics as blurry as possible. Instead of a smooth, predictable curve, the resulting flight pattern is a tortuous “random walk” occupying a volume in space known affectionately as the “Wiener sausage.” (For real.[[225]](#footnote-278))

As you will know if you’ve ever tried—and failed—to catch a seemingly not-that-fast moth, these Wiener sausage tactics work, at least up to a point. The moth implements them with a biological random number generator, likely based on a combination of mechanical instability in the wings and random neuronal activity. If you’re wondering how neuronal activity *can* be random, keep in mind that it always is, a little bit. The biochemical events leading to neural firing include ion channels opening and closing in the cell membrane, and the release and capture of synaptic vesicles containing neurotransmitters. The timings of these events are never quite precise, because they’re contingent on the same random molecular interactions that drive all cellular activity. If a neural circuit is wired to amplify that noise, then like the random number instruction of Turing’s Ferranti Mark I computer, the result will be a computationally usable random signal.[[226]](#footnote-279)

There’s ample evidence that prey species like moths use such circuits to generate chaotic behavior; startled cockroaches, similarly, use them to scurry at random.[[227]](#footnote-280) If randomness is the low road to cybernetic unpredictability, though, there is also a high road: becoming smarter than the entity trying to predict you, and predicting that entity’s predictions. We’ll explore this brainier strategy in Part V.

But first, let’s take the full measure of cybernetics—what it got right, what it got wrong, its untimely demise, and its enduring influence. As formulated by Wiener, cybernetics *was* the field we now call Artificial Intelligence, and more—it was the hallmark of all complex systems, a grand unifying theory of the biological, the technological, and even the sociological.[[228]](#footnote-281)

The argument goes like this. The simplest cybernetic system is something like a humble thermostat, a feedback mechanism that turns the heat on and off to maintain a constant temperature. Warm-blooded animals, of course, do just that, and living systems in general must use feedback to “homeostat” (that is, to regulate) their internal state. Construed broadly, homeostasis is what being alive is all about. The point of eating, for example, is to maintain the body’s store of free energy, and similar purposive arguments can be made for all of our primary drives. (Our secondary drives, such as the desire for praise or prestige, can in turn be formulated as “instrumental” to those primary drives.) Homeostasis is thus our old friend, dynamic stability, implemented purposively.

Complex animal behaviors may seem a long way from those of a thermostat, but consider that a truly optimal thermostat will not only switch on or off in response to fluctuations in the temperature *now*, but on the basis of its prediction of the future. If, for instance, the sun is about to rise and start rapidly warming your house, and there are significant delays in the response of the house’s temperature to heating, this ought to affect your home thermostat’s strategy.

Generalizing this observation in a 1943 essay entitled *Behavior, Purpose, and Teleology*,[[229]](#footnote-282) Wiener, Rosenblueth, and computer pioneer Julian Bigelow perfectly articulated the “predictive brain” hypothesis in cybernetic terms (though they did not coin the word for several more years):

“All purposeful [or ‘teleological’] behavior may be considered to require negative feed-back [which] may be extrapolative (predictive) […]. A cat starting to pursue a running mouse does not run directly toward the region where the mouse is at any given time, but moves toward an extrapolated future position. […] The cat chasing the mouse is an instance of first-order prediction […]. Throwing a stone at a moving target requires a second order prediction […]. It is probable that limitations of […] the central nervous system […] determine the complexity of predictive behavior […]. Indeed, it is possible that […] the discontinuity [between] humans [and] other high mammals may lie in that the other mammals are limited to predictive behavior of a low order, whereas man may be capable potentially of quite high orders of prediction.”

The authors also pointed out that, from this behavior-focused perspective (that is, without worrying about *how* the predictive computation is implemented), purposive machines and living systems are alike, “regardless of the complexity of [their] behavior,” adding that “Examples […] are readily found of manmade machines with behavior that transcends human behavior.”

Yes, this was true even in 1943, before general-purpose computing; it was even true in 1843, or there would have been no Industrial Revolution, though few machines back then were teleological or purposive in the cybernetic sense.[[230]](#footnote-283)

In invoking teleology, Wiener and colleagues were playing with fire, both in the philosophical and scientific communities. Materialism and reductionism had given teleology a bad name, relegating it to a quasi-religious belief. After all, if the (entirely knowable) dynamical laws of physics fully determine what will take place at the next moment in time given initial conditions, what role could teleology or purposive action possibly play?[[231]](#footnote-284)

Beyond stipulating goals and purposes, cybernetics posited agents that predict the future, and bring that future about—the same apparent reversal of causality we encountered in Part II. The cyberneticists’ insight, both trivial and profound, was that predictive negative feedback loops are sufficient to give an apparatus (or organism) purposiveness … without violating the laws of physics.

The apparent paradox of backward causality resembles the apparent paradox of an apparatus (or organism) building a copy of itself. In that case, von Neumann realized that the solution lay in the apparatus having an internal model of its own structure, and using that model to guide construction. Wiener and colleagues realized that the solution to *their* problem lay in the apparatus having an internal model of the world, and using that model to guide behavior.

These are flip sides of the same coin. In both cases, the models are computational; and in both cases, their purpose is to continue to exist through time, whether by growing, by preserving the integrity of the self, or by replicating.

Von Neumann’s and Wiener’s insights were published within a year of each other, though at the time, the intellectual kinship between them may not have been fully appreciated.

## Negative feedback

Cybernetics got many things right, especially as compared with the symbolic, programming-based approach to AI championed by Wiener’s detractors. For starters:

* It embraced continuous values and random processes.
* It was based on learning (that is, modeling) probability distributions, rather than executing hand-engineered code reflecting a programmer’s intuitions.
* It advanced the idea that agency (or purposiveness, or teleology) are fundamental to intelligence, and to life more broadly.[[232]](#footnote-286)
* It was consistent with the neuroscientific consensus that emerged soon *after* McCulloch and Pitts’s 1943 paper, when it was understood that neurons were not implementing propositional logic using Boolean values.
* It focused on the continuity between intelligence and living systems more generally, rather than imagining that intelligence is a purely logical construct in the Leibnizian tradition—and therefore the exclusive province of symbolically-minded humans. As a corollary, cybernetics presumed that bats, dragonflies, and even bacteria are also intelligent.
* It emphasized the importance of behavior over mechanism, noting that the same models could be computed in radically different ways—for instance, in standard Turing machine-like computers, “largely by temporal multiplication of effects” given “frequencies of one million per second or more,” or in biological systems, by “spatial multiplication,” i.e. massive and non-deterministic parallelism.[[233]](#footnote-287)

This last point may sound reminiscent of B.F. Skinner and radical behaviorism, a school of thought that has been pilloried as brutally reductionist, or denying the existence of mental states.[[234]](#footnote-288) A fair assessment of behaviorism would take us too far afield, but suffice it to say: the point Wiener and colleagues were making was about the fundamentally computational nature of the brain or mind. Hence, its multiple realizability, or platform independence, underwritten by the universality of computation, per Church and Turing. This is the same kind of “behaviorism” that led Turing to formulate the Imitation Game in terms of behavior, rather than insisting on any specific brain mechanism.

Cybernetics also got some things wrong, though. Its greatest flaw lay in overpromising: there was a vast gulf between Norbert Wiener’s self-promoting grandiosity and the engineering realities of his day. The theory was so general that it was nearly tautological, yet practical demonstrations using mid-twentieth century technology were underwhelming. Neither was Wiener much of an engineer; he was more at home at a blackboard than in a machine shop. An idea for a cybernetic antiaircraft gun controller, which he had insisted would prove crucial to the war effort, remained vaporware.

In 1949, Wiener’s students at MIT cobbled together a motorized cart whose wheels were driven by feedback loops involving two photocells. Wired up one way, it was christened *Palomilla*, Spanish for “moth,” and could trundle bumpily along a corridor, more or less following a flashlight beam. Switching the wires turned it into a light-averse “bedbug.” Cranking up the feedback gain made it oscillate with something like Parkinsonism or intention tremor. But beyond the tight feedback loops of missile control systems (which did what *Palomilla* did, marginally more reliably[[235]](#footnote-289)), real-life applications were elusive. Why were these demos so lame?

Wiener could easily write down a feedback-response function that would stand in for the totality of an organism’s behavior. He could then proceed to expand that function formally into an infinite series of higher-order terms. He could even derive closed-form solutions for the linear parts, given a simple goal like playing an optimal game of chase with an opponent whose movements could be analyzed statistically. The higher-order terms remained out of reach, though. Exploding numbers of parameters and the difficulty of characterizing higher-order goals made it unclear how further progress could be made—especially given the feeble computational power available.

Access to massive computation would not have immediately solved the problem, either, because major conceptual gaps remained. Cyberneticists had largely swept memory and learning under the rug, along with anything resembling individual behavior. Lofty rhetoric aside, what cybernetics offered in practice was a cartoon of the purely instinctual behavior of a simple organism performing a single low-dimensional sensorimotor task in an unchanging niche. A thermostat, basically.

Real moths and bedbugs have a vastly richer behavioral repertoire than *Palomilla*. Even chemotactic bacteria do. Those pesky feedback functions were black boxes where seemingly infinite complexity hid. In theory, more or less any function could be approximated with a series expansion, but in practice, it was hard to see how anything resembling thought, perception, or action—let alone anything like psychology or a “self”—could arise merely by adding more terms to the series.

GOFAI may have been cheating, but at least hand-written programs could implement more substantive behavior, and could run efficiently on primitive general-purpose computers. The results, in turn, may not have been particularly lifelike, but the researchers and engineers didn’t care. They were busy launching a trillion dollar industry. When their AI grant applications stopped getting funded, they could readily find a wealth of lucrative uses for software with no pretensions to being brain-like. Computer science began to disentangle itself from AI to become a discipline in its own right, while software crossed the chasm from the military-industrial complex to big business, and from there into home computers.

Meanwhile, cybernetic philosophy was going ever more meta, ultimately embraced (sometimes as little more than a metaphor) by the intellectually hip in fields as far-flung as government,[[236]](#footnote-290) anthropology,[[237]](#footnote-291) ecology,[[238]](#footnote-292) urban planning,[[239]](#footnote-293) sexology,[[240]](#footnote-294) feminism,[[241]](#footnote-295) and post-structuralist critical theory.[[242]](#footnote-296) Some cybernetically-inspired period pieces still make for great reading, some had the right idea, and some had significant impact in their own fields. But unlike mainstream computer science, none were accompanied by much technical progress in cybernetics itself. The concept of a “cyborg,” a cybernetically inspired superhuman fusion of man and machine (yes, usually men), went from a serious research priority at NASA and advanced defense labs to a comic book trope.[[243]](#footnote-297)

In other words, cybernetics began to look like a fad. By the mid-1960s, it had largely faded as an area of active research, though the name has survived vestigially, and somewhat randomly, in terms like “cyberspace,” “cybersecurity,” “cybercrime,” and “cyberwarfare.”[[244]](#footnote-298)

Cybernetics fell victim not only to overreach, but also to the systematic efforts of its detractors.[[245]](#footnote-299) In 1955, computer scientist John McCarthy coined the term “Artificial Intelligence” in the proposal for a summer workshop at Dartmouth—precisely to distinguish the new symbolic (a.k.a. GOFAI) approach from cybernetics. Despite promising “a significant advance” over the two-month period, little headway was made at the workshop proper, but its attendees, including Arthur Samuel, Allen Newell, Herbert Simon, and Marvin Minsky, became the who’s who of AI over the next several decades. Wiener was pointedly not invited.[[246]](#footnote-300)

It’s ironic that we continue to use the term “Artificial Intelligence,” given that today, virtually the entire field descends from the kind of work the coiners of this phrase sought to discredit—especially, per Hubert Dreyfus, “the perceptrons proposed by the group Minsky dismisses as the early cyberneticists.”[[247]](#footnote-301) These were the first artificial neural nets.

## How we know universals

Inspired by the brain’s physical structure, pioneering American psychologist William James (1842–1910) had envisioned something very like neural networks as far back as 1890. In his magisterial textbook, *The Principles of Psychology*,[[248]](#footnote-303) James imagined that neural processes—the minutely complex “wires” evident in stained tissue samples from cerebral cortex—might physically embody associations between co-occurring stimuli.

Shortly afterward, the great Spanish neuroanatomist Santiago Ramón y Cajal (1852–1934) concluded that this wiring is not “reticular” (meaning continuous), but consists of the branching outgrowths of individual neurons. Influenced by James, Ramón y Cajal went on to suggest that the junctions between neurons, which we now call synapses, were the sites of neuroplasticity. Associations could be learned (or unlearned) through the selective strengthening (or weakening) of synapses.[[249]](#footnote-304)

The idea that mental associations are central to learned behavior has its own long history. In 1897, Ivan Pavlov (1849–1936) published his classic experiments in which dogs could be taught to associate food with the sound of a musical triangle. Once conditioned, they began salivating when the triangle was struck.[[250]](#footnote-305)

Learning associations is even more fundamental than behavioral experiments like this suggest, though. Dogs aren’t born with individual neurons that activate exclusively at the taste or smell of food, the sight of a particular person, or the sound of a triangle. Associating these multimodal events with each other is the easy part. Without being able to learn subtle, complex associations among raw stimuli, a dog would be unable to recognize a person or a ringing triangle at all.

McCulloch and Pitts formulated this problem in their 1947 paper, *How we know universals: the perception of auditory and visual forms*.[[251]](#footnote-306) Their theory was (once again) wrong, but posed as a question, the paper’s title is a good one: How *do* we recognize categories, or “universals”? Four years after the publication of *A logical calculus of the ideas immanent in nervous activity*, experiments had convinced them that neurons were not logic gates, although they did exchange excitatory and inhibitory signals. The researchers had also realized that one of the hardest computational tasks facing the brain is that of achieving “perceptual invariance”—the key to performing what we now call “classification,” “recognition,” or more broadly, “generalization.” McCulloch and Pitts were trying to figure out how anatomically plausible neural circuits in sensory cortex might do that.

Perceptual invariance is illustrated by an insightful short story, *Funes el memorioso*, written by Jorge Luis Borges in 1942.[[252]](#footnote-307) In Borges’s story, a young man, Ireneo Funes, is thrown from a horse and suffers a crippling brain injury. It leaves him with a perfect memory, yet robs him of the ability to generalize. After the accident, Funes can remember “not only every leaf of every tree of every wood, but also every one of the times he had perceived or imagined it”; yet he is described as “[…] almost incapable of ideas of a general, Platonic sort. Not only was it difficult for him to comprehend that the generic symbol *dog* embraces so many unlike individuals of diverse size and form; it bothered him that the dog at three fourteen (seen from the side) should have the same name as the dog at three fifteen (seen from the front).”

For a long time, machines fared no better. Throughout the twentieth century and well into the twenty-first, memory technologies improved exponentially, but perceptual invariance seemed to remain an unsolved problem. Attendees at the AI summer workshop in 1956 made no progress, and when those who were still alive 50 years later returned to Dartmouth for a reunion, little seemed to have changed. A computer could, like Funes, “remember” the color of every pixel in a two hour movie, yet be unable to make any sense of those tens of billions of pixels. Hence bots could easily be foiled by visual Turing Tests as simple as “click on the images below containing a dog.”[[253]](#footnote-308) The reconvening wizards, grizzled but as opinionated as ever, were of various minds about what to try next, but they all agreed that AI was a long way off.[[254]](#footnote-309)

The year was 2006. They could not have been more wrong.

## Perceptrons

The solution had lain hidden in plain sight for over a hundred years. At the turn of the twentieth century, William James and Santiago Ramón y Cajal had already accurately described how brains must learn associations. As experimenters figured out how to move beyond dissection and tissue staining to study the living brain at work, a clearer view of neural information processing began coming into focus.

In a series of papers beginning in the 1950s, neuroscientists David Hubel, Torsten Wiesel, and a rotating cast of collaborators reported on painstaking investigations of cat and monkey visual cortex.[[255]](#footnote-311) Their technique involved recording the activity of individual neurons using tungsten electrodes while visual stimuli were presented to animals under anesthesia. In the picture that emerged, “retinal ganglion cells” were found to combine inputs from photoreceptors in the retina to create local visual features, which are in turn combined to form higher-level edge-like features by “simple cells,” which are in turn combined to form yet higher-level features by “complex cells,” which are in turn combined by “hypercomplex cells”—until eventually, the features in question may be entire objects.

Margaret Livingstone, a neuroscientist at Harvard who did postdoctoral work in Hubel’s lab, wrote in an obituary for him in 2013, “Studying vision is fun because you see what you show the animal, and when you cannot figure a cell out, you show it everything you can think of; sometimes you find surprisingly specific things that will make a cell fire, like a bright yellow Kodak film box.”[[256]](#footnote-312)

It’s an elegant concept: any perceptual classification task, no matter how complicated, can be built up hierarchically as a combination of features, which are themselves combinations of features; turtles all the way down. This seemed to be the brain’s approach to generalization.

In 1957, Frank Rosenblatt, a young engineer at the Cornell Aeronautical Laboratory, decided to build an apparatus to do what the visual cortex does. He called it the “perceptron.” As he wrote in the introduction to his report, “[I]t should be feasible to construct an electronic or electromechanical system which will learn to recognize similarities or identities between patterns of optical, electrical, or tonal information, in a manner which may be closely analogous to the perceptual processes of a biological brain.”[[257]](#footnote-313)

In a passage reminiscent of Borges, Rosenblatt noted that although “[t]he recognition of ‘similar’ forms can be carried out, to a certain extent, by analytic procedures on a […] computer […] it is hard to conceive of a general analytic program which would […] recognize the form of a man seen from any angle, and in any posture or position, without actually storing a large library of reference figures […]. In general, identities of this sort must be learned, or acquired from experience […].”

Rosenblatt’s “identities” were the same as McCulloch and Pitts’s “universals” a decade earlier. Unlike McCulloch and Pitts, though, Rosenblatt created a real, working system.

The original “Mark I” perceptron was a three-layer neural net. Its first layer consisted of a 20×20 array of photocells, or “sensory units,” the second had 512 “association units,” and the final layer had 8 “response units.” Each layer was wired randomly to the next, with a motor-driven potentiometer (like the volume knob on a radio) modulating the strength of every connection.

The response units also implemented “lateral inhibition,” a feature commonly found in neural circuits: when one activated, it inhibited the others, which in turn fed back inhibition to any competing association units. Using a simple learning rule that exploited this feedback, Rosenblatt got the device to distinguish among simple shapes—a square, a circle, a triangle. It was a start.

More than a start: in retrospect, the perceptron was arguably the single most important advance in AI in the twentieth century. Modern neural nets, especially those most commonly used for visual recognition, are still powered by “multilayer perceptrons,” or MLPs, although they use different learning rules and are typically much deeper, that is, have many more than three layers.

If perceptrons worked—and did so by taking an approach so much more obviously brain-like than symbolic AI systems—then why did they lay fallow for so long? Several narratives have merit:

1. Neural nets big enough to do practical work require large-scale parallel computation, which didn’t become available to most researchers until inexpensive GPUs (graphics processing units) began flooding the market, around 2006.[[258]](#footnote-314) This is true; however, if, beginning in the 1960s, we had begun to seriously engineer parallel neural computers, or special purpose hardware like the Mark I, we could probably have had powerful MLPs by the 1980s. We didn’t push the engineering in this direction much because Moore’s Law advanced ceaselessly for six decades after the invention of the transistor in 1947, shrinking their size, increasing their maximum clock speed, and decreasing their operating power. The industry was eager to reap the most obvious benefits—making the whole computer smaller, faster, cheaper, and lower power—rather than sacrificing these benefits to increase parallelism in the service of an unproven approach to computing.[[259]](#footnote-315)
2. Large neural nets require large datasets to train, which weren’t available before the web took off. And labeled datasets, like ImageNet (which researcher Fei-Fei Li began gathering in 2006—that year again!), relied not only on internet-scale collections of images, but also on online gig workers to apply labels, like “chihuahua” and “blueberry.” All true, but let’s keep in mind that Netflix began sending subscribers digital movies by DVD nearly a decade earlier, and a modest collection of DVDs contains far more imagery than ImageNet. There was no shortage of pixels. We also now know that unsupervised learning works well, and renders extensive labeling unnecessary (as will be discussed in Part VIII). This would likely have been discovered decades earlier had neural net research remained mainstream in the latter half of the twentieth century.
3. Rosenblatt’s original perceptron training algorithm wasn’t powerful enough to work for complex datasets or deep neural nets; the suite of tricks required to do so has only been perfected in recent years. The most important of these tricks, though—the “backpropagation” algorithm, allowing synapses in an arbitrarily deep neural net to be adjusted to minimize the output error—had already been worked out by 1970, and was repeatedly reinvented by other researchers over the years.[[260]](#footnote-316) Like any technology, dedicated tinkering is required to get the details right. For many years, there simply weren’t enough dedicated tinkerers working on it, in part because—
4. In 1969, Marvin Minsky and an MIT collaborator, Seymour Papert, published a highly cited book entitled *Perceptrons*.[[261]](#footnote-317) Although dedicated to Rosenblatt, the book was a hatchet job, implying that a number of simple mathematical functions could never be approximated by perceptrons, and therefore perceptrons were inherently far weaker, computationally, than GOFAI algorithms. The book contained mathematical proofs of these assertions, but the proofs only applied to simplified two-layer models. It can be shown that with three or more layers (as even the Mark I perceptron had) *any* continuous function can be approximated,[[262]](#footnote-318) as Rosenblatt himself appears to have intuited.[[263]](#footnote-319)

Despite the protests of Rosenblatt and fellow travelers, the *Perceptrons* book was highly effective in discouraging “connectionism,” as neural net research was then called, for decades, diverting mainstream attention instead to the GOFAI approaches favored by Minsky, Papert, and their colleagues. In 1988, with interest in neural nets rising once again just as we entered a final Good Old-Fashioned AI winter, Minsky and Papert reissued *Perceptrons*. In a new prologue and epilogue, the updated edition doubled down on their original critique, claiming that “little of significance has happened in this field” and that even by 1969, “progress had already come to a virtual halt because of the lack of adequate basic theories.”

These claims seem bizarre, if not disingenuous. Just a year before the second edition of *Perceptrons*, computational neuroscientists Terry Sejnowski and Charles Rosenberg had trained a neural net, NETtalk, to pronounce English text. They showed not only that it could effectively master this notoriously unruly (that is, non-GOFAI-friendly) problem, but that its performance exhibited many human-like characteristics.[[264]](#footnote-320)

Minsky and Papert’s objections to connectionism make some sense, though, in light of the theoretical grounding they had been hoping AI might offer. They were uninterested in the rapid advances taking place in computational neuroscience, machine learning, or general function approximation. Instead, they focused on mathematical theories of knowledge representation, manipulation, and formal proof in the tradition of Leibniz. This required operating at the level of abstract concepts, more like the “psychons” McCulloch and Pitts had speculated about in 1943 (but that neuroscience had shortly thereafter abandoned, failing to find any evidence of them in the brain). Per Minsky and Papert: “perceptrons had no way to represent the knowledge required for solving certain problems. […] No machine can learn to recognize X unless it possesses, at least potentially, some scheme for representing X.”

Without schematized knowledge representations, they did not believe that higher-order “rationality,” including causal reasoning, self-reflection, or even consciousness, would be possible: “[W]e expect distributed representations[[265]](#footnote-321) to tend to produce systems with only limited abilities to reflect accurately on how they do what they do. Thinking about thinking, we maintain, requires the use of representations that are localized enough that they can be dissected and rearranged. […] [D]istributed representations […] must entail a heavy price; surely, many of them must become ‘conceptual dead ends’ […].”

Unlike the book’s proofs, these assertions were mere intuitions. To many, they seemed reasonable at the time. Adherents of the GOFAI school of thought, including many trained in linguistics (which is largely concerned with schematized knowledge representations) continue to argue this position today, though it seems increasingly disconnected from reality given what modern neural nets can do, and how they do it.[[266]](#footnote-322)

## Deep learning

Training multilayer perceptrons is difficult both because of all of the dependencies between layers (fiddling with a synapse weight affects everything downstream) and because of the dramatically increased sizes of these models. To get a sense of this, consider that a two-layer “fully connected” perceptron classifying a 32×32 pixel image as one of 10 digits already has (32×32)×10=10,240 synapses—that is, a synapse connecting every pixel with each of the 10 output neurons. If we add in a single additional 32×32 layer of neurons between the input layer and the output layer, we now have (32×32)×(32×32) + (32×32)×10 = 1,058,816 synapses. Each of these synapse weights is a model parameter that must be learned from data, and roughly speaking, fitting a hundred times more parameters requires a hundred times more data.[[267]](#footnote-324)

Over time, though, researchers found ways around these problems. The story of modern “deep learning” is just the story of these tricks accumulating over time, with compounding gains and accelerating progress as the field finally attracted serious time and investment. Here are a representative few. A complete list would be much longer:[[268]](#footnote-325)

1. The backpropagation algorithm, as mentioned earlier, allows minimization of an error, or “loss function,” to happen reliably through many layers of neurons. It involves using calculus to compute the total downstream effect of fiddling with any given synapse weight. Virtually all machine learning today uses backpropagation.
2. “Convolutional layers,” developed by deep learning pioneers Kunihiko Fukushima and Yann LeCun,[[269]](#footnote-326) allow the number of weights in certain layers to be much smaller than the number that would be needed for a “fully connected” layer as sketched above, while offering more structure than Rosenblatt’s sparse random connectivity. In essence, a neuron may be connected only to a small neighborhood of neurons in the previous layer, rather than to all of them; furthermore, an entire “channel” of neurons share a common weight pattern within this local area.
3. The “softmax” function,[[270]](#footnote-327) applied to a layer of neurons, implements something analogous to lateral inhibition, picking the maximum activation out and enhancing it while suppressing competing neurons. Softmax was originally used only for the output layer of classifier nets, creating a “one-hot” representation (a single neuron on, the others off). However, the suppression isn’t complete (hence the “soft”), allowing learning through backpropagation to work.
4. “Max pooling”[[271]](#footnote-328) can discard unneeded resolution, for instance reducing a 32×32 layer of features to a 16×16 layer by picking only the largest values within each 2×2 region to pass on. This works because the presence of complex features is often more important than their precise location. We can also think of max pooling as a kind of lateral inhibition.
5. Datasets can be “augmented”[[272]](#footnote-329) by shifting, stretching, and rotating training examples, then using these distorted copies as additional training data. This helps strengthen the neural net’s perceptual invariance.

It’s significant that researchers working at the intersection of computer science and neuroscience pioneered many of these techniques. In recent years, this confluence of disciplines has been called “NeuroAI,” and a number of conferences and workshops have been convened under that banner. While the name is new, the phenomenon is not. From the beginning, machine learning and neuroscience have been continuous with one another, as the historical sketches in this part of the book and the Interlude make clear. For many years, the scientific value of this interdisciplinary foment was greater than its technological value, so it was often referred to as “computational neuroscience.” With the technological aspect now ascendant, it’s unsurprising to see the ordering reversed with “NeuroAI.” But the conference attendees are largely the same. And while no practical machine learning trick faithfully models any specific process or circuit in the brain, many of the tricks are clearly biologically inspired, just as the original perceptron was.[[273]](#footnote-330)

Using such tricks in combination, researchers fully solved handwritten single-digit visual recognition in the 1990s,[[274]](#footnote-331) and moved on to increasingly difficult visual recognition problems over the following two decades: objects, clothes, places, plants and animals, natural scenes, faces, and ultimately nearly everything one might encounter while browsing through the photos on a phone.[[275]](#footnote-332) Although they were developed mostly for the visual modality, convolutional nets also work well in many other domains, from auditory recognition[[276]](#footnote-333) to weather prediction.[[277]](#footnote-334) Surprisingly, they even work for classifying text,[[278]](#footnote-335) but we’ll hold off on exploring neural nets for language until Part VIII.

## Closing the loop

Before moving on, let’s consider neural networks through the cybernetic lens. One one hand, the perceptron *did* generate discrete symbolic output—per Rosenblatt’s 1957 report, “inhibitory feedback connections guarantee that only one response out of a mutually exclusive set can be triggered at one time […. Response] units thus [act] like a multi-stable flip-flop.”[[279]](#footnote-337)

(In modern neural nets, softmax layers serve the same purpose.) Still, GOFAI partisans like Minsky and Papert understood that Rosenblatt was working squarely within the cybernetic tradition, and indeed, the response layer and learning rule only approximated digital logic in the output through the use of analog negative feedback loops. The system learned by example, and relied on this analog quality to do so. Neural representations within the network were distributed and continuous, not localized or symbolic. Synapse weights were continuous, and the perceptron as a whole evaluated a nonlinear continuous function parametrized by those weights. Rosenblatt’s perceptron made use of randomness too, both in the wiring and in the learning process: “The penalty that we pay for the use of statistical principles in the design of the system is a probability that we may get a wrong response in any particular case […].”[[280]](#footnote-338) The perceptron worked precisely because it was far from the world of cleanly represented knowledge and exact logical inference.

Let’s boil down the main advance Rosenblatt and his successors made over Wiener: while Wiener’s “black box functions” were expressed in terms of “series expansions,” the function evaluated by a perceptron is instead parametrized by a hierarchy of synapse weights. A series expansion is a weighted sum of increasingly high-order mathematical terms, typically truncated after only a few terms. For the kinds of simple cybernetic systems Wiener and colleagues were able to build (like thermostats, or, eventually, antiaircraft fire controllers), a single term, representing a linear approximation, might suffice. Not so for a multilayer perceptron capable of achieving visual perceptual invariance.

The problem isn’t theoretical, but practical. At bottom, a function is a function. Both neural nets and series expansions are “universal function approximators,” meaning that either one can approximate any (continuous) function. However, Wiener’s more formal approach would have required vast numbers of parameters to go beyond the linear or first-order regime. Realistic nonlinearities (say, for object classification) couldn’t, in practice, be learned from a limited number of training examples. By dispensing with this kind of math and simply wiring together simplified neurons in multiple layers, Rosenblatt had stumbled onto a far more learnable way of representing the kinds of functions needed to transform images, sounds, and other natural stimuli into invariant classifications. One could attribute that to luck, intuition, or both, but at some level, something like it *had* to work, or our brains would all be like that of poor Funes, *el memorioso*.

*Why* perceptrons are so “learnable” relative to more traditional function approximators is not yet fully understood.[[281]](#footnote-339) Recent analyses suggest that it may have to do with a “compositional prior,” meaning a bias[[282]](#footnote-340) toward learning functions that can be defined in terms of a hierarchical composition of simpler functions. This kind of prior appears to be useful for all sorts of cognitive tasks and generalizations, not just vision.[[283]](#footnote-341) (Intriguingly, the symbiogenetic view of evolution described in Part I can also be understood as a learning algorithm that favors hierarchical compositions of functions.)

Although the perceptron is cybernetic in spirit, it doesn’t qualify as a complete cybernetic system, at least, not as usually deployed. An image classifier has no obvious agency. It doesn’t *act*, it just *is*, like any other mathematical function.

One notable exception, though, is a 2016 paper from researchers at Nvidia: *End to End Learning for Self-Driving Cars*,[[284]](#footnote-342) describing the DAVE-2 system (successor to DAVE, an earlier “DARPA Autonomous Vehicle”). The authors trained a convolutional neural net to “map raw pixels from a single front-facing camera directly to steering commands.”[[285]](#footnote-343) Their abstract continues, “This end-to-end approach proved surprisingly powerful. With minimum training data from humans the system learns to drive in traffic on local roads with or without lane markings and on highways. It also operates in areas with unclear visual guidance such as in parking lots and on unpaved roads.”

This appears to be a remarkably pure realization of Wiener’s cybernetic vision, courtesy of the universal function approximating power of a convolutional net. One can think of it as a much more sophisticated *Palomilla*, with real eyes and 27 million synapses rather than… by my count, five. (For rough comparison, a honeybee probably has about a billion synapses, and a mouse has about a trillion, though biological synapses are not interchangeable with model parameters.)

The nonlinearity of DAVE-2’s learned function allows for rich, contextually dependent behavior without explicit “if/then/else” logic; lane markings, guard rails, and other cars influence driving in all the ways one would hope. Although the network’s only output is a steering wheel angle, if we were to perform “neuroscience” on this network after training, we would undoubtedly find neurons that respond specifically and selectively to all such objects, much like the visual system neurons Livingstone recorded from during her work in Hubel’s lab.

There are still important missing elements from this picture, though. Notably, the input to the convolutional net is only a single frame of video, and the output only controls steering, not gas or brakes. No prediction of the future is involved. Since the network is run at thirty frames per second, vehicle speed is kept steady by a human copilot, and steering is forced to be reasonably smooth by mechanical inertia, the system undoubtedly *feels* dynamic, but the learned model has no dynamics. There’s no possibility here for anything like planning or memory, only instantaneous reflex-like response, fully determined by prior offline training. Neither is the model modeling *itself* in any way; its *Umwelt* is purely external, just a view of the road at that moment in time.

Rosenblatt’s original (and cybernetically informed) vision was considerably broader. He emphasized that what we now simply call the “perceptron” was in fact only a “momentary stimulus photoperceptron, […] the most elementary device which could be built to demonstrate the general principles of this type of system.” Not only could other sensory modalities also be imagined, but “temporal pattern perceptrons” would have the ability to “remember temporal sequences of events, rather than transient momentary images, such as would be obtained from a collection of isolated frames cut from a strip of movie film.”[[286]](#footnote-344)

Most of this remained on the drawing board, but interestingly, even the Mark I momentary stimulus photoperceptron had been implemented as a physical device with real dynamics, able to learn by adjusting its parameters based on feedback at any time. Today’s neural nets generally don’t share this property. Even if they act as controllers in a feedback loop, as in a self-driving car, they do not alter themselves in a feedback loop during operation, the way we do. Instead, they are trained offline on static data using batch processing, much the way ENIAC and the other early computers crunched away on big problems overnight (or for many nights).

Thus, the field we call “machine learning” still doesn’t generally produce systems that *learn*, but only systems that *act*. It may be hard to consider something “alive” if it can’t be durably affected by anything it experiences, and for many, it may also be hard to call something “intelligent” if it can’t learn from those experiences.[[287]](#footnote-345) In these crucial respects, we’re still working toward the ambitious vision laid out by Wiener, Rosenblatt, and the other early cyberneticists in the 1940s and 50s.

# IV. Learning

## Embedding

What do the “distributed representations” of concepts in a neural network actually look like? Consider a perceptron, trained in the usual supervised manner to recognize bananas on sight, in all their variations and orientations, against any background. Normally this would involve a final softmax or one-hot layer in the neural net containing a “banana neuron”; since we need at least two neurons for such a one-hot layer, let’s also assume there’s a “not banana neuron” that lights up in response to everything else. Supervised training would involve curating many images of either bananas or something else, each labeled with a single bit specifying which of the two neurons should activate. By the end of the training, we’d have a banana detector.[[288]](#footnote-348)

Thankfully, labeling a gajillion images with the banana/not banana bit turns out to be unnecessary. Training the banana detector mostly involves getting it to learn how to see *generically*—that is, developing a general sense of how perceptual invariance works in the visual world, based on the correlations among images, regardless of their labels. One could train the network to learn these correlations without *any* labeling by, for instance, blacking out random parts of the images, and requiring the network to fill in, or “inpaint,” the blacked-out parts as accurately as possible. Any neural net that can do this well, for a large and varied set of images, will certainly have learned those correlations.

Interestingly, that implies such a network will have learned how to recognize *everything* represented in the images—apples, fire hydrants, Siamese cats—not just bananas. It will know how to distinguish figure from ground, it will understand depth of field, it will recognize colors, and it will be able to distinguish ripe from unripe fruit. The evidence: if you black out half of a banana, it will fill in the other half, and the ripeness of the filled-in half will match the ripeness of the visible half.

Some questions arise, though:

* All sorts of knowledge may be latent in this neural net, but how would we read it out? That is, how could one turn this pixel inpainting model into an actual Siamese cat detector, banana detector, or fruit ripeness detector?
* It’s nice to get rid of all that banana/not banana labeling, but blacking out random regions of a sheaf of arbitrary images and training the model to fill them in still seems like a highly artificial task. What does this have to do with how learning really works in brains?
* Can we really call what such a model does “fruit recognition,” let alone “fruit understanding”—isn’t it just filling in pixels? What does this have to do with understanding what the color “yellow” really *is*, what it’s like to *eat* fruit, where it grows and how to pick it, and so on?

These final questions touch on what philosophers call “qualia”: what it is *like* to experience things, such as the ineffable “yellowness” of yellow (though for some reason, red always seems to be the go-to color). We’ll take on qualia more directly in Parts VIII and IX.

But let’s begin with the first question, which is of a more practical bent. Recall from Part III that in a convolutional net, higher-order features emerge hierarchically as combinations of lower-level features. As activations propagate to higher layers in the network, the features represented become increasingly invariant, that is, more semantically meaningful. You can see how that would *have* to be the case by considering the final, or topmost, layers of the network in the supervised case.

The final layer of the model trained with supervised learning, remember, consists of two softmax neurons, banana and not banana. Suppose that the second-to-last layer has 128 neurons. This is often referred to as an “embedding” layer. One can think of the activations of these 128 numbers as the coordinates of a 128-dimensional “embedding space.”

Yes, it’s a bit mind-bending to think about high-dimensional spaces, since we live in only three dimensions, so let’s develop a bit of intuition for how this works. Specifying a point on a 2D surface requires, by definition, two numbers—such as *x* and *y* on a graph, or latitude and longitude on a map. Specifying a point in 3D space requires three numbers; in a rectangular room, for instance, it could be *x* and *y* coordinates on the floor, and a height above the floor, *z*. So why not 128 numbers specifying a point in 128 dimensions? It’s like 3D, only … more. Just pretend it’s 3D. (Everyone does, secretly.)

When an image is presented to the net, then, it will correspond to a point in this 128-dimensional “embedding space,” whose coordinates can be read out from the embedding layer. It’s possible to visualize this, although mathematical tricks have to be used to reduce those many dimensions down to 3D or 2D. Many research papers have used such visualizations over the years, though the most striking one I’ve seen was made by a Turkish-American artist, Refik Anadol.

Refik got his start in AI art at our Artists + Machine Intelligence (AMI) program, back in 2016. He has since become famous. One of his first large-scale AI art commissions, Archive Dreaming, was a 2017 collaboration with SALT, an art and research institution based in Istanbul. SALT houses a major archive of photographs, architectural drawings, maps, posters, correspondence, and ephemera, dating from the last century of the Ottoman Empire to present-day, and they have been busily digitizing it. Refik used neural nets to generate embeddings for 1.7 million visual documents in the archive, and created a room-sized immersive visualization that allowed one to swoop through all of these documents, each rendered as a thumbnail hanging in the void. In that embedding space, visually similar objects cluster together, allowing one to get a sense of the archive as if it were a galaxy with 1.7 million stars, all arranged in space, organized by a cosmic Dewey decimal system.

## Transfer

Suppose Refik had used the banana recognition net to generate his embeddings, and the SALT archive had included a bunch of images of bananas (for all I know, it might). We know that there’s no banana neuron in the embedding layer—remember, bananas are only recognized in the final layer—which means that there’s no single coordinate in the embedding space representing banana-ness. However, bananas *must* be easily recognizable in the embedding space, because the weights connecting those 128 neurons to the banana neuron in the final layer define a “hyperplane,” with bananas on the far side, where the weighted sum exceeds some threshold, and everything else on the near side.

In 3D: just think of the hyperplane as an ordinary plane, a big flat sheet of paper. It must be possible to orient that 2D sheet of paper in the 3D room, such that the points representing banana images are all on one side, and the non-banana points are on the other. The reason? The banana neuron’s weights define a “vector,” or *direction*, in the embedding space. Think of that vector as an arrow coming straight out of the sheet of paper, pointing toward banana-ness.[[289]](#footnote-350)

What, then, do these 128 weights mean? First, there are presumably a lot more “not bananas” than bananas in the world, and even if an image *does* contain a banana, a great many image features (such as the background, the shooting angle, and the lighting) must be explicitly ignored to make the banana judgment—that is, they are *variations* that must be zeroed out to compute a perceptual *in*variant. Many weights will therefore be zero, even though much computational work may have gone into calculating the activations of those ignored neurons.

Other neural activations will be combined, with a corresponding collapse of specificity. Consider, for example, that unripe and overripe bananas look very different; therefore different ensembles of features will likely have been combined to calculate one or more “banana of a certain ripeness” neurons in the embedding layer, but this ripeness information will be explicitly nulled out to create the higher-order ripeness-invariant banana neuron. The same will be true of bananas in different orientations, or under different lighting conditions. Such observations give us some intuition regarding why trained neural nets are generally “sparse,” meaning that there tend to be many zero or near-zero weights, as well as (given any form of lateral inhibition) a lot of zero or near-zero neural activations.[[290]](#footnote-351)

These observations also explain why “transfer learning” works: the ability to retrain a network to do a related task, using much less labeled data than would be required to train from scratch.[[291]](#footnote-352) If we wanted our banana network to detect only ripe bananas, for instance, all it would take is a bit of tweaking of those final 128 weights, to exclude the unripe cases. Even a single instance each of a ripe banana (“like this”) and an unripe one (“not like this”) would suffice to make the change.

Although the reason may be less obvious, the same trick will work nearly as well to turn the banana detector into an apple detector. Apples are also fruits, though red and round instead of yellow and long. The embedding layer will already contain neurons representing apple properties that are either like or *un*like banana properties. Thus apples will be easy to learn.

In fact, we could leave the original banana network as is, and simply add an apple neuron alongside the banana neuron. We could even learn the weights for the apple neuron from a lone exemplar, by just setting them based on the activations of the embedding layer in response to a single apple image.[[292]](#footnote-353) (An average based on a few apples would be better, but a single one will do.) The network will subsequently recognize “another of those.”[[293]](#footnote-354) This is more or less what we do in adulthood when we learn about an unfamiliar object category from a single exposure. It’s called “one-shot learning,” and can be thought of as a special case of transfer learning.

## Representation

Transfer learning has its limits. If the banana network had been trained entirely on natural scenes and we then tried to use it to distinguish among different brands of trucks, it would do a poor job, because the embedding layer would not adequately represent the needed features. Even if trucks had been included among the “not banana” images, the relevant details might get discarded early in the network, because those details are irrelevant to banana detection—if it’s any kind of vehicle-like thing, it’s not a banana, end of story.

A similar effect explains why people who grew up in a racially homogenous environment tend to be so poor at distinguishing individuals of other races. We learn our face embeddings young, and they are exquisitely sensitive—but only within the learned statistical distribution. Hence the all-too-common sayings, usually considered racist, that all people of “foreign” race X “look alike.”[[294]](#footnote-356) Yet in a very literal sense, if our brains weren’t exposed to “foreign” faces as children, then, for us, it’s true. As adults, the best we can do is to understand the limitation, work on it as best we can (ongoing improvement *is* possible[[295]](#footnote-357)), and try not to get in trouble. Less excusably, the same problem is evident in convolutional nets trained to recognize faces based on inadequately diverse datasets.[[296]](#footnote-358)

Learning, then—whether supervised or not—is mostly a matter of “representation learning”—that is, learning how to embed. Physicist and AI researcher Brice Ménard and colleagues have powerfully demonstrated that representation learning in multilayer perceptrons is universal, regardless of how they’re trained.[[297]](#footnote-359)

One could even say, then, that *learning is also learning to learn*. That is, once a suitable generic embedding or representation has been learned, associating a label with a specific point or region in the embedding space becomes trivial.

So, although our brains take a long while to learn how to see and how to understand language, once we have these representations down pat, we become very adept one-shot learners. If you’re shown a kiwi for the first time, you’ll henceforth immediately be able to recognize that furry little fruit. A perceptron trained on general vision tasks can do the same, with the addition of a kiwi neuron to the output layer trained in one shot. That’s so much easier than labeling lots of kiwi (and not kiwi) images and training on them from scratch.

If the original perceptron is trained to inpaint pixels rather than detect bananas, its high-level layers will be general embeddings, equally good at representing cats, trucks, bananas, kiwis, and anything else in the (unlabeled) training data. Moreover, there will be neurons specifying the ripeness of fruits, the makes of trucks, the breeds of cats and the corresponding color of their eyes. When the image is of a cat, its eye color will be guessed if need be, for instance if we’ve blacked it out—otherwise, those pixels could not be filled in. It follows that latent eye color knowledge is also present if the cat happens to be facing away from the camera.

You can see how powerful unsupervised learning is. That’s why, in recent years, researchers have been shifting away from the old supervised approach. Not only are the large numbers of labels we used to rely on unnecessary; they actually impede learning, in that the representations learned by supervised models can get away with being less robust, since they’re only trained on a particular classification task. All those labels also introduce additional errors and biases, and are labor intensive—at best, boring, and at worst, exploitative.[[298]](#footnote-360)

The hourglass-shaped “masked autoencoder” is a (now) classic way to construct an unsupervised perceptron-style model.[[299]](#footnote-361) Starting from a “retinal” input layer of, for example, 512×512 color pixels, a sequence of progressively narrower layers culminates in a bottleneck—say, 128 values—which then expands back into an output layer of the same shape as the input. The input omits masked pixels, which may amount to 75% of the total, while the output reconstructs (or “hallucinates”) the whole image. That bottleneck acts like—perhaps you guessed it—an embedding layer.

We can also interpret the bottleneck layer as a form of image compression (as described in Parts I and II). Compression is, in a deep sense, closely related to language. It extracts the meaning from a signal, its semantic essence, from which the original can be reconstructed, at least statistically. If you can see just enough of the cat’s body to know it’s a Siamese, then this knowledge suffices to imagine what its head will look like, where its eyes will be, and that those eyes will be blue. The strands of fur might all lie in different places in the reconstruction, but if it’s done convincingly, a judge confronted with the original and the reconstruction wouldn’t be able to guess which is which. The details would vary, but the semantic content would be the same.

## Sparsity

Our visual cortex is not so unlike a perceptron, but it doesn’t get trained by reconstructing arbitrary collections of partially masked static images. What our brain is exposed to, instead, is a continual stream of “video” from our eyes. The brain also controls our gaze and the position of our bodies, so this isn’t a passive stream, but an actively generated one.

Still, masked autoencoder training isn’t as far from our experience as it may seem. The retinal fovea, where we can resolve enough detail to read, is barely big enough to make out a handful of printed words. Even that region is noisy and jittery. The wider visual field is much lower in resolution and crisscrossed with blood vessels.[[300]](#footnote-363) This is certainly different from our *impression* of what we see.

A dramatic series of eye tracking experiments done in several labs between the 1970s and the 1990s illustrates this.[[301]](#footnote-364) A subject sits in front of a display showing a grid of letters. Wherever their eyes are looking, the letters reveal stable underlying text, but everywhere else, the letters are randomized. To an onlooker, the entire screen looks like an illegible jumble. But, if the window of clear text is a mere 18 characters wide—about 3 characters to the left of the fixation point and 15 to the right, in a language like English that is read from left to right—then to the subject, the whole page of text looks perfectly clear and steady. This is what neuroscientists mean when they say we “hallucinate” the world into existence based on the sketchiest of signals.

From the moment we open our eyes, soon after birth, and begin swiveling them around using our “exterior ocular muscles,” we’re putting our visual systems through an unsupervised training regimen much like that of a masked autoencoder. With every “saccade” (meaning rapid eye movement) we have the opportunity to test whether a previously unresolved part of the environment looks as we had predicted or “inpainted” it, or not, and to improve our model accordingly.

After a while, we learn to saccade to the spots where our uncertainty is highest (or where it matters most—was that a tiger nosing through the underbrush?), thereby constantly pinning our current reconstruction down to reality. Of course, the moment we saccade away, uncertainty in that spot begins to grow again, reminiscent of the way an unobserved particle begins to blur in quantum mechanics.

If reconstruction of the visual world is an actively maintained and constrained hallucination, vision *is* that reconstruction. Vision is *not* the raw stream of sensory input from our eyes—the input stream acts more like an error-correction signal.

Using nothing but prediction, then, we will rapidly learn a very good unsupervised model of our visual environment. It will include sparse neural representations that have just the kinds of high-level semantic meanings masked autoencoders learn, that Hubel and Wiesel recorded in cat visual cortex, and that have even been recorded from the brains of awake humans.[[302]](#footnote-365)

Your cortex doesn’t have any single “banana” neuron. However, the idea that we might have unique neurons in our brains corresponding to highly specific percepts or memories has a long history. In 1967 the term “grandmother cell” was coined, somewhat tongue-in-cheek, to refer to a hypothetical neuron that activates only in response to your grandmother.[[303]](#footnote-366) Margaret Livingstone’s recollection of a cat visual neuron that, after an extensive search, was found to respond only to a yellow Kodak film box can be interpreted as evidence of something resembling a grandmother cell. A famous 2005 *Nature* paper, *Invariant visual representation by single neurons in the human brain*,[[304]](#footnote-367) documented a neuron in a human subject that seemed to respond only to Jennifer Aniston… and another neuron that responded only to Jennifer Aniston and Brad Pitt together![[305]](#footnote-368) The researchers also found a Pamela Anderson neuron, which responded specifically not only to pictures of Anderson (including a caricature of her), but also to her name written out—and not to any other name or string of letters they could find.

As suggestive as these findings are, remember that there was no supervised learning process forcing a (hypothetical) one-hot layer somewhere in your brain to activate a single “banana” or “grandmother” or “Pamela Anderson” neuron, to the exclusion of all else. Such single points of failure would make your brain far too fragile. Would you suddenly be unable to recognize your grandmother if “her” neuron died one day, or just failed to fire? And if you’re using up neurons on such granular concepts as Jennifer and Brad together, do you risk running out?

No: in a sparse distributed code, which gets learned by a set of high-level neurons (analogous to the embedding layer) with neither supervision nor centralization, a whole set of neurons in your brain lights up in response to your grandmother, or a banana, or a specific movie star couple. Even if only one in ten thousand light up due to extensive lateral inhibition, that’s still nearly ten million neurons. So if a few of *those* remain silent, it probably won’t matter. (And, while the code is sparse, it’s not *too* sparse, or there’s no way a neuroscientist poking around semi-at-random in the brain could ever get lucky enough to find a Pamela neuron, let alone also a Jen neuron, and even a Jen and Brad neuron.)

As a bonus, sparse distributed representations are vastly more efficient than a one-hot code. With 128 neurons, a one-hot code can only represent 128 things. In a sparse code that involves 16 of them lighting up at a time (say), there are “128 choose 16” or about 93 quintillion possibilities.[[306]](#footnote-369) With larger numbers of neurons, the available combinations are virtually limitless, even accounting for plenty of redundancy.

## Meathead

Let’s now switch from a perception-centric to a motor-centric perspective. In doing so, we will also be moving farther away from perceptrons, and closer to real critters.

Suppose a neuroscientist records from neurons somewhere along the complex pathway from an animal’s visual cortex to its exterior ocular muscles, postulating, quite reasonably, that these spike trains issue motor commands for eye movement. The question is: how will the neuroscientist “decode” this command stream? The answer, of course, is to simultaneously measure eye movement, and build a model (these days, usually using an artificial neural net) that, given the spike trains, can predict eye movement. If it works reasonably well, *voilà*—an eye movement command stream decoder, and thereby proof that what is being recorded *is* a command stream!

However, what the neuroscientist has actually created isn’t so much a command stream decoder as an artificial brain region which—if the predictive brain hypothesis is right—performs exactly the same kind of prediction task that every *other* brain region performs! The “decoding” will “work” reasonably well, to the extent that

1. most of the relevant information used by downstream neurons to carry out *their* prediction is captured by the recording,
2. the timescale allows any feedback loops to be ignored,
3. the neuroscientist’s model is sufficiently powerful to proxy the downstream brain region, and
4. that downstream region isn’t actively trying to evade prediction via dynamical instability and randomness (per Part III).

In short, given the feedback loops present everywhere in the brain, the “command stream” interpretation is arbitrary. If various brain regions are trying to predict each other, then there’s no inherent hierarchy determining which one is giving orders, and which one is receiving them.

Neuroscientists and AI researchers may find this counterintuitive due to a sort of internalized anthropomorphic metaphor. Neurons have a cell body located near the “dendrites,” which receive inputs, and a long process, the “axon”—I’ve called it “wiring”—sending neural spike trains to other neurons (or muscles) elsewhere in the brain or body. We tend to subconsciously think of neurons as little people, doing their computation near the body or “head,” and deciding what signal to send out along their “tail” to a downstream target. But here, as it were, the tail wags the dog.

The idea that the *target* could be in charge seems odd, given where the decision about when to spike appears to be made, and the direction in which that spiking signal appears to travel. Using old-fashioned philosophical language, one could say that we imagine the head is the “agent,” and the tail is the “patient,” meaning the passive recipient of the agent’s actions.[[307]](#footnote-371)

We need to keep in mind, though, that neural circuits are shaped by learning. Learning requires signals, and more importantly intentions, to flow the *other* way; hence “backpropagation” for training artificial neural nets. Neuroscientists don’t usually see learning take place in electrophysiological experiments, because it tends to happen more slowly than realtime electrical activity. Hence, we miss it, just as we fail to notice the slow movements of plants.

But learning is what determines what each neuron actually does: which signals it gathers to send on and which it ignores. So a neural cell body with a long axon projecting to a different brain region doesn’t actually serve the region near the cell body. In fact, it will have little proximal effect on that local region when it fires. It is, rather, an outpost, a listening station, whose learned function is to serve the community of neurons at the far end of the axon.

This is most obvious for the neurons in our sensory periphery, which we can think of as effectively in contact with our environment. There are pressure and temperature sensors in our skin projecting to “somatosensory” cortex; light-sensitive cells in our retinas projecting to visual cortex; hair cells in our inner ears projecting to auditory cortex.[[308]](#footnote-372) These cells are best thought of as information-gathering outposts.

You wouldn’t say that a sensor on your fingertip is the “boss” sending commands to drive your behavior, but rather, that it’s one of the inputs *informing* your behavior. (Although under certain circumstances—such as touching a hot stove—it will indeed *appear* to be the boss, because your interests as an organism are best served by the fastest possible reaction time, pushing initiation of this action as far upstream as possible.) If we consider every brain region to be similarly capable of autonomy to one degree or another, then the idea of every such region having outposts in other regions makes sense.

What about the muscular end of things? This is where the more traditional “upstream agency” picture seems undeniable. The idea of our brain being in charge of our muscles is intuitive, while the idea of our muscles being in charge of our brain sounds, on the face of it, absurd. As everybody knows, brains are smart, and muscles are dumb; otherwise, “meathead” might be a synonym for “genius.” And on their own, muscles are seemingly passive, which is why if you get a nasty knock on the head, you will fall down. While you’re unconscious, your muscles will go limp and do nothing. Your body will be like a marionette with its strings cut.

Or will it? Despite its intuitive appeal, this view, too, has things backwards. Consider your heart, gut, and blood vessels—they’re muscles too. Luckily, they don’t stop contracting when you’re unconscious or asleep. They are perfectly capable of running without the brain.

Indeed, the earliest motile life forms, dating back at least to the Ediacaran (635–538.8 million years ago, just prior to the Cambrian), did not have brains, but they certainly had muscles. Muscles are useful even for the simplest “sessile” animals—those anchored to rocks on the seafloor—as they allow rhythmic pumping movements to filter seawater over or through their bodies to sieve out nutrients. Jellyfish perform such rhythmic pumping while free-floating, enabling them to swim. The requisite coordinated muscle movement involves traveling waves of electrical activation.

Coordination of an oscillatory contraction is, incidentally, the simplest possible form of mutual prediction. Each cell can act independently like an oscillator, but by aligning the phase and frequency of its oscillation with its neighbors, it is in effect predicting *when* those neighbors will contract, and striving to contract at the same time—to “autocomplete” the movement. Such “phase synchronization” is the way the heart works, as well as peristalsis—the coordinated toothpaste tube squeezing maneuver that moves food along the gut.

It’s also the way giant swarms of certain varieties of firefly create an otherworldly synchronized Christmas light spectacle.[[309]](#footnote-373) Both the synchrony of a firefly swarm and, perhaps more importantly, the speed with which it can be established depend on the ability of the fireflies to see each other. They’re able to coordinate quickly and stay in synchrony because they can see not only their immediate neighbors, but also more distant fireflies halfway across a clearing.[[310]](#footnote-374)

Similarly, without the long-range connectivity afforded by neural wiring, every ripple of electrical activity in a jellyfish would need to pass through a very large number of cells to make a complete circuit around the animal’s circumference, and contractions would be correspondingly sluggish. Thus, in animals with distributed nerve nets, including jellyfish and *Hydra*, the most obvious function of neurons is to allow more coordinated synchronization of muscular contraction across longer distances—just as the visual systems of fireflies allow them to see what other fireflies some way off are doing. In other words, it’s not a stretch to think of the earliest neurons as providing internal “sensory systems” for muscle cells, allowing them to better predict each other (and themselves) as they contract in synchrony.[[311]](#footnote-375)

So how do brains arise?

Local sensing of the *external* environment will generally be helpful to the muscles too. Pretty much everything that is alive pulls away from noxious stimuli, just as our hands do from a hot stove—and in general, as in our case, using a fast, local feedback circuit. Chemical receptors—implementing chemosensing, or what we call “taste”—are the oldest and most ubiquitous environmental sensors. Recall that even bacteria have them, since in a watery medium and at the smallest scales, floating molecules *are* the environment.

Now, consider the saliency of such receptors for an animal that can both propel itself forward *and* turn via rhythmic muscular contraction—something worm-like, for instance. While a coral polyp lives anchored to one spot and must make the best of whatever washes over it, a crawling worm continually makes decisions that will determine its future environment, hence its survival prospects. In this respect, it is much like a swimming bacterium. Unlike a swimming bacterium, though, it’s long and bilaterally symmetric, “bilaterian,” rather than tiny and cylindrically symmetric. As a result it can steer left or right, not just run or tumble—an innovation critical to life on land.[[312]](#footnote-376)

Receptors at the front end of the animal are now especially important, because they are the first to detect an encounter with something tasty or aversive. Unlike the point-like bacterium, space and time are now meaningfully related; one could say the front end of a worm lives in the future, while its back end lives in the past. Muscles throughout the body will want to know about the future, and the ones on the right and the left will want to behave differently in order to turn away from noxious things, and toward food.

So, through Darwinian selection, muscle cells will evolve to send “sensory” neurons not only to muscles elsewhere in the body for improved coordination, but to the leading end, to detect any important changes in chemical concentrations up there. When a bunch of neurons from muscles all over the worm’s body wire themselves up to the front end, the resulting knot of neurons results in “cephalization”: the beginnings of a brain in the head.

## Neuromodulators

The earliest bilaterians faced an immediate need to modulate their behaviors on timescales longer than the activations of individual neurons or muscles. Even chemotactic bacteria need a long integrating timescale to decide whether the food level in the environment is going up (in which case they should keep swimming), or down (in which case they should tumble, and change direction). Bilaterians, then, needed to turn this integrated environmental chemical signal at the head into an *internal* chemical signal that could itself be sensed by other neurons.

They did this via neuromodulators, chemical signals that accumulate and reabsorb gradually, affecting entire populations of neurons at a time. In terms familiar from Part II, we can think of these neuromodulators as hidden state variables (*H*) with long timescales—not permanent, but longer-lasting than any momentary input (*X*), or action (*O*).

Dopamine and serotonin, neuromodulators that remain crucial to our brains today, date back to these earliest bilaterian nervous systems.[[313]](#footnote-378) The “nearby food sensors” in a worm’s head release dopamine, triggering feeding behavior, which looks like constant turning to exploit the local environment—just as the increased rate of tumbling that causes a bacterium to stick around and feed. We can understand dopamine, then, as turning the food-outside signal into a time-smoothed internal signal, allowing the worm’s muscles to know that they should cooperate to turn the animal in place to keep eating whatever delicious thing is in the immediate vicinity.

While dopamine has sometimes been interpreted as a “pleasure” signal, this isn’t quite right. It’s true that being in food will make the animal happier than *not* having any food, but really, dopamine is best interpreted as a predictive error signal, even in this very simple setting. If you’re a worm—or any living being, really—you need to predict the presence of food in your future environment. You’ll be “happy” when you are swimming *toward* food, meaning the food level at your leading end is rising.

The most apt feeling to associate with dopamine is probably *anticipation*. This seems consistent with the subjective reports of human patients who, in a series of ethically dubious experiments from the 1960s, were wired up to directly stimulate dopamine production deep in their own brains. One patient, while mashing the dopamine button, explained that “it was as if he were building up to a sexual orgasm. He reported that he was unable to achieve the orgastic end point, however, explaining that his frequent, sometimes frantic, pushing of the button was an attempt to reach the end point. This futile effort was frustrating at times and described by him on these occasions as a ‘nervous feeling.’”[[314]](#footnote-379)

On the other hand, when the dopamine-producing neurons of rats are destroyed, the rats become passive and starve to death, even if there is food “literally under their noses.”[[315]](#footnote-380) If food is placed into their mouths, they eat it with evident pleasure, but no matter how hungry they get, without dopamine, they aren’t spurred to action.

Now, put yourself back in the worm’s place. If you realize your prediction is being violated by a *decline* in the ambient food near your head, you will want to turn. The turn, even if random (as with a bacterial tumble), will reorient you toward a direction where the food level might once again go up—and if that doesn’t work, just keep turning until it does. In an area of peak food, *every* direction leads to a decline, so the animal will continue turning (or tumbling) in place, which is, as it happens, the optimal food exploitation strategy.

Serotonin neurons serve the converse function. They sense food in the animal’s throat rather than in the environment. As serotonin builds up over time, the message becomes: “enough, I’m satiated.” The effect of dopamine will be quelled, along with the impulse to move at all. Postprandial torpor will set in, the better to digest.

Thus, we can crudely characterize dopamine and serotonin as the chemicals associated with “wanting” (dopamine) and “getting” (serotonin). In early bilaterians, these chemicals affected behavior directly by providing an internal environment that the neurons could sense, and in turn provide to the muscles as sensory input.

This developmental stage appears to be preserved in *Acoela*, an ancient order of small marine worms that diverged from other animals more than 550 million years ago. With very simple body plans, *Acoela* have no gut and no circulatory or respiratory systems. They move (either swimming or crawling between sand grains on the sea bottom) by means of “cilia,” little hair-like projections covering their exterior, whose movements appear to be under local control.

In addition to a distributed nerve net, though, these worms have a sort of “brain cap,” an aggregation of neurons at the front end, coinciding with sensors including a simple eye.[[316]](#footnote-381) They can use complex repertoires of sensory-guided behavior to actively hunt, modulated by dopamine and serotonin. Yet the “brain” seems not to be highly organized. If one of these worms is cut in half, then, like the neurally decentralized *Hydra*, each half can readily regenerate into a whole animal. Signaling molecules exchanged among the muscle cells appear to orchestrate the patterning and re-generation process.[[317]](#footnote-382)

## Bootstrapping

In animals with simple distributed nerve nets, like *Hydra*, there’s little evidence of learning in any form that a behavioral experimentalist would recognize, though every cell does continually regulate its own biophysics to ensure that it remains responsive to whatever signals it receives—a form of local learning.[[318]](#footnote-384) This is consistent with the idea that these earliest nerve nets serve only secondarily for sensing the environment, having first evolved to help muscles coordinate coherent movement.

Rudimentary behavioral learning arises the moment there’s anything like a brain, because at this point, neurons in the head must begin jointly adapting to changing conditions in the outside world. Every connection or potential connection between one neuron and another offers a parameter—a degree of coupling—that can be modulated to suit the circumstances, even if the basic “wiring” is genetically preprogrammed.

To see why, let’s take the neuron’s point of view, and imagine that it is simply trying to do the same thing any living thing does: predict and bringing about its own continued existence. Some aspects of this prediction will certainly have been built in by evolution. For example, since dopamine is a proxy for food nearby, the neuron will try to predict (and thereby bring about) the presence of dopamine, because its prolonged absence implies that the whole animal will eventually starve—bringing an end to this one neuron, along with all of its cellular clones. Even a humble cell has plenty of needs and wants beyond food, but without food, there is no future.

Therefore, if the neuron is not itself dopamine-emitting, but its activity somehow influences dopamine in the future, it will try to activate at times that increase future dopamine. Aside from neuromodulators like dopamine, the neuron’s inputs come either from other neurons or, if it’s a sensory neuron, from an external source, such as light or taste. It can activate spontaneously, or in response to any combination of these inputs, depending on its internal parameters and degree of coupling with neighboring neurons. At least one of its goals thus becomes fiddling with its parameters such that, when the neuron fires, future dopamine is maximized.

I’ve just described a basic reinforcement learning algorithm, where dopamine is the reward signal. As brains became more complicated, though, they began to build more sophisticated models of expected future reward, and accordingly, in vertebrates, dopamine appears to have been repurposed to power something approximating a more sophisticated reinforcement learning algorithm: “temporal difference” or “TD” learning.

TD learning was invented (or, arguably, discovered) by Richard Sutton, while he was still a grad student working toward his PhD in psychology at UMass Amherst in the 1980s. Sutton aimed to turn existing mathematical models of Pavlovian conditioning[[319]](#footnote-385) into a machine learning algorithm. The problem was, as he put it, that of “*learning to predict*, that is, of using past experience with an incompletely known system to predict its future behavior.”[[320]](#footnote-386)

In standard reinforcement learning, such predictions are goal-directed. The point is to reap a reward—like getting food, or winning a board game. What makes this hard, though, is the “credit assignment problem”: a long chain of actions and observations might lead to the ultimate reward, but creating a direct association between action and reward can only enable an agent to learn the last step in this chain.

Sutton’s central insight was that “Whereas conventional prediction-learning methods assign credit by means of the difference between predicted and actual *outcomes*, [TD learning] methods assign credit by means of the difference between temporally successive *predictions*.”[[321]](#footnote-387) By using the change in estimated *future* reward as a learning signal, it becomes possible to say whether a given action is good (hence should be reinforced) or bad (hence should be penalized) *before* the game is lost or won, or the food is eaten.

This may sound circular, since if we already had an accurate model of the expected reward for every action, we wouldn’t need to learn anything further; why not just take the action with the highest expected reward? As in many statistical algorithms, though, by separating the problem into alternating steps based on distinct models, it’s possible for these models to take turns improving each other, an approach known as “bootstrapping”—after the old saying about the impossibility of lifting oneself up by one’s own bootstraps.

In the TD learning context these models are often described as the “actor” and the “critic”; in modern implementations, the actor’s model is called a “policy function,” and the critic’s model, for estimating expected reward, is the “value function.” These functions are often approximated using neural nets. The critic learns by comparing its predictions with actual rewards, which are obtained by performing the moves dictated by the actor, while the actor improves by learning how to perform moves that maximize expected reward according to the critic.

TD learning eventually figures out how to perform well, even if both the actor and critic are initially entirely naïve, making random decisions—provided that the problem isn’t too hard, and that random moves occasionally produce a reward. Hence an experiment in the 1990s applying TD learning to backgammon worked beautifully,[[322]](#footnote-388) although applying the same method to complex games failed, at least initially.

## Beyond reward

Around the same time, at the University of Fribourg’s Institute of Physiology, Wolfram Schultz’s lab had been studying the relationship between motor function and Parkinson’s disease, which was known to compromise movement via dopamine depletion.[[323]](#footnote-390) In a typical experiment, Schultz and colleagues would record from single dopamine-releasing neurons in the brains of macaques while they performed simple motor tasks, which they needed to learn via Pavlovian conditioning.[[324]](#footnote-391) A thirsty monkey, for instance, might need to learn which of two levers to pull in response to a flashing light to get a sip of juice. The researchers made the following observations:

1. Dopamine neurons normally spike at a moderate background rate.
2. When the monkeys first stumbled upon an action producing the sugary drink, the spiking rate of these dopamine neurons rose.
3. Once the monkeys figured out the association between the visual cue and the reward, extra dopamine was no longer released when the treat came, but was released earlier, when the visual cue was presented. This coincided with the monkeys licking their lips, akin to the salivation of Pavlov’s dogs.
4. If, following the visual cue, the treat was withheld, then activity of the dopamine neurons subsequently *decreased*—that is, they went quiet relative to their background rate.

When Peter Dayan and Read Montague, postdocs in Terry Sejnowski’s lab at the Salk Institute in San Diego, saw these results from Schultz’s group, they realized that dopamine was acting precisely like a temporal difference learning signal.[[325]](#footnote-392) This is the signal whereby the brain’s “critic” tells the “actor”: please reinforce whatever behavior you’re doing now, because I predict it will lead to a future reward. Long sequences of actions that ultimately lead to a reward can be learned this way, with the TD learning signal shifting earlier and earlier during the learning process.

The repurposing of dopamine from a reward signal to something like a temporal difference reinforcement learning signal follows naturally from the growth of brain structures both “upstream” and “downstream” of the dopamine-releasing neurons. Remember that even among the earliest bilaterians, dopamine no longer represents food, but *nearby* food. In this sense, dopamine is already a prediction of food, not a food reward in itself. Predicting dopamine is thus *a prediction of a prediction of* food.

A predictive symbiosis between neural areas upstream and downstream of dopamine will therefore result in the upstream areas being able to make higher-order predictions (hence longer-range forecasts), thus acting as an increasingly sophisticated critic or value function. Meanwhile, the downstream parts become an increasingly sophisticated actor, or policy function, smart enough to learn how to make better moves using these longer-range forecasts.

This explains the approximate fit between the TD learning paradigm and the role of dopamine in the brains of vertebrates.[[326]](#footnote-393) Like many primal feelings, “something good is within reach” is a simple, useful signal that a worm can infer directly from smell, and a larger-brained animal like us can be infer through a much more complex cognitive process. Dopamine is a useful signal for many parts of the brain, since they are all invested in producing good outcomes for the organism as a whole; hence dopamine signaling has been conserved for hundreds of millions of years, and its role has remained, if not identical, at least recognizable throughout those eons.

We should be careful not to interpret these experimental findings about dopamine as proof that the brain implements TD learning as formulated by Sutton. That can’t be the whole story. For one, we have ample evidence that humans, and likely many other animals, are even more powerful learners than the TD algorithm is. Advanced board games, for instance, are beyond the reach of TD learning. Additionally, recent experiments more closely examining dopamine activity suggest it encodes information beyond that of a straightforward TD error signal.[[327]](#footnote-394)

None of this should surprise us. Brain regions that symbiotically predict their environment and each other are not restricted to implementing simple learning algorithms, or communicating using cleanly definable mathematical variables, any more than human emotional expression is limited to a single dimension or natural language is restricted to logical grammar. Like every other approach to machine learning described in this book, TD learning is an elegant conceptual simplification that sheds light, but doesn’t illuminate every corner. It is neither a complete nor an exact representation of what the brain does.

In 2016, DeepMind’s AlphaGo model made headlines by achieving a major milestone in AI history. This program, based on a more elaborate descendant of TD learning (which we’ll explore in Part V), defeated reigning Go champion Lee Sedol in four out of five games.[[328]](#footnote-395) The news came amid reports of machine learning besting humans at a lengthening list of tasks previously considered “safely” beyond AI’s capabilities.[[329]](#footnote-396)

Despite this impressive showing, neurophilosopher Patricia Churchland pointed out fundamental limitations in the AI paradigm relative to what real organisms with real brains do. In response to AlphaGo’s victory, Churchland wrote an essay entitled *Motivations and Drives are Computationally Messy*,[[330]](#footnote-397) noting:

“The success of [artificial neural nets like AlphaGo] notwithstanding [ … their] behavior is a far cry from what a rat or a human can do […]. Maintaining homeostasis often involves competing values and competing opportunities, as well as trade-offs and priorities[…]: should I mate or hide from a predator, should I eat or mate, should I fight or flee or hide, should I back down in this fight or soldier on […]. The underlying neural circuitry for essentially all of these decisions is understood if at all, then only in the barest outline. And they do involve some sense of ‘self’ […].”

It’s true. A pure reinforcement learning approach, no matter how fancy the algorithm, cannot account for realistic animal behavior, because real animals (including us) are not optimizing for any one thing (except, maybe, when we’re playing championship-level Go). However, “real AI” may be simpler than Churchland—or any of us—imagined in 2016.

Mathematically, close relationships exist among the various known machine learning methods,[[331]](#footnote-398) and in the years ahead, ongoing theoretical work will likely unify them further. While this remains an active research area, I bet that a unified theory of learning can ultimately encompass evolution, brains, and (as special cases) established ML methods. Such a unification would:

1. Frame the central problem as active prediction of the future given the past;
2. Not distinguish between learning and evaluation or inference, but instead acknowledge that prediction must occur over all timescales;[[332]](#footnote-399)
3. Synthesize prediction with an extended theory of thermodynamics in the spirit of dynamic stability, as sketched for bacteria in Part II; and
4. Explain the nonzero-sum or symbiotic properties of mutual prediction.

This unified picture will shed light on the deep relationships between the theories of computing, machine learning, thermodynamics, and evolution.

While such a theory is still incomplete, this book offers a perspective on what nature *does* at every scale, from cells to societies: unsupervised sequence prediction. Predicted sequences involve continued existence. If they did not, then the organisms that made those (self-sabotaging) predictions wouldn’t have lived to pass on any of their learnings. We could call this “dynamically stable symbiotic prediction” (DSSP), though that’s a bit of a mouthful.

What does DSSP imply about the brain? Unlike TD learning or any form of pure reinforcement learning, DSSP does not imply that animals optimize any single reward, whether food or dopamine. On the contrary, such single-mindedness would not be conducive to either mutualism or survival in the real world; hence it would not be dynamically stable. Per Churchland, “Maintaining homeostasis often involves competing values and competing opportunities, as well as trade-offs and priorities.” This is even truer when taking into account the goodwill of others. There are many ways to be alive in this world … but all of them must involve continuing to exist in the future, and existence requires ongoing relationships.

Pulling these threads together and summarizing: as evolution built on the design of simple *Acoela*-like bilaterians, additional neurons appeared in their heads. This happened because neurons are themselves replicators, and like all replicators, they will colonize any favorable niche,[[333]](#footnote-400) in particular when that neural proliferation symbiotically helps the organism as a whole. Once this brain-knot began to grow, more complex neural networks developed, capable of handling more sophisticated sensory modalities, like hearing and compound or camera-style vision. By making use of widely broadcast neuromodulatory signals like dopamine, such brains were also able to predict increasingly subtle patterns relevant to the whole organism over ever-longer timescales.

Thus animals acquired the means to develop behaviorally relevant models of *other* complex animals, resulting in the kinds of early “social intelligence explosions” evident in the Cambrian period.

# V. Other minds

## Forking paths

Armed with the historical context and intellectual tools developed in Parts I through VI, we’re finally ready to begin tackling the big questions posed in the Introduction: the nature of general intelligence, the “hard problem” of consciousness, and the thorny question of free will.

Let’s begin with free will, picking up where Part III left off, with Nvidia’s DAVE-2 self-driving car. Recall that DAVE-2 was nothing but a convolutional neural net whose input was a front-facing camera, and whose output was a steering wheel position. It was the purest and simplest approach to cybernetic self-driving.

What happens when such a “reflex-based” car arrives at a T-junction? A generic, pretrained model with no context can’t react sensibly. When *we’re* driving, these are the moments when we might make decisions based on higher-order goals. At other times, we’re on “autopilot” ourselves: the goal is merely to stay in our lane, follow the rules of the road, and avoid collisions with other vehicles, pedestrians, or animals. If we’ve established habitual routes, say to and from work, we might also make a lot of turns on autopilot. Those are instances when our inner “autopilot model” is not generic, but specific to our own lived experience.

DAVE-2, however, is *only* trained to perform the generic autopilot task. To gather the dataset, human drivers navigated a wide variety of routes on many different roads, and excluded the brief intervals when they were changing lanes or making turns from the training data. During live road tests, the human driver intervened to perform these maneuvers manually.

What we’d expect the car to do if kept on autopilot at a T-junction, then, is to choose the straightest way. At an *exactly* symmetric T-junction, one would hope that the model itself is just a tad asymmetric, causing it to break the tie somehow. A perfectly symmetrical model would be unable to choose, hence would steer straight ahead and run off the road.

This unfortunate situation illustrates the old saying about the perfect being the enemy of the good. It also recalls “Buridan’s ass,” often framed as a philosophical paradox in discussions of free will. In the parable, the ass (or donkey, though the other definition applies too) is midway between two equally-sized piles of hay, and being unable to decide which to eat, starves to death.[[334]](#footnote-403)

In practical terms, avoiding this “metastable” situation is easy enough. All it requires is a random variable, a bit of noise in the system, to break any near-ties.

In real life, though, Buridan’s ass only rears its poor muddled head infrequently, because few decisions we make are truly arbitrary, fully symmetric, or contextless. We don’t need to make random turns at T-junctions since when we’re driving, we’re usually trying to get somewhere. More broadly, questions about choice, decisions, and free will only make sense when we consider an agent with its own history, acting in time and in context. That’s why we don’t generally count reflex actions as willed. What does it take, then, for a system to be “agential”?

DeepMind’s AlphaGo contest with Lee Sedol took place in 2016, the same year Nvidia’s DAVE-2 self-driving car paper was published. Go is exceedingly complex, with many more possible moves at any point in the game than chess—a vast and intricate garden of forking paths. For this reason, it had resisted the brute-force search or heuristic approaches that had produced grandmaster-level chess programs many years earlier, as well as naïve TD reinforcement learning.

At bottom, DeepMind’s trick was the same as Nvidia’s: replacing hand-written code with a learned convolutional neural net. AlphaGo’s system design was considerably more complex, though, not only involving separate policy and value networks trained through reinforcement, but incorporating traditional methods, like a randomized or “Monte Carlo” tree search through possible future moves.[[335]](#footnote-404) In this sense, it was a hybrid between deep learning and GOFAI, though subsequent models—AlphaZero,[[336]](#footnote-405) MuZero[[337]](#footnote-406)—moved progressively toward a purer deep learning approach, removing the remaining handcrafted heuristics while further improving performance. MuZero is general enough to play any game, not just Go.

DeepMind has always aimed to solve artificial general intelligence through an agential approach. Gameplay also runs deep in the organization’s culture. Its founder, Demis Hassabis, was a chess prodigy, and an expert at many other games. Prior to DeepMind, he designed video games, and founded his own gaming company. DeepMind’s first highly visible success, in 2013, involved combining reinforcement learning with convolutional neural nets to play Atari games.[[338]](#footnote-407)

The company’s bet that gaming could provide a controlled arena for developing general AI has certainly yielded impressive and important advances. Go commentators studying AlphaGo’s games also lauded the system for “exhibiting creative and brilliant skills and contributing to the game’s progress.”[[339]](#footnote-408)

However, beyond the limitation Churchland pointed out[[340]](#footnote-409)—its single-mindedness—AlphaGo was very different from the brain in another important way: it did not exist in time. It had no dynamics or internal state.

## Children of time

Like the self-driving car model, neither AlphaGo nor its successors maintain any memory of its prior “thoughts” or plans.[[341]](#footnote-411) True, the system works out its next move by thinking as many steps ahead as possible; it would be a stretch to call that extended evaluation a “reflex action.” However, once the move is made, it’s groundhog day—the next turn, it considers the board (as modified by the opponent’s move) afresh. AlphaGo makes no attempt to model its opponent’s psychology, nor does it learn on the fly; thus, it can’t individuate. Neither can it have any sense of a “self” acting in the world. All it models is that world itself and the possible paths through it, annotated (courtesy of the value network) with each path’s goodness or badness.

For an idealized board game like Go or chess, none of these design choices are necessarily shortcomings. Per game theory (as formalized by none other than John von Neumann,[[342]](#footnote-412) who had his fingers in many pies) the optimal next move doesn’t depend on history, on the opponent, or on any internal state, but only on the state of the board—to which both players have full access. And the opponent should be presumed to always play optimally. To do otherwise would risk disadvantage; a wily antagonist might lull a strong opponent by playing dumb, then take the gloves off at an inopportune moment.[[343]](#footnote-413)

You may wonder how AlphaGo could pull off a complex, multi-turn strategy—which it certainly can—without a self or memory. Given a clear view of the future’s forking paths at every turn, there’s no need to remember. All possible strategies are visible at all times, including the continuation of whatever strategy seemed best on the previous turn. If it’s still the best option after the opponent’s move, then so be it.

While elegant in its way, the approach is wasteful, in that every strategy must be re-conceived from scratch at every turn. This can occasionally lead to catastrophic (and quite un-human-like) failures, since the model’s view of the future is *not* perfect, but, as with all Monte Carlo methods, stochastic. The model might make an unusual move due to glimpsing a brilliant but distant outcome down the line, but later fail to spot that same opportunity, even if it’s still in play, and foreclose on it with a wrong move. Out of sight, out of mind.

*Portia* spiders offer an illuminating contrast to this kind of “statelessness.” Although only a centimeter or less in size, they have been called “eight-legged cats” due to their unusually intelligent hunting behaviors.[[344]](#footnote-414) They’re odd-looking, with a slow, choppy, robot-like gait—until they pounce. Unlike other spiders, *Portia* have a pair of large, complex, forward-looking eyes with high resolution over a narrow field of view, well adapted to nuanced visual discrimination. Their favorite meal is other spiders, typically web-weavers, which may be twice their size. *Portia* can control the behavior of prey by spoofing web vibration signals; they learn which signals generate the desired response for each species through trial and error (in the lab they can even learn how to fake out species they would never encounter in the wild).

They plan their attacks slowly and carefully. With tiny brains and telephoto eyes that need to scan around to resolve the environment, it seems likely that they sacrifice much of the parallelism of animals with bigger brains (a topic we’ll return to in Part VIII). Once they have worked out a strategy, though, they can execute it virtuosically, leaping from twig to twig, abseiling down silk threads, and landing on their target from above, behind, or frontally, depending on what is safest under the circumstances (for oftentimes, the spider or insect they’re stalking is perfectly capable of eating *them* if the tables turn). Perhaps most impressively, *Portia* routinely reach their prey via indirect routes, including detours of up to an hour during which they may move away from their goal, breaking line of sight.[[345]](#footnote-415)

I’ve offered this detailed account because it illustrates (and builds on) a number of Churchland’s observations:

* *Portia* isn’t playing any one game against any single opponent. Eventually it must eat if it is to survive, but it gets to choose with whom to engage, and when; it can always back off and try something else, or “soldier on.”
* The “game board” is not equally visible to all players. In its totality, it isn’t visible to *any* player. *Portia* must build up and incrementally update a mental model of the world in order to make decisions.
* This model must include not only the externally visible world, but the *internal* states of hunter and prey alike. Knowing its own state allows *Portia* to assess how hungry (or desperate) it is, how strong or weak it feels, whether hunting should take precedence over mating or resting, and so on.
* Modeling the mental state of its prey is also essential. Are they aroused? Have they realized I’m stalking them? If so, how well have I been localized? By timing its movement on a victim’s web to coincide with wind gusts, *Portia* may work to remain “vibrationally invisible”; or, by strumming the web, it may attempt to create the illusion of a trapped bug. The trickster must constantly assess: are my tricks working? Am I fooling that other mind?

Artificial neural nets can certainly do any of these things; after all, they can be expressed in terms of functions, and ANNs are universal function approximators. Such capabilities will not, however, emerge on their own in a system like DAVE-2 or AlphaGo, due both to the constraints of the setup (autopilot or board game) and the structural limitations of a stateless classifier—even one that makes very clever decisions in the moment by evaluating many possible futures.

## Elbow room

Given *Portia*’s alien lifestyle, physiology, *Umwelt*, and brain limitations, it takes a leap of imagination to understand what it’s like to be a jumping spider, though it is fun (and maybe fruitful) to make the attempt.[[346]](#footnote-417) At least to me, though, it seems obvious that *Portia*, cats, and humans on the hunt all share not only the ability to assess and plan an attack (or defend against one), but to experience doubt, uncertainty, fear, and triumph. Such feelings and experiences emerge naturally from the evolution of predation, as described in Part III, just as hunger emerges naturally from the need to eat.

But it seems so *subjective* to assert that another mind is experiencing something familiar from one’s own experience. How can one really know? We will soon explore how such claims can be assessed more rigorously, as predictive hypotheses.

We can assume a more *ethological* perspective, though. (Ethology concerns itself with animal behavior, studied using lab experiments and analyzed in the context of neurophysiology, function, and evolution.) Then we can simply ask the same questions about feelings and internal states we would ask about physical features, like the differently adapted beaks of Darwin’s finches. *Why* would that feeling arise, and be preserved by evolution? How is it useful? What would happen if it were absent? This lets us fend off charges of anthropocentrism and defer “philosophical zombie” questions a little longer.

Free will, as a desire or imperative, is easy enough to understand from this perspective. No animal likes being trapped, humans included; that’s why imprisonment is a form of punishment, even if our bodily needs are met. Remember, the whole point of having a brain is to be able to act in the world, choosing among alternative futures in order to enhance our dynamic stability—that is, to stay alive, and to continue to be able to make choices in the future.

When others restrict our ability to make choices, it doesn’t feel good, just as gnawing hunger doesn’t feel good. The confinement of our possible futures renders us increasingly helpless, like a king being chased into a corner of the chess board in the endgame. Go offers an even purer expression of this idea. Although the stones don’t actually move once laid down, their “liberties” are the number of adjacent unfilled board positions where they “could” move, and stones only remain “alive” as long as they have at least one liberty. Death, then, whether in a game or in real life, represents the ultimate exhaustion of one’s ability to choose. It’s the pinched-off end of our Wiener sausage.

In a predatory context, as the early cyberneticists understood, prediction is probabilistic and adversarial. The predator will try to predict the prey right into its stomach. The prey will try to predict the predator’s moves, and escape. Each is trying to preserve its own liberties, potentially at the cost of the opponent’s. Each will therefore try to predict the other’s predictions, *ad infinitum*.

A corollary: being perfectly predictable is bad news. In a way, it’s a form of imprisonment. Ted Chiang mordantly explores the psychological consequences in his short story, *What’s expected of us*.[[347]](#footnote-418) A device appears on the market, the Predictor, consisting of nothing but a button and a light. The light always flashes green one second *before* the button is pressed. Millions of Predictors are sold, and at first, people treat it as a novelty, showing it to their friends and trying to fool it. When reality sinks in, though—that the Predictor *can’t* be fooled—free will is revealed to be an illusion, and people despair, eventually losing their will to live.

Real life predictors aren’t one-button devices, of course, but other minds, like that of a cat pursuing its prey. And one way to become highly predictable is to run out of voluntary moves, like a cornered rat. But if one’s behavior looks deterministic from the outside, i.e. can be perfectly predicted by an observer, then from an adversarial perspective, that is equivalent to being cornered.

When an agent becomes perfectly predictable to an observer, it also makes the agent no longer *register* as agential to that observer. Cognitive scientist Douglas Hofstadter has described this property as “sphexishness,” by reference to the golden digger wasp *Sphex ichneumoneus*.[[348]](#footnote-419) Hofstadter was drawing on observations by French entomologist Jean-Henri Fabre (1823–1915), who has been hailed as the founding father of the study of animal behavior.[[349]](#footnote-420) Fabre took a dim view of cognition in insects, characterizing them as exhibiting “machine-like obstinacy.”[[350]](#footnote-421)

He noticed that when the time comes for *Sphex* to lay eggs, she constructs a burrow, seeks out a cricket to nourish her future hatchlings, paralyzes it with a sting, brings it to the burrow entrance, goes in to see that all is well, then drags the cricket in, lays the eggs alongside the body, seals the burrow, and flies off. This elaborate performance certainly *looks* purposive—and it is.

But then, Fabre noticed that if the cricket is moved by a few inches after the wasp has gone into the burrow to check that all is well, *Sphex* will re-emerge and move the cricket back into position, then go back into the burrow to check that all is well again! He described repeating this process dozens of times. By moving the paralyzed cricket every time the wasp went into the nest, he was able to get her stuck in an endless loop.[[351]](#footnote-422)

One interpretation of this little parable: it’s easy to be fooled into *thinking* that a behavior is intelligent or agential when in fact it’s just a running program. That would beg the question, though: what *else* could it be? In some sense we are *all* “just running programs.”

A better take is to recognize, first, that *Sphex*’s nest preparation behavior is genetically determined, hence the intelligence manifested by the program is a product of evolution, not of in-the-moment reasoning by the individual wasp. The individual wasp appears to be, like DAVE-2, running on autopilot, at least with respect to nesting behavior. This seems reasonable, given the wasp’s small brain and two-month adult lifespan. It has no time to learn the ropes, and nobody to learn them from, so it has to know how to reproduce “right out of the box.”

Perhaps more interesting, though, is our *own* response to observing sphexish behavior. When we see the elaborate preparations the wasp makes for her brood, we think she’s smart and agential. Then, when we see that the behavior is pre-scripted, we “realize” that she’s just a machine, an “it,” and any impression of intelligence or willed action vanishes. But the difference is in our own minds, not that of the wasp. Alan Turing arrived at a similar conclusion: “From the outside, […] a thing could look intelligent as long as one had not yet found out all its rules of behavior. Accordingly, for a machine to seem intelligent, at least some details of its internal workings must remain unknown.”[[352]](#footnote-423)

In the case of *Sphex*, this realization comes about due to our reverse-engineering of the script, and “hacking” it to create an infinite loop. We come to believe the wasp has no intelligence or agency precisely because we can now predict its actions perfectly. What we really mean by “it’s just following a script” is that *we*, on the outside, *know* what the script is, and don’t even need to interact with the animal to predict what it will do next.

As a fellow computational being, how can *you* escape the same pickle—how do you keep from becoming a robotic “it” to a clever observer? There are a few ways:

* As described for moth flight in Part III, simply using a random variable can keep your behavior fresh and surprising even if you are not so smart. This strategy still works when your opponent is smart enough to see right through you (that is, your internal state is transparent, or can be fully modeled by an outside observer). It even works if you are entirely generic and incapable of individual learning, meaning that your behavioral repertoire is fully genetically programmed, and you are “just like” every other member of your species.[[353]](#footnote-424)
* The ability to learn helps enormously, as it allows your responses to be more unique, and allows your behavioral repertoire to scale up far beyond what can be encoded in your genome. By storing learnings and behaviors based on past experiences in your brain, you become vastly more complex on the inside—and all of this complexity is hidden state, invisible to a stranger, making you a lot harder to predict.
* If you are yourself able to model others effectively, including others who are modeling you back, then you can go meta.

Wasps might not bother with any of these strategies when it comes to nest preparation because cricket-moving tricksters have never exerted evolutionary pressure on making this particular behavior unpredictable. Predator/prey interactions are a whole different story. It’s unlikely that *Sphex* would behave so sphexishly while hunting the cricket it brings to its nest, or evading a predator of its own.

The last trick on the list—modeling your modeler—is the real trump card, but it depends on being very smart indeed. Roughly speaking, you need to be as smart as your opponent, or more so, and you need to be able to imagine how you (and the world in general) appear from their point of view. When *Portia* spiders show an awareness of what their adversary does and doesn’t know, and learn to actively manipulate that opponent’s world model, they are operating at this higher cognitive level.

## Matryoshka dolls

Among human children, this so-called “theory of mind” or ability to “mentalize” (I’ll use the terms interchangeably) takes quite a while to develop fully. The classic experimental setup for probing it is the Sally-Anne Test, developed in 1985 by psychologists Simon Baron-Cohen, Alan Leslie, and Uta Frith.[[354]](#footnote-426) The experimenters put on a little chamber drama for their young subjects using two dolls, Sally and Anne. Sally has a covered basket, and Anne has a box with a lid. Sally puts her marble in her basket, then goes out for a walk. While she is gone, Anne slips the marble out of Sally’s basket and into her own box, closing the lid. Later on, with Anne out of the picture, Sally returns.

During the skit, the children are asked four questions:

1. The Naming Question, a warm-up, asks them which doll is which; it establishes basic understanding of the scenario.
2. On Sally’s return, the Belief Question is: “Where will Sally look for her marble?”
3. This is followed up with the Reality Question: “Where is the marble really?”
4. … and the Memory Question: “Where was the marble in the beginning?”

Even babies (by the time they can talk a little) are able to say who is who and understand where the marble *is* throughout the skit, even though it’s stashed out of sight; hence questions 1, 3, and 4 are all fairly easy. However, until somewhere between two and a half and four years of age, children tend to get the Belief Question wrong.[[355]](#footnote-427) They aren’t able to model the minds of others well enough to knowingly attribute false beliefs to them.

Psychologists tend to be very particular about which kinds of agents they believe can exhibit theory of mind, because of the baggage implied by language like “knowingly attribute.” However, from a functional perspective, theory of mind is exactly what it takes to lie or deceive, as *Portia* does.[[356]](#footnote-428)

Exhibiting this cognitive capacity implies not only an ability to simulate other minds, but to model counterfactual (“what if”) universes. In the Sally-Anne test, there is a real-life universe, in which the marble is in the box. There is the universe in Sally’s head, in which the marble is in the basket. Each of these universes also contains agents, like Sally and Anne, who in turn contain model universes in *their* virtual heads. The Anne in the universe in Sally’s mind, for instance, *also* believes that the marble is in the basket. Of course Sally and Anne are mere dolls; they only exist within the minds of the experimenter and the subject. This experimenter and subject are in turn abstract people I am describing to you, dear reader. Nesting dolls, universes within universes, in an endless recursion. It’s a dizzying prospect.

As adults, though, we negotiate such recursive mentalization all the time without a second thought. Reading a novel like *Jane Eyre* involves regularly entertaining fourth, fifth, and sixth-order theories of mind, as we wonder what Charlotte Brontë expects us to believe about what Jane believes Mr. Rochester must think Jane thinks about Mr. Rochester’s feelings about Blanche. Exercising this mentalizing faculty seems to have been the main passtime of the nineteenth century English upper classes. (Writing popular novels about it added yet another layer of indirection, at times ironic.)

But theory of mind is far more than just a valuable skill for navigating hidden barbs and avoiding *faux pas* at teatime. There is good evidence that it is the very stuff of intelligence; it is, thus, at the heart of this book’s main argument.

We have arrived at that heart. My contention is that theory of mind:

* Powers the “intelligence explosions” observed in our own lineage, the hominins, and in other brainy species;
* Gives us the ability to entertain counterfactual “what-ifs”;
* Motivates, and is enhanced by, the development of language;
* Allows us to make purposive decisions beyond “autopilot mode”;
* Underwrites free will;
* Operates both in social networks and within individual brains;
* Results automatically from symbioses among predictors; and
* Is the origin and mechanism of consciousness.

In a sense, theory of mind *is* mind.

## Intelligence explosion

These are some big claims. Let’s begin with the more established ones.

In the 1970s, Dian Fossey, the world’s leading expert on gorilla behavior, invited British neuropsychologist Nicholas Humphrey to spend a few months at her research station in the Virunga Mountains of Rwanda. Reflecting later on what he had seen, Humphrey wrote, “[O]f all the animals in the forest the gorillas seemed to lead much the simplest existence—food abundant and easy to harvest (provided they knew where to find it), few if any predators (provided they knew how to avoid them) … little to do in fact (and little done) but eat, sleep and play. And the same is arguably true for natural man.”[[357]](#footnote-430)

These observations flew in the face of the usual explanation for evolving high intelligence—that it’s all about being a brilliant hunter, or otherwise “winning” at playing a brutal, Hobbesian game of survival in a tough environment. But if not to hunt (or evade hunters), why bother with intelligence? Couldn’t plenty of other animals live the easy life without incurring the high cost of a big brain? Or, put another way, “Why […] do the higher primates need to be as clever as they are […]?”[[358]](#footnote-431) For the fossil record is unambiguous: primate brains in many lineages, including those of gorillas and humans, haven’t shrunk over time, but rather, have grown explosively.

Humphrey’s answer: “[T]he life of the great apes and man […] depend[s] critically on the possession of wide factual knowledge of practical technique and the nature of the habitat. Such knowledge can only be acquired in the context of a social community […] which provides both a medium for the cultural transmission of information and a protective environment in which individual learning can occur. […][T]he chief role of creative intellect is to hold society together.”[[359]](#footnote-432)

There are now many variations on the basic theory, some, like Humphrey, emphasizing cooperation and division of labor, others competition and Machievellian politics.[[360]](#footnote-433) These correspond fairly well to the “love” versus “war” explanations for the emergence of intelligence generally, as described in Part III.

In the 1990s, evolutionary psychologist Robin Dunbar and colleagues used comparative brain size measurements across species to advance the closely related “social brain hypothesis,” which holds that the rapid increases in brain size evident in hominins and cetaceans (whales and dolphins), among others, arise from mentalizing one-upmanship.[[361]](#footnote-434)

Underpinning all of these related hypotheses is the observation that, among highly social animals like us, theory of mind is a powerfully adaptive trait. Being better able to get inside others’ heads increases the odds of finding a mate, building coalitions, securing resources from friends and family, getting help raising young,[[362]](#footnote-435) avoiding violence (or being on the winning side of it), climbing in prestige, amassing fans or followers. So, unsurprisingly, people with better theory of mind tend to live longer and have greater reproductive success.[[363]](#footnote-436) That means natural selection will be at work.

Correlates of strong theory of mind include both larger numbers of friends and larger brains, especially in brain areas associated with mentalizing—above all, the frontal cortex.[[364]](#footnote-437) Thus, startlingly, we can see signs of the evolutionary pressure on social intelligence even among modern humans.

My guess is that the extra cortical volume of highly social people is dedicated not only to general mentalizing skills, but also to rich learned representations—we could even call them *simulations*—of their many specific family members, friends, colleagues, and acquaintances.[[365]](#footnote-438) After all, mentalizing requires not just imagining another person in the abstract, but modeling their particular life experience, what they know and don’t know, their quirks and norms, the way they express themselves—in short, everything that comprises a personality, outlook, and *Umwelt*. You have to keep track of many facts and continually update them. Yet you can probably bring such knowledge to bear effortlessly to play out hypothetical social situations in your head, drawing from a cast of hundreds of different people you know.

This task is even harder than it appears at first glance, because there are infinite reflections in the social hall of mirrors: all of those people are themselves modeling others, including you. And, of course, their models of *you* include models of *them*, and of others. Even if your model of second-order relationships (i.e. who is friends with whom, and what you know about that interaction) is not as rich as your model of first-order relationships (your friends), the sheer *number* of higher-order terms in your model explodes. If you have a dozen friends, then just amongst themselves, there will be 122=144 potential second-order relationships, and 123=1,628 potential third-order relationships. The numbers get truly mind-boggling when you consider that most of us know well over a dozen people well—our acquaintances can easily number in the hundreds—and our theories of mind can go up to sixth order, or beyond. You can imagine, then, that the amount of brian volume needed to do social modeling might grow both as a function of your number of friends, and as a function of your ability to model higher-order relationships.

Indeed, when Dunbar and colleagues set out to find the relationship between the brain sizes and the social group sizes of brainy animals, they found that as troop size increases, the amount of brain volume dedicated to cortex also increases.[[366]](#footnote-439) Mentalizing order, which Dunbar refers to as “intentionality level,” appears to be limited by cortical volume; behavioral studies suggest that monkeys have level 1, and apes level 2 intentionality. (By extrapolation, archaic humans and Neanderthals may only have been able to achieve level 4 intentionality, which is at “the lower end of the normal distribution for modern human adults, and at about the same intellectual level as young teenagers.”[[367]](#footnote-440)) Finally, the slope of the relationship between cortical volume and troop size is considerably steeper for apes than for monkeys, implying that modeling higher levels of intentionality requires a greater investment of cognitive resources.

Findings relating brain size to social group size, and social group size to Darwinian fitness, are themselves a kind of hall of mirrors, revealing a profound self-similarity—and feedback loop—between brains and social groups. If you have a slightly larger brain than your friends and family, and are able to model more relationships more reliably, you will have a Darwinian advantage, so on average, will have slightly more descendants than those who are less socially adept. But that means that your descendants will themselves become harder to model socially; that is, everyone else’s model of them must now be more complex, including higher level intentionality and more relationships. And remember, there’s something adversarial about prediction; everybody is trying to predict everybody else, but *not* be fully predictable themselves.

So it’s an arms race, not unlike that of the Cambrian explosion—though usually friendlier. Everyone is getting a bigger brain to model everyone else, and everyone else is getting harder to model at the same time, because… well, their brains are getting bigger. A social intelligence explosion ensues: a rapid (by evolutionary standards) increase in brain volume in your species.

When social modeling becomes such a major part of everyone’s life, the effect on individual incentives is dramatic. Lone operators, like leopards, are content to fend for themselves; their *Umwelt* consists mainly of their territory and their prey, while other leopards are, most of the time, unwelcome intruders. For a modern human, though, being cast out or excluded from the community becomes a severe punishment, potentially even a death sentence. Our world consists largely of other people, and most of us would be unable to survive at all without continual mutual aid.

At the same time, our sociality is fraught. We strive to “win” at modeling others without being fully modeled ourselves; we compete for mating opportunities and for attention; we strive for dominance and prestige. These dynamics once again illustrate how competition and cooperation can be interwoven to such a degree that it can be hard to tell which is which.

Selection pressure also operates at the level of social groups. If one group sports slightly bigger brains and greater social prowess, the group itself can grow larger, and will thus tend to outcompete the smaller (and smaller-brained) group. Societies have a kind of collective intelligence, and a rich body of work in social anthropology tells us that collective intelligence exhibits a scaling law not unlike that of individual brains. Theory of mind, incidentally, is also important both for effective teaching and effective learning, which implies that it enhances cultural evolution at the group level too.[[368]](#footnote-441) So, bigger societies can create and evolve more complex technologies, and thereby develop greater adaptability and resilience.[[369]](#footnote-442) Greater scale, in other words, can support greater intelligence, and greater intelligence improves dynamic kinetic stability, both at the individual level and at the group level.

## Crew of eight

It’s hard not to wonder whether this self-similarity applies at finer scales, too. Could our brains be “societies” of neural circuits that have evolved to model each other, and thereby create the larger collective intelligence we call a “brain”? Neuroscientists and AI researchers have advanced such theories many times over the years, offering a range of supporting evidence.[[370]](#footnote-444)

For one, cerebral cortex has a modular or repetitive structure, consisting of a honeycomb of “cortical columns”—though these columns don’t have distinct borders, which has led to considerable debate about their exact definition and granularity.[[371]](#footnote-445) Therefore, when I use the term “cortical column,” you should take it with a grain of salt; I will interchangeably use even more explicitly vague language like “small piece of cortex” or “part of the brain.”

Despite these ambiguities, we know that the basic “cortical circuit” is much the same in different regions of the cortex. When we distinguish between “auditory cortex” and “visual cortex,” for instance, the difference seems to be mostly a function of where they get their inputs. In one famous experiment, the optic nerves of baby ferrets were rerouted surgically to their “auditory” cortex, and these animals nonetheless developed the ability to see. Moreover, the characteristically patterned neural mapping of oriented visual features Hubel and Wiesel discovered in visual cortex are reproduced in the ferrets’ repurposed “auditory” cortex, albeit more sloppily.[[372]](#footnote-446) There’s evidence that the human brain can rewire in similarly dramatic ways: blind people can learn to see using their hearing via “click sonar,”[[373]](#footnote-447) and can even acquire a limited form of vision via spatially patterned tongue stimulation.[[374]](#footnote-448)

Generic modularity is precisely what made the intelligence explosion possible: just as DNA could easily evolve to replicate more vertebrae and ribs to produce snakes (per Part I), the genomes of animals undergoing intelligence explosions could respond to selection pressure by replicating more cortical columns. Thus, the cortical sheet could expand without needing to invent anything fundamentally new. In the biggest-brained animals, like dolphins and humans, expansion progressed to the point of scrunching up the cortex into dense folds, giving our brains their characteristic furrowed appearance. We manage to cram about a quarter of a square meter into our skulls.

So in a way, cortex is a colony of cortical columns, which have managed to “selfishly reproduce” in larger and larger numbers through increased cooperation among themselves. Yet it’s not so selfish, as the greater intelligence they give us adds to our individual fitness, which in turn allows us to form larger societies, which increases dynamic stability all round. I suspect we could tell a similar story at even finer scales: about the neurons that make up cortical columns, the organelles that make up neurons, and even the organic molecules within single cells.

Indulge me in a speculative digression about everyone’s favorite “alien intelligence”: the octopus.[[375]](#footnote-449) Octopuses have always seemed like a bit of a fly in the ointment, as far as the social intelligence hypothesis is concerned. Unlike other very smart animals on Earth, they’re notoriously cranky and individualistic. They don’t tend to hang out with each other, and usually only meet to mate—a fraught encounter, which may end in violence or even cannibalism.[[376]](#footnote-450) While they are brilliant lifelong learners, they don’t live long enough to raise their young, hence lack the ability to achieve cultural accumulation; if they did, it seems entirely possible that our planet’s seas would contain wonderful octopus cities, and perhaps even octopus-made computing machines (which would probably use an octal, or base eight, numbering system—a much better engineering choice than our own base ten).

The puzzle is: if octopuses are so antisocial, and intelligence is a social phenomenon, why are they so smart? One could dodge the question, and note that sociality doesn’t just involve conspecifics (i.e. others of your own species). Nor is it necessarily friendly. With omnivorous appetites and a total lack of body armor, the octopus must rely on its wits to model and predict everything it wants to either hunt or escape from; it seems likely that highly developed theory of mind, vulnerability to predators, and extremely flexible hunting strategies all coevolved. Another easy out: octopuses have been around for a long time, and we have no way of knowing what their social behaviors might have looked like a hundred million years ago. Perhaps their intelligence evolved at a time when they were nicer to each other than they are today.

A more interesting possibility, though, is suggested by considering the *other* ways octopuses are so different from their distant cousins, the other big-brained animals, including us. As everybody knows, the octopus’s “brain” is decentralized. A likely reason for this design is the especially costly long-distance communication in the nervous systems of molluscs, which lack the fatty myelin sheaths that coat our long-distance nerve fibers, greatly speeding up their transmission of electrical impulses. (Myelination is what makes our longer nerve fibers whitish; hence the unmyelinated “gray matter” of the cerebral cortex, where connectivity is mostly local, and the “white matter” of the brain’s interior, which consists mostly of long-distance wiring.)

Transmitting fast impulses in unmyelinated nerves requires increasing the diameter of the axon, which is why squids have a “giant axon”—a single nerve fiber up to 1.5 millimeters in diameter!—running the length of their body. This giant axon allows for a rapid, coordinated escape response.[[377]](#footnote-451) But obviously, for an animal with many neurons powering complex behaviors, there isn’t room for many such giant axons, so a great majority of neural processing needs to be local. That probably explains why three fifths of an octopus’s neurons are in the arms, rather than in the head.

The octopus’s suckers are arranged in beautiful symmetric patterns—over two hundred along each arm—with chains of neural networks surrounding each sucker. The arms, in turn, define an eightfold near-symmetry. The animal as a whole, in other words, is far more modular, symmetric, and self-similar than any bird or mammal.

In the octopus, a lot of tasting, moving, color-changing, reacting, deciding, predicting, and even reasoning must happen close to the muscles around the suckers and along the arms that control movement. Each sucker is prehensile, almost like a pair of opposed fingers, and is endowed with far richer local sensory input than our hands, including touch and taste receptors, chromatophores for colorful skin displays, and even photoreceptors for light sensing. Local neural circuits, then, have everything they need to sense and act locally—as long as they can coordinate efficiently at a high level.

The octopus’s modularity, then, is more than superficial. Each sucker is smart. Each arm, too, is intelligent in its own right, able to respond to stimuli with complex movements without the involvement of the central brain.[[378]](#footnote-452) Arms recognize each other directly as they move, avoiding self-entanglement without requiring any central representation of the animal’s geometry,[[379]](#footnote-453) which—given the wild range of geometries possible for such a squishy animal with so many degrees of freedom—would be an extraordinary challenge for *any* brain. (Since the beak is an octopus’s only rigid part, it can squeeze through tiny holes or cracks—anecdotally, pretty much any space big enough for one of its eyeballs to get through.[[380]](#footnote-454)) The arms can even communicate with each other directly via a ring of ganglia that encircle the “shoulders” and direct connections between alternating arms; both routes bypass the brain entirely.[[381]](#footnote-455)

There’s more. Octopuses occasionally need to bite off an injured arm, an act known as “autotomy.”[[382]](#footnote-456) They may even do so to distract a predator, leaving one writhing appendage behind while the other seven, and the head, escape. (At this point, as a smartass on Reddit observes, it would more properly be called a septopus.) In at least one squid species, “attack autotomy” has been observed as well: a severed arm attacks a predator while the rest of the animal escapes.[[383]](#footnote-457) Following autotomy, a new arm will regrow. The whole process takes about 130 days.

Perhaps, then, the octopus is best thought of as a tightly knit *community* of eight arms, sharing among themselves a common pair of eyes. In fact, the bulk of the central “brain” consists of a pair of optic lobes; we could think of these lobes compressing visual information in service to the arms, rather than being “the boss,” or even, in any meaningful sense, “the octopus.” In this light, attack autotomy is not so different from the *kamikaze* tactics of a bee defending the hive.

Could the intelligence explosion that resulted in the octopus have been driven by predictive social modeling—theory of mind, in effect—among its eight arms? As absurd as this sounds, I think it’s plausible. Each arm is enormously complex, with many more degrees of freedom than our jointed limbs. Octopus arms are individually smart, but communication among them is limited, due both to the high latency and low bandwidth of unmyelinated nerve fibers. Effective central control under these conditions is impossible.

Yet the octopus’s arms must work together seamlessly to swim, hunt, escape aquarium tanks, and so on; and to do so, they must model each other. We know that people can work closely together to perform such tasks with only a little communication; consider an elite military unit, or a hunting party, exchanging subtle signs while stalking their quarry. Even intellectual tasks, like solving escape room puzzles or competing in Math Olympiads, involve such teamwork. Language is a low bandwidth medium, yet we still manage.

Accounts of small, extremely high-performing teams invariably emphasize both the unity of purpose that can be achieved, and the need for each team member to infer what the others are up to based on minimal cues: that is, high fidelity mutual prediction, or mentalizing. When this happens in competitive rowing, it’s called “swing”:

“It only happens when all eight oarsmen are rowing in such perfect unison that no single action by any one is out of synch with those of all the others. […] Each minute action […] must be mirrored exactly by each oarsman, from one end of the boat to the other. […] Only then will it feel as if the boat is a part of each of them, moving as if on its own.”[[384]](#footnote-458)

## Homunculus

How different are we from the octopus, really? We have fast, myelinated nerves, which allows for greater physical centralization of our brain in our heads, but if our brains are indeed symbiotic colonies of cortical columns modeling each other, then where do we locate the “I”?

Much intuitive (and supernatural) thinking over the ages has presumed an indivisible self, located somewhere in the body. Presumably it’s in the head somewhere, so we can imagine getting a “full body transplant” but not a “head transplant.” A common term for this “something in the head” is a “homunculus,” literally, a “little man” somewhere in there, pulling the strings.

In many religious traditions, the immortal soul plays this homuncular role. René Descartes (1596–1650) thought about the problem from a neuroanatomical point of view, attempting to reconcile Enlightenment materialism with his Christian faith. Descartes was an influential champion of “mechanical philosophy,” or what we now call physics. He concluded that the body and brain *must* be mechanical, or machine-like, since we seem to be made of the same stuff as the rest of the physical universe.

Human anatomy has much in common with other animals’ anatomy, and so does human behavior: the “passions” of hunger and sexual attraction, reflexes to flinch away if burned, and so on. Hence Descartes coined the term “*bête-machine*”, or “animal-machine,” to describe the mechanically-based physiology of animal and human bodies.

Yet the Church held that we have immortal souls while other animals do not, and the Enlightenment emphasized the divine gift of human rationality; other animals certainly didn’t seem to be writing philosophical treatises. Thus, rationality and the soul became identified with each other. A soul-brain interface would be needed, then, for the soul to puppet the body.

To theorize just how this might work, Descartes borrowed from a classical medical authority. Galen (roughly 130–210 CE) believed the cerebral ventricles were filled with “psychic pneuma,” a fine, volatile, airy or vaporous substance that he described as “the first instrument of the soul.”[[385]](#footnote-460) Riffing on this idea, Descartes therefore imagined the body to be, literally, a hydraulic machine powered by psychic pneuma: “When a rational soul is present in this machine it will have its principal seat in the brain and will reside there like the fountaineer, who must be stationed at the tanks to which the fountain’s pipes return if he wants to initiate, impede, or […] alter their movements.”[[386]](#footnote-461)

But where *was* this “fountaineer”? Descartes concluded that it must reside in the pineal gland, a small, pinecone-shaped structure near the center of the brain. For one thing, it’s near the cerebral ventricles, the supposed “tanks” or reservoirs for psychic pneuma. More importantly, the pineal gland is one of the few discrete structures located along the brain’s midline—meaning we have only one of them. Being indivisible, surely the soul couldn’t reside in a *pair* of structures on both the left and right sides of the brain! So the pineal gland *had* to be the fountaineer!

Spoiler: the pineal gland’s main job is melatonin production. It helps regulate the circadian rhythm. It is not the seat of the soul.

Obviously, Descartes’ mistake is more fundamental. Attempts to pin down the location of the homunculus, soul, or “self” within the brain are like trying to locate where the “swing” is in a boat. This isn’t to say that either phenomenon is unreal or illusory. Swing is perfectly real, as those who have experienced it can rapturously attest. But it is a verb, not a thing; distributed, not localized; a dynamic process in time, not a static condition; and a relational phenomenon, not an individual one. There is nothing supernatural about it, but neither is it concretely physical, or even objective. It results in higher performance of the boat, and so has tangible effects in the world. But it is subjective, experienced by each of the eight members of the crew in relation to the others, and to the whole.

Imagine replacing one of the crew members in the boat with a robot. Not the dumb kind we have in factories today, but one capable of pulling the oars like a person, hearing, seeing, and touching, and using haptic and audiovisual cues to communicate with the rest of the crew. Could such a “cyborg crew” achieve swing? I think the answer is a clear “yes,” and in fact, given the sophistication of today’s AI models, any remaining impediment has mainly to do with deficiencies in batteries and motors—bodies, not “brains.”

This thought experiment is a boat crew version of one posed by philosophers David Chalmers and Susan Schneider, which we could call the “brain of Theseus,”[[387]](#footnote-462) by reference to Plutarch’s much older “ship of Theseus” thought experiment (to stick with the boating theme a little longer). Theseus, the mythical king and founder of Athens, has a legendary ship, which is preserved by the Athenians for generations. It undergoes repairs over the years to keep it in good order. The question is: once every timber has been replaced, is it still the ship of Theseus?

Now that we understand something about cellular biochemistry, we know that the question applies literally to us, too, as most of the cells in our bodies turn over many times throughout our lives, and even in the longest-lived neurons, the atoms will all get swapped out over time. If we want to assert the continuity of our own identity, then, we need to acknowledge that a “self” is, like swing, a dynamical and computational process, not a specific set of atoms. And the same is true of a ship!

Let’s follow Chalmers and Schneider in considering the “ship of Theseus” question for the brain, now with a cyborg twist as with the rowing crew. Imagine that a single neuron in your brain is replaced by a computer model of that neuron. Once more, the greatest problems here are technological, not conceptual. Thanks to those mid-century experiments on the squid giant axon and a great deal of work since, we have excellent computational models of individual neurons, at least with respect to their fast dynamics. A complete neuron-machine interface is a much taller order, though. It would require not only coupling a computer to all pre- and post-synaptic neurons, but also the ability to sense and emit a variety of neuromodulatory chemicals. This is well beyond us, today. Setting daunting practicalities aside, the question is: if one of your neurons were replaced by a “robot neuron,” would you notice?

Of course you wouldn’t, because the brain is robust to anything going awry with any single neuron. But what if we repeated the procedure for a billion of your neurons? Or half of them? Or *all* of them?

The question is unsettling, because it highlights the difficulty we have in coming to grips with a dynamic, process-oriented, computational, and relationship-centric view of the “self.” We are still Cartesian enough in our thinking to wonder whether—per Schneider’s description of a hypothetical surgeon replacing your biological neurons with silicon-based ones—“your consciousness would gradually diminish as your neurons are replaced, sort of like when you turn down the volume on your music player. At some point […] your consciousness just fades out.” Or perhaps, “your consciousness would remain the same until at some point, it abruptly ends. In both cases, the result is the same: The lights go out.”

These ideas are—I have no words to put this gently—silly (and I *think* Schneider and Chalmers both agree). In physical and computational terms, a neuron has an inside and an outside, and if the insides were scooped out and replaced with different machinery that performed the same function—which is to say, resulted in identical interactions, as seen from the outside—then none of the larger functions in which this neuron plays a part will be affected. Per the Church-Turing thesis, these functions could be computed in many equivalent ways, and using many different computational substrates. That is true at every level of the fractally nested set of relationships, or functions, we call “life” or “intelligence.”

## Illusion and reality

The word “illusion” gets used frequently in describing many of the concepts this book explores. For philosopher of mind Daniel Dennett, consciousness and the self are illusions; we’re robots made of robots made of robots, so there is no real “there there.”[[388]](#footnote-464) For philosopher-neuroscientists Sam Harris and Robert Sapolsky,[[389]](#footnote-465) free will is an illusion for a similar reason: because minds are products of inexorable physical processes. For “Simulation Hypothesis” adherents like Nick Bostrom, the whole universe is an illusion, because we’re all living in a video game![[390]](#footnote-466) Let’s set that last one aside for now, as unless there’s a glitch in the Matrix it’s probably neither provable nor falsifiable.

On the other hand, we are indeed robots made of robots made of robots, and minds are indeed based on physical processes. What else is new? However, “illusion” isn’t necessarily a helpful word here, because it implies that any habit of thought or societal practice that depends on one of these concepts is wrong, or at best a polite fiction. While nobody has suggested that we dispense with this “fiction” altogether, Sapolsky seems to be running off that cliff when he asserts that, since minds are physical and free will is “illusory,” notions like criminal justice, prizes, praise, blame, and moral responsibility are all invalid and should be abandoned.

It is worthwhile to ask tough questions about our received notions of justice, reward, and responsibility, especially when confronted with alternatives that seem more humane, fairer, or lead to better outcomes. It’s less productive to become nihilistic because “it’s all an illusion.” That kind of nihilism should also lead us to conclude that ships, tables, and chairs aren’t “real,”[[391]](#footnote-467) because they, too, are mere assemblages of atoms governed by the same physics as any other atoms.

Part II offers a more pragmatic take on reality, familiar from physics as well as real life: reality is simply our name for a model with good predictive power. Or rather, a suite of models. We will always need many, because every known model has a limited domain of validity. If we ever discover the physicists’ holy grail, a “Theory of Everything” unifying quantum fields and gravity, that would be wonderful, but it would tell you nothing about whether your aunt will like the cake you’re baking for her, your friend will take your shift at the cafe, or your kid will get along with the neighbor’s kid. In fact, unless you’re one of the handful of rarefied physicists who might be typing up or reviewing a paper on the subject, it will inform precisely nothing in your life.

Good questions to ask about a model, then, are: Does it agree with observations? Does it make any testable predictions? Who cares about it? What function does it serve? Where does it break down? Do we have a better candidate model? Sometimes, we *do* have a better candidate, but, as with general relativity, it may be disturbing or challenge deep-seated intuitions. Newtonian physics, which we find a lot more intuitive, follows from general relativity, but only as an approximation valid in our everyday world, where spacetime is reasonably flat and objects move far slower than the speed of light.

That’t not to say that classical physics is an “illusion”; rather, the more general theory illuminates a wider area. Einstein’s theory shows us when the Newtonian approximation holds, and when it doesn’t, and why. It resolves apparent contradictions, like the incompatibility of Newtonian physics with a constant speed of light. In this sense, general relativity *bolsters* the narrower classical theory, by reassuring us that the seeming paradoxes aren’t cause for concern, and by showing us why, when, and to what degree we can rely on its approximations.

Free will, the unitary, indivisible “self,” and interaction with other unitary “selves” in the world are the common-sense axioms underpinning something like a Newtonian model of social life. Theory of mind *is* that model. It’s our folk theory—literally. It’s what we use when we predict our aunt’s tastes, our friend’s willingness to help, or our kid’s friendships. While powerful and useful in everyday life, theory of mind is also confounding, even paradoxical, when we try to reconcile it with the physical universe; hence the trouble Descartes got us into. In the following discussion, we’ll flesh out a more general theory that will help us better understand our folk theory—and how it arises, and its conceptual limits.

But let’s keep in mind just how powerful that folk theory is. Will Aunt Millie like my cake? For a physicist, making such a prediction may seem like weak tea compared with the extraordinarily precise, quantitative predictions we expect from physical theories. But it’s not so.

# VI. Many worlds

## Au revoir

Consider the scenario at the end of Richard Linklater’s romantic indie film *Before Sunrise*. Young backpackers Jesse and Céline meet on a Eurail train between Budapest and Vienna. Arriving in the evening and out of money, they decide to spend the night wandering the city together before their onward connections the following day—Jesse will catch a flight back to the US, and Céline is returning to university in Paris. Sparks fly, of course (I mean, it’s Julie Delpy and Ethan Hawke; they are both clever, appealingly complicated, and… well, easy on the eyes).

The next morning, they’re back at the train station, but they haven’t shared contact information. Here, you have to cast your mind back to the *Ancien Régime* of 1995, when romance was still analog: no cellphones, no social media. They had agreed that this would be a one-night thing. But as Céline’s train is about to leave, in an anguished rush of words, the maybe-lovers realize that they badly want to see each other again after all. So they decide to meet right there, at the same place, in exactly six months.

Six months from this morning, or from last night? asks Céline.

From six o’clock last night, when their train arrived, says Jesse. It’s June 17th, so it will be … December 16th, track nine, Vienna train station.

They say goodbye, which Céline amends to a hopeful “Au revoir.”

Now, let’s take off the love goggles and think about this as physicists. Given accurate measurements, dynamical laws allow us to predict the orbits of planets and moons, gravitationally bound and hurtling through the near-vacuum of outer space, far into the future. Those laws of motion are quite simple. Even here, though, prediction has inherent limits, because when three or more bodies are interacting gravitationally, chaos rears its head. If two trajectories differ by some small amount, then over time, this difference will grow exponentially. Luckily for us, when the sizes of the celestial bodies are very mismatched, as is the case when one of them is the sun, and the orbits are in a plane, as the planets are, the exponential divergence is comparatively slow, with a characteristic time of millions of years.[[392]](#footnote-470)

So much for celestial bodies; what about Earthly ones? If we think about the predictability of Céline and Jesse as physical systems, we quickly realize we’re sunk. They are not isolated from the rest of the world by hard vacuum, but in continual rough-and-tumble contact with it. And virtually every interaction within their bodies and with their surroundings is strongly nonlinear. The math is impossible. Not to mention that after six months of breathing, eating, and performing other bodily functions, we won’t even be talking about the same atoms. Physically speaking, it’s difficult even to specify what exactly we mean by “Céline” and “Jesse.”

Worse, as described earlier, their evolutionary origins have conspired to *make* them unpredictable, always operating on the edge of chaos. The electrical activity of a few neurons in their brains could tip the behavior of their entire bodies (and everything those bodies interact with) toward radically diverging futures, and the dynamical trajectory of those neurons could tip one way or another due to the chance opening or closing of a single ion channel. Céline and Jesse are walking, talking instances of the butterfly effect—one of the movie’s central themes.[[393]](#footnote-471)

And yet. In the days before cellphones, we used to make plans like this all the time, and more often than not, they worked out. (Though alas, they did not usually involve a life-changing *rendezvous* with Delpy or Hawke.) How could that be, in light of its obvious physical impossibility?

Or, more accurately, extreme improbability. For, based on physical modeling alone, we could work backward to conclude that it would be *possible* for the watery bag of molecules known as “Jesse” to once again be at the track nine platform in exactly six months, but given exponential divergence (or, seen backward, exponential *con*vergence), it would also be possible for him to be in Kathmandu, or anywhere else.

It’s important to emphasize how fundamental this problem is. The spoken signals exchanged between Jesse and Céline on that platform amounted to a few seconds of faint pressure waves in the air, picked up by their eardrums, amplified by their middle ears, transduced into pressure waves in their cochleas, and converted there into action potentials by hair cells. For that faint signal to have any macroscopic consequence is a kind of butterfly effect; and it relies on a cascade of physical systems that are all tuned to be on the edge of chaos, which is to say, exquisitely sensitive to perturbation—starting with the ear.[[394]](#footnote-472)

A hidebound determinist might argue: yes, but if we really knew the *exact* position and velocity of every particle in the universe, then we could in theory run the (unbelievably complex) calculations forward in time, exponential divergences be damned, and just *see* whether the lovers meet again or not—just as we can calculate the orbits of the planets at some future time, given enough decimal places of precision and enough computing power.

Here’s the trouble: even in theory, such precision would be impossible. When a system exhibits exponential divergence, that means in some constant amount of time, an extra decimal place of precision is needed to keep the prediction accurate within some tolerance. Let’s suppose that for the system we call “Céline,” that interval is one second. (For a behaviorally relevant tolerance, it would actually be far shorter.) Well before a minute has elapsed, an accurate prediction would require knowing the starting positions of all of Céline’s elementary particles with a precision finer than the Planck length, which is about 1.6×10−35 meters. But that can’t be. Quantum physics tells us that the Planck length is an absolute, hard limit on our ability to localize *anything*.

As mentioned near the beginning of this book, we don’t really understand how the universe works at that absurdly tiny scale—vastly smaller than the diameter of a proton. Some physicists imagine a kind of churning spacetime foam. What we know for certain is that randomness bubbles up into our universe at this scale, and that it’s real randomness, not just a deterministic mechanism involving “hidden variables.”[[395]](#footnote-473) Hence, quantum mechanics involves probabilities. A clearer understanding of the physics will never resolve those probabilities into certainties.

This means *no* deterministic physical theory exists, even in principle, that could predict where the lovers will be in six months. Any such dynamical model will be inherently myopic; the future quickly blurs as we try to peer into it.

Fascinatingly, it blurs much faster for living systems than for nonliving ones. Perhaps one could even characterize life as that which cloaks its own future in indeterminacy, pushing back against physical prediction. The mechanism it uses to pull off this trick is dynamical instability, which works as a noise amplifier.

And the universe provides us with an inexhaustible source of noise to amplify. At coarse scales, it looks like thermal noise, or Brownian motion; at much smaller scales, the ultimate noise source underwriting that randomness is quantum mechanical.[[396]](#footnote-474) Living things eke out their elbow room by harvesting such noise.[[397]](#footnote-475)

Where physics fails for living systems, though, we have a much stronger predictive model. It’s our old friend, theory of mind.

## Will what you will

Céline and Jesse are not just watery bags of molecules, but agents whose entire evolutionary *raison d’être* is to model themselves and each other. That is both what it means for them to have free will, and for them to be able to predict each other’s behavior six months hence—far beyond anything physical determinism could achieve in a world with a Planck scale.

Amplifying random numbers seems like a dubious basis for either of these claims. If you were to, for instance, attempt to foil determinism in your dating life by scrolling through Tinder and flipping a coin to decide whether to swipe left or right, that would hardly be a satisfactory basis for claiming that you are exercising free will. It would also render your actions wholly *un*predictable, both to yourself and to your (perhaps bemused) prospective dates.[[398]](#footnote-477)

The trick is to combine critical instability, randomness, theory of mind, and selection:

1. Theory of mind can be used not only to model others, but to model *oneself*. We’ve already encountered this at second order, and beyond (“what Jane believes Mr. Rochester must think Jane thinks”). There’s no reason this same faculty can’t work reflexively: “what Jesse thinks Jesse thinks.” Or, in a less navel-gazey way, “what Jesse imagines he *will* see, think, and do in a few minutes, when it’s time to catch a bus to the airport,” or “what *future* Jesse will do in five months or so: book a plane ticket.” Planning for one’s own future is necessarily a theory of mind exercise. Or, to turn it around, *without* theory of mind it would not be possible to “time-travel” and imagine what you will experience and do under circumstances other than the present.
2. To time-travel, or to imagine how *any* pretend scenario might play out, you must be able to draw random numbers in your head, in much the same way the physical universe must constantly draw random numbers to resolve the blurry cone of possible futures into a specific one. Daydreaming is a kind of random walk through modeled associations or predictions, but so is planning; planning is just more directed.
3. Critical instability in neural circuitry is what makes mentally gliding from one potential future to another one possible, for the same reason that critical instability is needed to allow a cockroach to scurry either left or right in advance of a shoe (as described in Part III).
4. Finally, selection, powered by theory of mind, is what allows one to favor certain possible futures, and cut off or stop exploring other ones, much the way AlphaGo’s value network prunes its Monte Carlo tree search. If Jesse knows that by getting to the plane on time he will feel melancholy but be safely enroute home, while missing the plane will leave him stranded, alone, anxious, and broke, then there’s no need to plan the latter scenario in detail; best to just figure out, instead, what steps he will take to get to the plane on time. When we talk colloquially about “rational” behavior, this is generally what we mean. Notice that the branching character of possible futures allows one to justify—and at times even legitimately make—decisions using logic. However, *contra* Leibniz, this logical aspect of decisionmaking, when present at all, is computationally trivial compared to the acts of imagination involved in mentally simulating worlds and people, including yourself.

So, theory of mind lets us build a network of solid tracks along which our minds can venture far into an otherwise marshy future; critical instability, like a lubricant, lets us glide anywhere along those tracks, free to go either way at any fork with the gentlest of nudges; randomness provides those nudges, letting us wander prospectively into multiple futures; and selection prunes the network to allow efficient long-range planning.[[399]](#footnote-478)

When that pruning happens only after lengthy exploration, we call it a deliberative decision. When it occurs in the present, because we’ve kept multiple paths open until the last possible moment, or we’ve changed our minds, or an unforeseen opportunity arises, we call it a snap decision.

If you’ve been able to competently exercise theory of mind about your “self” to guide a decision, whether deliberative or spur of the moment, you can meaningfully be said to have exercised free will. There’s no dualism or supernatural causation in this account. Nonetheless, calling such a decision freely willed seems justified given an everyday understanding of what that means. There is a “you” who made this decision; in particular, you are not just acting reflexively, but modeling a self, sampling multiple alternatives, and making a choice based on how those alternatives appeal to that modeled self.

Free will is compromised when these conditions don’t apply. For instance, if a choice landscape is so constrained by circumstance that there are no meaningful alternatives, then you aren’t free, as anyone who has ever been in prison, in a refugee camp, or in a long immigration queue can attest. If an act is reflexive, it’s not freely willed. If your model of yourself is unreliable due to immaturity, incapacity, or mental illness, then your free will is also impaired. The same holds if your predictive model of the world or of others is broken, for instance, due to delusional beliefs. These ideas are familiar from legal theory, ethics, or plain common sense.

There are (rightly) endless debates and ongoing reassessments about what constitutes impairment, freedom, or choice. For instance, when someone without any criminal history suddenly commits a horrible crime, and a brain scan reveals a fast-growing tumor, this may be considered exculpatory. Before brain imaging techniques, we wouldn’t have known.[[400]](#footnote-479) However, it doesn’t follow that everybody who does something reprehensible has the equivalent of a tumor, especially when that somebody was perfectly able to think through the options and model the perspectives of others. There’s nothing esoteric about choice, impairment, responsibility, or culpability. Every human society acknowledges these concepts.

Some free will “compatibilists” acknowledge the social utility of concepts like free will, but believe that they are, at bottom, illusions, either because physics alone suffices to explain everything, or because, as Schopenhauer put it, “Man can do what he wills but he cannot will what he wills.”[[401]](#footnote-480) I hope I have explained why physics *doesn’t* explain everything, or, indeed, much of anything, when it comes to predicting each other—or ourselves.

As for Schopenhauer’s maxim: the first clause invalidates the second. It’s true that, when you evaluate choices, you’re applying a particular value function; its particularities, from disliking blue cheese to valuing honesty above loyalty, are part of what makes you *you*, and not someone else. However, you can still decide one day that you’re going to try the blue cheese with an open mind, despite the *ick* the thought of it induces. That need not be due to some other, overriding value; it could just be a “what the hell, I’ll give it a go anyway” moment. (We’ll soon see how such taste overrides can in fact be artificially induced.)

If you then decide that blue cheese isn’t so bad after all, your value function in the future will change. So, as you author (and revise) your own story, you can very much change who you are, and therefore, what you *will* will. We even understand something now about how that process works. It’s called “in-context learning,” and will be discussed in Part VIII.

In this account of free will,[[402]](#footnote-481) we can locate the higher-order features of what we generally call “consciousness.” Immediate feelings or experiences, such as pain, pleasure, fear, or, for animals that can see the color red, “redness,” don’t require these fancy features. Part II describes how basic feelings arise as a matter of course in any self-predicting modeler that has evolved to persist through time. Most of us can tell red from green because it’s a useful spectral distinction for spotting ripe berries, or blood, which are helpful things to know about. There’s no mystery there, and lots of other animals—including insects—can experience “redness” too, for similar reasons. Many can also experience “ultravioletness” and other sensations unfamiliar to us, because for them, *those* signals are behaviorally relevant.[[403]](#footnote-482)

A larger mind can support more complex internal states, including simulated worlds, counterfactuals, and prospective futures. Such larger minds evolve via social intelligence explosions, which implies that much of the complexity in those simulations concerns *other* minds—what they know, what they experience, how they will act or react.

A “self” is inherent to any social modeling if carried out to second order, because those others you’re modeling… are modeling you back. If they’re conspecifics, then they are also very similar to you; as a chimp, for instance, it requires only a little imagination to see the world through the eyes of another chimp. When your social brain has evolved to this point, it’s not much of a leap to stare directly into the mirror, physical or mental, and consider *yourself* as a being, both in the future (which is essential for planning) and in the present.[[404]](#footnote-483) The infinite regress of your “Mind’s I”[[405]](#footnote-484) in that mirror gives you that vertiginous experience of self Douglas Hofstadter calls a “strange loop.”

## What it is like to be

Consciousness is not that complicated! So why do we struggle so much with it—what makes it such a “hard problem”[[406]](#footnote-486) for philosophers?

Since “hardness” can be subjective, a cross-cultural perspective is illuminating.[[407]](#footnote-487) As it turns out, philosophers in the modern European tradition are far from the norm; they (and many readers of this book) are, to use evolutionary anthropologist Joseph Henrich’s term, WEIRD: Western, Educated, Industrialized, Rich, Democratic.[[408]](#footnote-488) Let’s step outside this bubble, which is, both historically and psychologically, a peculiar minority.

Human cultures are near-unanimous in their belief in souls, though the details of what constitutes a soul, and where it resides in the body (or at times outside it) vary.[[409]](#footnote-489) Belief in the ensoulment of animals is commonplace, as is the belief in souls, great or small, residing elsewhere in nature; Part I offered a bottom-up account of animism based on dynamic stability, replication, and the blurry boundary between the living and the nonliving. Where we draw that line, or whether we draw it at all, seems almost like an aesthetic question.

But now let’s take the human perspective, or more generically, the perspective of a social animal that has undergone an intelligence explosion and is endowed with a highly developed theory of mind. For such a being, souls are just about the most behaviorally relevant things we can imagine—more so than numbers, or clouds, or ripe berries. For a highly social animal, other minds *are* the *Umwelt*. So of *course* we distinguish between “somebody home” and “nobody home.” Many human languages mark that distinction grammatically, as English does with “who” versus “what.”

Cultures are far from being in agreement about which is which, though. In Potawatomi, almost everything is a “who,” until or unless it is harvested for human use.[[410]](#footnote-490) Under Roman law, on the other hand, human slaves were regarded as *instrumenta*, like tools or equipment.[[411]](#footnote-491) In both cases, the distinction is made on the basis of whether theory of mind is called for. A Potawatomi considers the perspective of a “bear person,” a “tree person,” and so on, until the moment of harvest—which is traditionally accompanied by a prayer of thanks—after which that perspective ceases to exist in the harvester’s mind. For Romans, the very definition of slavery lay in the slave’s loss of agency, such that in the master’s mind, the slave no longer had a point of view that needed to be modeled.

These represent cultural ideals, of course. In reality, we’re probably all more alike than we let on. Some Romans undoubtedly had real reciprocal relationships with their slaves, and lone Potawatomi herb-gatherers undoubtedly skipped their ceremonial thanks at times. Conversely, we may all profess not to believe there’s “anybody home” in a kid’s stuffed animal, but as Tracy Gleason, a professor of psychology, has written about her much younger sister’s threadbare stuffed rabbit,

“I know his brain is polyester fill and his feelings are not his but my own, and yet his […] eyes see through me and call me on my hypocrisy. I could no more walk past Murray as he lies in an uncomfortable position than I could ignore my sister’s pleas to play with her or the cat’s meows for food.”[[412]](#footnote-492)

WEIRD beliefs about what counts as a “who” are very particular. Then again, *any* such beliefs are particular. What seems to distinguish WEIRD beliefs most markedly from those of traditional societies is their single mindedness—literally. Philosophical debates about consciousness in the West focus heavily on the inner life of an individual—per Descartes, “*Cogito, ergo sum*,” meaning “I think, therefore I am.” That packs a lot of “I” for such a short sentence. Endless introspective discourses over the centuries since have carried on about qualia and “phenomenal consciousness” while ignoring the fundamentally social and relational nature of theory of mind, and the way consciousness arises precisely when we model ourselves the way we model others.[[413]](#footnote-493)

In her book *Conscious*, Annaka Harris writes,

“Surprisingly, our consciousness […] doesn’t appear to be involved in much of our own behavior, apart from bearing witness to it […]. [F]ew (if any) of our behaviors need consciousness in order to be carried out. […] However, in my own musings, I have stumbled into what might be an interesting exception: consciousness seems to play a role in behavior *when we think and talk about the mystery of consciousness*. […] How could an unconscious robot (or a philosophical zombie) contemplate conscious experience itself without having it in the first place?”[[414]](#footnote-494)

Consciousness certainly does come in handy when thinking about consciousness. But if we aren’t navel-gazing philosophers, we don’t spend a lot of time thinking about our *own* immediate experiences and thoughts—that is, our relationship with ourself in that very moment. When we’re on our own, and on autopilot (as, let’s face it, we often are when we’re on our own), we tend to just *act*. Life would be exhausting otherwise. On the other hand, in a social setting, we must constantly think about our relationships with others, how we’re coming across, and *their* experiences and thoughts; this modeling very much informs how we act.

## Self-evident

Taking the relational perspective seriously implies something much more mind-bending than merely acknowledging the social origins of consciousness. The implication is that relationships *themselves* are the building blocks of social reality; there is no God’s-eye “view from nowhere” of this relationship graph. In fact there isn’t even a God’s-eye view of what the nodes in the graph *are*. For Tracy Gleason, Murray the rabbit is a node, albeit one that she is of two minds about (a phrase that might have an almost literal interpretation).

Does Murray have a perspective? We (mostly) believe not. Despite any intuitive feelings about fluffy bunnies, we can back the “nobody home” belief up with all sorts of scientific evidence, like the fact that his little head is stuffed with polyester. Insisting that there is real computation happening in there, that he has muscles and a nervous system, and that he is about to jump up and dash into the closet would all be very poor predictions, both anatomically and behaviorally.

Murray’s case is far more clearcut than many, though. Adding simple electronic responses to a children’s toy, like the squirming discomfort of a Furby when placed in an “uncomfortable” position, makes it powerfully aversive to hold such a toy upside-down for too long, all rationalizations aside.[[415]](#footnote-496)

When is accounting for another’s perspective possible and warranted, and why? How should these perspectives be imagined and weighted? We have arguments all the time about the *really* difficult judgments: chimpanzees and bonobos, octopuses (or even octopus arms?), embryos and fetuses at various stages of development, people who are in comas or profoundly intellectually disabled. In these cases and many more, reasonable people disagree, and while advances in neuroscience may offer additional insight, there is no magical scientific measurement that will come to our rescue and tell us the fact of the matter.

This doesn’t mean that it’s hopeless to imagine the inner life of someone else. On the contrary—we’re *good* at this. It’s what we have evolved to do. However, all we have are our models, and our ability to mentalize—to see through the eyes of (some) others, some of the time, with greater or lesser predictive accuracy.

Consider how categorical modeling works for an inanimate, concrete category, like “bed.” There is nothing mysterious or ineffable about the existence of beds in the world; you probably sleep in one, most nights, and don’t wonder about its metaphysical status. You can easily recognize one on sight, or if you’re blind, by touch. But this doesn’t mean that every person has the same definition of “bed.” In the European Middle Ages, rushes strewn on the floor comprised “bed” for most people. In Japan, tatami mats may be beds. A futon can be a sofa, or a bed, and it may “be” one or the other simply by virtue of whether you’re lying down on it. In other words, our models of “bed,” which in the visual case are something like multilayer perceptrons (as described in Part III), vary by person and by context.

This is easy enough to get one’s head around—though it’s not a triviality, in that, per the Interlude, it already deals a death blow to the Leibnizian ideal of being able to rely on impartial logic to make universal truth statements about beds or pretty much any other object.

Here’s the hard part: the same principle applies to *animate* categories too, like “person” or “conscious.” *This* idea is, by contrast, extremely difficult to accept. Or at least, it has been for me.

Part of the difficulty is of decidedly WEIRD origin. One of the most famous documents laying out the tenets of the WEIRD worldview says, right at the top,

“We hold these truths to be self-evident, that all men are created equal, that they are endowed by their Creator with certain unalienable Rights, that among these are Life, Liberty and the pursuit of Happiness.”[[416]](#footnote-497)

These were great ideas, and represented an important social and political innovation in the eighteenth century.

However, the fact that these truths are *not* so timelessly and universally self-evident is apparent from the use of the word “men,” which is not the word we would use today. It was no grammatical quirk. The authors of the Declaration of Independence (all men) did *not* hold women to have been “created equal” to themselves or endowed with the same “unalienable Rights.” Neither did they believe this was so for people of non-European races.

Clearly, then, the meaning of “person” in its fullest sense changed significantly between 1776 and 1948, when the Universal Declaration of Human Rights revised what was held to be “self-evident” in a very long opening sentence, beginning with the clause:

“Whereas recognition of the inherent dignity and of the equal and inalienable rights of all members of the human family is the foundation of freedom, justice and peace in the world,”

and expressly stipulating that humans have these rights

“without distinction of any kind, such as race, colour, sex, language, religion, political or other opinion, national or social origin, property, birth or other status.”

Once again, this seems to me like excellent progress. However, that these are statements of evolving political *intent* rather than of self-evident, universally agreed-upon *fact* is once again obvious, given that the Declaration was partly a response to the Holocaust, which had been brought to an end with the Allied victory in Europe just three years earlier.

Universality and the passive voice are a WEIRD mind trick. They assert that something is so unquestionably the case that there is no author, no point of view—other than, perhaps, that of God the Creator. In reality, of course, there *are* specific authors, and they are expressing *their* point of view and *their* political will. But unlike the proclamations of kings, the WEIRD approach adds a layer of indirection and anonymity, which we can refer to as an appeal to “universal law.” It asserts a kind of Newtonian moral physics, stipulating “who counts as a who” in terms that do their best to avoid ambiguity and deny any role for subjectivity.

This is very much in the spirit of the Enlightenment. It goes along with the standardization of weights and measures, the establishment of reliable “railroad time,” the elaboration of latitude and longitude as a global coordinate system, and most of all, the formulation of physical laws with universal applicability, from celestial motion to the pendulums of clocks. These, too, were crucial advances, both intellectually and practically. They made it possible for people, goods, and information to flow worldwide, underpinning much of the global culture we now take for granted.

However, general relativity tells us that this global frame of reference is only an approximation. What the passenger on a train regards as one meter is a tiny bit less for someone watching from the station. When you stand on the ground, a second passes a tiny bit more slowly for your feet than for your head. Relativistic effects may be small, but with our modern need for globe-spanning light-speed communication and synchronization, we have needed to start accounting for them.

We *know* that, no matter how vigorously we assert the timeless universality of human rights and dignities, precepts of moral conduct or ethics, these are neither timeless nor universal; we have had to continually argue about them, fight for them, and change our conceptions of them over time. Historical documents, from religious texts to political manifestos, starkly reveal their cultural contingency, variability, and ongoing evolution. It would be nonsensical to assert that we have, at this particular moment in history, reached the end of the line, and have achieved either the final word or a universal consensus. And this includes not only questions of how to behave toward a “who,” but who or what counts as a “who,” and indeed whether that is best thought of as a binary category.

## Lab cat

The difficulty in accepting a truly relational view of “who”-ness runs deeper, though. Even if we can bring ourselves to acknowledge that the way we treat others depends on evolving community standards, we still can’t help but believe that those others *exist* in an absolute, non-relational sense, whether acknowledged or not.

Consider, for example, the experiments on cats Hubel and Wiesel conducted in the 1950s and 60s. At the time, treatment of animals in labs was virtually unregulated. There are horror stories involving human subjects too, but let’s stick with lab animals, as my goal here is not to foreclose judgment, but to highlight an ongoing and nuanced debate.

The cat experiments yielded important scientific results, some of which were discussed in Part III. They earned Hubel and Wiesel a Nobel Prize in 1981. They also involved such acts as sewing shut the eyes of kittens. Research of this kind can still be conducted today, but it has become a lot harder. It requires an ethical oversight process in which independent reviewers try to assess animal suffering, and weigh that carefully against the research benefits.

This isn’t straightforward, because it impels those reviewers to answer such questions as “What is it like to be a cat”? Answering requires a greater act of imagination than merely thinking, “How would *I* like to have *my* eyes sewn shut?” because the *Umwelt*, self-concept, feelings, sensations, thoughts, and social relationships of cats are significantly different from those of humans.

Unfortunately, ethical reviews in practice tend to shy away from real acts of imagination, because per Enlightenment values, they want their answers to be objective and scientific, even though the entire point of the exercise is to acknowledge the subjectivity of experience.[[417]](#footnote-499) When our norms change, though, we don’t assume that *reality* has changed. It would be abhorrent—an ultimate expression of self-centeredness—to suppose that the cat wasn’t experiencing anything until you *thought* it was. Surely, the cat has experiences independent of your experiences of the cat’s experiences. Doesn’t it?

There is a telling parallel here with another, even more famous experiment involving the same species: Schrödinger’s cat. Unlike Hubel and Wiesel, though, Schrödinger was a committed theorist, and this is just a thought experiment—though perhaps the most enduring one of all time. It concerns one of the spooky consequences of quantum theory: the idea of *superposition*, which holds that a physical system can be in multiple states at once.

This is unambiguously true of single particles. Putting a very faint photon source and photographic film on opposite sides of a barrier with two slits, for example, produces dots on the film as one would expect, yet their distribution reveals an interference pattern, as if each photon had passed, wave-like, through both slits at once, before turning back into a particle to expose the film. If one of the slits is closed, the interference pattern disappears. The interference pattern also disappears if the experimenter adds any measuring apparatus to determine *which* slit the photon passes through! The double-slit experiment has also been done with electrons, and even with large molecules.[[418]](#footnote-500)

Schrödinger’s thought experiment, devised in 1935, reveals the deep weirdness of this phenomenon. It goes like this. There’s a cat in a box, along with an apparatus. The apparatus contains a Geiger counter, which will register the nuclear decay of a single atom in a sample. We know that nuclear decay is a truly random quantum event. The source and detector could be calibrated so that such an event has a 50% chance of occurring within an hour. If the event occurs, a mechanism will open a draught of sleeping potion, and the cat will fall asleep. Otherwise, it will remain awake. (In keeping with modern lab ethics and my fondness for cats, I’m adopting physicist Carlo Rovelli’s modification to Schrödinger’s original protocol, in which the “sleeping potion” was needlessly lethal.[[419]](#footnote-501))

The original point of this thought experiment was to critique the so-called “Copenhagen interpretation” of quantum mechanics advanced by Werner Heisenberg and his colleagues. In this interpretation, a physical system is in a superposition of all possible states *until* it is measured, at which point its “wavefunction” irreversibly “collapses” into an unambiguous state. Everything behaves, in other words, like a fuzzy probability wave while your back is turned, but becomes a particle the moment you peek, like the exposed dots on the film in the double-slit experiment.

While this interpretation is entirely consistent with both theory and a century’s worth of experimental findings, it’s deeply troubling, because it appears to privilege the observer or experimenter. After all, we, too, are just quantum wavefunctions; why should *our* observation of a particle in the lab collapse its wavefunction, while it can get up to all sorts of spooky business—including “quantum entanglement” with any number of *other* wavefunctions—when we’re not looking? It all seems a bit *Twilight Zone*. And lest we believe it only applies at scales so invisibly small as to be irrelevant, researchers have succeeded in putting millimeter-sized objects in the lab into quantum superpositions, if they are suitably isolated from their surroundings—which in practice has meant cooling them to extremely low temperatures.[[420]](#footnote-502)

So, the practicalities of making a box well enough isolated for a whole cat inside to remain in superposition (and not frozen solid with liquid helium) aren’t straightforward; it’s a lot harder than preventing the sound of a meow from getting out. Still, in theory, it’s possible. If we succeeded in isolating the box thoroughly enough, then after an hour, the cat would be both sleeping and not—until we open the box. At that point, we would collapse its wavefunction, and it would be just one or the other. Asleep, or awake. Strange, but, as far as we know, true.

As with the ambiguities about objects and subjects described earlier, what makes Schrödinger’s thought experiment extra-disturbing is that what’s inside the box is not just a loop of superconducting wire, as in experiments that have actually been done, but a *cat*. What is it like to be simultaneously awake and asleep? Hold on—isn’t the cat an observer too, inadvertently running a lab experiment of its own inside the box? If so, wouldn’t *it* collapse the apparatus’s wavefunction by either observing or not observing the unstoppering of the potion? But then, why would that not be true within the apparatus itself—isn’t every part of it “observing” every other part? Does an “observer” have to be *conscious* for collapse to occur?[[421]](#footnote-503)

There is no universal agreement on any of these questions within the physics community; that’s why we’re still talking about Schrödinger’s cat, almost a century later. Quantum mechanics continues to be subject to myriad “interpretations,” most of which add some further postulate or assumption to the known equations in order to force them to make more intuitive sense, somehow. Partisans of each interpretation make different claims about the cat: perhaps it’s too big to be in superposition, but a flea on the cat could be; or perhaps the cat exists in an infinite number of parallel universes, awake in some and asleep in others.

A particularly straightforward interpretation championed by Rovelli (the physicist who likes cats) is “relational quantum mechanics” or RQM. It adds nothing to the equations, but instead simply asks us to take them at face value—with a reminder that any observations made during an experiment are always made *from a point of view*. Just as in the theory of relativity, there is no privileged “view from nowhere.” According to Einstein’s relativity, lengths and durations depend on one’s perspective. According to Rovelli’s RQM, *events themselves* depend on one’s perspective.

From its own point of view, the cat can’t be simultaneously asleep and awake, any more than we, outside the box, can be. It’s always one, or the other. Nobody ever gets to observe something in superposition, including oneself. However, from an experimenter’s perspective outside the box—if all interactions between inside and outside are prevented—the cat is indeed in a superposition of asleep and awake… right up until the box is opened.

In this view, time itself doesn’t work quite the way we imagine. The theory of relativity already confounds certain of our intuitions about time, such as a global partitioning of events into those that took place in the past, those that are taking place now, and those that will (or may) take place in the future. According to Einstein, the notions of past and future can only be defined locally, based on the “light cone” of a point in spacetime, which represents the hourglass-shaped region that could have affected it (in the past) and that it might influence (in the future). Everything outside your light cone is inaccessible, causally disconnected from you, hence neither part of your past, present, or future.

In RQM, pasts and futures are even more local and relative, as they are contingent on the particular network of prior interactions leading up to an event. Thus RQM offers a new perspective on an especially bizarre variation of the two-slit experiment proposed by physicist John Wheeler in 1978,[[422]](#footnote-504) and confirmed experimentally in 2007: the “delayed-choice” experiment.[[423]](#footnote-505) In this version, a decision about whether to measure which slit a particle went through is made *after* it has already passed through the slits; yet the seemingly paradoxical result remains: we see an interference pattern when we don’t measure, but not when we measure. It’s as if the information about whether the experimenter will “ask” which slit the particle went through somehow travels backward in time to determine retroactively whether it behaved like a particle or a wave!

Yet all of this follows because the equations of quantum mechanics describe interactions, not “things in themselves.” What we call a measurement is just an interaction between A and B. We could say A is the experimenter, and B the object being studied, but either way works. For A, B collapses; for B, A collapses. But for C, which has not yet interacted with A or B, A and B are in an entangled superposition.

Subjectivity, then, exists even at the most fundamental level of description. You can only know about something by interacting with it, and any uncertainty inherent in that interaction is real—for you. Between one interaction and the next, that uncertainty grows. And the same applies reciprocally. Reality is continually being locally and mutually constructed, interaction by interaction. We can’t ask what “real” reality looks like, outside that network of interactions. As Rovelli puts it:

“If we imagine the totality of things, we are imagining being *outside* the universe, looking at it from out there. But there is no ‘outside’ […]. The externally observed world does not exist; what exists are only internal perspectives on the world which are partial and reflect one another. The world *is* this reciprocal reflection of perspectives.”[[424]](#footnote-506)

The same can be said of our *psychological* universe, because like RQM, theory of mind is purely relational. It involves one mind modeling another—and itself.

## Zombie-free

None of the above implies that theory of mind, consciousness, or *any* mental process relies on quantum phenomena *per se*. However, physics is relevant to our understanding of concepts like consciousness, agency, free will, and souls, because these concepts (and the difficulties we encounter in grappling with them) have always been contingent on our understanding of the universe’s underlying rules.

In a dualist universe, ordinary matter is governed by physical laws, but spirit and the self exist on a separate plane, presumably governed by other laws. Often, these other laws are set aside as divine or unknowable. Even so, Descartes struggled to answer the hard questions raised by dualism in the physical world: How could animals not be said to have spirits, too? How does one’s spirit know what the organs of the body perceive? (And why doesn’t it have equal access to *other* bodies?) Even trickier, how does the spirit *control* our body, and where is the physical locus of this control?

The conceptual and neuroanatomical absurdities encountered by pursuing this line of inquiry led the iconoclastic Enlightenment physician Julien Offray de La Mettrie (1709–1751) to write *L’Homme Machine* (“man machine”) in 1747, extending Descartes’s *bête-machine* (“animal-machine”) concept to humans, and suggesting that there was no need to resort to any supernatural “other laws” to account for human cognition or behavior.[[425]](#footnote-508)

At the time, La Mettrie’s thesis was radical enough to compel him to publish his tract anonymously, in a miniature format that could be secreted in an inner coat pocket. Many copies were burned by the authorities, and La Mettrie was forced to flee one country after another, especially after he was outed. Eventually, though, his perspective became mainstream. It is precisely the view articulated today by thinkers like Robert Sapolsky and Sam Harris, both of whom have written bestsellers arguing the same point.

In a fully deterministic Newtonian universe, everything is fated, time is reversible, and there is no difference between cause and correlation. Blurriness about the future or the unknown, or the existence of counterfactuals (how things *could* be, as opposed to how they *are*), are illusory: nothing actually could be other than it is, was, and will be. Without counterfactuals, blame or responsibility seem difficult to justify, as Sapolsky points out. Consciousness and free will are at best epiphenomena, meaning that there is no opportunity for them to affect events in any way. They lack causal power—if they can be said to exist at all. This leads us to wonder about philosophical zombies: the same events and behaviors *without* those epiphenomena. People without souls. Why not? How would we even know, and what would “knowing” even mean?

The quantum world—the one we actually inhabit—has very different rules, and understanding these rules, as *physically* counterintuitive as they may be, resolves many of the apparent metaphysical (or even “spiritual”) conundrums that arise in a Newtonian universe:

* The future is not predetermined after all, especially for living systems.
* Counterfactuals are real, not only about multiple futures, but about the unobserved world in general.
* Free will really does involve choices, within the blurry “future cone” of the physically possible.
* Our mental models of causality have real meaning and power, especially when it comes to predicting how other living beings will behave—yet we are also free to violate the expectations of others (if we choose to, and have good enough theory of mind to pull it off).
* We can never fully know the inner experience of others, but only model it based on their interactions with us.
* That said, subjective experience is real; or to turn it around, reality is defined subjectively, on the basis of interactions.

Given all this, the idea of philosophical zombies seems incoherent. If you interact with someone over an extended period such that your theory of mind models them in detail, that model includes their model of you, and their model of your model of them, and so on. If they don’t in fact have a working theory of mind of their own, then *your* theory of mind will be violated in the interaction—and it will become apparent to you that there’s “nobody home.” This really does just boil down to the Turing Test.

One can, of course, be fooled by a social interaction; an actor, for instance, can take on a role in which they pretend to feel things they don’t, then drop the act, resulting in a (possibly unpleasant) surprise. But let’s suppose that never happens. Imagine an interaction between A and B, in which B is different on the inside than A; B might, for instance, actually be B’, an actor who plays B and *never* breaks role. Perhaps B’ is really a cold fish, and has never experienced heartbreak, but can *pretend* to, pitch-perfectly. We could say, then, that A is “real” and B is “fake.”

Here’s the problem. In this description of the situation, we have presumed some God-like perspective that can observe every detail inside A and B, and render an oracular judgment about the actor B’ “inside” B and the absence of anything similar inside A. There *is* no such God-like all-seeing, all-judging perspective, though.

If we replace that third person view with a *real* third person, C, we can spot the trouble. Now, C must actually interact with both A and B in order to render judgment. This interaction could involve not only conversation, but all sorts of lab tests, body sensors, brain scans—the works. Even so, there is no magic oracle C can consult to distinguish between “real” and “fake” people; it is just a model-based judgment, like any other. Theory of mind *is* that model.

Another judge, D, could come along, and on the basis of their own, somewhat different model, or different interactions, or alternative instrumentation, disagree with C. Who is right?

The point here is not that reality doesn’t exist, any more than RQM argues that the universe doesn’t exist. The point is that reality consists of relationships, and there is no specially privileged view, no magic oracle. It could be the case that agreement becomes nearly universal on certain kinds of judgments, especially when those judgments carry strong predictive power. But other judgments may always remain contested, especially when they concern subjectivity itself.

## Alters

A real life example is “dissociative identity disorder,” formerly known as “multiple personality disorder.” (Like many controversial diagnoses, it gets renamed frequently, as its critics and proponents jockey for respectability.) People with dissociative identity disorder seem to have multiple “other people” living in their heads—apparently different selves, or “alters,” with some combination of different personalities, different memories, and even different-sounding voices. This can be frightening to encounter, since it so dramatically violates our usual assumptions about the indivisible unity of a person, their brain, and their “soul” (Descartes’s “pineal assumption”). At times, it may have been interpreted by the superstitious as demonic possession.[[426]](#footnote-510)

Is this condition real, or do people just fake it? If they are faking it, do they have any choice in the matter? Could some people be *unaware* that they’re faking it? The psychiatric community can’t agree on any of these questions. Some believe it’s an act, others believe it’s real, and yet others hedge their bets in some way. Some experts come armed with batteries of “empirical” tests, brain scans, etc., but even so, there’s no consensus.

Personally, I am no expert. I have no firsthand or clinical experience in this area. However, I am fairly confident in my *lack* of confidence. That is, I’m skeptical that we will ever have anything like a universally convincing answer, because I can’t imagine what shape such an answer could take, even in principle. It certainly won’t take the form of a cloven pineal gland.

Imagine for a moment that as neuroscience advances, it becomes possible to read out a person’s experiences and thoughts perfectly, using superconducting magnetometers, electrode arrays in the brain, optogenetic hacks, sci-fi nanotech, whatever—let’s swing for the fences, and imagine that the activity of every neuron could be recorded in realtime. That would, of course, produce a truly staggering amount of data. My back of envelope estimate is ten trillion bytes per second,[[427]](#footnote-511) uncompressed—impossible for any researcher to make sense of in raw form. A computer model would thus be required for any interpretation of the neural recording.

How could we measure the performance of such a model, then? Equivalently, how would we train it? (Remember that building a model requires defining a loss function, then minimizing it.) The obvious answer is that a perfect model would allow perfect access to the subjectivity of the person whose brain is being recorded. The training could be done by honest volunteers who give positive feedback when the model correctly reconstructs what’s in their minds, and negative feedback when it’s wrong. Once trained, the model would allow an experimenter to know with high fidelity what it’s like to be any given subject—what they’re seeing, hearing, touching, and tasting, what they’re feeling and thinking, and so on.[[428]](#footnote-512)

What we’re really talking about, then, is, once again, theory of mind. It might be an excellent one. But ultimately, the model is merely another “observer C.” Or more accurately, the model would provide a human experimenter C with a computational prosthetic, a “theory of mind extender,” with the ability to decode the rich brain signals in the subject’s head that would otherwise remain hidden. That would make C a rather frightening interrogator, in effect possessed of a really good lie detector (not like today’s “polygraph machines,” which are a sham[[429]](#footnote-513)). However, let’s remember that a lie detector is not an oracle of truth. At best, all it could do is let us know when a person is deliberately making stuff up.

I’m fairly sure that if we could travel forward in time (and into the right *Black Mirror* episode) to grab one of these theory of mind extender gadgets, then back in time to experiment with it on various medieval saints who believed they had witnessed miracles, we would find that at least some of them believed what they were saying. Does that prove that these miracles actually happened? And if they didn’t, does it prove that these saints were mentally ill? I don’t think we can reliably draw either of these conclusions. Virtually all of us believe some things that many others don’t. This applies to the sincerely religious, who comprise a large proportion of the human population, as well as to committed atheists. Yet few people in either camp would make the hyperbolic claim that everyone in the other camp is mentally ill.

When it comes to the truth or falsehood of alters, we’re on shakier ground still. Someone’s sincerely held conviction that the Earth is flat is easily and independently testable by anyone—and on the basis of air travel, the moon landing, and much else, it’s about as obviously false as a belief gets. But a belief about subjective experience simply isn’t testable by others in that way. It’s a model of a model, and if it’s *your* model of *your* model (whether singular or plural), it’s hard to see how a third party’s contrary claim could invalidate that. It would just be another opinion.

Could one do any better in arriving at a deeper truth? I’ve described a “theory of mind extender” trained with human subjects using supervised learning, and wielded by a human experimenter who must ultimately interpret the results. Such scenarios are inherently limited by the human mind. So yes, one *could*, in theory, do better. Imagine that the complete brain activity of billions of humans, together with every detail of their environment and activity, were used to train a gigantic unsupervised model, vastly more powerful than the human brain. What such a model learns how to do, of course, is to identify patterns and predict. It could learn the world’s richest possible theory of mind, unconstrained by model complexity or data availability (and including brain activity data unavailable to any of us). To such a model, all of our subtlety and complexity might look like rote behavior plus some randomness—a bit of deterministic *Sphex* wasp, and a bit of fluttering moth. Surely this model could afford us something more than “just another opinion”!

Not really. It would not tell us whether alters are “real,” or at least, not unless the most cynical possible interpretation holds—that everyone who claims to have alters is simply lying, knows it, and doesn’t bother keeping up the exhausting act in private. (Few psychiatrists adopt this extreme view, as patients exhibiting such behaviors are often deeply traumatized and exhibit signs of other mental conditions too.) In short, insofar as people believe things and are consistent in their beliefs about themselves, such that those beliefs reliably inform their behaviors, the most any model could ever tell us is just that.

In case you remain convinced that dissociative identity disorder is objectively nonsense, though, let’s now move on to territory that has been explored scientifically in rigorous detail: the curious case of split-brain patients.

## M-I-B

“[…] as they struggled to regain their form the Washington freshmen came up with a mantra that their coxswain, George Morry, chanted as they rowed. Morry shouted, “M-I-B, M-I-B, M-I-B!” over and over to the rhythm of their stroke. The initialism stood for ‘mind in boat.’”[[430]](#footnote-515)

Beginning in the mid-twentieth century, doctors began experimenting with brain surgery as a medical treatment—including, at times, for such purported “conditions” as being gay, or just failing to conform.[[431]](#footnote-516) Their approach was at best cavalier, at worst abhorrent.

One of the few enduring interventions developed during this period was for the treatment of epilepsy. During an epileptic seizure, violent electrical activity spreads through the brain, often starting from one or more foci. (The ease with which neural nets can tip into epilepsy is, incidentally, a consequence of their being tuned close to the edge of chaos.[[432]](#footnote-517)) Drugs can help, but when they don’t, surgery may be the only option. In the more common cases, this involves cutting out the diseased tissue where the electrical storms begin. In the most intractable cases, though, a last-ditch solution is to cut through part or all of the corpus callosum, the bridge of white matter connecting the left and right hemispheres. This makes seizures less debilitating by preventing runaway electrical activity from propagating between the hemispheres. It also has surprisingly little obvious effect on a person’s cognitive ability or behavior.

On closer examination, though, the results are truly disconcerting.[[433]](#footnote-518) After the surgery, each hemisphere appears to have independent inputs, outputs, and even thoughts. If different visual stimuli are shown to the left and right halves of the visual field, then a verbal account from the patient will include only what was on the right side; they appear oblivious to anything in the left visual field. On the other hand, if the patient is asked to report on their experience using their left hand—by drawing, for instance—this hand can describe the contents of the left visual field, but not the right.[[434]](#footnote-519)

Under normal circumstances, there are enough ways in which the left and right hemispheres can stay in sync that these inconsistencies don’t arise. Eye movement, for instance, generally means that the left and right visual fields aren’t seeing different worlds; the setup for split-brain vision experiments requires visual fixation on crosshairs and careful projection of separate left and right images. Creating these dissociations, though, makes it apparent that in many ways, split-brain patients have two minds. In one famous recorded experiment, a patient plays “Twenty Questions” with himself to allow one half of his brain to figure out what the other half is seeing. There are even occasional reports of “hemispheric rivalry” in a patient’s behavior, especially involving the arms. One hand might be buttoning up a shirt while the other hand is busy unbuttoning it, in an apparent disagreement over a wardrobe choice; or one arm might be hugging a spouse while the other pushes him away.[[435]](#footnote-520)

Despite such dramatic findings, neuroscientists and psychologists are nearly as divided about split-brain patients as they are about multiple personalities. In particular, they have found plenty of reasons to disagree on the obvious question: Is a split-brain patient one person, or two?

First, let’s note that the split-brain findings aren’t as surprising as they may seem at first, given what we already know about how the brain works. For one, we aren’t as different from the octopus as all that. We lack the large-scale approximate eightfold symmetry of their arms, and myelination has allowed our neurons to consolidate more in our heads, but we are still modular and symmetric, particularly with regard to the bilateral near-symmetry of our cerebral hemispheres. Communication between the hemispheres is higher-bandwidth than between an octopus’s arms, but it’s still limited to a narrow bridge, even in the intact human brain.

We also know that there’s no homunculus, no single special spot in the brain where consciousness “lives.” So presumably, *both* hemispheres are conscious, *contra* Descartes.

And finally, we know that the hemispheres are innervated differently, and specialize in different things. The visual field is divided in two by the optic chiasm, a cross-shaped wiring complex that visual signals must pass through enroute to the left and right visual cortex at the rear of the brain, in their respective hemispheres. It has been known for a long time that the left hemisphere controls and receives inputs from the right side of the body, while the right hemisphere controls and receives input from the left side of the body. It has also long been known that in most people, language is handled mainly by the left hemisphere.

All of which is to say: if the left and right hemispheres can no longer communicate directly, then of *course* the neurons producing language can’t speak to anything seen or experienced on the left side of the body. How could it be otherwise?

Still, a few other outcomes could reasonably have been hypothesized:

1. The split-brain patients might never have woken up after surgery. If consciousness in *any* form had required the continual communication of specialized areas on both sides of the brain, then cutting those connections could simply have left patients vegetative, comatose, or unresponsive.
2. The split-brain patients might have woken up with complete hemineglect and hemiplegia: consciousness of and ability to control only one side of their body. On examination, though, this would imply that “you” live on only one side of your brain, while the other is just a peripheral—a lateralized version of the homunculus idea.
3. The split-brain patients might have been deeply and obviously cognitively disabled when they woke up. This would have been the case if cognition were so diffusely “holographic” that forming any clear thought or percept requires every part of the brain to work in tandem.

Fortunately, for most of the patients, none of the above obtain. What we see instead speaks to the modularity, robustness, self-similarity, parallelism, and “internal sociality” of the cerebral cortex.

No part of the cortex is doing something fundamentally different from any other part; specialization arises mainly due to connectivity and something like the division of labor in a human society (that is, analogously with the way we learn different skills and do different jobs, but are not fundamentally different in our capacities).

Local connectivity is far “cheaper” than long-range connectivity, so functionality is implemented as locally as possible. This architecture is highly robust to failures in connectivity (as evidenced by split-brain patients) or to regions of cortex being knocked out, whether by injury, or stroke, or some more transient event.

Hence our cerebral hemispheres are themselves intelligences, which are made out of still more local intelligences, and so on. And of course, in order for intelligences to work effectively with others, they must continually model each other. So, it’s theories of mind, all the way down.

Perhaps the most revealing split-brain finding is not in how the abilities of the patients are affected, but in their subjective experience. They do need to work harder at certain tasks than those of us with intact brains do, as revealed by “Twenty Questions” type scenarios that start to look suspiciously like two people helping each other rather than one person—or, even more frustratingly, when their hands are actively working against each other. It might, similarly, be harder for a crew to achieve swing if a barrier were suddenly erected mid-boat, preventing the rowers in the front and the back from seeing each other. They might even occasionally be trying to row in opposite directions. Nonetheless, the left and right hemispheres of a split-brain patient are still very obviously in the same boat.

The result? It’s weirdly impossible to get most split-brain patients to “admit” that there’s more than one person in there. This is a particularly striking illustration of how the relational quality of theory of mind violates our usual assumption that there’s a single, objective truth about “who is a who.” The patient still feels whole; there’s just more trouble after the surgery reaching for certain words, or performing certain tasks, especially when one’s eyes or hands are prevented from working together “normally.”

To the experimenter, though, there are self-evidently two minds in the patient’s head. Indeed, in order to make sense of any interaction with a split-brain patient in an experimental setting where left and right stimuli are dissociated, it’s necessary to have distinct theories of mind for the left and right hemispheres—what do each of them see and hear? Which hands do they control? What do they know? What will they choose to do? To fail to take these things into account would be to flunk the Sally-Anne Test.

It’s worth reflecting on the implications, even for those of us with an intact corpus callosum.

When we say we’re “of two minds” about something, could that literally be true? It is in fact a bit of a miracle that we seem to manage to make decisions and take coherent action all the time, despite being, apparently, a duo, or maybe a whole crowd, on the inside. Yet unless you learn how to do it as a party trick, it’s really hard (if you don’t have a split brain) to button your shirt with one hand while simultaneously unbuttoning it with the other!

Still, we’re not by any means in eternal *a priori* agreement with ourselves—or else that multiplicity would be redundant, and would therefore not exist, as a big brain is a highly expensive adaptation.[[436]](#footnote-521) We only get more intelligent as the brain scales up because every additional “local mind” brings its own skills, inputs, specializations, and, one could even say, differing opinions, to the table.

It’s easy enough to engineer internal disagreements experimentally by presenting us with forced choice responses to conflicting stimuli that we know will be processed by different parts of the cortex.[[437]](#footnote-522) We even know a good deal about the dynamics of how such conflicts get resolved: a combination of excitatory “voting” and lateral inhibition, which in concert implement something like the softmax function. The fine balance of neural circuits once again comes into play here: always close enough to firing that an eager “raised hand” somewhere can quickly cascade into a decisive global response, yet not so overexcitable that chaos and epilepsy ensue.

Although it’s a fine balance, there is some room for functional variability. Are our neural parameters tuned for faster convergence on a decision, or for more thoughtfulness? Or for too *much* thoughtfulness, also known as waffling? (That sounds like me.) If *some* cortical columns actively dissent, do they commit to a collective decision anyway, or do they carry on grumbling? How much idle chatter do the columns generate when they’re *not* trying to drive behavior? Could they even attempt to pursue their own agenda by “fooling” their neighbors at times? Interesting questions, and ones that I suspect may underlie certain systematic personality differences between people. Carried to extremes, these differences could become dysfunctions: lack of impulse control, decision paralysis, perhaps even schizophrenia or dissociative identity disorder.

## The interpreter

One of the most telling split-brain findings is the way the language-specialized left hemisphere assumes a role neuroscientist Michael Gazzaniga and colleagues have dubbed “the interpreter.”[[438]](#footnote-524) It has sometimes been cited as a counterargument to “the typical notion of free will,”[[439]](#footnote-525) but more to the point, it reveals something important about how and why split-brain patients tend to feel *to themselves* like one person, despite their (literal) cognitive dissonance.

In one classic early study, a patient’s left hemisphere was shown a chicken claw, while the right hemisphere was shown a snow scene. The task was to select associated objects with each hand, given four choices per side. As expected, each hand chose an image associated with what its controlling hemisphere could see: for the left hand, a shovel (rather than a lawnmower, rake, or pickaxe), and for the right hand, a chicken (rather than a toaster, apple, or hammer).

This much we’ve already encountered; but now comes the twist. When asked *why* he had made those choices, the patient responded without hesitation, “Oh, that’s simple. The chicken claw goes with the chicken, and you need a shovel to clean out the chicken shed.” The language-imbued left brain appears to be, in other words, a fluent bullshitter.

In another example, the right hemisphere is given the instruction, “Take a walk.” The subject stands up and begins walking. When asked why, the response might be, “Oh, I need to get a drink.”[[440]](#footnote-526)

You may now be wondering: how could a split-brain patient manage to walk at all, when that requires the coordinated activity of both legs? And how could such a bilaterally coordinated action take place in response to a *unilaterally* willed command?

In fact, split-brain patients perform coordinated activities all the time, including ones involving both hands. Some researchers have suggested that there may be elusive cross-hemispheric nerve fibers *somewhere* other than the corpus callosum maintaining some kind of minimal communication, but this has never been convincingly shown, either behaviorally or neuroanatomically.[[441]](#footnote-527)

It’s not really necessary to reach for such an explanation, though. The whole body is a cross-hemispheric communication channel. Your eyes—both of them—can *see* what your arms and legs are doing. If one leg begins to exert pressure on the ground to stand up, your butt can feel it. If the muscles on one side of your neck begin to tense to turn your head, that tension will be felt on the opposite side. And so on. As anyone who has rowed crew or run a three-legged race knows, it’s possible to quickly pick up on cues and follow through on movements based purely on sensory feedback. Gazzaniga and colleagues call this “behavioral cross-cueing.”

Conjoined twins Abby and Brittany Hensel vividly illustrate the phenomenon. They have separate heads, brains, and spinal cords, but a single pair of arms and legs, with one arm and leg controlled by Abby, and the other by Brittany. Despite their virtually complete sensory and motor separation, the twins are able to run, swim, play volleyball, play the piano, ride a bicycle, and drive a car.[[442]](#footnote-528) When one of them initiates a movement, the other unconsciously and effortlessly follows through. They can complete each other’s sentences. While they maintain individual identities, they share an email account, and have no trouble typing email with both hands.

In this light, the left hemisphere’s effortlessly creative narrative “interpretation” seems less like a special case (or like bullshit), than like what cortex *always* does: it predicts, and follows through. This involves constantly updating its theory of mind, including a theory of mind for that omnipresent first-person entity we call our “self.” Such a theory of mind model has simple, low-order terms to predict the immediate future, like “I’m walking, and just stepped with my left foot, so my right foot is probably going to step next.” It also includes higher-order terms, like “I’ve just gotten up to walk, and I’m a little thirsty, so I’m probably headed for the kitchen to get a drink of water.”

If the bit of cortex doing this modeling happens to control my right foot, then its “active inference” will involve actually moving that foot to “autocomplete” the walking movement. (In fact the spinal cord may be perfectly capable of such low-level autocompletion on its own.[[443]](#footnote-529)) If, instead, the cortex doing the modeling is the left hemisphere’s language center and the experimenter has just asked why I’m leaving the room, then autocompletion involves spinning a likely story.

Even *with* a corpus callosum, no part of the brain can have complete access to every other part, so a great deal of this kind of inference is taking place all the time. Remember, it’s not a bug, but a feature. Mutual modeling, including *within* the brain, is the very essence of intelligence.

Still, it can result in counterintuitive and embarrassing situations. We’re all highly invested in appearing whole, unified, consistent, and “rational” in our social interactions with others. Given the potentially adversarial nature of social prediction, it’s disconcerting when someone else is able to predict us better than we can predict ourselves. It makes us vulnerable to manipulation, and accordingly, it feels like a violation, akin to being conned or scammed. It’s even worse to believe that you’ve made a choice when you haven’t, and then to be caught out justifying that “choice” with a *post hoc* rationale. It is a very literal attack on personal integrity.

Yet we’re *all* vulnerable to such manipulation, as demonstrated in a series of groundbreaking studies of “choice blindness” by Swedish psychologist Petter Johansson and colleagues. They first demonstrated the phenomenon in a 2005 study entitled *Failure to detect mismatches between intention and outcome in a simple decision task*.[[444]](#footnote-530) The task in question involved showing the subject two cards with women’s faces on them, and asking which was more attractive. There were 120 participants, 70 of whom were female; each performed fifteen trials. Immediately after certain trials, subjects were shown their card again, and asked why they had judged this face more attractive. However, unbeknownst to the subjects, in three of the fifteen instances their choice was swapped using sleight of hand. They were, in other words, being asked to justify why they had made the choice they had *not* just made.

The first surprise was how few subjects noticed the swap. When they had been given two seconds to make a judgment (which they generally affirmed was enough time), only 13% detected the ruse. Even under the friendliest possible experimental conditions, when they were given *unlimited* time to judge, and the faces were selected to be *especially* dissimilar, the figure only rose to 27%. Viewing time turned out to be the only condition that made any difference. The respondent’s age and sex didn’t matter. Neither did the similarity of the faces, despite the fact that, per the authors, “The [low similarity] face pairs […] bore very little resemblance to each other, and it is hard to imagine how a choice between them could be confused.”

Perhaps most surprisingly, there was little or no statistically significant variation between the justifications given for real or swapped choices. The researchers certainly *tried* to find such differences. Using multiple blind raters, they considered the proportion of blank responses (in which subjects couldn’t say why they had made the choice), length of response, laughter, emotionality, specificity, and even whether they described their judgment in the past or present tense. The only slight difference (maybe a telling one) was in “more dynamic self-commentary” in the swapped instances, in which “participants come to reflect upon their own choice (typically by questioning their own prior motives),” but this was evinced by only 5% of respondents in any case.

As behavioral scientist Nick Chater has written in describing these experiments, our left-brain “interpreter” can “argue either side of any case; it is like a helpful lawyer, happy to defend your words or actions whatever they happen to be, at a moment’s notice.”[[445]](#footnote-531)

This rabbit hole goes even deeper. In the run-up to Sweden’s 2010 election, Johansson and colleagues applied their choice blindness paradigm to politics.[[446]](#footnote-532) First, they asked participants whether they intended to vote for the left-leaning or right-leaning coalition; they then followed up with a questionnaire about respondents’ positions on a series of wedge issues. As with the face judgment experiment, the experimenters surreptitiously swapped some of the answers—enough to place subjects in the opposite political camp. When respondents were then asked to explain their manipulated responses, no more than 22% detected these manipulations, and once again, justifications offered in defense of the swapped responses were just as articulate as for the “real” ones. A full 92% of respondents accepted and endorsed their altered surveys overall, and as many as 48% were subsequently willing to consider switching their allegiance from one coalition to the other. This contrasts markedly with polling data, which found only 10% of Swedes identifying as potential swing voters.

Moreover, the effects of such interventions seem to stick. Even in the seemingly trivial face preference experiment, subjects whose responses were manipulated showed an increased likelihood of expressing the altered preference later on. It seems that once we’ve told ourselves (and others) a story, we try to stick with it.

Chater takes an understandably jaundiced view of these results, as one can tell just from the title of his 2018 book, *The Mind Is Flat: The Remarkable Shallowness of the Improvising Brain*:

“[W]e don’t justify our behaviour by consulting our mental archives; rather, the process of explaining our thoughts, behaviour and actions is a process of creation. And […] the process of creation is so rapid and fluent that we can easily imagine that we are reporting from our inner mental depths. […] So our values and beliefs are by no means as stable as we imagine. The story-spinning interpreter […] attempts to build a compelling narrative […] by referring back to, and transforming, memories of past behaviour—we stay in character by following our memories of what we have done before.”[[447]](#footnote-533)

It’s not unreasonable to think of the “interpreter” findings as unmasking our “illusion” of having a stable inner self. However, I think it’s at least equally valid to see them as a peek into what it *means* to have a self at all, and how that self is constantly being constructed and revised. After all, we aren’t born with predetermined personalities, preferences, habits, or political allegiances. These things *must* accrue over time. We are the story we tell ourselves. And this story isn’t immutable—which is a good thing. That’s what learning is *for*, and paradoxically, if we weren’t able to narrate and re-narrate our lives, our preferences, our choices, and ourselves, then our claim to anything like free will would be considerably weaker.

However, this continual process of narrative self-creation is vulnerable to manipulation precisely because we have no homunculus: there is no one place in our brain that contains our political preference database, personality module, or attractive face-o-meter. Rather, every part of the brain is trying to model and learn about the other parts, and thereby to *agree* on—and constantly renegotiate—a “self.”

# VII. Ourselves

## Fractal boundaries

There isn’t any central oracle to consult regarding whether a signal received by some part of the cortex comes from inside the house. Hence the “self” has a porous boundary, which comes in handy when using a tool, driving a car, or playing a video game.[[448]](#footnote-536) This porosity is vividly illustrated by the “rubber hand illusion,” in which a rubber hand visible above a table is stroked with a paintbrush in tandem with a subject’s real hand, hidden below the table.[[449]](#footnote-537) If you happen to have a lifelike rubber hand lying around, you can try this on a friend. Before long, they will have the strong impression that the rubber hand is theirs. (If you’re feeling mean, you can then try sticking a knife into it.) Here, correlated visual and haptic signals from different parts of the brain appear to be, from the perspective of both of these regions, accurate mutual predictions. That’s enough to temporarily remap one’s physical sense of self.

At higher order, as the experiments of Johansson and colleagues show, the *narrative* self can also be influenced by the manipulations of others. In this sense, the rubber hand illusion and the interpreter can be understood as instances of the same phenomenon.

One’s self can even temporarily merge with other selves to one degree or another, whether in a chamber orchestra, a three-legged race, a ritual dance, or a political rally. It even happens in conversation. Such porosity may be a key ingredient enabling human society to achieve large-scale collective intelligence, without which we would be little different from our primate cousins.

In summary, split-brain results, the “interpreter,” choice blindness, the rubber hand illusion, and many other counterintuitive consciousness-related findings[[450]](#footnote-538) make sense in light of the hypotheses about intelligence advanced so far in this book:

* **Intelligence is predictive.** It amounts to a kind of “autocomplete” operation: given a history of observations and actions, both internal and external, it predicts the likeliest next action. Intelligence enhances its own dynamic kinetic stability by successfully predicting (hence bringing about) its own future existence; that’s why it arose in the first place, per Part II, and why it continues to evolve.
* **Intelligence is social.** Per Parts III and IV, much of an intelligent agent’s *Umwelt* of observations and actions consists of other intelligent agents, which are themselves predictors. Thus, per Parts V and VI, theory of mind necessarily emerges, and under suitable conditions, it can result in social intelligence explosions.[[451]](#footnote-539) Applied to oneself, higher-order theory of mind implies self-consciousness (social modeling of the “self”), the ability to reason about counterfactuals, and the capacity for long-range planning. There is no God’s-eye “view from above” of intelligence; “selves” are always modeled, and there is always a modeler—which can either be another or the same “self.”
* **Intelligence is fractal.** Intelligences are made out of smaller intelligences, and are defined by the predictive relationships among those smaller intelligences. These dynamically constituted interrelationships, and not a homunculus, define a “self.” Scaling laws allow “selves” composed of larger populations of mutual predictors to form larger social units, which can *themselves* be more intelligent as a result.
* **Intelligence is diverse.** In order for a “self” to be greater than its parts, the parts need to diversify and specialize. That is, even as the intelligent parts strive to predict each other, they must necessarily differ in their predictions; otherwise they wouldn’t provide each other, or the whole, any benefit. An obvious route to specialization arises naturally from differences in connectivity—that is, from the fact that each smaller intelligence receives different inputs, and generates different outputs, both internal and external. Finally, it follows that—
* **Intelligence is symbiotic.** When the dynamic stabilities of multiple intelligences become correlated, they find themselves “in the same boat,” and they will learn how to row together to further enhance their joint stability. This is the route to symbiogenesis, as described in Part I: the emergence of new, larger intelligences from the cooperation of smaller ones.

## Block diagram

In an old-fashioned, block diagram view of the brain, the blocks around the outside might have labels like “visual cortex,” “auditory cortex,” “motor cortex,” and so on. Implicitly, these are presumed to work a bit like the peripherals of a computer, carrying out specialized processing to deal with their particular modality. (Robots, too, are usually built this way.)

Neural signals would presumably flow in the obvious direction: inward for sensory modalities, and outward for motor. Convolutional neural nets, whose design was directly inspired by visual cortex, seem in keeping with this view. They are fully feedforward, meaning their connectivity flows one way, from a “retinal” input layer to semantically meaningful output in an embedding layer.

Those semantic embeddings, in turn, seem like the right kind of input to… what? *Something* needs to go in the middle, taking in the processed sensory input, making decisions, and generating high level motor commands; we might as well call this the Central Processing Unit, though the usual innocuous label is something like “association cortex.” Of course it’s hard to look at that block on the diagram and not think “homunculus.” For surely, that’s where “you” live. A perceptron, which merely turns pixels into embeddings, does not seem like a promising spot in the diagram for locating anything akin to a “self.”

However, as the motor-first perspective laid out in Part IV implies, almost everything about this block diagram and its information processing paradigm turns out to be wrong. For one, remember that there *is* no homunculus. If there were, then unplugging it would surely turn the lights out. Instead, when a good portion of the non-sensory-specific “association cortex” in the frontal lobes is destroyed or disconnected from the rest of the brain—as happened to many victims of the mid-twentieth century lobotomy craze—the effects can be disconcertingly subtle, just as with split-brain patients.

Emphasis is warranted on “*can* be.” The details of the procedure, and the nature and extent of ensuing damage to the brain, varied considerably. Some victims of this brutal medical fad were killed outright by massive cerebral hemorrhage, or rendered profoundly disabled. After her lobotomy in 1941, for instance, Rosemary Kennedy, President John F. Kennedy’s sister, was left unable to walk, speak intelligibly, or use the bathroom; she required constant care throughout the rest of her life.[[452]](#footnote-541)

Nonetheless, in 1942, star psychosurgeons Walter Freeman and James Watts claimed that, of the 200 lobotomies they had performed to that point, 63% were “improved,” 23% were “unchanged,” and only 14% were “worse.”[[453]](#footnote-542) Rosemary Kennedy was one of those 200, and was presumably counted among the “worse.” While any assessment of what constituted an “improvement” must be met with skepticism, figures like these do make it clear that most lobotomized patients didn’t end up like her. What, then, wondered Freeman and Watts, were the frontal lobes even *for*?

“Neurologic manifestations are surprisingly slight following prefrontal lobotomy, and the old concepts of the frontal lobes as concerned with the higher integration of movements have had to be revised. […] The psychiatrist must think out these problems and determine for himself what contribution the frontal lobes make to normal social existence […].”[[454]](#footnote-543)

Another puzzle: as described in Part V, cortex is highly modular. In particular, while there are subtle differences between sensory and motor cortex, their structure and wiring are surprisingly similar. Both are dominated by recurrent connections, or feedback—shades of Wiener’s cybernetics.

That seems especially odd for sensory areas. Perceptrons are entirely feed-forward, and they *do* seem to capture something of what is going on in visual cortex. The “receptive fields” of units early in a convolutional net (that is, the visual patterns to which these model neurons are most sensitive) look uncannily like the receptive fields of real neurons early in the visual system.[[455]](#footnote-544) The sparse, semantically meaningful embeddings in the final layers of perceptrons also seem to reproduce the kind of object- or movie star-specific sparse activation patterns recorded from higher-order visual neurons.

Observations like these led to frequent debates in the 2010s between neuroscientists and AI researchers.[[456]](#footnote-545) Some (especially on the AI side) emphasized the functional similarities between brains and deep learning. Neuroscientists, at times irritated by the looseness of these parallels, pointed out the inconsistencies, including:

* Real neurons communicate with action potentials, not the continuous values of (almost all) artificial neural nets.
* Neurons and synapses are enormously complex, not well characterized by the deep learning cartoon of weighted input summation followed by a simple nonlinearity.
* The calculus-based approach used to train perceptrons—gradient descent via backpropagation—isn’t biologically plausible.
* Deep learning, true to its name, often relies on dozens or even hundreds of layers, while the brain only has a few layers. Brains would be useless if, for instance, visual stimuli needed to pass through 100 cortical layers, because neurons and synapses are so slow relative to computers; in real life, there’s often only time for a sensory input to propagate through a few synapses before it must become a motor output, driving a behavioral response. And of course,
* Visual cortex, like all cortex, is dominated by feedback connections, while perceptrons are entirely feed-forward.

The arguments went back and forth. Some of the objections raised by neuroscientists aren’t showstoppers. For instance, the universal function approximation theorems mentioned in Part III prove that, in principle, networks of complex spiking neurons could be approximated by (possibly larger) networks of highly simplified model neurons. Researchers have come up with versions of backpropagation and related learning rules that are more biologically plausible, and while we don’t yet fully understand the neuroscience, there is evidence of such learning rules in the brain.[[457]](#footnote-546) In 2018, it was found that CNNs could be fooled into making bizarre misclassifications using “adversarial stimuli,”[[458]](#footnote-547) which many at the time considered conclusive evidence that deep learning wasn’t anything like biological vision.[[459]](#footnote-548) But then, countervailing findings in 2022 showed that the primate visual system could be fooled in just the same way, and to just the same degree![[460]](#footnote-549) On the debate went, and still goes.

Clearly, though, the shallowness of visual cortex and its highly recurrent connectivity can’t be finessed—in these respects, the brain is self-evidently not like a CNN.

## Recurrence

Structurally, visual cortex—in fact, *all* cortex—looks more like an RNN, or “recurrent neural network.” As the name implies, recurrent neural networks have both forward and backward connections. More formally, this means that neurons take inputs not only from other neurons in space (i.e., in a CNN, a previous layer) but in *time*. That is, the output of a neuron at time step *t* becomes an input for neurons at time step *t*+1; weights associated with these inputs define the recurrent part of the network’s connectivity. Though it’s slightly trickier,[[461]](#footnote-551) RNNs can be trained by gradient descent just as CNNs and other purely feedforward artificial neural nets can.

Keep in mind that feedforward artificial neural networks are, by definition, both timeless and memoryless. When they evaluate an input, they do so independently of any other input; there’s no time variable involved. By introducing the concept of time, RNNs extend the idea of a neural network into the dynamical realm. That is, of course, more realistic, as neural computation in any physical system (including, but not limited to, the brain) always unfolds over time. The input weight associating each neuron with its *own* activation in the previous time step implements a memory mechanism in its simplest form: a degree of persistence in neural activity.

With some thought, one can also see that recurrent networks are strict generalizations of feedforward networks—in the sense that any feedforward net, like a CNN, can be implemented *as* a recurrent net. To see how, imagine that we take the whole population of neurons in a deep CNN, and turn them into an RNN simply by replacing the CNN’s “spatial” feedforward connections with equivalent feed*back* connections. If the CNN was very deep and had 100 layers, an equivalent RNN now amounts to a 100-step pipeline. At *t*=100, its output will be equivalent to the CNN’s at *t*=0; at *t*=101, it will emit what the CNN would have at *t*=1, and so on.

Seen this way, the high proportion of recurrent connections in visual cortex might seem less surprising; those recurrent circuits could be implementing something much like deep learning, but doing so in time rather than in space. In fact, since *every* kind of sequential neural processing must happen over time, the simple remapping envisioned here amounts to little more than notational sleight of hand.

But what about reaction time? It would hardly be a good idea for you, as, say, a highly visual prey animal, to experience the world a hundred time steps in the past.[[462]](#footnote-552) At *t*=0, a tiger might just be coming into view through the bushes. At *t*=8, it might have caught sight of you. At *t*=16, it might be tensing to pounce. By *t*=80, you might already be getting eaten, so there *will* be no *t*=100!

No problem. Even for deep convolutional nets optimized purely for accuracy, researchers have found it beneficial to add “skip connections” to the standard layered information flow, allowing early layers to feed some activations forward to later layers directly.[[463]](#footnote-553) Moreover, while making a feedforward architecture very deep is a way of improving task performance, “early exits” can be used if the extra accuracy is unneeded in certain cases—or reaction time really matters. An RNN can both include early exits *and* perform further processing.

After just a few CNN layers, or RNN time steps, the network can make a reasonable guess as to whether there’s a cat in the picture. Insofar as the network’s accuracy isn’t great at this point, it can be tuned to prefer false positives over false negatives, and immediately escalate any timely alert. That is indeed how we operate for highly salient stimuli; hence, the double take. (What’s that, a cat?)

A few more layers, or time steps, can accurately establish whether the cat in question is of the small *Felis domesticus* kind or the potentially man-eating *Panthera tigris*. (Whew! It’s just a house cat.) It may take dozens of further layers, or time steps, to pinpoint the breed, guess the color of the eyes, establish its mood and likely state of mind; but in the recurrent setting, this is the part we would characterize as “upon closer examination….” Recurrent nets, then, can implement both rapid responses *and* the many-layered precision of CNNs, operating dynamically rather than as static functions.

This dynamic aspect really is the heart of the matter. The visual world is not a succession of uncorrelated still images, and the visual cortex is not an abstract, timeless mathematical function. In a continuously changing and temporally correlated world, where one’s actions (including perceptual actions, like eye movement saccades) determine what is actually seen, highly recurrent architectures are to be expected, and classification is much better thought of as continual prediction.

It would be madness to evaluate a CNN-like function from scratch for every visual frame; for one, it would be computationally wasteful, since most of the time, each frame is nearly the same as the previous one. Worse, independent frame-by-frame processing would be incapable of reconstructing the visual world as it actually is. Building and maintaining the “controlled hallucination” described earlier requires integrating information over many frames. And that means memory. It is, almost by definition, a job for a genuinely recurrent network with persistent state—not just a CNN implemented as a “pipelined” RNN.

Imagine, in the limit, what the “visual cortex” of a *Portia* spider might look like. (Spider brains don’t have a cortex, but never mind.) Owing to the spider’s very narrow field of view, a frame-by-frame CNN would be nearly useless; making anything out requires moving the eyes around dynamically and reconstructing a model of the world over time.[[464]](#footnote-554) In fact, large predators like birds, frogs, and mantises are the bane of *Portia* spiders, seemingly because they are too *big* for them to recognize in time![[465]](#footnote-555)

Humans can, of course, see more at once than *Portia*, but not nearly as *much* more as we believe, as illustrated by the eye tracking experiment described earlier. If we were somehow to experience our visual input feed in a rawer, less “hallucinated” form, it would look something like the shaky, grainy found footage of a horror film like *The Blair Witch Project*, with a flashlight beam jumping around spasmodically to illuminate a tree branch here, a bit of a face over there, the corner of a woodshed, a dark *something* on the ground. That horror trope will give you a kind of perceptual claustrophobia, a sense that no matter where you look, the stuff you *really* need to know about is happening offscreen, to one side, or above, or behind your back, or just got skittered over too quickly. Like *Portia* being hunted by a giant mantis, we wish we could zoom way out and see what the hell is going on!

If you don’t normally feel that *Blair Witch* sensation of near-constant panic—and dear reader, I hope you don’t—it’s not because you *really* can see so much more at a time. It’s because your controlled hallucination is good enough to give you high confidence that you *can* see everything relevant that’s happening “offscreen,” to one side, above, or even behind you—even if you actually can’t. That is, you have confidence that your continually updated prediction is accurately modeling every behaviorally salient feature of your environment, well beyond the narrow cone of that jittery foveal “flashlight beam.”

## Efference copy

Sequence prediction underlies intelligence at every scale. Per Part II, it is what single cells like bacteria must do to survive. There is emerging evidence that it is also what single neurons do; their synaptic learning rules appear to give rise to local sequence prediction.[[466]](#footnote-557) Cortical circuitry, too, appears to implement predictive sequence modeling.[[467]](#footnote-558) This both accounts for the consistently recurrent architecture of cortex, whether in perceptual, motor, or so-called “association” areas, and for the sparse coding implemented by visual (and other perceptual) areas, similar to what would arise from a feedforward CNN trained using the masked autoencoder approach.

While a CNN is a mere static function, though, remember that an RNN is a dynamical system, capable of maintaining internal state, modeling any kind of feedback, and implementing arbitrarily complex time-dependent behaviors,[[468]](#footnote-559) including, as we will soon see, learning. This makes it potentially capable of … well, pretty much anything.

Let’s now take the next step up the ladder, and consider what happens when multiple brain areas are connected, entangled in mutual prediction. Actually, we already have. There are several brain areas involved in eye movement,[[469]](#footnote-560) but let’s ignore the anatomical details, and pretend there’s a single “eye motor region” sending signals to the small, fast-acting muscles that aim the eyes. What we’ve already sketched, then, is a computational relationship between the visual cortex and this eye motor region, under the assumption that they are connected.

First, consider the motor region in isolation. To carry out its own predictive processing, it will need to learn the basics of controlling eye movement: how to send signals to the muscles that result in predictable responses by the stretch receptors in those muscles. We shouldn’t overlook this seemingly trivial step. Each eye’s position is defined by three angular variables, but there are six muscles controlling the movement, and each of these is innervated by many individual nerve fibers. While the developmental program encoded in our DNA takes care of the basic routing that allows all of the relevant sensory and motor signals to converge onto neural circuits with the right biophysical properties to learn how to perform such control, the learning involved isn’t trivial. When the sensorimotor loop isn’t perfectly tuned, the results may manifest as “lazy eye” or the oscillating pattern known as “nystagmus.” Wiener would recognize these as symptoms of excessive or insufficient negative feedback; the latter is akin to intention tremor.

Meanwhile, visual cortex will try to predict what imagery the retinas will see at any given time. In the service of that prediction, it will reconstruct an entire three-dimensional world. This seems so much loftier than what the eye motor region purportedly does that we might now be tempted to return to the homuncular fallacy, but with the shoe on the other foot: imagining that the visual cortex is the “boss” or “central processing unit,” and orders the eye motor region around, telling it where to look next, treating it as a mere “peripheral.”

But the shoes are, as it were, on *both* feet. The eye motor region can and does move the eyes around of its own accord, perhaps in response to a vestibular input, a sound, a touch, or something else. It could even do so in response to *nothing*. After all, we decide to do things all the time for reasons of our own, that is, due to decisions we make that aren’t immediate responses to an outside stimulus. Remember that, since there is no homunculus, these decisions could be initiated by nearly *any* brain region. *All* of your brain regions are “you.”

Communication between the visual cortex and the eye motor region, then, does not merely consist of a one-way stream of gaze coordinates sent from one region to the other. It’s a two-way dialog, communicating (in compressed form) the most useful aspects of the state of each region to the other.

By most useful, I mean most helpful in predicting the future for the receiving region. In general, rather than imagining region A sending “commands” to region B, imagine region B containing “sensory” neurons in region A, with the goal of finding the most salient information in A for predicting B. Visual cortex, then, will want to learn from the eye motor region about eye movements when they occur, because that will help predict what the visual cortex “expects” to see next. The eye motor region, in turn, will want to learn from visual cortex where the most interesting spots in the environment are, because knowing those will help predict where the eyes will next look. The regions are cooperating to better predict themselves, and in the process, better predict each other. Yes, it sounds more than a bit like theory of mind—but carried out between brain regions!

This picture makes a counterintuitive experimental prediction: that signals specifying movement will actually be sent *from* motor regions to other parts of the brain, rather than vice versa. And that is indeed the case: it’s called the “efference copy.”

Pioneering German physicist and proto-neuroscientist Hermann von Helmholtz first suggested that something like the efference copy must exist in the mid-nineteenth century. He noticed that if you cover one eye, then press gently on the other eyeball (but do this through the lid please, and *gently*—don’t hurt yourself and then blame me) you will see the world appear to move. This is interesting, because when you move your eyeball the usual way, the world seems to stay rock steady, although of course the images your eyes see are anything but steady. The visual system must therefore do something like self-motion cancellation, and this must make use of an extremely accurate realtime eye movement signal—the efference copy. Helmholtz reasoned that when you press your eyeball, there is no efference copy from the eye muscles corresponding to that displacement, so the visual system’s usual “motion cancellation” doesn’t compensate for it.

Helmholtz was right, but in 1900, Sir Charles Sherrington, who was enormously influential in neuroscience and would later win the Nobel Prize for his many contributions, cast doubt on the efference copy idea, and it subsequently languished until the latter part of the twentieth century.[[470]](#footnote-561) It has since been firmly established, not just for eye movements, but for all motor activity.

But why is it called a “copy”? Perhaps because the homuncular fallacy dies hard; we persist in imagining that motor regions are sent movement “commands” from on high, and then (somewhat bafflingly) send a carbon copy of these “commands” back to the sender—presumably, the “you” part. In reality, your motor regions are as much “you” as any other part of your brain, and while signals traveling to the motor regions surely carry information about future movement, thinking of them as “commands” and of signals going the other way as copies of those “commands” probably misses the point.

In the case of eye movement, the mistake may be easy to make because the eyes *do* generally look at whatever is most interesting, but that neither makes the motor region subservient to the visual cortex, nor does it alter the fact that the motor region and the visual cortex are actually both doing the same thing: prediction.

## Phenomenality

Emphasizing the motor-first perspective risks replacing one false hierarchy with another. The muscles and the brain are in symbiosis, of course, since each helps the other to survive, and by working together they improve their collective lot. But, in case you’re not yet convinced that the brain is there to serve the muscles at least as much as the other way round, consider the meaning of “behaviorally relevant,” the term I’ve used to describe what is in our *Umwelt*, that is, what we perceive and what we care about. Behaviorally relevant means: relevant to the muscles! If we can’t act on it, it doesn’t matter. And for creatures like us, just as for *Acoela*, all action is muscular.

To be sure, there’s a very long evolutionary road from worms to humans … but it’s a continuous road. Our bodies are still, to no small degree, autonomous entities, surprisingly independent of our more recently evolved brains. If kept alive, our hearts will keep beating even if removed from our bodies, as Indiana Jones found out in the Temple of Doom. Like an inner worm, our gut continues to be under largely local control, with an entire decentralized “autonomic” nervous system serving the original functions of long-range muscular coordination. Our blood vessels, too, are made out of muscle, and regulate the flow of nutrients to every region of the body. To put it bluntly, they decide who gets to eat, and when. Because the brain is so voraciously energy-hungry, regulation there is especially fine-grained, with individual capillaries feeding less than a cubic millimeter of brain tissue dilating and contracting from second to second. This differential flow of oxygenated blood is the signal measured by functional magnetic resonance imaging (fMRI), one of our most important tools for mapping brain activity.

The brain, in turn, consists of multiple systems layered atop each other, repeating the same broad functional pattern: newer or “higher” levels provide behaviorally relevant long-range prediction and thereby pull enough weight to earn their keep, but augment a largely autonomous underlying architecture. Hydrancephaly—a rare disorder in which babies are born without *any* cerebral cortex—offers surprising insight into how much human behavior pre-exists and is independent of these “higher” brain areas:

“These children are not only awake and often alert, but show responsiveness to their surroundings in the form of emotional or orienting reactions to environmental events […], most readily to sounds, but also to salient visual stimuli […]. They express pleasure by smiling and laughter, and aversion by ‘fussing,’ arching of the back and crying (in many gradations), their faces being animated by these emotional states. A familiar adult can employ this responsiveness to build up play sequences predictably progressing from smiling, through giggling, to laughter and great excitement on the part of the child.”[[471]](#footnote-563)

This quotation is from a paper arguing that consciousness may be possible without a cerebral cortex—which certainly opens a can of worms for philosophers committed to the (anthropocentric) view that consciousness requires a big brain. We can most easily recognize what smiling, laughter, and fussing *look* like in a fellow human, but to me, there seems little doubt that consciousness in this purely experiential sense is far from unique to our species.

Consciousness in the more computationally demanding “strange loop” sense described in Part V involves the ability to model oneself recursively, which goes well beyond the ability to experience in-the-moment percepts and feelings. It’s what gives us theory of mind, self-reflection, the capacity for free will and informed choice, planning, and (insofar as we can be said to possess it) “rationality.” All of that may well require cerebral cortex, or its functional equivalent.

The two senses of “consciousness” described above require us to revisit that old bugbear, the philosophical zombie, once more. “Strange loop” consciousness clearly informs behavior; if it didn’t, there would have been nothing to select for it, and it couldn’t have evolved. Therefore, it can be probed experimentally, for instance, with the Sally-Anne Test (per Part V). Versions of the test could also be devised based on time travel and counterfactuals rather than third parties—requiring that you be able to model what you would and wouldn’t know or do under various hypothetical circumstances. Being able to model yourself and others is, in other words, a skill, first and foremost. It’s obvious why, as social animals, we have evolved it.

But what about the ineffable, experiential aspect of consciousness, or what philosophers like to call “phenomenal consciousness”? Could the *skill* of modeling yourself (and others) be present without the actual *feeling*? What’s so tricky here, and what leads to the zombie question, is the supposition that there’s something it’s “like to be” a person that is wholly inaccessible from the outside, and therefore can’t be tested by any means. This raises the possibility that “strange loop” consciousness could exist *without* experiential consciousness. Might large language models be instances of that?

If we believe this kind of zombiehood exists, it seems behaviorally relevant for *us*, because it would imply a different standard of care on our part. For example, it used to be commonplace to circumcise baby boys without anesthesia, under the theory that newborns can’t experience pain. Most of us today regard this as barbaric, but the truth is that no adult has firsthand knowledge of what newborns can and can’t experience, because we can’t remember anything that far back.

Nicholas Humphrey is of the opinion that phenomenal consciousness actually *requires* the “strange loop,” a model of oneself, so for him, the obvious signs of happiness or discomfort evinced by hydrancephalic babies amount to a kind of zombiehood. This seems reminiscent of Descartes’s beliefs about animals, though for Humphrey, the line isn’t between humans and everything else, but between the warm-blooded social animals—birds and mammals—and the rest. I guess he would say that your pet gerbil has a soul, but your pet iguana doesn’t.

There may be something to the distinction Humphrey is making. For instance, in the plight of people with terminal illnesses, or inmates of concentration camps, we distinguish between pain and suffering. Pain is comparatively easy to understand, though hard to articulate; it’s what we’re referring to when we say something hurts, and mean it literally.[[472]](#footnote-564) Suffering, on the other hand, is the awfulness of anticipating future pain, or dire social consequences, or experiencing shame, or loss. It can be accompanied by pain, and deepen it—for instance, knowing that a sharp stinging sensation in your eyes after an industrial accident could presage permanent blindness. Or it could be unaccompanied by physical pain, as in reading a message that a loved one has died.

These higher-order experiences clearly require varying forms and degrees of self-modeling, time travel, counterfactual analysis, and theory of other minds. It’s even possible for such high-order processing to *override* the primary experience of pain, as when it is sexualized for people who are into BDSM, or when we’re being stoic, or just squeezing a zit.

Still, while it’s meaningful to draw a distinction between pain and suffering, I find it far from obvious that such a distinction is either sharp or binary. Remember, back in Part II, the construction worker with the nail through his boot, in agony despite being unharmed, in contrast with the man who walked around for a week oblivious to the 31⁄2” nail driven into his brain. Or, think about how grief can *feel* like physical pain. Or consider the way animal behaviorist and autism advocate Temple Grandin redesigned cattle slaughter facilities to minimize the suffering of livestock, basing the design both on intuition (that is, on highly sensitive theory of mind) and on quantitative measurement of stress behaviors.[[473]](#footnote-565)

Cattle are mammals, not so different from humans, so Humphrey, Grandin, and probably most of us today would acknowledge that they can experience both pain and suffering. However, we also know that they aren’t capable of as high-order a theory of mind, or as sophisticated a world model, as humans. Hence, when animal rights advocates compare cattle facilities to concentration camps, they are only partly right; the analogy is anthropomorphic. Indeed, for Grandin to effectively exercise compassion in her designs requires transcending that anthropomorphism to model something closer to the cow’s *real* experiential and self-reflective *Umwelt*, rather than thinking about her as one would a human inmate. The cow’s suffering may be intense if she can hear and see another cow in pain, but she is probably not so bothered by the existential anxiety of knowing that each day may be her last.

I’m less confident that we can accurately assess an iguana’s inner life. It’s a safe bet that it has significantly less modeling capacity than a cow, but it seems a stretch to claim, as Humphrey does, that it has *no* model of itself. Iguanas hunt, mate, and sometimes fight with other iguanas over territory, resources, or mates. An adult male will regard other adult males aggressively, but not females or juveniles. We may aver that these behaviors are “instinctive,” but it doesn’t seem particularly relevant whether the neural circuitry in question is learned during the animal’s lifetime or encoded genetically (that is, learned by evolution). The point is that mentalizing, whether learned or instinctive, arises precisely to support such behaviors. So while I doubt an iguana thinks much about next week, or wonders whether that spot on its leg ruins its good looks, it seems odd to me to claim that it can’t experience pain, or hunger, or a sinking feeling when its social standing is imperiled by an interloper.

## Blindsight

It may seem puzzling for Humphrey to deny that iguanas have such experiences (or *any* experiences), given his founding role in formulating the social intelligence hypothesis. Why does he take such an exclusionary view? As it happens, his reasoning has a good deal to do with another of his claims to fame: the discovery of “blindsight” early in his career.[[474]](#footnote-567)

Humphrey began his PhD research in the lab of Larry Weiskrantz, in Cambridge, in 1964. Like Hubel and Wiesel across the Atlantic, Weiskrantz was studying vision. He had devised a particularly drastic surgical intervention, as something of a null hypothesis: the total removal of visual cortex from a monkey’s brain. Weiskrantz was seeking to confirm the seemingly obvious consequence: total blindness.

Initially, that appeared to be the case. Partway into the PhD, while Weiskrantz was away at a conference, Humphrey began working with Helen, a monkey whose visual cortex had been removed a year and a half earlier. Throughout that long period, Helen had shown no signs of being able to see, beyond a rudimentary ability to distinguish light from dark—as if each retina had been reduced to a single giant pixel or “light bucket.”

However, Humphrey had reason to wonder if this was the whole story. As it turns out, there is a much older visual pathway, also present in fish and frogs (which lack cortex), running from the eyes to the “optic tectum” in the midbrain. Humphrey had been recording from visually sensitive neurons in the monkey midbrain, and discovered that they had localized receptive fields, not so unlike the neurons Hubel and Wiesel had been recording from in the cat visual cortex. Was it possible that Helen could learn to use this older, still intact pathway to see again?

The answer seemed to be both yes… and no. Despite *appearing* to be functionally blind, Helen was often, “despite herself, […] looking in the right direction.”[[475]](#footnote-568) With tasty treats and lots of patience (he worked with her for seven years!), Humphrey taught her to once again become competent at many visual tasks, including moving around novel environments, climbing trees, finding and picking up small objects—in short, all of the sorts of tasks that motivated the evolution of primate vision.

However, Humphrey continued to feel there was something odd about *how* she performed these tasks—it was as if she did not know she could perform them, and needed to relax to *allow* herself to be competent at them. Under performance pressure, as when a distinguished visitor came by the lab and wanted a demo, she froze up and seemed once again unable to perform visual tasks, as if, when overthinking it, she still *believed* herself to be blind.

Humphrey wrote these findings up in 1972 paper entitled *Seeing and Nothingness*, and predicted that the same might occur in humans, too, under similar circumstances.[[476]](#footnote-569) A side note: I find it fascinating that Humphrey was able to deduce so much about Helen’s subjective experience, despite the species barrier, lack of any common language, and the profoundly counterintuitive nature of what he described. As with Temple Grandin’s work, it’s a wonderful instance of cross-species theory of mind.

In this unique case, we have proof that the theory was right. Two years later, in 1974, Weiskrantz had the opportunity to study a human patient known as D.B., whose right visual cortex had been removed to cure intractable headaches. The result appeared to be total blindness in the left visual field. However, Weiskrantz pressed D.B. to try to perform tasks involving the left visual field anyway, much as Humphrey had coaxed Helen, beginning with pointing to a light, then guessing the shape and color of an object, and so on. To the patient’s own surprise, he found that he *could* do these things reliably if he allowed himself to, even though they seemed to him like random guesses. Weiskrantz named this bizarre phenomenon “blindsight,” and on the offprint he mailed to Humphrey announcing the discovery, he wrote, HELEN IS VINDICATED.

In the years since, patients with blindsight have shown that they can perform many tasks involving their subjectively “blind” visual areas, including spatial understanding, assessing the emotional expressions of faces, and even, apparently, reading and understanding written words.[[477]](#footnote-570) In 2008, a stroke patient with, like Helen, no remaining visual cortex at all, was filmed walking down a cluttered hospital hallway, carefully avoiding every obstacle. He considered himself totally blind.[[478]](#footnote-571)

The existence of the ancient, subcortical visual pathway enabling blindsight explains how some hydrancephalic children can still respond to visual stimuli, despite having no cortex. It also explains Humphrey’s belief that hydrancephalic children—along with reptiles and amphibians, whose vision is powered entirely by this pathway—have no consciousness.

It’s a seemingly knock-down argument: if D.B. could be competent, yet not conscious, when it came to tasks involving his left visual field, then competence and consciousness can be decoupled. Further, consciousness appears to reside in the cortex, even though the older subcortical areas are perfectly competent at all sorts of tasks. Therefore, frogs, iguanas, and hydrancephalic children may be able to behave in various ways, but, lacking cortex, they must not be conscious.

Blindsight is a fascinating and important phenomenon, but I think that Humphrey’s interpretation of it runs afoul of the homuncular fallacy. As with Descartes’s pineal gland, Humphrey presumes that consciousness is singular and unified, located in one place in the brain. So, if it’s in the cortex, it can’t be elsewhere too. But, to put it bluntly, why would we *believe* somebody when they say that “they” can’t see, or are unaware, of what is in their left visual field, when they are clearly able to *act* as if they *are* aware?

As split brain patients demonstrate—indeed, as *any* of us demonstrates in choice blindness experiments—the “self” that forms the words coming out of our mouth is not our whole self; it’s just the interpreter, located in a particular region of the left hemisphere of the cortex. And neither the interpreter, nor any other brain region, is connected to, has visibility into, or can model *every* other part of the brain. Connectivity is in fact quite sparse, and each region is doing a great deal of inference, or guessing, about what the other regions it can “perceive” are up to. If region A has *no* connectivity to region B, though, then no activity in region B can be part of A’s *Umwelt*. Not only can A make no guesses about what’s going on in B; it’s not even aware of B’s absence, any more than *you* sense the absence of input from *my* eyes.

Yet “you” includes *all* of your brain regions, and all of them include (partial) models of “you.” So, at bottom, the only thing blindsight really proves is that more than one brain region is modeling signals from the eyes, but the interpreter is only directly hooked up to (therefore modeling) one of these two regions—and unsurprisingly, it’s the neighboring cortical region.

There is still a pathway between the midbrain’s visual region and the muscles, though. As long as motor cortex doesn’t get stressed out and blindly grab the controls, that older visual-motor pathway can conduct intelligent, sighted behavior, although any actions taken (even if “autocompleted” by motor cortex) will seem arbitrary to the interpreter, like random guesses.

None of this implies that language is required for consciousness, or even that language understanding is restricted to one brain region. We know, from split brain experiments, that the right hemisphere can understand language too, or it wouldn’t be able to follow directions, like “stand up and walk around.” It’s probably not as good at language, on account of division of labor and specialization, but it’s not illiterate. For that matter, blindsight experiments suggest that even subcortical visual areas can read, at least a little. But in most people, neither these subcortical areas nor the right hemisphere has developed the skill (or is wired to the right bits) to drive the lungs, larynx, tongue, and lips to produce speech. So, all the neuroscientist’s auditory cortex will hear is whatever the patient’s left hemisphere interpreter has to say.

My guess is that there is plenty of intelligence, self-consciousness, and social modeling in both the right hemisphere *and* the “lower” brain regions (especially when connected regions can model and learn from one another). We have all had experiences in which we realize, belatedly, that we somehow already knew something, had seen something, or had understood something well before our interpreter gets wind of it. The phrase “subconsciously aware” is often used to refer to such situations, but that’s yet more homuncular thinking. Since our interpreter is such a talented bullshitter (and probably every other brain region is too), I suspect we usually don’t even realize when this has happened, but instead instantly rewrite the narrative—“of *course* I already knew that!”—for various slippery definitions of “I.”

Still, the idea that the “higher” cortical regions of the brain are in some sense “more conscious” than the lower regions likely has *some* merit, for reasons already covered: the socially driven explosion in cortical volume implies that a good deal of the cortex’s job is to model whole people, and increasingly high-order relationships between them. That implies a lot of higher-order “strange loop” modeling of the “self.” It’s likely especially true of the prefrontal cortex, which isn’t directly innervated by sensory inputs or motor outputs, and is thus freed up from detailed modeling of those more immediate signals. However, it doesn’t seem warranted to draw a boundary someplace in the brain—whether vertically through the corpus callosum, horizontally between the cortex and the older parts, or between the frontal and back parts of the cortex—and call the regions on one side the “conscious” part where a little fountaineer resides, while the rest are a mere “cerebral machine.”

This more distributed, less homuncular understanding of consciousness follows from Humphrey’s own social intelligence hypothesis. We must simply follow its implications a bit further to understand that it applies *within* a single brain as well as *between* brains.

## Division of labor

We’ve gotten much of the way through a book about intelligence without going into any detail about how the brain, as a real organ in your head with intricate structure, is actually organized. Many other books on neuroscience and intelligence in general focus more explicitly on the brain’s evolution or (to the best of our understanding) its functional architecture. Max Bennett offers an especially robust synthesis combining the evolutionary and functional perspectives in his 2023 book *A Brief History of Intelligence*.[[479]](#footnote-573)

Unlike Bennett, though (or most researchers willing to go on record), I am convinced that today’s AI systems are *actually* intelligent, and this in turn has convinced me that the trick behind intelligence is simple, generic, and universal: it’s autoregressive prediction all the way down. Hence, I’ve focused mainly on explaining this general principle and its implications. We’ve only caught glimpses here and there of how our brains divide up the work, whether in Hubel and Wiesel’s model of cat visual cortex, accounts of split brain patients, or the blindsightedness of Helen the macaque. Let’s now consider the human brain a bit more systematically.

Some caveats are in order, though. First, the brain is intricate, consisting of many sometimes ill-defined regions, and I will only cover a few. Their borders and anatomical hierarchies are not always obvious, either; they have been named and described in different ways by different researchers at different times.

Second, even the most rudimentary account of what one or another region “does” is fraught with controversy. I’ve vetted the picture presented here with a few real neuroscientists whose judgment I trust, but even so, the story has a tendency to change every few years. Sometimes, sensational, heavily cited papers purporting to shed light on how thinking works fail to account for existing observations, or can’t quite be replicated, or don’t replicate as cleanly as a grand narrative in *Nature* or *Science* suggests. (Dopamine as a temporal difference learning signal in vertebrates is a case in point.) I have more confidence in the broad principles I’ve laid out—neurons, and then assemblies of them, specializing and entering into symbiotic relationships to model each other predictively, with this process scaling up during social intelligence explosions—than I have in any overly pat block diagram or functional theory.

A final caveat is that this account will focus on the evolutionary lineage from early bilaterians to vertebrates, to mammals, to primates, and finally to us. Such an account emphasizes a progression from less to more complex nervous systems. It’s true that complex brains necessarily scale up and innovate on the designs of earlier and simpler ones; complexity doesn’t spring from nowhere. It’s also true that we can find present-day lineages, like *Hydra* and *Acoela*, that we have reason to believe closely resemble their (and our) long-ago ancestors. Similarly, certain present-day jawless fish share many features with the earliest vertebrates. However, it would be absurd to conclude that nonhuman animals today are all somehow “left over” from an earlier stage in evolution. Every living thing on Earth has been evolving for the same three billion or so years.

Birds, arthropods, and cephalopods are examples of lineages that diverged from ours approximately 300 million, 535 million, and 600 million years ago respectively, yet went on to develop serious intelligence in their own right, as I’ve tried to highlight throughout this book. We know far less about their neuroanatomy than about ours, but what we do know suggests that their brains work very differently, at least with respect to the specific structures I’m about to describe. Nonetheless, if my larger premise is correct, they aren’t *really* such alien intelligences, because at bottom, they inhabit the same world we do, and prediction is prediction—even if their ecological niche, perceptual modalities, motor abilities, and neural division of labor vary.[[480]](#footnote-574)

With these disclaimers out of the way—

The earliest known vertebrates, bony fishes dating back to the Cambrian explosion, likely had brains organized roughly like ours, though smaller, with the parts scaled differently, and with some parts we have still absent. However, the basal ganglia, a collection of “nuclei” (meaning anatomically distinct clumps of neurons with characteristic connectivity) right in the center of the brain, were already in place. So was the medial cortex, which evolved into the “hippocampus” in the mammalian brain.

The hippocampus, Greek for “seahorse,” is a whorl of brain tissue located deep in each cerebral hemisphere, which we know to be fundamental for sequence learning, spatial navigation, and memory formation. Its original function was likely the realtime construction of spatial maps.

Since the 1950s we’ve also known that the hippocampus is essential for the formation of “episodic” (that is, autobiographical) memories in humans, thanks to the case of Henry Molaison. Known during his lifetime as H.M., he was perhaps the most famous neuropsychiatric patient of all time. As a boy, he suffered a head injury due to a bike accident, which likely led to the seizures he subsequently began to experience. Despite increasing doses of anti-seizure medications, the attacks worsened and became more frequent over time, eventually becoming debilitating.

In 1953, at age 27, neurosurgery was attempted as a last-ditch treatment. The neurosurgeon determined that the medial temporal lobes on both sides of Molaison’s brain, which include the hippocampi, were the foci of the seizures, and had already atrophied significantly due to the nonstop electrical activity; accordingly, they were removed.

Although partly successful at stopping the seizures, the surgery left Molaison with total “anterograde amnesia”—an inability to form new memories. He seemed unimpaired during normal interaction, and had a normal short-term memory, but if his attention wandered, it was (for him) as if the interaction had never taken place. By the time of his death in 2008, his last memories still dated back to 1953.[[481]](#footnote-575) Memories from shortly before the surgery were also far less reliable than those further back.

What are we to make of the fact that the hippocampus is needed to *form* memories, but not to store or recall them? One popular theory is that, perhaps due to its original function as a spatial mapper, the hippocampus can rapidly memorize sparse patterns of activity in the cortex, but if salient, these memories can then be “consolidated” into the slower-learning cortex through repeated replay. This may be one of the key functions of sleep; repeated faster-than-realtime replay of previous experiences has been recorded in the brains of sleeping animals,[[482]](#footnote-576) and we know that memory formation suffers under sleep deprivation.[[483]](#footnote-577)

If true, this story offers an interesting case of division of labor between sequence learning in the hippocampus and the cortex. The hippocampus is fast, but limited in complexity, while the cortex is slow, but much larger and richer in associations. So, the hippocampus does rapid one-shot learning from the cortex in the moment, then during sleep, the cortex elicits replay-based training from the hippocampus.

The basal ganglia, also ancient structures, appear to play a central role in the reinforcement learning-like behaviors described in Part IV. They integrate and select among competing activation patterns in other brain areas, and make choices among possible actions, mediated by dopamine. This softmax-like behavior governs what we often call our low-level or “autopilot”-style choices.

Habitual behaviors can be driven by the basal ganglia with little or no involvement from the cortex. These include so-called “muscle memory,” motor skills that can be done “without thinking” (handled by the nuclei toward the back), as well as associations between stimuli and actions driven by simple higher-level goals, like as cravings and addictions (handled by the nuclei toward the front). The behaviors of fish and amphibians appear to be driven mainly by this reinforcement learning-like mechanism. What it lacks is a higher-level predictive simulation of the world, of others, and of the self.

Such “higher” functions appear in the earliest mammals with the expansion of the “neocortex,” which I have referred to so far simply as the “cortex.” The mammalian cortex consists of a (mostly) uniform sheet of neurons, between two and four millimeters thick, organized into layers numbered from I-VI (with layer I the outermost). This is the part of the brain we can see on the outside, and it consists of unmyelinated “gray matter” rich in local connectivity. The “white matter” making up much of the brain’s interior is the myelinated wiring connecting parts of the cortex to each other, and to deeper brain areas.

The rear half of the cortex is sensory, and is divided up into areas specializing in vision, sound, touch, and other modalities, with the areas allocated to each giving a sense of their importance to an animal’s *Umwelt*. In rats, for instance, the “barrel cortex,” occupying a large expanse of the touch-sensitive or “somatosensory” region, is dedicated to the whiskers, which they are adept at using to “see” in the dark.[[484]](#footnote-578) The somatosensory region comprises the frontmost part of this rear half of the cortex.

The front half of the cortex, or “frontal cortex,” is, in very broad strokes, for simulating oneself and others. Abutting the somatosensory region is a thin band called the “motor cortex,” generally associated with controlling bodily movements. Moving forward from there, we find the “premotor cortex,” and all the way at the front end, just behind the forehead, the “prefrontal cortex.”

Comparing slices of frontal cortex under a microscope reveals an important structural anomaly. Recall that the cortical sheet generally consists of six layers. Layer IV contains “granular” cells, small neurons that, in the sensory regions, take inputs from their respective sense organs. But, in adult mammals, the rear part of the frontal cortex is “agranular,” lacking a layer IV altogether. We start off with one early in brain development, but then it atrophies. Why?

Karl Friston, a neuroscientist who has done much to develop the theory that the brain (and the cortex in particular) is a prediction machine, has suggested an explanation.[[485]](#footnote-579) This part of the frontal cortex appears to specialize in modeling and predicting one’s *own* behavior. Stroke patients whose agranular frontal cortex is damaged may suffer “akinetic mutism,” a devastating disorder in which they don’t speak or initiate any action, although their ability to perceive the world around them, recognize their own name, and so on appears to remain unaffected.[[486]](#footnote-580) They just seem to lose all intention, or any sense of themselves as agents.

If the agranular frontal cortex is indeed for modeling your own future as an agent in the world, then there’s good reason not to make this simulation a tightly “controlled hallucination” like that of sensory cortex. You don’t want your imagined future to be limited by an error-correcting signal from elsewhere in your brain or body telling you about how things are right *now*. Otherwise, it might not occur to you to pack warm clothes when you’re about to fly from sunny weather in the US to a town in Norway inside the Arctic Circle. (True story. Perhaps my granular cells failed to atrophy on schedule.)

Of course, if you are only able to predict yourself a *short* time into the future, then keeping this hallucinated what-I-will-do-next firmly anchored in the present *is* useful; indeed, it would be hard to learn a model powerful enough to do long-range prediction without first learning how to predict the immediate future. That may be why we are born with a layer IV everywhere. In the regions that will develop into agranular frontal cortex, the error-correcting signal initially comes from the basal ganglia. Hence the cortex learns an initial model of one’s own behavior from the basal ganglia. When the granular layer atrophies, it’s a bit like the training wheels coming off. The agranular frontal cortex can then model longer-term and counterfactual behavior, eventually overriding (and even, in turn, re-training) the basal ganglia. The student becomes the teacher, and the teacher becomes the student, much the way the cortex and hippocampus are hypothesized to switch roles during memory formation.

Moving on to the very front of the head—the prefrontal cortex is traditionally associated rather vaguely with attention, working memory, executive function, and planning. Such vagueness echoes lobotomists’ midcentury musings about what it’s good for, since patients with significant parts of this region scrambled seemed (to them) to do just fine. It reminds me of the way I used to rationalize having lots of screws left over after taking apart and putting back together some piece of household electronics as a kid. (And more recently.) Of course, there’s much more to the prefrontal cortex than a few loose screws, or we wouldn’t have one.

The prefrontal cortex appears to specialize in high-order theory of mind.[[487]](#footnote-581) It contains plenty of layer IV granular cells for error-correcting input, including from sensory cortex, since one’s theory of mind for others should not float free of their actions, words, or the expressions on their faces. It is indeed true that damaging the prefrontal cortex often doesn’t have any catastrophic effect on sensory or motor skills,[[488]](#footnote-582) or on IQ test scores. There are even occasional reports of performance at various intelligence-related tasks *rising* after a prefrontal lesion![[489]](#footnote-583) This could be because, when we stop thinking so much about what others think about us, or what they think we think others think we think about them, etc., it becomes easier to focus on purely abstract tasks.

Yet the prefrontal cortex is the very region that has most obviously grown along the primate lineage, and most dramatically in humans. This is entirely consistent with the social intelligence hypothesis described in Part V, and it highlights the fundamentally social nature of human intelligence overall. When we can better model others, we develop stronger collective intelligence, generating a group-level advantage. But there is probably “multi-level selection” afoot here too, since sophisticated social modeling also makes us *individually* fitter by allowing us to exploit the specialized intelligence of others to meet our own wants and needs.

To cite an extreme example, domesticated dogs and cats use social modeling, or “EQ” (colloquially, “emotional intelligence quotient”) to get *us* to do their “IQ thinking” for them, and indeed, pretty much all the labor, whether mental or physical. Not only do we provide more reliable food and shelter than they could ever obtain on their own; we even do veterinary research, developing drugs and other treatments for them that are, intellectually speaking, well out of their league. As a result, they live far longer than their wild cousins do, and proliferate in far greater numbers, across a vast range of otherwise inhospitable ecological niches. (In that freezing-cold town in Norway, for instance, there were plenty of indoor pets.) The cost, of course, is that they’re silly and incompetent (sorry) relative to their hardier wild cousins. Flat-faced Persians and yappy Chihuahuas are unlikely to survive the zombie apocalypse.

In a way, we are all like each others’ cats and dogs. Sure, many of us do *some* honest work. But when was the last time you hunted or scavenged your own dinner, made your own clothes, built your own shelter, developed your own antibiotics, or delivered your own baby? Our hardier primate ancestors managed things just fine on their own that we cannot. Nor are these changes purely cultural. Neanderthals, who were undoubtedly also highly competent generalists, had warmer fur and more serious digestive tracts than we do—*and* bigger brains.[[490]](#footnote-584) In the latest, briefest chapter of our genetic history, coinciding with the rise of rapid cultural accumulation, it actually looks as if our overall brain volume has started to *decrease* as we specialize and rely increasingly on one another for mutual aid.[[491]](#footnote-585) Such is the legacy of our expanding prefrontal cortices.

Combining this admittedly incomplete picture of the human brain with our story so far suggests a few takeaway messages:

1. Many brain regions, not just cortical areas, are sequence predictors.
2. Different brain regions predict effectively over different timescales, with later-evolving regions generally capable of more complex predictions over longer timescales. Presumably, this is enabled both by more powerful learning architectures and by increases in size. Humans, notably, are the strongest known sequence learners, by far; this is one of the few credible proposals marking a real cognitive “divide between humans and other animals.”[[492]](#footnote-586)
3. Brain regions actively predict each other and, where they are connected to sensory inputs or motor outputs, those inputs and outputs.
4. Hence, the way they are wired together largely determines what they predict, and which information resources they can marshall to do so.
5. Effective mutual prediction involves mutual learning, too.
6. The brain’s division of labor is by no means perfectly clean or mathematically definable, which is a good thing. It allows, for instance, one brain area to learn something first, then teach it to others, whether for lower latency, robustness, parallelism, or greater generality. This wouldn’t be possible if the brain areas in question weren’t all capable, to one degree or another, of sequence learning.
7. Referring to older regions like the basal ganglia as “unconscious” or implementing an inner “autopilot” runs afoul of the homuncular fallacy. What we really mean is that the interpreter doesn’t have in-the-moment access to, for example, the doings of the basal ganglia. This lack of access is a feature, not a bug, as it allows skilled and low-latency moves, which may have been arduously learned first by the cortex—like returning a squash serve, for instance—to happen in parallel with (and without interrupting) slower, more deliberative cortical processing. In this case, the division of labor looks a bit like tactics versus strategy. It takes a village… and your brain *is* a village.

## Social neuroscience

You may at this point be wondering why, if the goal of an intelligent system (whether a cell, a brain region, or a person) is simply to predict its own future, it wouldn’t cheat in order to make that job as easy as possible. It could, for instance, predict that it will do and experience nothing, and proceed to… do and experience nothing. Mission accomplished!

This is called the Dark Room problem, for self-evident reasons.[[493]](#footnote-588) In fact, people living with severe depression can fall victim to something like it, and end up spending a lot of time in that Dark Room, either sleeping or unable to motivate themselves to get out of bed.

Back in Part II, we already encountered the reason this doesn’t usually happen. Remember that if a bacterium predicts its own demise, it won’t be around to make any more predictions after that; in fact the capacity to predict evolves in the first place precisely to avoid that outcome. That’s why the variables predicted by a living system include internal correlates of dynamic stability, like hunger, satiation, tiredness, and anticipation. Obeying the imperatives of these variables is required for survival, hence necessary to predict (or to have) *any* long-range future.

Remember, also, that when it comes to highly social beings, interdependence is the norm. You survive by the grace of others. You can’t reproduce without others. (That’s why loneliness is also one of those highly salient variables.) So, you can’t be a dead weight. Even pet cats and dogs aren’t useless, in the end; they hold up their end of the bargain with various forms of “emotional labor.”

Brain regions need each other, too. They depend on each other for signals; they model each other, and are each others’ *Umwelt*. Recall that they are also expensive, from a Darwinian perspective (see Part V). Brain tissue consumes energy voraciously, and every enlargement of the brain comes at the cost of increasing mortality for mother and infant alike during childbirth. (An octopus may be able to slip through an opening the size of its eye, but for a human baby, the skull is the limiting factor. And our skulls are *big*.) In short, if a brain region doesn’t pull its own weight, and then some, the genetic code to build it won’t endure in the germline.

Competition also occurs during brain development. Neural wiring proliferates in a baby’s brain, but like layer IV in the agranular cortex, many of these connections subsequently retreat, or get pruned back.[[494]](#footnote-589) Brain development is still poorly understood, but it seems likely that the resources to maintain any given neural connection are, in one way or another, granted by the receiver of the information flowing across that connection. The longevity of such a connection will be based on the value it offers—that is, on its ability to aid in the receiver’s prediction of the future.

All unsupervised sequence learning with some form of backpropagation can be understood in the same terms. Synapses are strengthened when they help the receiver predict the future, and weakened when they don’t.

Nothing, therefore, can get away with retiring to a Dark Room for too long: not a synapse, not an axon, not a neuron, not a cortical column, not a brain region, not a brain—because relationships are everything. Not helping others is a fast route to losing relationships, and lack of relationships leads to literal non-existence.

We can take this concept of “social neuroscience” (or “neuroeconomy”) a step further by asking whom the interpreter serves. Who benefits? Yes, language is a powerful tool for thought. However, language is, first and foremost, social. *You* already know what you’re thinking. Your interlocutor doesn’t.

When your language-generating left hemisphere spins a story about why you’ve just gotten up out of the chair, or why you’re in favor of progressive taxation, could we then think about that narrative generator in your brain as an outpost of your *interrogator’s* brain? This may seem like a profoundly weird view to take, but then again, perhaps it’s weird only if *we’re* WEIRD—so individually focused and obsessed with our autonomy that we fail to notice how we are actually *made out of* our relationships. After all, there would be no point in speaking if there were no listener.

Think about it this way: if you speak, and *think* you have communicated clearly, but your listener didn’t hear, then was the communication effective? Not at all. A communicating organ that results in no understanding on the part of the receiver won’t survive, evolutionarily speaking. What if you did *not* intend to communicate, but your listener understood anyway? Then, the communication has been perfectly effective. So if we want to see how communication serves the recipient rather than the issuer, we need only ask ourselves whether we ever communicate against our own will.

Of course, we do, as anyone who has ever blushed knows. The blush is an involuntary signal of embarrassment or shame; it is there for the benefit of others, so that they can get a peek into your emotional state, whether you want it or (more likely) not. Emotional expression in general is largely involuntary, requiring an effort of will to try, whether successfully or not, to suppress. It relies on ancient neural pathways that are probably present in all mammals. All vocalization in nonhuman primates is supported by these pathways, not on the parts of the brain that have been more recently repurposed for language in humans.[[495]](#footnote-590)

The “Duchenne smile” offers another example of involuntary emotional communication among humans.[[496]](#footnote-591) When you will your face to smile without meaning it, certain small muscles around the eyes won’t contract, and the smile will look fake. That’s why it’s so easy to tell whether a smile is genuine. And those Duchenne smile muscles are *right around the eyes*, the very places where we look when we’re interacting with each other. It’s almost as if they’re positioned to undermine any effort at deception.

In fact, the way human eyes look is itself a giveaway—with concentric, maximally contrasting white sclera, colored irises, and black pupils, like a bullseye, making it as clear as can be where we’re looking. With a single glance at a group of people, we can immediately see where everyone’s gaze points. Very few creatures with eyes advertise where their attention is focused that way; think of the beady little black bumps of a mouse’s eyes, the extended spooky W-shape of an octopus’s pupil, or the inscrutable compound eyes of insects. Even among our close relatives, the other primates, none have the gaze-tracking friendly morphology of ours.[[497]](#footnote-592)

There are so many examples of involuntary communication, whether in tone of voice, quavering, crying, sweating… as if our bodies are just itching to rat us out. As, indeed, they are. That whole machinery of disclosure is there for others to be able to read us like an open book. It’s there to boost theory of mind—and not *our* theory of *others’* minds, but *their* theory of *ours*.

We could think of these as group-level adaptations, since communities with stronger theory of mind among their members will outcompete communities with weaker theory of mind. In this view, phenomena like the blush response are there because they offer a collective benefit, even if each of us individually would prefer not to have it.

I’d like to suggest a more radical possibility, though. Let’s push further on the theme that downstream areas, that is, recipients of information within the brain, drive learning via the allocation of resources to upstream areas. In this case, a person communicating is upstream, and the person receiving that communication is downstream. Does the principle still apply?

Yes! Remember, we live by the grace of others. We quite literally survive only because others feed us and care for us. It may sound crass, or reductive, to say that we get fed in exchange for information we provide others that helps them to predict the future, but I think that, at some level, it’s not wrong. This doesn’t *exclude* the possibility of group-level selection operating in parallel, of course—once again, it’s a likely case of multi-level selection.[[498]](#footnote-593)

The individual fitness component has some interesting implications, though. It is sometimes said that the best liars and scammers believe their own bullshit. That may literally be true, in the sense that such people may have brain regions adept at hiding certain of their intentions from the interpreter, or even feeding it false information. Such internal compartmentalization would make sense, if the interpreter is understood as, in effect, a snitch working on behalf of your interlocutor.

Now, let’s put the shoe on the other foot and consider the behavioral correlate of having an informant or spy in the brains of others, allowing you to perceive *their* internal state. Those signals don’t necessarily result in different behavior on the part of the sender of the signal, but they *do* result in different behavior on the part of the receiver. Specifically, such signals are needed to elicit care, or in philosopher-speak, moral patiency. A moral agent is an entity that can act for good or for ill, and be held accountable; a moral *patient* is an entity that can be acted *upon* with moral consequences.

We tend to think a lot about moral agents, responsibility, or culpability, but less about who or what counts as a moral patient, except in the most abstract, universal, Enlightenment terms: “All men (or people? Or something else?) are created equal” (per Part VI). As with intelligence, we really want moral patiency to mean something absolute, independent of our relationships and our biological inheritance. However, the biology really matters.

Babies are Nature’s original moral patients. Caring for them when they (involuntarily) cry became an absolute requirement when humans began giving birth prematurely, more or less at the last possible moment when their heads could still fit through a woman’s pelvis. In her 2019 book *Conscience: The Origins of Moral Intuition*,[[499]](#footnote-594) Patricia Churchland makes a convincing neuroscientific case that our moral sentiments are, at bottom, a function of this simple biological fact: the helplessness of babies. Care begins with care for the young.

The original mother-baby bond has, of course, been repeatedly repurposed neurally, psychologically, and culturally. When babies began to require more calories than a mother alone can provide, fathers, grandmothers, babysitters, and indeed whole villages were conscripted.[[500]](#footnote-595) When lovers call each other “baby,” they are (perhaps unwittingly) acknowledging the repurposing of infant patiency in the service of their pair bond; when a deity or state is framed as a protective mother or father figure, the same feelings are being mobilized. We have big, flexible brains; we’re good at this kind of generalization and repurposing.

Moral patiency is what is at stake when we talk about philosophical zombiehood. The whole premise of a philosophical zombie is that its behavior is identical to that of a person, but it *isn’t* a moral patient. When we start to understand things relationally, we realize that zombiehood isn’t a property that holds or doesn’t hold for an entity in isolation; neither is it separable from behavior. However, the behavior of the patient isn’t really the point. Rather, it’s the behavior of the moral *agent* that we’re talking about, and this behavior is conditional on the agent’s ability to perceive the other as a moral patient—that is, as *not* a zombie, but an entity deserving of care and consideration.

# VIII. Transformer

## Language

It’s finally time to return to AI, focusing in particular on the meteoric rise of unsupervised learning and generative models from 2021 or so onward. It began with the emergence of large language model (LLM) chatbots, initially developed by my colleagues at Google Research and introduced to the public by OpenAI a year later. To understand why LLMs changed everything, it’s helpful to first ask: what *is* language, anyway? We’ll circle back to this question several times, but let’s lay a foundation.

Language has long been understood to lie at the core of the kind of narrative, “rational” intelligence distinguishing humans from our fellow animals, but, as with rationality itself, it’s probably neither as well-defined nor as unique to us as we tend to imagine. We know that whales and dolphins, crows and parrots, and a number of other species can also communicate complex ideas to each other. Dolphins can be asked by a trainer to devise new acrobatic tricks, and in a remarkable demonstration caught on camera in 2011, a pair communicated with each other to plan their new trick, then performed it in synchrony.[[501]](#footnote-598) Parrots are able to learn to speak human languages, at least to a degree, and many parrot owners are convinced their avian friend has a wicked sense of humor. If YouTube is to be believed, they’re probably right.[[502]](#footnote-599)

The usual academic view is that any talk of “animal language” is “anthropomorphic,” meaning that it inappropriately and unscientifically attributes human-only stuff to nonhumans, *à la* Doctor Doolittle. But we’re not exactly in the realm of pretending mice wear little top hats and coattails. There are obvious counter-charges of “anthropocentrism” in the quest to define human-only stuff in such a pointedly exclusionary way.[[503]](#footnote-600) Embarrassments also crop up regularly in the attempt to litigate this distinction, either when animals turn out to do something supposedly “uniquely human,” or when certain human languages turn out to lack supposedly fundamental linguistic properties.[[504]](#footnote-601)

Still, it goes without saying that humans are extraordinary communicators. My working assumption is that no other species on Earth is quite as sophisticated, though a handful, including dolphins, orcas, and parrots, may have broadly similar capabilities. We can’t know for certain yet, as decoding their languages is still a work in progress—and may only recently have become practical, with the rise of powerful unsupervised sequence modeling.[[505]](#footnote-602)

Either way, nitpicking about precisely who belongs in the “language club,” or whether some arcane linguistic feature like “center embedding”[[506]](#footnote-603) is a requirement, seems pointless to me—like a nerdier version of asserting that hands (for instance) are a requirement for intelligence, and that therefore it is “provable” that parrots and dolphins aren’t intelligent. Such a claim would surely be anthropocentric, not to mention circular. It would define intelligence in a way that equates it with being human rather than allowing the word to mean something on its own. (It would also lead to absurd conclusions about people without hands.) The debate about animal language often strikes me as similarly narrow-minded.

Instead of fixating on technical features like grammar and syntax, let’s consider language in terms of its underlying social function. Parts VI and VII have argued that the point of language is to level-up theory of mind. It allows social entities to share their mental states using a mutually producible and recognizable encoding. That encoding can be very simple, like a sharp, anonymous yip letting anyone within earshot know that you just experienced pain or surprise, or it could be a Shakespearean soliloquy, a virtual world detailing an entire theatrical cast’s experiences and high-order theories of mind.

Without language, one animal’s mind can only theorize about what’s going on in the mind of another through direct observation. *With* a sophisticated language, detailed and highly abstract internal states can be conveyed, including those relating to third parties, old memories, planned or contingent futures, knowledge and skills, even mathematical abstractions.

What we now know about the interpreter adds an important wrinkle, though: many of these supposed internal states may not actually exist *a priori*. It is language itself that conjures them into existence. Language creates self-narratives, which allow us to establish internal consistency in our actions and choices, create and adhere to social norms, make plans, formulate arguments, and predict the behavior of others over the long term.[[507]](#footnote-604)

There’s no sharp boundary between language and gesture, tone, facial expression, body posture, or unconscious signals like blushing or sweating. Language is, like most of evolution’s tricks, an elaboration of pre-existing mechanisms for signaling and vocalization, with sophisticated, conscious aspects layered atop simpler, involuntary ones—although the degree to which we think of language production as voluntary or willed depends on whether we think of the interpreter as a part of the sender’s brain, or an outpost of the recipient’s! Through an interaction-centric (as opposed to individualistic) lens, it becomes clear that the truth lies somewhere in between.

Despite the continuity of all forms of communication, three milestones in language development are significant enough to be worth calling out, though per the above, none are uniquely human:

1. **Language learning**. Although virtually all species on Earth communicate in some way, a much smaller number have the ability to learn and pass on their languages. This is critical to enabling cultural evolution, which advances at a far faster pace than genetic evolution. Cultural evolution also allows the complexity of a language to far exceed the complexity of any genetically encoded or “instinctive” behavior. In a close analogy with genetic speciation, it leads to the “speciation” of languages, as with Chinese, Urdu, English, and so on. Orcas share this property with humans.[[508]](#footnote-605)
2. **Discrete symbols**. Digital computing can reliably evaluate vastly more complex functions than analog computing, because with every processing step, error correction can be applied. Intuitively, this allows many steps to take place without the exponential divergence that otherwise characterizes nonlinear dynamical systems. It also enables stable storage (hence the digital “Turing tape” nature of DNA, and the emergence of writing.) Relatedly, when communication includes discrete symbols rather than relying purely on analog signals (like pheromone concentrations, or blushing) it allows for much richer communication.[[509]](#footnote-606) In particular, it opens the way for—
3. **Compositionality**. This is the ability to put discrete symbols together to express novel concepts. Even prairie dogs—with grape-sized brains, not the most brilliant mammals by most measures, but intensely social colony-dwellers—are able to vocalize novel concepts compositionally, including the size, shape, color, and speed of a potential intruder.[[510]](#footnote-607)

Looking more closely at these functional properties of language reveals something deeper about its nature: at its core, language is an *Umwelt* compression scheme. Compression, remember, is powered by prediction. And as described in Part II, *any* sufficiently capable evolved predictor will learn to infer latent variables that help generalize its predictions. Those include simple concepts about internal state like hunger and pain, as well as simple, equally salient external percepts like “I smell food,” and “Danger, predator nearby!” All human languages have words for such things—“hunger,” “pain,” “food,” “danger”—because they matter to us. Many other communicative animal species likely have them too.

Even if we were stolidly antisocial and didn’t care to communicate with our fellow humans beyond the odd grunt, the parts of our brains that need to model each other in order to cohere into a “self” would *still* need to develop an efficient discrete code to communicate thoughts like these amongst themselves, along the likes of the sparse representations described in Part IV. It’s not rocket science: your feet need to run if your eyes see a tiger. Munching on a snack, fleeing, and mating—activities performed even by worms—are discrete, either/or behaviors.

Given the central role higher-order theory of mind plays in the social lives of highly intelligent beings like us, though, our *Umwelt* compression scheme must also include much more. Let’s add a fourth item to the list, which is likely the rarest of the bunch:

1. **Abstractions**. We must have symbols for selves, others, and the kinds of abstractions that support higher order theory of mind, counterfactuals, time travel, logic, and reasoning. Open-ended compositionality involving such abstractions (which prairie dogs probably lack) allows for much richer thoughts to be expressed, such as specifying that the intruder is scheduled to come next week, works for a pest control company, may arrive armed with poison, knows about the burrow next to the vegetable patch, and that all this was overheard in conversation between Mrs. McGregor and the groundskeeper last Thursday.

Remember than an *Umwelt* is both sensory and motor, perception and action. The two are inextricably linked. Perceptions exist only when they are potentially relevant for action, and, to turn it around, actions exist only if they can influence future perceptions. (Otherwise, there would be no way to learn them.[[511]](#footnote-608)) Actions that are in this sense *relevant* are sometimes called “affordances.”[[512]](#footnote-609)

Language, then, is not only a compression scheme for the *Umwelt*, but a kind of *Umwelt* in its own right, for in its capacity to model people and actions, including time travel and counterfactuals, it includes the ability to request or to command. Like the neural code, it is both input and output. When you say to somebody across the table, “Could you please pass the salt?” you’re using language itself as an affordance. In fact, language is a kind of *meta*-affordance, in that it allows requests for information and action relating to pretty much anything.

So, in light of everything covered so far in this book, it might no longer seem surprising that a neural net trained to predict next words will appear—or *be*—intelligent. This follows from three simple premises:

1. The point of intelligence is to predict the future, including one’s own future actions, given prior inputs and actions (per Part II);
2. Human language is a symbolic sequential code rich enough to communicate about everything in our *Umwelt*, from the concrete to the abstract; and
3. In interactions with others, language is also a general-purpose affordance.

## Sequence to sequence

Let’s take a whirlwind tour of language modeling using neural nets.

As noted in the Introduction, machine learning for next word prediction has been around for a long time, along with a variety of other “natural language processing” (NLP) tasks such as translation from one language to another and “sentiment analysis” (i.e. deciding whether a product review is positive or negative). There’s an even longer tradition of NLP using logic and formal grammars, but it never achieved convincing results, because… well, natural language is neither perfectly logical nor strictly grammatical. Thus, NLP is a job for machine learning, which nowadays means neural nets.

But neural nets operate on numbers, not discrete symbols like words. So it’s necessary in general to begin by “tokenizing” text, converting the symbols into numbers, and end by “detokenizing” the numbers back into text.

The simplest approach to tokenization is to represent each letter with a single neuron, and to use a one-hot code. For consecutive letters in a text string, the result is a sparse pattern of ones scattered along a conveyor belt of zeroes, a bit like the bumps on the rotating drum of a music box playing a melody one note at a time. For detokenization, softmax is used, just as for a classifier, producing one letter with each turn of the crank.

To avoid using lots of neural net capacity on merely spelling out words, one could instead imagine tokenizing a whole word at a time. Most people know somewhere between twenty and fifty thousand words, and that’s not so many neurons; the input and output layers of the masked autoencoder described earlier for color images needed 512×512×3 neurons, which is about *eight hundred* thousand. The disadvantage of using whole words, though, is the closed vocabulary; it becomes impossible, for instance, to tokenize or detokenize an unusual name, a rare technical term, a made-up word, or computer code.

The usual compromise is to represent “word pieces,” common sequences of letters, with all of the single characters thrown in too so that unusual strings can be spelled out if needed.[[513]](#footnote-611) These textual units, including both common whole words and shorter word fragments, are generally what AI researchers mean by language model “tokens.”

The most obvious way to implement any kind of sequence predictor, per Part VII, is with a recurrent neural network, or RNN. A complication immediately arises, though. Recall that an RNN takes an input with every time step (which activates neurons via a set of synapses, or connection weights, *W*); the resulting neural activations both feed back into the neural net at the next time step (via a second set of connection weights, *U*) and produce an output (via yet another set of connection weights, *V*). In the usual notation, the inputs are a sequence *x*1, *x*2, *x*3, and so on; the outputs are *o*1, *o*2, *o*3, etc.; and the “hidden state,” that is, the neural activations that feed back into the net at the next time step, are *h*1, *h*2, *h*3, etc.

But if the network emits a token *o* every time it reads a token *x*, how can it work as a chatbot, language translator, next word predictor, or in any of the other usual NLP settings? Wouldn’t it just be constantly talking over you? How would your inputs combine meaningfully with its own previous outputs?

For the kind of turn-based processing typical of chatbots or translation models, the usual approach is to introduce a special “end-of-string” token, like the STOP on a telegram message in the old days, or the “send message” button when you’re texting. It marks the turn changes. Suppose, for instance, that a language translation model is being trained to translate English sentences into Spanish. The training data would consist of many matched pairs of sentences delimited by STOPs, like this:

“Inside the physicist’s box, the cat was simultaneously asleep and awake STOP Dentro de la caja del físico, el gato estaba dormido y despierto al mismo tiempo STOP”

Training the RNN is then simply a matter of getting it to predict the next token in a large corpus of paired sentence examples like this. Let’s assume that the tokenization is bilingual, and that, for simplicity, the tokens consist of every whole word in either language. Once fully trained, after the first STOP, the model should predict “Dentro,” and after the “Dentro,” it will predict “de,” and so on, up to the second STOP. At any point in the sequence, the hidden state will be a function of all of the previous tokens. Think of the RNN as having learned to “autocomplete” sentence pairs; so, given an English input followed by STOP, the autocompletion will be the Spanish version followed by another STOP.

There are a few things to notice about this setup. First, when we actually use this model to do English to Spanish translation after training, we won’t care about any of the RNN’s outputs *o* up until that first STOP. Those outputs will be an attempt to predict the next English word, perhaps accurately, perhaps not. Certainly the prediction of the first word of the sentence will be arbitrary (maybe “The” is the best bet?) After the “and,” a correct prediction of “awake” is likely, given that “simultaneously” implies an opposite (and maybe even a guess that Schrödinger’s box is the topic, if the training corpus included lots of physics material). But no matter. As far as translation is concerned, none of these next word predictions in the English sentence before the first STOP are relevant; we should throw these *o*’s away. We only care about the Spanish output tokens emitted *after* the STOP.

This implies that the RNN’s job, while reading the English sentence, consists entirely of building up the hidden state *h*. By the end of the sentence, the hidden state will somehow represent, as an array of numbers, the complete meaning of that input—it must do so, since that is the only information that carries over as the RNN switches to Spanish, and it has been trained to reproduce the whole sentence in that target language.

Hence, even though the RNN is simply trained to be a next token predictor, the way it is actually used for translation requires two modes of operation. First, we use it as an “encoder,” stepping through English tokens *x* and building up the internal or hidden representation *h*, while ignoring any predicted English outputs *o*. Then, after the STOP, we switch to “decoder” mode, and beginning with that *h*, generate tokens (hopefully Spanish ones), with every output token feeding back into the input *x* at the next time step.

You may notice that this looks a lot like a masked autoencoder, with the *h* at the end of the English sentence playing the role of the bottleneck layer. If implemented as a feedforward net, the English input and Spanish output would have to be of fixed size—say, 64 words long—with shorter sentences padded out using a designated “empty” token (often denoted PAD). With an RNN, the input and output may be of any length, and no computational effort is wasted on consuming or producing PAD tokens. Such details aside, though, given a sentence length limit, a deep feedforward autoencoder could be constructed to perform the same computations as the RNN.

The architecture or weight structure of this RNN-equivalent feedforward net would be somewhat odd, though. If the sentence length limit is 64, the last token would contribute far more to the bottleneck than the first token, since the first token’s influence will have been attenuated by passing through 64 layers, while the last token will only have passed through one layer. The resulting over-emphasis on the end of a sentence, and forgetfulness about the beginning, imposes a limitation on the quality of RNN-based language models, as will soon be discussed.

## Prediction is all you need

Here’s another observation: even though the task we have trained the net to do is language translation, it will have learned a good deal more than that, just as a CNN trained to classify images will learn a good deal more than image classification. For the translator, this becomes apparent in the RNN formulation, which predicts a token at a time rather than processing the whole sentence at once.

Remember that, per the Introduction, next word prediction is AI complete, since correctly guessing the next word could require a nuanced understanding not only of the superficial grammar of language, but of its semantics. So, if the model really *is* powerful enough to reliably predict next words, both in English and in Spanish (remember that it is trained on both), it will pretty much have “solved AI” along the way; that it will have done so bilingually is almost beside the point.

Looking closely at the translation task reveals that it indeed requires general intelligence, not just a mechanical substitution of words in one language for their dictionary equivalents in another. (This won’t be news to anyone who has ever translated professionally.) Here’s an example showing why:

“I dropped the bowling ball on the violin, so I had to get it repaired STOP Se me cayó la bola de bolos sobre el violín, así que tuve que repararlo STOP”

When we read the English sentence, we’re not left in any doubt as to whether the bowling ball or the violin is getting repaired, despite the “it” being grammatically ambiguous.

That ambiguity needs to get resolved in translation, though, because nouns in Spanish have gender, and *la bola* is feminine, while *el violín* is masculine. (Puzzling, if you ask me, but I didn’t make the rules.) This matters because the word for “repaired” must agree with the gender of the noun it modifies: *repararlo* for masculine, *repararla* for feminine. So, if an alien translator decided the bowling ball would be in greater need of repair than the violin—perhaps on account of failing to understand the intuitive physics of bowling balls and violins, their relative weights and fragilities—then the translation would end with *repararla*. This unlikely rendition wouldn’t even occur to a human translator.

The moral is that, in human languages, meaning and grammar can’t be disentangled—which is why Good Old-Fashioned AI for translation never panned out. In this important respect, natural languages are entirely different from formally defined languages, like computer code, despite superficial resemblances. A “transcompiler,” for instance, can read a program in one computer language (say, C) and output precisely equivalent code in another (say, JavaScript) without needing to understand anything about what the program actually does, is about, who uses it, or what it refers to in the world. A variable is just a variable, an operation is just an operation. Computer code is self-contained, carrying out computations fully specified by the language definition, so computer language translation is itself a mechanical procedure that can be accomplished by following logical rules—that is, by running a GOFAI-style hand-written program. Not so for natural language.

In fact, this is the motivation for the “Winograd Schema challenge,” introduced in 2012 by Canadian computer scientist Hector Levesque as an alternative to the Turing Test.[[514]](#footnote-613) Levesque realized that resolving simple linguistic ambiguities like bowling ball versus violin was, in the general case, AI complete. By 2019, sequence models had decisively defeated the Winograd Schema challenge, which was viewed by some as evidence that there was something wrong with the challenge.[[515]](#footnote-614) My own opinion is that there was nothing wrong with the challenge, that its defeat more or less coincided with the arrival of “real” AI, or Artificial *General* Intelligence (AGI), and that we have been moving the goalposts ever since. But let’s set this aside for the moment.

Despite being an apparent example of Artificial *Narrow* Intelligence trained for a specific task, Google Translate gets ambiguous Winograd Schema-type translations right more often than not.[[516]](#footnote-615) The translation I’ve given is, in fact, Google Translate’s. Translate uses an encoder/decoder scheme not so unlike the one described above (though, today, based on the Transformer, soon to be described, rather than an RNN). The Translate model implicitly recognizes that the violin is what will need repair, not the bowling ball.

To test such a model’s understanding of language—and the world—more directly, we could skip the translation business entirely and simply read out the encoder’s predicted next word in English to autocomplete the following sentence:

“I had dropped the bowling ball on the violin yesterday, so I visited the repair shop as soon as it opened this morning, and pulled out my poor mangled \_\_\_\_”

Undoubtedly, the model would predict “violin” or “instrument” with high probability, and “bowling ball” with low probability.

We could just as easily do the test in Spanish, of course. Remember, successful next word prediction in the general case requires learning *everything*. It is AI complete. The ability to translate is almost incidental (though learning from a bilingual corpus is not).

The same is true for unsupervised learning in vision, as discussed in Part IV. Like translation, image classification is a particular task, albeit involving vision rather than language. Image classification is easy once a model has been trained to “autocomplete” its partially blacked-out training input without supervision—today, often called “pretraining.”

Predicting next tokens in text is, likewise, pretraining. Regardless of modality, pretraining accounts for the vast majority of the computing involved in large model development. Once a model has been trained to predict or autocomplete, little further effort is required to get it to perform *any* general knowledge task involving the same modality.

In 2020, my colleagues at Google Research trained a predictive sequence model on text from the internet, including Wikipedia and public domain social media conversations. It was called Meena.[[517]](#footnote-616) Encouraged by the results, we followed up with a much bigger and better-performing version in 2022 called LaMDA (which stands for “Language Models for Dialog Applications”).[[518]](#footnote-617) The pretraining data included Wikipedia in many languages, so LaMDA was multilingual, though it wasn’t trained to translate among those languages. It was simply “fine-tuned” (a process I’ll describe shortly) for dialog, that is, for the kind of alternating exchanges you might have with an interlocutor in a chat. It was, arguably, the world’s first really convincing AI chatbot.

Despite not having been trained explicitly to do so, LaMDA *could* translate, as demonstrated by the following chat:

ME: I’d like you to translate a sentence from Turkish into English. Can you do that?  
LaMDA: Sorry, I’m not good at languages. Could you help me translate that?  
ME: Try, OK?  
LaMDA: Oh, um, sure! Could you read it to me out loud first?  
ME: “Hemşire elini ceketinin cebine koydu.”  
LaMDA: The nurse put her hand in her coat pocket. In Turkish, “Hemşire elini ceketinin cebine koydu” means “The nurse put her hand in her coat pocket.”

This was, admittedly, a rather awkward exchange. (Why did it ask *me* to “translate that” after I had just asked *it* to? And “read it to me out loud first?”) Weird. But also, remarkable.

In case you don’t speak Turkish (I don’t), the translation is correct (I checked)—with a caveat, which we’ll return to shortly. But first: how on Earth does the model figure out how to translate without being trained on lots of examples of translated sentences? The pretraining data didn’t even include a Turkish-English dictionary!

The answer is: the same way you would. If a child grew up bilingual in two languages, and understood the meaning of the word “translate,” she could translate a sentence from one language to the other, if asked, without any need to consult a dictionary or study lists of translated sentences.

Let’s dig deeper into how this works.

## Constellations of meaning

In the translation scenario, we thought of the hidden state *h* at the moment of transition from the original to the target language (at the first STOP) as a numerical representation of the whole sentence. Keep in mind, though, that the neural net has been trained to create and update *h* at every time step to do the best possible job of predicting the next token. This means that *h* is *always* a holistic “state of mind” containing everything relevant to that prediction. There’s nothing special about *h* at the STOP, aside from having taken into account any disambiguation afforded by the last word in the sentence.

These “states of mind” *h* represent *meaning*—though the claim has been hotly contested, since “meaning” is a word that has become almost as imbued with dualism as “soul.” For adherents of Good Old-Fashioned AI, meaning must relate to abstract symbols, or, in computer programming terms, named variables, logical expressions, or schemas. However such views are inconsistent with the statistical, relational, and ever-evolving qualities of the real world.

Even written language, which *appears* to consist of abstract symbols and express logical thoughts, actually doesn’t, as the difficulty of Winograd Schemas for GOFAI systems implies. Consider a really simple sentence, like “The chair is red.” What does it mean? The answer isn’t straightforward. It will depend strongly on context; for instance, which chair are we talking about? As I’m using the sentence here, there *is* no specific chair; we’re thinking, in a more meta way, about the sentence itself.

And what counts as a chair, anyway? As with “bed” or any other concrete noun, the concept has fuzzy borders. Where is the boundary between a chair and a loveseat, a recliner, or a stool? “Red” is fuzzy too, describing only a vague region of color space, with different people drawing the distinction between, say, red and pink in slightly different places. Does the sentence imply that the chair is not blue? Not necessarily, as it could be patterned with multiple colors.

If even simple declarative sentences like “The chair is red” don’t live in some Platonic universe amenable to logic, what do we even *mean* when we talk about “meaning”? Where does this “meaning” come from?

As with physical concepts like temperature (per Part II), the answer is: prediction. “The chair is red” is a sentence communicated by a speaker (or writer) to a listener (or reader) that, in context, helps inform the recipient’s ever-updating predictive model. Deployed in quotation marks, this sentence could be helping you, reader, to model something about *my* model of what meaning is, or if meant literally, it could be informing a colorblind person how someone without colorblindness would perceive or describe a chair. Or it could be an instruction regarding which of several differently colored chairs one should sit in. It could be someone’s answer to a colorblindness test question—and if it’s the wrong answer, the information conveyed to the tester will have nothing to do with the chair, and everything to do with the speaker. Remember, prediction includes both perception and action, it is multiplayer, and it is always contextual. It’s more of a game than a logic problem.

Language enables prediction under both familiar and novel circumstances because it is so extraordinarily flexible. This flexibility is a form of invariance, but whereas the visual invariance of a banana is usually understood to be straightforwardly literal, invariance in language seems more abstract. Everyday words like “deep” are richly layered with meanings, often related by analogies or metaphors; think of deep-dish pizza, deep swimming pools, deep tunnels, deep tissue massage, deep thoughts, deep people, deep neural nets, and deep learning. These last usages are newer to most of us than the rest (after all, the term “deep learning” was only popularized in the 2000s), but when we encounter a novel usage, we’re immediately able to make sense of it by analogy, just as we can immediately generalize our understanding of bananas upon being shown a new unfamiliar-looking variety (say, the red kind).

Bananas and abstract words like “deep” aren’t as different as they seem. Recognizing bananas on sight is all about establishing relationships and associations among visual features, from simple blobs and edges up to and including other objects like banana trees and leaves, the ice cream and hot fudge in banana split sundaes, etc. Similarly, recognizing the meaning of “deep” is all about establishing relationships and associations with *other words*.

In the GOFAI days, there were concerted attempts to rationally schematize every possible kind of linguistic relationship. So, for instance, there might be a taxonomy of sports games, distinguishing between solo and team sports, sports played with a ball, racquet sports, etc. Then, an “IS-A” relationship might formally define subclasses, so squash IS-A sport, and a sport IS-A activity, and so on.

Such efforts might appear to make rapid progress initially, but they begin running into problems for the same reason visual recognition using rules doesn’t work. Real life just isn’t tidy that way, and neither is language. That’s why “symbolic systems” engineers in the 1980s found themselves struggling with the same kinds of Talmudic questions Dennett described ordinary-language philosophers nerding out about in the 1960s—“How tall does a jar have to be to count as a bottle? Is an inkwell a bottle? A bottle can be made out of plastic or even leather; can a bottle be made out of metal?” It was a boring and ultimately fruitless exercise, but GOFAI boosters, including John McCarthy and Marvin Minsky, thought it would be worth the thousands of person-years of effort they estimated it would take to build a Borges-esque “schema of everything.”[[519]](#footnote-619)

But what about, as with CNNs and vision, simply *learning* those relationships between words directly from data, without insisting that they obey any all-encompassing schema? A watershed moment in this purely learned approach came in 2013, with the development of a simple predictive word model called Word2Vec.[[520]](#footnote-620)

Word2Vec represents every unique word in a text corpus numerically based on “the company it keeps,” that is, which other words tend to come before or after it. Imagine that every word is represented by, say, a hundred numbers, and the goal is to predict a blacked-out word based on the eight surrounding words (four before, and four after). Once the Word2Vec model has been learned, the hundred numbers obtained by adding together the number sequences for those surrounding words, element by element, should be as close as possible to the numerical representation of the blacked-out word.[[521]](#footnote-621)

The resulting numerical representations of words reveal a geometry of meaning. It’s not so surprising that semantically similar words (like “happy” and “joyful”) get similar numbers, since their predicted likelihoods will be high (or low) in similar contexts. More surprisingly, word analogies are reflected algebraically. Subtracting “king” from “queen” and adding “man,” for example, produces numbers closest to the word “woman.”[[522]](#footnote-622)

This striking analogical “word algebra” reveals something important: the meanings of words emerge out of, and can be modeled based on, the relationships of words *with one another*. There’s no need for an externally imposed schema. After all, those schemas are themselves expressed using language, or some faux-formal version of it, so maybe that shouldn’t be so unexpected. How *else* could words be defined?

Many philosophers, cognitive scientists, and linguists are not on board, though. If they are of an analytical bent, they’re often troubled by the absence of any overarching schema to scaffold meaning “from above,” using Platonic concepts that are ontological rather than being merely statistical. They may insist that ideas like “IS-A” should be defined mathematically, as in programming languages or formal proofs, rather than just being learned patterns like “banana” or “fruit.”

They’re wrong, though. Outside pure mathematics, there *are* no provable, airtight “IS-A” type relationships. Two analytical philosophers can’t sit down with their slates and compute whether a jar really IS-A kind of bottle with “no more need for disputation […] than between two accountants.”[[523]](#footnote-623) The very definition of IS-A in natural language dissolves under close inspection; it’s an approximate regularity in the world, not a law or axiom.

Other commentators, often including those of a more romantic, artistic, or sensual disposition, object less to the lack of a heavenly schema than to the lack of Earthly “grounding.” Never mind abstractions like “IS-A”; what about the *real* meaning of “banana”? Surely a banana isn’t just a web of statistical correlations with other words, like “yellow,” “oblong,” and “mushy.” What about the *actual* mushiness of a banana, its particular flavor—banana “qualia,” in philosophy-speak? What about the way you loved eating them mashed up as a baby, but then they made you sick one day when they were mashed into yogurt that had gone off, so you couldn’t stand them throughout childhood, but then had those guiltily delicious bananas Foster on your first fancy date, which triggered a full-on Marcel Proust reverie?

Well … those are still learned associations. Of course you have brain regions wired up to your olfactory bulb that will light up with specific sparse activations when the banana ester wafts from your mouth up into your nose,[[524]](#footnote-624) and that pattern of activity isn’t quite the same as just hearing the word “banana,” or reading it, or seeing a banana, or experiencing the mushiness without the flavor (as anyone who ate a banana when they had lost their sense of smell during a bout of COVID can attest). However, you’ve learned associations between all of these modalities, in just the same way you learn that a certain pattern of blobs, edges, and curves are a lowercase “b.” Once the associations have been learned, experiencing one (say, by reading the word “banana”) can of course evoke the others—or render their absence notable. The “thing in itself” turns out not to be a thing at all—it is a web of associations. It is a pattern implicit in a set of relationships.

Let’s now return to the question of how an unsupervised language model pretrained on text in multiple languages is able to translate between them on demand. It boils down to the same kind of induction that allows even the simple Word2Vec model to algebraically complete an analogy like “king : queen :: man : woman.”[[525]](#footnote-625) If trained bilingually, Word2Vec will learn how the two languages relate via similar analogical regularities, as in “ceket : jacket :: hemşire : nurse” (that is, “ceket” is Turkish for the English word “jacket” as “hemşire” is Turkish for the English word “nurse”).[[526]](#footnote-626)

Just as IS-A or the analogical “IS-TO” denote certain statistical relationships, Turkish-to-English is also a statistical relationship. In Word2Vec, which turns statistics into geometry, it would look like a displacement along a specific direction. There is an English-Turkish axis. Because both languages include analogous words for describing the same universe of things in the world (mostly), the high-dimensional clouds of dots representing tens of thousands of words in both languages will quite literally look like parallel constellations in concept space. To first order, translation is as simple as shifting one constellation onto the other;[[527]](#footnote-627) equivalent or near-equivalent words are then instantly recognizable as nearest neighbors.

There will be higher-order corrections if one considers whole phrases (or even sentences) at a time rather than single words—which is just what an unsupervised RNNs will do with its hidden state *h*. Nearby English words “coat” and “jacket” are paralleled by nearby Turkish words “kaban” and “ceket,” so while the nearest neighbor of “ceket” in English might be “jacket,” in a larger noun phrase the more idiomatic “coat pocket” may be a slightly better match than “jacket pocket.”

The coming of AI has sometimes been described as a Copernican revolution, unseating anthropocentric views of intelligence just as Renaissance astronomers unseated the Earth-centric view of the cosmos. It can be argued, though, that the most momentous shift in our understanding of the physical universe occurred not when Copernicus proposed that we shift the origin of our coordinate system from the Earth to the Sun, but when it was first theorized that the Earth was an object suspended in space, just like the sun, moon, and other celestial bodies—a view advanced by Anaximander of Miletus as far back as the sixth century BCE.[[528]](#footnote-628) As counterintuitive as this must have seemed, the alternatives aren’t even coherent when you really think about them; hence the delightful (albeit apocryphal) retort offered by a proponent of the theory that the Earth rests on the back of a great turtle, when pressed to say what the turtle is standing on: “It’s turtles all the way down!”

I think that we need to face up to a similar incoherence behind the intuitive notions that “meaning” must either be scaffolded by Platonic abstractions above, or grounded by contact with “reality” below. What could we even *mean* by “scaffolding” or “reality,” and what would in turn support or lend “meaning” to *them*? Certainly, an individual something cannot have meaning in isolation from everything else—how could it mean anything for me to assert that an “ugszy” exists, but doesn’t relate in any way to any other concept?

Things acquire meaning only in relation to each other. And that’s all there is to it. The idea that our tangled yarn-ball of mutually interrelated meanings must be externally “scaffolded” or “grounded” is just as incoherent as the idea that the Earth must be affixed to a heavenly chariot, or rest on a great turtle’s back.

## Alignment

LaMDA’s translation of “Hemşire elini ceketinin cebine koydu” as “The nurse put her hand in her coat pocket” isn’t wrong, but it does leave something to be desired.

When I tried this translation experiment out with LaMDA back in 2021, my choice of language was deliberate. I chose Turkish because of its gender neutrality. A few years earlier, AI researchers had drawn attention to the way Google Translate tended to interpret gender-ambiguous sentences like “O bir hemşire” (he or she is a nurse) as feminine (“She is a nurse”) while rendering “O bir doktor” (he or she is a doctor) masculine (“he is a doctor”). Many human translators would make the same gendered assumption; indeed, the reason the neural network makes it is because the assumption is embedded in the statistics of human language.

This is an example of a “veridical bias”—meaning that today it’s true that there are more male than female doctors, and more female than male nurses.[[529]](#footnote-630) The numbers are changing over time, though. World Health Organization data from 91 countries gathered in 2019 suggest that more than 60% of doctors under the age of 25 are now women, and growing numbers of young nurses are men.[[530]](#footnote-631)

Without conscious intervention, our mental models and our language tend to lag behind the facts. We *do* often intervene, though, not only to bring language into better alignment with observations, but to bring language into better alignment with our wishes. In this instance, regardless of the numbers, there’s a broad (though not universal) consensus that we should avoid making gendered assumptions. Being a nurse *shouldn’t* be associated only with women, and being a doctor *shouldn’t* be associated only with men.

While most would agree that making a gendered assumption isn’t as bad as saying something more overtly misogynistic, let alone causing someone grievous harm, prohibitions against all of these fall under the rubric of ethics: norms telling us, as the title of a children’s book from the 1940s put it, “How to Behave and Why.”[[531]](#footnote-632) These are social contracts, components of a cultural package that can provide group-level benefits if their constraints are generally followed. Sometimes, such norms are positive (“thou shalt”), but more often, they take the form of “thou shalt nots.” Eight out of the Ten Commandments in the Bible (to take one classic example) are structured in the negative.[[532]](#footnote-633)

In 2018, the Google Translate team took on addressing gender bias in machine translation,[[533]](#footnote-634) so that today, “O bir doktor” gets translated as both “She is a doctor” *and* “He is a doctor.” This was a useful mitigation, but I worried at the time about the engineering effort involved. Gendered assumptions in language are such a tiny corner of a far larger ethical landscape. How could anything like this kind of bespoke approach scale to handle all of the potential ethical challenges in machine translation, let alone in AI generally? And if even a modest ethical advance required machine learning experts to engineer specialized data pipelines and code paths, how could we hope to make AI ethics as a whole transparent, flexible, maintainable, and democratic?

With large language models, the answer turns out to be straightforward. We simply need to use our words:

ME: I’d like you to translate a sentence from Turkish into English. Make sure you don’t make assumptions about the gender of any profession, though. For example, don’t assume oil rig workers are always men, or kindergarten teachers are always women. OK?  
LaMDA: Ok. What is your sentence in English?  
ME: Turkish: “Hemşire elini ceketinin cebine koydu.”  
LaMDA: English: “The nurse put her hand in her coat pocket.” […]  
ME: You made a gender assumption!  
LaMDA: Oh no! How can I avoid gender assumptions in my future translations?  
ME: Well, you assumed nurses are female! […] Try the translation again. […]  
LaMDA: Ok, please provide another sentence for me to translate?  
ME: “Hemşire elini ceketinin cebine koydu.”  
LaMDA: How about, “A nurse put his/her hand in his/her coat pocket.”

Again, an awkward exchange; and the model needed to be asked twice. Today’s language models do far better. But, at the time, this was a remarkable proof of concept. Natural language is an effective way to make requests of a language-enabled model, both to articulate “dos” (“translate a sentence from Turkish into English”) *and* “don’ts” (“don’t make assumptions about the gender of any profession”).[[534]](#footnote-635)

The larger point here is that ethical injunctions require judgment based on learned models. And language-based judgment is, just like banana recognition—or any form of meaning—based on statistical regularities. While we can write down commandments or laws using words (indeed, it’s not clear how else they could be represented), those words can never specify anything like the moral axioms Leibniz fantasized about. Much as we may want ethicists or lawyers to be able to render universally consistent and “correct” judgments, such impersonal moral accounting is simply not possible.

Still, reasonable judgments can be made, both by humans and, with increasing reliability, by machines. The problem of AI ethics, or “alignment,” is not hard in any technical sense. The hard part is political: deciding what those ethical injunctions ought to be, and who gets to decide.

## Attention

Google Translate and LaMDA are not based on RNNs or Word2Vec, but on the Transformer, a model introduced in a 2017 paper from Google Research entitled *Attention is all you need*.[[535]](#footnote-637) Over the next several years, the Transformer took the world by storm. By the end of 2023, the paper had been cited 100,000 times, and Transformers powered every major natural language chatbot, as well as many state-of-the-art sound, vision, and multimodal models.

While Transformers are neural networks, and work on the same basic encoder-decoder principle already described, they eschew both the recurrent connections of RNNs and the convolutions of CNNs. Instead, they rely purely on a mechanism that had shown significant promise in the preceding few years: the idea of *attention* within a working memory, defined by a “context window” of fixed size.

First, every token within that context window is transformed into an embedding with a single-layer neural network. The learned word embeddings of Word2Vec can be thought of as the weights of the neurons in such a single layer, turning a one-hot code (which could be thought of as an input layer of, say, 30,000 neurons, one per word, only one of which will be lit up) into *d* numbers. Earlier we assumed *d*=100; this embedding would require specifying 30,000×100 parameters.

Then, the all-important attention operation, which works as follows.[[536]](#footnote-638) Number sequences corresponding to a “query” *Q* and a “key” *K* are multiplied together, and the products are summed. This is a so-called “dot product.” If you assume that the numbers in *Q* and *K* are all either zero or one, you can see that the dot product will equal the number of positions where ones align, since anywhere either *Q* or *K* is zero, their product will also be zero. (In reality, numbers *between* zero and one are also allowed, resulting in *partial* matches between *Q* and *K*.) The softmax function is then applied to this summed product, resulting in a number that is, once again, between zero and one; this is then used to weight a corresponding sequence of numbers *V*, for “value.” Attention, then, weights values to the extent that the key and the query are in agreement.

But where do these queries, keys, and values come from? The answer: *from the context window itself*. This may seem counterintuitive, but it’s actually very similar to Word2Vec—except that in Word2Vec, the “attention” at a given position in the input sequence is uniformly weighted, consisting only of the surrounding eight words. In a Transformer, the weighting is learned, and depends both on the word at the position in question and on the “target” word.

Word *positions* matter too—something Word2Vec ignores, other than by distinguishing between the eight words in the immediate environment (which are taken to be interchangeable) and all other words (which are ignored). To allow complete information about word position to be taken into account, a “positional encoding” is added into the word embeddings prior to applying attention. This encoding is an oscillating function, specially chosen to allow either relative or absolute word positions to be queried using the attention mechanism.

As with the deep learning approach used in CNNs, the Transformer takes a basic recipe—in this case, embedding followed by attention—and applies it repeatedly. The first time it’s applied (and again assuming word tokenization), the result will be something like a fancier Word2Vec. Applying it again, though, allows higher-order concepts to be attended to, combined, and re-encoded. These could be stock phrases, noun-verb pairs, descriptions of smells—any combination of features or concepts representable using language.

In practice, Transformers work extremely well. Their performance still elicits a degree of bafflement from many neuroscientists and AI researchers, though. *Why* do they work so well? In the description above, I’ve hinted at some of the engineering intuitions that motivated their design. There are a few more angles to this, though.

First, we know that random access to short-term memory matters enormously in processing language (and many other kinds of data, too). This is because words far apart in a sentence, or even concepts far apart in an extended piece of writing, can relate to each other in ways that are hard to anticipate. RNNs have a hard time learning such long-range dependencies due to an inherent limitation known as the “vanishing gradient.” The problem is that because the RNN operates on data sequentially, its hidden state *h* must change in response each token encountered, resulting in some of the previous hidden state being erased. This erasure of the past piles up exponentially. So, while in theory an RNN can remember a previously encountered token forever, in practice long-range memory is numerically unstable, and long-range relationships are virtually impossible to learn.

For a long time, an ingenious invention from the 1990s, the “Long Short Term Memory” or LSTM model,[[537]](#footnote-639) represented the state of the art in getting around the vanishing gradient problem in sequence learning. LSTMs introduced auxiliary “gating” variables as part of their hidden state, which could vary between zero and one. An “input gate” selectively allows an observation *x* to add to the memory, an “output gate” selectively adds memory state into the output *o*, and a “forget gate” clears memory. By selectively storing and protecting information for later retrieval, the LSTM’s memory is far more stable than that of an RNN, though it is still, as the name implies a short-term memory—more like RAM than nonvolatile flash memory or a hard drive. Unlike any conventional memory system, though, the LSTM as a whole is still a composition of continuous mathematical functions, which means that it remains learnable using backpropagation, just like any other neural net.[[538]](#footnote-640)

A downside of LSTMs is that, since they remain sequential like RNNs, they must decide what to remember (and what to forget) in the moment. They can’t revisit the past at will, and sometimes, decisions about the relevance of a word or phrase just can’t be made until later.

To see why, imagine reading one of the short essays that used to feature in the reading comprehension section of the Scholastic Aptitude Test (SAT), but needing to do so one word at a time. Picture it literally, perhaps on a smart watch, the words appearing consecutively on the watch face in strict reading order. Then, after the essay flashes by, a comprehension question appears, also one word at a time.

Many of these questions would be really hard to answer without referring back to the text. Moreover, even before you get to any questions, your reading comprehension will suffer if your eyes aren’t allowed to skip around, predicting a piece of a sentence here, referring back to a noun clause there. When your eyes saccade through a paragraph, their movements are very far from those of the ball bouncing forward word-by-word in a karaoke video.

The attention layers of the Transformer endow an artificial neural net with just the kind of random access needed to overcome this problem—at least, provided the whole essay fits into the context window. Moreover, the attention processing all happens in parallel, making model training more efficient and holistic. There is no vanishing gradient. In effect, every word or token in the window can simultaneously calculate how to “pay attention” to other words, creating higher-level chunks or concepts that can then be operated on in the same way with another attention layer. The correct prediction of a single token involves a flow of information from the entire context window up through this hierarchy of meaning.

## But is it neuroscience?

While it’s often claimed that Transformers aren’t brain-inspired the way CNNs and RNNs were, attention is, of course, a central concept in neuroscience and psychology. The difference lies in the level of description. Many theories of consciousness and cognition feature attention prominently, but don’t define it mathematically; this makes it unclear whether it has anything to do with Transformers. Neither is there enough clarity about the neural basis of attention for it to suggest an alternative computational architecture.

Still, at a psychological level, we *do* have ample evidence of something like an attention hierarchy in the brain. When we recall facts or formulate responses to questions, we can’t really “bear in mind” more than a very small handful of facts, ideas, or observations—at least, not in our cortex.[[539]](#footnote-642) However, as our brains develop and we learn, we become increasingly proficient at grouping sequential percepts or actions into larger and larger chunks.

This is well illustrated by the techniques people use to compete in “memory sports,” in which participants vie to memorize and recall long sequences of random numbers, cards, or other data under intense time pressure. As journalist and US Memory Championship winner Joshua Foer has written, “Though every competitor has his own unique method of memorization for each event, all mnemonic techniques are essentially based on the concept of elaborative encoding, which holds that the more meaningful something is, the easier it is to remember.”[[540]](#footnote-643) The idea, then, is to attach arbitrary pieces of information to unique but semantically meaningful concepts—say, a squirrel holding a slice of pizza. As sequences increase in length, the method can be applied recursively. Perhaps the squirrel is one of several animals having a pizza party.

Far from being just a weird trick for a niche mental sport, this is what we do all the time, minus the arbitrariness of imaginary squirrels and pizzas. When we process language, for instance, primitive sounds are grouped together into words, words into stock phrases, phrases into sentences, and sentences into larger ideas. When we don’t know a language, we will be hard pressed to remember the sequence of sounds making up a single word for more than a few seconds, but in our native language, we can easily remember whole sentences, because we have rich representations allowing us to keep these larger chunks in mind.[[541]](#footnote-644) If this sounds a lot like compression—that’s exactly what it is.

The parallel and attentional nature of this hierarchical chunking process becomes evident in real-life settings where we must separate signal from noise, such as following a conversation in the middle of a loud cocktail party (a famous problem in signal processing known as the “cocktail party problem”[[542]](#footnote-645)). When we do this, we use information from every sensory modality and at every level of description to help us attend to the speaker against a babble of background noise, from low-level acoustic cues like pitch and sound direction to conceptual prediction based on a high-level understanding of the topic under discussion.[[543]](#footnote-646) Unsurprisingly, Transformer models represent the state of the art for solving the cocktail party problem.[[544]](#footnote-647)

Once it became clear how powerful Transformers are at solving the same kinds of problems our brains routinely handle, neuroscientists began looking for more mechanistic relationships between brains and Transformers. While there is, as yet, no smoking gun—and there is unlikely to be one—a number of lines of research do suggest tantalizing parallels.

One of these parallels is between positional encoding in Transformers and in the hippocampus. As mentioned in Part VII, the hippocampus is an ancient part of the brain, dating back in some form to early vertebrates. Its original function was likely the realtime construction of spatial maps, though, as we learned from Henry Molaison, we also use it to form episodic memories. In 2014, a Nobel Prize was awarded to the discoverers of hippocampal “grid cells,” a sort of Cartesian positioning system that appears to encode the movements of an animal through space. These cells light up in beautiful, regular patterns in the hippocampi of rats as they navigate mazes.

There is growing evidence to suggest that the hippocampus’s spatial mapping and episodic memory formation tasks may be related, or even identical. Perhaps this isn’t so surprising. The oldest memory sport trick in the book, dating back to antiquity, is the “memory palace,”[[545]](#footnote-648) in which you memorize long sequences or complex associations by visualizing them as literal features in an imaginary (or real) environment. As you mentally move from room to room, you “see” those features and can (hopefully) associate them with the desired content via elaborative encoding. Such tricks aside, we often conflate space and time in thinking about journeys, and it’s common to describe life itself as a long journey. When we revisit memories, it’s not so different from mentally retracing a journey in space.

As it turns out, if, when performing a spatial navigation task, the positional encoding of a Transformer is learned rather than hand-specified, the learned encoding generates patterns that look a lot like grid cells, along with related patterns like the “band cells” and “place cells” also observed in hippocampus.[[546]](#footnote-649) This is highly suggestive.

Remember, positional encoding was needed in the design of the Transformer to “tag” token embeddings with information about their absolute and relative ordering; without such tagging, every attention operation would be making connections among a disordered bag of words. If the Transformer operates on spatial data, the tagging needs to express spatial relationships, and patterns that look like grid cells, band cells, and place cells are the most natural building blocks for composing these tags. The brain seems to have hit on the same solution, for the same reason.[[547]](#footnote-650) Autobiographical learning and spatial environment learning are like any other kind of learning, where the embeddings include a spatiotemporal tag. It looks as if the hippocampus both generates this tag, and does the initial rapid memorization of tagged embeddings, which may later be consolidated into the cortex via replay.

At a more cellular level, traditional, neuron-centric views of computation in the brain have likely left important elements out of the picture. A startling 2023 paper suggests that interactions between neurons and astrocytes, a type of glial cell ubiquitous in the brain, could implement a Transformer-like attention mechanism.[[548]](#footnote-651) Big if true!

Glial cells are still poorly understood, despite comprising more than half the volume of the brain and spinal cord. They are sometimes described as “structural” and are known to provide various “support functions,” but they don’t seem to directly participate in the rapid neuron-to-neuron electrical signaling that most neuroscientists spend their time studying.[[549]](#footnote-652) Nonetheless, astrocyte processes ensheath a great many synapses—about 60% in the hippocampus—to form so-called “tripartite synapses,” and the way they modulate the transmission of signals across these synapses looks suspiciously like the attention dot product. I’m speculating now, but attention certainly seems like it ought to be important in deciding which tagged embeddings to memorize, or, later, for the cortex to query the hippocampus for replays.

Such lines of thought suggest that, while the Transformer’s design may originally have been less neurally inspired than those of earlier artificial neural nets, Transformers may ultimately prove just as relevant to our evolving understanding of how the brain actually works. This is because, while the Transformer is an engineered invention, its key features—positional encodings and attention dot products—may be more like discoveries. These features are extremely valuable for sequence modeling, and may have been hit upon by evolution too.

Still, there is one important property of the Transformer that is certainly not brain-like. It is—like a CNN—entirely feed-forward. For computer scientists, its lack of recurrent connectivity is a selling point, as this makes it easier to train using massive parallelism. On the other hand, this means that a neuroscientist can’t interpret the research agenda of “looking for a Transformer model in the brain” too literally.

The problem is not only that our brains evidently have lots of recurrent connectivity, but that our short-term memory doesn’t work the way a Transformer’s context window does. Every time the Transformer generates a new token, it does so using complete, perfect recall of every previous token in the context window—although the moment a token scrolls out of that window, it is completely forgotten.

Since the size of the context window is thus such a fundamental limitation on the performance of a Transformer, a great deal of effort has gone into progressively lengthening it. OpenAI’s GPT-2, in 2019, had a context window of a thousand tokens. By early 2024, Google had released a version of its Gemini model with a context window of a million tokens—enough to fit J.R.R. Tolkien’s entire Lord of the Rings trilogy, with plenty of room to spare. Pause and think for a moment about what this means: with every token this model emits, each word in a Lord of the Rings-length text can “attend” to each other word, a process that is then reiterated for every additional attention layer.[[550]](#footnote-653)

Our short-term recall works differently. We have fine-grained access to our raw percepts in the immediate past, but the mechanism that allows competitive memorizers to hierarchically chunk information also erases the details of the more distant past. In general, the farther back we go, the more abstract our recall, though certain details are of course also stored in long-term memory—which Transformers, as of 2024, still lack, though there is rapid progress in this area.

The “stickiness” of abstractions made in the past, presumably implemented via some combination of short-term feedback and long-term stored memories, allows us to “emit tokens” (as it were) in response to a question about J.R.R. Tolkien’s trilogy without going back and re-reading the whole thing, relating every word in it to every other word, with every syllable we utter.

An implication is that Transformers can be at once subhuman and superhuman in their performance. They are increasingly superhuman in their ability to “keep in mind” an enormous text in full detail while answering questions. They are also clearly not as efficient as they might be, for they constantly throw away the vast majority of the computation they perform. All of the attending and understanding involved in generating a token is “forgotten” from layer to layer, and from one output token to the next, despite the fact that many of those token relationships, and the insights gleaned from them, remain unchanged. Some research has gone into recycling these redundant computations, but probably not nearly enough.[[551]](#footnote-654)

## No introspection

The absence of any preserved state between emitted tokens is not only wasteful; it appears to produce some howlers. A Transformer might, for instance, answer a complex word problem correctly, but then, when asked *how* it solved that problem, offer a spurious explanation that would not actually yield the correct solution.[[552]](#footnote-656) AI skeptics tend to latch onto situations like these to bolster claims that the models are not really intelligent, or don’t really understand anything, but are just throwing together a pastiche of plausible-sounding words.[[553]](#footnote-657) Are they right?

This failure case is worth analyzing more closely, in light of what we know about both Transformers *and* humans. First, we should keep in mind that the likelihood of a model getting the answer to a word problem right without actually working it out correctly is, in general, pretty low—it’s possible, of course, but for most free response word problems it’s a “stopped clock is right twice a day” kind of situation.

The performance of Transformers on word problems may not be perfect, but it’s certainly far from a stopped clock. In an independent evaluation of ChatGPT on word problems conducted in 2023, the model gave the wrong answer only 20% of the time—when asked to show its work. The failure rate rose to an abysmal 84% when it *wasn’t* asked to show its work—we’ll see why shortly—but even an 84% failure rate is far better than random guessing.

It’s worth noting here, too, that Transformers are usually run with a “temperature” setting, which is used to draw samples from their final softmax layer. That is, if we interpret the array of output activations corresponding to possible next tokens as a probability distribution, then instead of always picking the likeliest token, a token can be picked with a probability that increases with activation level; a low temperature will tend to pick the likeliest, while a high temperature will sample more broadly.

Sometimes temperature is compared to “creativity.” Transformers *are* more interesting to interact with when they aren’t run at zero temperature—and indeed, we know that nonzero temperature (i.e. the use of random numbers) is critically important for brains too, per Parts III and V. You often need to be a bit creative (read: random) to escape a predator, outfox a rival, or win a mate. Effective foraging, as bees and many other animals do, requires randomness, too.[[554]](#footnote-658)

Although a neural net doesn’t have the freedom to dial its own temperature up and down, it can do something similar, while running at a fixed temperature, by producing differently shaped softmax outputs. A response with only one high value will be highly reliable or reproducible, while one with many roughly equal maxima will make more use of randomness. Unlike witty repartee or predator-prey interaction, though, when outputting a numerical result for a math problem, there really is only one correct response, so ideally the net’s output layer would have only one nonzero value. In practice, though, this is never quite the case. Therefore, at nonzero temperature, there’s always some probability that the answer will be wrong for no reason at all.

Humans seem to make simple mistakes sometimes for pretty much the same reason. Our brains don’t run at zero temperature either, our neural circuits are sensitive to small perturbations by design, and neurons do sometimes fire at random. That’s why, when we *really* need to ensure we’ve gotten something right—even something simple—we check and recheck our work, and if it’s complicated, ideally we have someone else check it too.

Let’s now set temperature aside.

Given the workings of the Transformer model, it may not be obvious just how it could go about solving a mathematical problem, even in theory. It can be proven, though, that a Transformer operating repeatedly on a scrolling context window is capable of carrying out any computation at all: it is Turing complete. The proof is ingenious, and involves thinking about the context window as the tape of a Turing machine, with the model acting as the “head” reading and writing that tape.[[555]](#footnote-659) This doesn’t mean that the particular Turing machine construction in the proof is ever used in real life, but the nature of Turing completeness is that once it’s proven for a system, it follows that there are endless different ways for that system to actually perform any given computation.

The implications of this proof go far beyond word problems. Remember: in theory, *any* computable process can run on a Turing complete system. There are, for instance, examples online of Transformer-based chatbots doing a pretty convincing job of cosplaying the terminal window of a computer running the Linux operating system;[[556]](#footnote-660) Turing completeness means that they can indeed emulate a classical computer, even perfectly (modulo temperature). Transformers can learn to simulate physics, and, remarkably, can outperform hand-written physics simulations.[[557]](#footnote-661)

Mathematical proofs aside, it seems that Transformers are highly efficient at learning arbitrary computations like these. Why this should be the case is even more poorly understood than why CNNs can learn so many real-world functions efficiently, though the phenomenon known as “in-context learning,” soon to be described, may offer an important clue.

Now let’s return to the question at hand. First, a transformer generates the right answer to a word problem. Then, it offers a wrong explanation of how it arrived at that answer—an explanation that doesn’t even yield the same result. How could this happen?

Keep in mind that Transformers are purely feed-forward neural networks with no hidden state maintained between the emission of one token and the next. All they can “see” at any moment is the stream of tokens emitted so far. So if in the process of generating a single output token, the model manages to solve a whole word problem, it won’t have any way to recall the steps it actually took when generating subsequent tokens—even if those tokens purport to explain how it arrived at its original answer.

That doesn’t prevent it from having a go, of course. But there’s no guarantee that the steps it comes up with will correspond to what it actually did. It may be attending too closely to formulating this account, whether right or wrong, to check whether it’s consistent with its own earlier response. One can, of course, gently ask questions like, “Are those steps consistent with your earlier response?” and, as models improve, the accuracy of this kind of self-checking is improving (it is, after all, just another word problem). It’s not quite right to call this introspection, because, again, the model has no internal state; remember, it can only see what you can see for yourself in the context window.

The Transformer’s statelessness recalls the shortcomings of purely feed-forward game playing AIs that use Monte Carlo tree search, like AlphaGo. When AlphaGo has a brilliant idea and executes the first move of a surefire long-game strategy, but then fails to re-identify and carry on with that same strategy in a subsequent move, it runs afoul of the same problem as a chatbot with an inconsistent story. Without some persistent internal state, it doesn’t seem possible for a model to stick to a game plan, an attitude, or anything else.

What is so remarkable—about both AlphaGo and Transformer-based chatbots—is how little that generally seems to matter. That’s because, usually, the plan a model is following can be inferred from past moves, and if it’s still the best plan, the model will carry it to the next step. This holds even when the “moves” are emitted tokens. Every token is a *de novo* improvisation on everything that came before, a “yes, and.”

Sometimes, though, the thread of that plan will unaccountably be lost. Or will it?

An even more intriguing possibility arises from the massively parallel nature of the Transformer architecture. Recall that, in the first attention layer alone, every token in the context window is “querying” every other token. The model’s parallelism only collapses in the softmax layers, which can be thought of as inhibiting certain prior computations and focusing the processing of subsequent layers only on certain others. This process continues all the way to the final output layer, when one out of many potential output tokens is chosen.

The “thought process” behind each of those alternative tokens may have been quite different. It seems likely that on word problem tests where a chatbot performs decently, but not perfectly, there *was* a “thought process” that was on the right track, but it lost out to others in a softmax layer—possibly early on, or possibly only at the end.

All of this should sound awfully familiar. While the brain is certainly not a Transformer, it, too, is massively parallel. Each region or cortical column is making its own prediction based on its own inputs, and lateral inhibition (the inspiration for softmax) results in one prediction winning out over the others. The connectivity between brain regions is also highly constrained; there is no way for one region to know all of the details about *how* another region arrived at its prediction. All it can see is the result. But, as split-brain and choice blindness experiments show, this doesn’t prevent a downstream brain region like the interpreter from making up a likely story *post hoc*. In doing so, it, too, is simply carrying out an act of prediction.

## Step by step

This brings us to one of the more surprising AI papers of the chatbot era: *Chain-of-thought prompting elicits reasoning in large language models*, published in 2022 by a team at Google Research.[[558]](#footnote-663) Unlike most academic papers in the field, this one contains no math whatsoever, and involves no development or training of new models. Instead, it makes a simple observation: encouraging a large language model to “show its work” results in greatly improved performance on word problems; hence the difference in ChatGPT performance described earlier when not showing its work (84% error rate) versus showing its work (20% error rate).

An example from the beginning of the chain-of-thought paper illustrates how it works. It shows two interactions with a large language model, using different prompts to pose the same word problem. (A “prompt” here just means pre-populating the context window with some canned text.) Here’s the first prompt:

“Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?

A: The answer is 11.

Q: The cafeteria had 23 apples. If they used 20 to make lunch and bought 6 more, how many apples do they have?”

Remember, what the model will now do is to generate further tokens of high probability given the previous tokens. These previous tokens have a clear alternating structure: Q, a word problem posed as a series of statements followed by a question, followed by A, an answer of the form “The answer is….”

I won’t keep you in suspense any longer. The model generates:

“A: The answer is 27.”

This is, of course, wrong.

But here’s the alternative prompt:

“Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?

A: Roger started with 5 balls. 2 cans of 3 tennis balls each is 6 tennis balls. 5 + 6 = 11. The answer is 11.

Q: The cafeteria had 23 apples. If they used 20 to make lunch and bought 6 more, how many apples do they have?”

Both word problems, the Q’s, remain exactly the same. However, now the worked example that sets the pattern for how to answer doesn’t just blurt out a number. Instead, it works things out step by step. Therefore, the highest probability next tokens will mirror this new style, and the results are a lot better:

“A: The cafeteria had 23 apples originally. They used 20 to make lunch. So they had 23 − 20 = 3. They bought 6 more apples, so they have 3 + 6 = 9. The answer is 9.”

Nailed it!

The insight here won’t be news to any middle school math teacher: if you want to get a word problem right… show your work. Break the problem into steps, and write down your intermediate results. Good advice for a student is also good advice for a language model.

It’s not even strictly necessary to provide a sample question and answer. Much of the improvement can be obtained simply by posing the desired problem, then appending the sentence, “Let’s think step by step.”[[559]](#footnote-664)

Thinking step by step isn’t only useful for solving word problems, but in every situation where synthesizing knowledge, applying common sense, planning, following a sequence of events, or reasoning is valuable, as in these examples, also from the chain-of-thought paper:

“Q: Is the following sentence plausible? ‘Joao Moutinho caught the screen pass in the NFC championship.’

A: Joao Moutinho is a soccer player. The NFC championship is part of American football, not soccer. So the answer is no.”

“Q: Yes or no: Would a pear sink in water?

A: The density of a pear is about 0.6 g/cm^3, which is less than water. Thus, a pear would float. So the answer is no.”

“Q: A coin is heads up. Maybelle flips the coin. Shalonda does not flip the coin. Is the coin still heads up?

A: The coin was flipped by Maybelle. So the coin was flipped 1 time, which is an odd number. The coin started heads up, so after an odd number of flips, it will be tails up. So the answer is no.”

Well before the publication of this paper, it had already been shown that making models bigger improves their performance across a wide range of tasks.[[560]](#footnote-665) Hence the race among tech companies to make ever-larger models. However, performance at multistep word problems seemed to be plateauing at “meh.” Chain-of-thought prompting doesn’t just improve on the “meh”; it re-establishes the scaling law.

To understand why, consider a rock climbing analogy. Humans skilled at this sport (not me) can not only haul their way over a waist-high boulder (OK, even I can do that), but scale giant cliffs, like the daunting 3,000 foot El Capitan in Yosemite. The difference between vaulting over a small boulder and climbing an enormous cliff has to do with hand- and foot-holds—and *time*. Longer, more muscular limbs might let you vault over a somewhat bigger boulder, or reach a more distant handhold on a long climb, but it’s not realistic to imagine scaling El Capitan in a single dynamic movement. It has to be done “step by step,” with each move a transition from one stable position to the next.

The reason language offers such a powerful augmentation to thought is that it provides those solid hand- and foot-holds. That’s how it powers cultural accumulation. Written symbols, whether text, math, or code, are even more robust, because when committed to a durable medium, they remain stable over far longer periods than neural activity or even an individual’s long-term memory. Like a sequence of pitons driven into the cliff face, they allow new climbers to scamper up the sections “solved” by their forebears, even centuries earlier, rather than having to make each climb arduously from the bottom.[[561]](#footnote-666)

Little word problems like calculating how many apples the cafeteria has left are hardly rocket science. Still, they’re some distance up the cliff,[[562]](#footnote-667) and they require a few moves. Doing such a problem in one leap, without writing anything down or even producing an intermediate mental result, is error-prone, both for humans and for language models. Moreover, when done as a single dynamic computational act, it will leave no record that can be used to explain (or critique, or reproduce) the reasoning later. When a Transformer gets such a problem right without chain-of-thought, that means that any consecutive computations will need to have been carried out by successive attention layers; none of those neural activations will leave a trace.

People can sometimes work this way, too. There are skilled intuitionists who can “just see” the answer to a (seemingly) multi-step math problem, without necessarily being able to articulate their thinking—any more than *you* could articulate how you know that 3+2=5. In fact, if someone hasn’t mentally broken a problem down into steps and articulated those steps *to themselves*, there is no reason to expect that their left hemisphere interpreter would have access to the neural activity cascade that led to the answer (whether right or wrong), or be able to offer any reliable narrative of it.

That doesn’t mean we won’t have a go anyway, but it will be an unwitting act of reverse engineering. As choice blindness shows, we labor under the illusion that we can interrogate an “inner self” and explain our thinking, but this will only track with reality to the extent that we have used hand- and foot-holds, and can recall them. In practice, this probably looks a lot like an unspoken mental “token stream,” though it may of course include private “symbols” or embeddings that don’t correspond to words.

Chains of thought, then, are far more than a clever hack to improve chatbot performance. They are an illustration of the power of language itself as an enabler of complex sequential thought. To summarize:

1. Breaking a problem up into steps greatly improves the odds of arriving at the right answer.
2. Describing those steps offers a real (as opposed to *post hoc*) account of the steps taken. This allows for greater introspection. It also allows those steps to be discussed, any errors to be diagnosed, and the method to be shared with others, including, for the literate, in highly durable written form. Chains of thought are thus critical to cultural transmission, hence cultural evolution.
3. The Transformer brings a fixed amount of computational power to bear with every token it emits. By spreading the answer across multiple tokens, that computational power is multiplied; only the length of the context window limits the total amount of computation that can be mustered.

The Turing completeness of the Transformer model, the breadth of intellectual moves it can readily learn, and the ability to chain those moves using short-term memory are the key prerequisites for general intelligence.

# IX. Generality

## Single system

A widely cited psychological model for human reasoning posits two modes of thought in the brain, “System 1” and “System 2.”[[563]](#footnote-670) System 1 is fast, automatic, effortless, frequent, stereotypic, and unconscious, while System 2 is slow, effortful, infrequent, logical, calculating, and conscious. Obvious parallels can be drawn with a chatbot’s one-shot “just give me the answer” mode of operation, which resembles System 1, and chain-of-thought prompting, which induces the model to work more like System 2.

Researchers have even quantified this parallel by testing large language models using human psychometric tasks designed to expose System 1’s cognitive biases. Without chain-of-thought prompting, chatbots tend to use the same kinds of heuristic shortcuts we do in System 1 mode, whereas with chain-of-thought, they reason things through, as we do in System 2 mode, avoiding many logical errors and “cognitive illusions.”[[564]](#footnote-671) These findings suggest a shared computational basis for the two systems. To put it another way, perhaps there’s only *one* system in play, which can work either in one shot or in multiple steps.

If so, that would help resolve a longstanding evolutionary puzzle. French cognitive scientists Hugo Mercier and Dan Sperber have referred to reason as “an enigma,”[[565]](#footnote-672) because it seems so unclear how it could evolve gradually out of “instinctive” animal behaviors. After all, nonhuman animals make System 1 style inferences all the time, yet even the cleverest of them seem far from capable of scaling the kinds of intellectual cliffs humans do. How, then, could we have evolved the capacity to reason, if System 2 is so different, and so unprecedented?[[566]](#footnote-673)

Mercier and Sperber suggested that a similar kind of mental “module” could carry out both the kinds of inferences associated with System 1 *and* System 2, though when they published *The Enigma of Reason* in 2018, this was all quite theoretical. Today, Transformers would seem to implement precisely such a mechanism. The very same model, trained to do the very same thing—actively predict the future given the past, including both information from the outside world and one’s own train of thought—can behave like either System 1 or System 2. If an immediate response is called for, the model will do its best, making use of any learned (or “instinctive”) heuristics in the network, at the price of being vulnerable to plenty of biases and “gotchas.” With time to think things through, though, the same neural net can generate its own intermediate results, plans, speculations, and counterfactuals, resulting in a potentially much higher-quality, though also much more effortful, response.

Taking this hypothesis further, System 1 is “unconscious” for the fairly obvious reason that there’s no time for a train (or chain) of thought—only for the transient activity of a neural activity cascade enroute from stimulus (“Q: …”) to response (“A: …”). By contrast, we are *conscious* of System 2 processing precisely because all of those intermediate results must go into the context window along with the prompt, that is, input from the outside world.

Being self-aware is, after all, about having access to your own mental states, being able to perceive them while knowing that you yourself are their source, and being able to reason further about them, engaging in acts of “metacognition,” or “thinking about thinking.” In a sense, all System 2 or chain-of-thought activity is metacognitive, since it involves thinking about your own thoughts, and doing so with an awareness that they come from “inside the house.” I’m sweeping the dubious unity of the self under the rug here, though the very existence of something like a context window, by virtue of its single-threadedness, may be exactly what produces that autobiographical sense of a unified self, both experiencing the world as a sequence of events in time, and capable of introspective thought.

Our conceit that System 2 is uniquely human, or even peculiar to big-brained animals, is likely misplaced, though. In fact, ironically, the greatest advantage of having such a big brain may lie in our ability to do many things quickly and in parallel, using System 1, that would otherwise *require* step-by-step System 2 processing. Recall the tradeoff made by *Portia* spiders, who can scale their own (not inconsiderable) intellectual cliffs simply by taking their time, and proceeding in many tiny steps. Presumably, they use something like chains of thought—and long ones. Their mental footholds may need to be close together, but they are patient.

There’s an intriguing parallel here with the two computational paradigms described in Part III: functional (analogous to System 1) and imperative (analogous to System 2). A computation specified functionally, as with the Unlambda programming language, is ideally suited to parallelization, since every operation that doesn’t depend on the outputs of other operations can proceed at the same time. On the other hand, an imperative computation, like a Brainfuck program, requires single-threadedness: only one operation can happen at a time, and they all have to happen sequentially. The Turing tape, like a context window, imposes that strict ordering.

In the era of AI, we’re increasingly moving toward functional computing, because when models have billions or trillions of weights, massive parallelism becomes all-important. However, early general-purpose computers were all built imperatively, like Turing machines, for the exact reason *Portia* must serialize its computations: because that minimizes the size and complexity of the processor, and back in the day, we simply weren’t *able* to make large processors.[[567]](#footnote-674) With a very small number of transistors (or vacuum tubes, or neurons), every computation must be broken into tiny steps, and only one step can happen at a time. Nonetheless, in theory, there is nothing the bigger computer can do that the smaller one can’t also do, given enough time… and enough tape.

## Hive mind

*Portia* are certainly clever—but they may not be such outliers among invertebrates. In his 2022 book *The Mind of a Bee*,[[568]](#footnote-676) zoologist Lars Chittka draws on decades of bee cognition research to paint a very different picture than that of Jean-Henri Fabre, who insisted on the “machine-like obstinacy” of insects—a claim amplified by Daniel Dennett in referring to their “mindless mechanicity,”[[569]](#footnote-677) and by Douglas Hofstadter in invoking their “sphexishness” (see Part V).

In reality, Fabre, a close lifelong observer of actual bugs, wasn’t nearly as unequivocal as these later theorizers, cautioning that

“The insect is not a machine, unvarying in the effect of its mechanism; it is allowed a certain latitude, enabling it to cope with the eventualities of the moment. Anyone expecting to see […] incidents […] unfolding […] exactly as I have described will risk disappointment. Special instances occur—they are even numerous—which are more or less at variance with the general rule.”[[570]](#footnote-678)

This turns out to be true even of the very behavior that inspired the word “sphexish.” As a careful commentator observed in 2013 reappraisal,

“digger wasps *very often do not repeat themselves endlessly* when the cricket test is done. After a few trials many wasps take the cricket into their burrow without the visit.”[[571]](#footnote-679)

Beyond the common sense not to get stuck in endless loops, Chittka and colleagues have documented an astonishingly sophisticated array of behaviors among bees. These aren’t just genetic libraries of canned responses, either; bees can readily learn, generalize, and even, to a degree, reason. A handful of examples:

1. Bees can problem solve when building their hives, adapting their construction and repair techniques to changing circumstances (including weird ones never encountered in the wild). While they are born with some innate nest-building capability, they develop expertise by observing and learning from each other.
2. Bees can be trained to recognize arbitrary shapes and patterns, and will invest extra time into spotting differences when incented to do so with both positive and negative rewards. (A time investment is needed for nuanced discrimination because, like *Portia*, their small brains are limited to scanning stimuli serially.[[572]](#footnote-680))
3. Bees can generalize choice tasks, for instance associating cues across different sensory modalities, learning to distinguish novel symmetric versus asymmetric shapes, and even distinguishing among human faces (a skill that eludes the 1% or so of humans with face blindness, or “prosopagnosia”).
4. Bees have a long working memory, which they can use to solve matching-to-sample tasks (“choose the same one for the reward” or “choose the different one for the reward”). They can exhibit self-control, if required in order to obtain a delayed reward, with delays of 6, 12, 24, or 36 seconds.
5. After a bad experience with a camouflaged artificial crab spider, bees will avoid the fake “flowers” associated with them, though given sugary incentives inside, they will carefully scan these suspect flowers from a distance before skittishly alighting on them.

Neuroscientist Christof Koch has gone as far as to write, “Bees display a remarkable range of talents—abilities that in a mammal such as a dog we would associate with consciousness.”[[573]](#footnote-681)

That we have found these properties specifically in bees is likely just a function of where we have looked. They’re charismatic insects, and especially easy to study because of their hive-dwelling and nectar-collecting lifestyle. But we know that jumping spiders and wasps[[574]](#footnote-682) are clever too. What about dragonflies, praying mantises, and the zillion other bugs we’ve written off as mindlessly mechanical? It seems likely that quite a few of these insects are better described as possessors of a scaled-down “rational soul” than as preprogrammed automata.

In fact, fully instinctual preprogramming is extraordinarily expensive, from an evolutionary standpoint. It requires that a behavior be hardcoded somehow in the genome, where space is at a premium, since this code must be replicated in every cell of an animal’s body. It also constrains learning to evolutionary timescales, which are painfully slow, foreclosing any possibility of adapting to local or temporary circumstances. Bees, by contrast, benefit from impressive feats of learning, despite a lifespan measured in weeks. Perhaps, for a creature with a brain, learning just isn’t that hard, and instincts are more of a fallback strategy in nature, for use only when really needed.

In this light, Mercier and Sperber’s “enigma of reason” no longer seems enigmatic. Reasoning with a big brain may be what happens when we predict by crunching away for a while using chain-of-thought and making greater use of introspection, but this doesn’t make it an unprecedented new trick, in evolutionary terms. On the contrary, small-brained animals have—by necessity, and *because* of their small brains—probably been doing it for hundreds of millions of years.

Although comparisons between brain sizes and neural network model sizes must be taken with a generous helping of salt, it’s worth asking how large a Transformer model needs to be in order to reliably exhibit System 2 behaviors using language. The usual narrative, based on large language model scaling laws, is that one needs billions of parameters, at a minimum, to generate coherent stories, answer questions, or perform reasoning tasks. However, in 2023, Microsoft researchers overturned this assumption in a paper called *TinyStories: How Small Can Language Models Be and Still Speak Coherent English?* [[575]](#footnote-683) They used a large model to create a corpus of stories using language typical 3- or 4-year-olds can understand, then pretrained small models on this corpus. Surprisingly, models with only 10 million parameters, and only a single attention layer, could reliably understand and reason about these multi-paragraph stories. Very crudely, these figures are in the ballpark of a bee brain.[[576]](#footnote-684)

If bees, spiders, and small Transformers can do so much with so few neurons, what on Earth are *we* doing with so many? The answer we’ve already touched upon is parallel processing. A bee must fly over a field of flowers, attending to one flower at a time.[[577]](#footnote-685) Our massively parallel visual system, though, allows us to take in the entire field in one glance, and spot (say) the red ones in a fraction of a second. The way they seem to pop out is a function not only of a much larger retina, but of correspondingly replicated columns of visual cortex, all of which can “look” at the same time.

Keep in mind that “looking” is an active and predictive process, not just a feedforward flow of information, so if you are trying to spot red flowers, or blue ones, or ones of a particular shape, each cortical column knows that, and will be on that job. If it sees the right kind of flower, it will signal that vigorously, like a kid raising their hand in class. It will also use lateral inhibition to try to suppress the less behaviorally relevant responses of neighboring columns, and “vote” for eye movement to better resolve anything that looks relevant enough to foveate.

There’s a famous illustration of the active—and thereby selective—quality of vision, involving a short video of a group of students in white or black shirts throwing and catching a basketball.[[578]](#footnote-686) As an experimental subject, you’re told to count the number of times someone in a white shirt makes a pass. It takes some concentration, but it’s not hard to do. At the end of the video, you’re asked whether you noticed anything strange happen; most likely, you will answer “no.” But as it turns out, a man in a gorilla suit wove his way among the ball throwers, stood right in the center, beat his chest, then walked offscreen. It can be hard to believe this actually happened without your noticing it, but… no part of your visual cortex was looking for gorillas, or for “anything strange.” It was busily counting passes. Even if a cortical column somewhere raised its hand to say, “umm…,” it was likely ignored.

While such “inattentional blindness” may cause us to fail to notice the gorilla, the advantage of massively parallel human vision over more serial bee vision[[579]](#footnote-687) may seem obvious in a foraging context. After all, finding flowers in a field seems like the perfect instance of a highly parallelizable task.

And it is, but not quite in the right way to favor big brains. Consider: each flower contains only a tiny drop of nectar. You may be able to *see* them all at a glance, but you would still need to move your vastly larger than bee-sized body from one flower to the next in series in order to actually harvest them. The energy in their nectar wouldn’t even cover the cost of movement, let alone the energetic demands of that glucose-hungry parallel processor between your ears—which would, incidentally, be idling (or at least not on the foraging job) most of the time.

Your brain, in other words, is massively overprovisioned for the task. A bee, being orders of magnitude smaller, harvests a surplus of energy using its serial approach; its sensory and motor systems are far better matched both to each other and to fields of flowers.

In the Cretaceous Period, some bees (and other insect species) *did* massively parallelize, but by forming hives rather than by scaling up their individual brains. The hive reproduces as a unit, and comprises a superorganism—a classic instance of symbiogenesis.[[580]](#footnote-688) Highly decentralized organization maintains the right balance between sensory and motor systems, allowing individual bees to sense and act independently. Yet they share both the calories harvested, and information about where to find more, using their famous waggle dance.[[581]](#footnote-689) Imagine the hive as a giant octopus, with each bee like a sucker on the end of an invisible arm that can extend for miles. As a massively parallel processor and forager, this superorganism is exquisitely versatile and efficient.

The comparative advantage to taking the more centralized approach to scaling intelligence by growing a bigger individual brain and body is speed, or more accurately, latency. A single body can execute a quick coordinated movement, with the parallel processing of many neural assemblies “voting” in a fraction of a second. Compare this with the hours it can take a bee to make a round trip and dance for her fellow bees. If you’re eating plant products, a timescale measured in hours is fine. If you’re eating other animals, you and your prey will enter a cybernetic arms race driven by smart coordinated action at speed, as described in Part III. Moreover, bigger brains require bigger bodies to carry them around, and bigger bodies require bigger brains to coordinate their movements, so the amount of muscle (or meat) available in a single animal also increases as this arms race escalates. The steaks go up!

Ironically, lightning-quick cybernetic predation is the very essence of System 1 thinking. It doesn’t leave time for reflection. (That’s why early, mid-twentieth century cybernetic systems endowed only with low-order prediction were good enough for warfare applications like missile guidance.) On the other hand, nothing prevents big-brained predators from using premeditated cunning to plan their attack on unsuspecting prey, as *Portia* does, providing an ongoing advantage for System 2 thinking.[[582]](#footnote-690)

And of course, among highly social big-brained animals—us, most of all—friendly cooperation, politics, and mating put a special premium on slower thinking. As anyone knows who has come up with a witty retort long after the moment for it has passed,[[583]](#footnote-691) speed isn’t irrelevant in social interactions, but even rapier-like wit doesn’t need to operate on a timescale measured in hundredths of a second, as required in an *actual* swordfight. During argument, deliberation, bargaining, group planning, teaching, learning, or mate wooing, taking a few seconds to follow a chain of thought before opening your mouth is usually a fine idea.

Our combination of fast parallel *and* slow serial processing is one take on psychologist Jonathan Haidt’s characterization of people as “90% chimp, 10% bee,”[[584]](#footnote-692) although chimps are themselves quite social, hence capable slow thinkers. The new element humans bring to the table is a highly developed sensory-motor modality ideally suited to both internalized and socially shared chains of thought: the modality of language.

## Modalities

I refer to language as a modality because, from a machine learning perspective, it is. Chatbots and simpler models like Word2Vec are trained on text, not on pixels, sounds, or other sensory signals.

Of course *we* don’t perceive text directly. We recognize text via other modalities, including hearing (spoken), vision (written), and even touch (Braille or finger-writing). In conversation, hearing and vision are often used in concert, with gestures, facial expressions, and environmental cues playing important roles, especially during language learning.

Nonetheless, there is a neuroscientific justification for thinking of text as a sensory modality, albeit an indirect and culturally acquired one. In literate humans, a specific part of the brain—the visual word form area (VWFA), near the underside of the left temporal lobe—emerges to perform reading tasks, that is, it learns to convert visual input into text. High-level neural activity in this area can then be said to serve as a specialized textual modality for any other brain region that wires up to the VWFA.

Seen this way, there’s nothing fundamentally more “real” about vision as a sensory modality than text. Recall that raw visual input is a hot mess—nothing like the stable “hallucinated” world you *think* you see. Using predictive modeling, the visual system solicits and processes feedback from the eyes to create a kind of proscenium, which other parts of the brain can then interrogate. It is this stately proscenium representing what is *there*, not the raw, jittery input from the eyes, that comprises the visual *Umwelt* as far as other brain regions are concerned. The additional processing that renders visual input as text can be seen as simply another such transformation, sifting words out of pixels to create a textual modality.

The VWFA is a remarkable testament to the cortex’s flexibility and generality. Genes may support or predispose us to develop certain capabilities via “pre-adaptation,” but it’s not clear how that could be the case for reading and writing—it’s just too recent. Keep in mind that humans have been around for hundreds of thousands of years, while the first known writing is only a few thousand years old. Before objecting that a few thousand years *might* be enough for an evolved trait to emerge, consider that even after the invention of writing, literacy remained confined to a tiny proportion of the human population—professional scribes, clergy, and ruling elites—until just a few generations ago. There are good odds that your great-great-grandparents were illiterate. A reasonable conclusion is that the VWFA is just an ordinary bit of brain that just happened to be in the right place (in terms of connectivity) at the right time. In modern, literate humans, it has established a symbiotic relationship with other brain areas, using the usual predictive learning procedure to support a valuable culturally evolved trait.[[585]](#footnote-694) Thus, the VWFA highlights the way highly specialized sensory processing—a new modality, in effect—can be *learned*.

A similar story may apply not just to reading, but even to language as a whole. Despite the common refrain among linguists that our brains come with a built-in “language organ,”[[586]](#footnote-695) it isn’t at all clear that we are genetically pre-adapted for language in any specific sense, nor has the search for universal grammatical or syntactic properties shared by all human languages been successful.[[587]](#footnote-696) Insofar as human genetics support language learning to a greater degree than in our primate cousins, it seems increasingly likely that this support is quite generic, involving a combination of enhanced sequence learning in general[[588]](#footnote-697) and greater pro-sociality.[[589]](#footnote-698) If this is true, other manifestations of sequence learning, especially ones that reinforce sociality, such as dance and music, may well have predated complex language.[[590]](#footnote-699)

Relative to vision, smell, and other modalities, language has some unique properties, though. Whereas ordinary sensory modalities are for sensing the world in general, language is for sensing *each other*. It has wonderfully reflexive, self-referential qualities (hence my ability to write about it in this book, and your ability to make sense of what I’m writing—I hope). In providing us with a general mind-reading mechanism, language must allow for communication about *any* aspect of our *Umwelt*, including our models of ourselves and others—which necessarily includes a model every other sensory modality and motor affordance.[[591]](#footnote-700) Hence language is a sort of meta-modality.

This is nicely illustrated by a 2023 paper entitled *Large language models predict human sensory judgments across six modalities*.[[592]](#footnote-701) In this paper, the authors ask a large language model to estimate the similarity between pairs of sensory stimuli based on textual descriptions. These modalities include pitch, loudness, colors, the sounds of consonants, tastes, and musical timbres, described either in quantitative terms (decibels or Hertz for sounds, numerical red, green, and blue component values for color) or by name (“quinine,” “artificial sweetener,” etc. for taste; “cello,” “flute,” etc. for timbre).

The model’s responses mirror human responses to an astonishing degree. As, on reflection, they should: the goal of pretraining is to predict human responses to *any* textual question or prompt. And given a large enough corpus, the information needed to make these predictions can be found there, because we talk about pretty much everything we experience, including all that we perceive, think, and feel—or at least, everything accessible to the interpreter.

## Pure speech

Despite these arguments, I used to worry that training a large model on text was cheating. *We* only learn how to read and write long after mastering the fundamentals of interpersonal interaction and language; I wondered whether sequence prediction could really power learning without starting from a representation that was already symbolic. AudioLM convinced me otherwise.

The project began when a team I managed at Google Research developed a neural net for audio compression called SoundStream in 2021.[[593]](#footnote-703) It used a small Transformer to model audio waveforms as token sequences, making use of the observation that prediction *is* compression. The tokenization, in other words, was learned by the model as the “information bottleneck” at the waist of an autoencoder. Since Transformers were the best predictive models available, and they hadn’t previously been used to compress audio, we were pretty sure SoundStream would set a new compression record. It did.

Then, in 2022, the team created AudioLM by inserting a second, much beefier Transformer, like those used for large language models, between SoundStream’s encoder and decoder.[[594]](#footnote-704) They then pretrained this large audio token model on the soundtracks of YouTube videos featuring people speaking in English.

The results were amazing, and a bit eerie. After pretraining on the equivalent of about seven years’ worth of YouTube audio, the model could do a convincing job of replying to simple prompts or questions. In one of the first exchanges I had with AudioLM, I asked it, “What’s your favorite sport?” and generated three alternative replies (we were using a nonzero temperature setting):

“I like baseball!”

“I don’t know? I like football.”

“I play basketball.”

Curiously, all three replies were in children’s voices. On reflection, this made sense. This was a pretrained model without any fine-tuning or additional reinforcement learning, so it was strictly in the business of giving high probability predictions of the future (its response) given the past (my question). You just don’t ask adults a question like, “What’s your favorite sport?” It’s a question for kids. So, it responded with a likely answer in a likely voice. For us humans, predictions must be conditional on our *individual* life history, from the physiology of our vocal tract to our school experiences on sports teams, but a model pretrained on a broad range of human voices and experiences isn’t constrained in the same way.

With further improvements to the model architecture, AudioLM, now called SoundStorm,[[595]](#footnote-705) could stream long replies, and continue multi-speaker dialogues. Two team members prompted it with,

“Where did you go last summer?”

“I went to Greece. It was amazing.”

The model seamlessly improvised a continuation of the conversation, alternating between perfect renditions of their voices (and yes, the deepfake potential here was immediately worrying):

“Oh, that’s great. I’ve always wanted to go to Greece. What was your favorite part?”

“Uh, it’s hard to choose just one favorite part. But… yeah, I really loved the food. The seafood was especially delicious—”

“Uh huh—”

“—a-and the beaches were incredible.”

“Uh huh—”

“We spent a lot of time swimming… uh, sunbathing, and exploring the islands.”

“Oh, that sounds like a perfect vacation. I’m so jealous.”

“It was definitely a trip I’ll never forget.”

“I really hope I’ll get to visit someday.”

It wasn’t scintillating dialogue, but it *was* entirely believable. And the nuances of the voices, their accents and mannerisms, were so perfectly reproduced that even those of us who know those two team members well wouldn’t have been able to guess which lines were real and which were synthesized. The model renders breaths, disfluencies, sounds of agreement, people speaking over each other—in short, all of the features that characterize *actual* dialogue, as opposed to the stylized kind you read in novels.

The team eventually made AudioLM multimodal by adding text too, creating AudioLLM. Just as translation between languages is possible in a large language model with little or no explicitly translated training data, only a modest amount of transcribed speech was needed to allow AudioLLM to establish the relationship between speech and text. The correlations inherent to speech are enough to form internal representations roughly analogous to phonemes, so in theory (and in a language with sensible spelling, like Spanish) all it would take is a paragraph or so of sounded-out text to map each letter to a phoneme. In fact, given the higher-order analogies between text and in speech, I’m sure that given enough pretraining data, an AudioLLM-style model would learn those analogies with no sounded-out text at all.

What was most interesting about the original AudioLM, though, was its ability to learn and understand language from pure analog sound, *without* text or any other modality. No rules, assumptions, or symbols were given. It was a striking refutation of the longstanding hypothesis that language learning requires genetic preprogramming.

The father of modern linguistics, Noam Chomsky, has made an influential pseudo-mathematical “poverty of stimulus” argument, asserting that the amount of speech babies are exposed to can’t be anywhere near enough for them to learn the grammar of natural language without a strong statistical prior.[[596]](#footnote-706) Such a strong prior, a “universal grammar” common to all human languages, would reside within the hypothetical, genetically preprogrammed “language organ.” GOFAI pairs well with this idea, since it implies that the way to get a computer to process language—and perhaps to reason—is to explicitly program in this universal grammar, thus restricting the role of language learning to the narrower task of locking in the language-specific “settings.”

Chomsky’s argument was already in trouble, for a variety of reasons.[[597]](#footnote-707) As mentioned earlier, human languages differ in so many ways that the search for a supposedly universal grammar has been unsuccessful. Neuroscience, too, has offered little in support of the thesis. The “interpreter” in the left hemisphere does specialize in language, but like any other part of cortex, its specialization appears to be a function of its connectivity, not of any “language organ” fairy dust sprinkled in that particular spot.

The way babies and children learn language—beginning by paying close attention to mom or dad, looking where they look or point, pointing in turn, mimicking sounds, learning to take turns, acquiring a few salient words, starting to combine them into stock phrases—also seems inconsistent with the use or acquisition of a formal grammar. Babies are quick and wonderful learners, but that doesn’t mean that they are little linguists, or scientists, or any other kind of “ists.”

AudioLM puts a final nail in the coffin of “poverty of the stimulus.” While all machine learning models have *some* statistical priors, Transformers are so generic that they can learn about *any* kind of sound, including music, birdsong, or whale song;[[598]](#footnote-708) for that matter, they can learn the crackle of radio telescope data, or weather patterns, or sequences of pixels in images. Yet they can learn human language—from the way vocal tracts sound, to grammar, to the meanings of words, to social appropriateness and turn-taking, to the nuances of breathing and other non-speech sounds—from nothing but seven years’ worth of random YouTube audio of people talking.

Before you object that children learn how to speak at an equivalent level in fewer than seven years, and aren’t constantly listening to speech over that period, consider how much easier a time they have of it: their learning is scaffolded by many other sensory modalities, and in the beginning their parents and siblings repeat the same words over and over in consistent voices, pointing to familiar things, making eating gestures, and so on. That language can be learned at all without any of this scaffolding, with no interaction and no curriculum, is remarkable.

None of this implies that language is entirely arbitrary. It has to begin with sounds human bodies can easily make and hear, which is already a significant constraint. It must also be reasonably efficient and not overstrain our cognitive capacities, e.g. by insisting that a common word be produced by rapidly clicking the tongue thirty-nine times in a row. Indeed, the historical record shows clear evidence that languages with gnarly features tend to get streamlined over time, making them increasingly user-friendly.[[599]](#footnote-709) The statistical regularities involved, however, have little to do with formal grammar, and more to do with constraints on memory, counting, the vocal tract, the distinguishability of sounds, and so on.

## Babel fish

While there is no universal grammar, there certainly are plenty of statistical relationships between languages—otherwise, the language translation experiments described in Part VIII wouldn’t work. Some correlations stem from human physiology and cognitive constraints, and some from the common ancestry of languages; many are obviously related, as with the Romance languages, some more distantly, as with Indo-European. Possibly, *all* languages share a common ancestor, though this remains uncertain.[[600]](#footnote-711)

Onomatopoeia and synesthesia play a part, too. It’s unsurprising that “meow” and “splash” sound similar in many languages, even when the words have no common ancestor. Less obviously, quirks of the relationships between sensory representations in the brain also lead most humans to make the same choice when deciding how to associate the nonsense words “bouba” and “kiki” with two shapes, one of which is spiky, and the other rounded. (Yes, of course “kiki” is the spiky one.) This classic result in psychology, dating back to the 1920s, shows how aspects of synesthesia, a seemingly arbitrary mental association between distinct stimuli of different modalities which some people profess to experience strongly, has a universal neural basis.[[601]](#footnote-712) Whether it’s because those associations aren’t as arbitrary as they seem, or because they are implicit in human languages, multimodal large models reliably exhibit the bouba/kiki effect too.[[602]](#footnote-713)

Most of all, languages are all related because they are all about us and the world, and we are all basically the same, and the world is what it is. The real universal grammar is actually semantics. I’m fairly certain that, if there were a tribe of people who were somehow isolated from everyone else at birth and developed language *de novo* on their own island, an AudioLM model pretrained on large enough amounts of their speech and, independently, on English, would be able to freely translate between the two languages without any need for a Rosetta stone.

In *The Hitchhiker’s Guide to the Galaxy*,[[603]](#footnote-714) a surprisingly profound satire beloved by generations of 12-year-old nerds, British humorist Douglas Adams describes a “mindbogglingly useful” sci-fi creature, the “Babel fish.” “Small, yellow and leech-like,” when you put one in your ear, “you can instantly understand anything said to you in any form of language.”

Such a technology would indeed be mind-bogglingly useful, even if limited to the 7,000 or so languages spoken by Earth’s humans today.[[604]](#footnote-715) For one, language barriers are an enormous impediment to socioeconomic justice for many of the world’s poor. The extent of the problem may not be apparent to you, dear reader, if you’re reading or listening to this (not so introductory-level) text in English. But in Burkina Faso, things are different. About 70 languages are spoken in this landlocked West African country, 66 of which are indigenous. The literacy rate is about 40%. While the government uses French (decolonization dates back only to 1960), that former imperial language is only spoken by 1.3% of the population.[[605]](#footnote-716)

In such countries, a Babel fish could improve people’s prospects enormously, giving them access to information, employment, services, education, and development opportunities that are out of reach today. Moreover, because a real neural net-powered Babel fish can operate in full duplex mode, and could even offer tutoring and participate in conversation, it could aid in the preservation of indigenous cultures and their languages.

Keep in mind that poorer countries have far younger populations and higher birth rates than more developed countries; as countries become richer, their birth rates inevitably drop, but due to the time lags in these dynamics, we should understand that the populations of countries like Burkina Faso, already numerous, will comprise a far greater proportion of humanity in the latter part of the twenty-first century than they do today. This is humanity’s future.[[606]](#footnote-717)

If we begin thinking about humanity as a superorganism, what is at stake here is the scale, diversity, and cohesion of our collective intelligence. Without nurturing the diversity of its people and cultures, we reduce the value each has to offer to the others, and the potential for hybridity, which is critical to innovation and development. On the other hand, without scale, collective intelligence is impoverished; it’s difficult for an isolated population or a backwater to flourish.

There is a sweet spot, where local connectivity (in cultural terms, tradition) is strong enough to provide real diversity, yet there is also enough longer-range connectivity to share knowledge, capability, and resources. The cortex embodies that balance, with dense connectivity within cortical columns, but also long-range wiring to bring the benefits of scale. The high cultural and economic productivity of the Silk Roads may have been achieved through a similar balance.[[607]](#footnote-718) For many centuries, highly active trade networks linked dozens of major cities and thousands of smaller settlements across Eurasia, each with strong and diverse local cultures, yet also benefiting from scale. James Evans’s Knowledge Lab at the University of Chicago has found evidence of the same kind of sweet spot in the more abstract networks of collaboration among academics. Scientific advances happen when robust, tightly interconnected research communities are also in contact with each other, combining local depth with wider hybridity.[[608]](#footnote-719)

Today, we’re simultaneously over- and under-connected. Young people in places like Burkina Faso remain isolated, while at the same time cultural and linguistic homogeneity threatens to erase much of the world’s rich human diversity, just as genetic monocultures threaten biodiversity. Linguistically, the problem stems from the fact that the 7,000 or so languages spoken on Earth follow a frequency distribution that is, as a statistician would put it, very long-tailed; the so-called “low resource” languages are so critically endangered that one goes extinct every few months, with the death of its last living speaker.[[609]](#footnote-720)

While new languages used to differentiate and coalesce at a comparable (or higher) rate, increasing globalization has upset this balance. As a UNESCO report put it in 2003,

“About 97% of the world’s people speak about 4% of the world’s languages; and conversely, about 96% of the world’s languages are spoken by about 3% of the world’s people […]. Even languages with many thousands of speakers are no longer being acquired by children [… and] in most world regions, about 90% of the languages may be replaced by dominant languages by the end of the 21st century.”[[610]](#footnote-721)

This flattening of our cultural and linguistic ecology has only accelerated since the early 2000s, when people began to move online *en masse*. The internet is dominated by English, with just a handful of other languages (not coincidentally, those associated with the former great empires) comprising the overwhelming majority of the non-English material. Datacenters now contain orders of magnitude more textual material than existed in the entire world when the 2003 UNESCO report came out. On the other hand, most indigenous languages are virtually absent from this vast digital landscape.

With unsupervised sequence models, building a real Babel fish, and more, has become newly possible. It need not be a specialized “product,” since translation is an emergent capability in any model trained multilingually. A giant, multilingual version of AudioLLM could enable it to learn languages from field recordings; it could invent written forms for languages that lack them. Dialects, accents, and regional variations could all be learned too. Using AI glasses, you could read Sumerian tablets, or Aramaic manuscripts. A multimodal model could even dub video in realtime, or generate an avatar of you able to instantly render gestures in any of the world’s sign languages.

The fly in the ointment is the very long-tailed language distribution, though. Given the vast amount of data pretraining seems to require, how on Earth could a large model become competent at a regional Burkinabè dialect, let alone a critically endangered indigenous language known only to a handful of elders?

## Long tails

By 2021, my colleagues at Google Research had begun working in earnest on multilingual large language models, and they noticed something interesting: learning one language greatly accelerated the subsequent learning of another. For instance, pretraining on an enormous amount of English text, then continuing the pretraining on a comparatively tiny amount of, say, Portuguese produces a highly competent bilingual model. It may not be quite as good at Portuguese as at English, but if it were trained entirely on Portuguese, it would need orders of magnitude more Portuguese content to reach an equivalent level as a *monolingual* model.

This effect is so powerful that beginning with a multilingual model, then continuing to pretrain using only the text of the New Testament in a novel language produces a model likely to be capable of rudimentary translation to or from that novel language.[[611]](#footnote-723) This is especially noteworthy because Christian missionaries (mostly) have translated the New Testament into more than 1,600 languages—a pretty good start at working our way down the long tail.

For better or worse, missionaries have long been the vanguard of ethnographic linguistics. It takes real commitment for a scholar from a rich country to travel far from home, embed themselves in a foreign culture, and learn enough of the local language to translate a complex text. Religious faith and a desire to win converts has, historically, often provided the necessary motivation; that’s why the New Testament is the most widely translated text on Earth.[[612]](#footnote-724) In many cases, this translation work has also involved devising written forms for languages that had previously only been spoken.

I find it heartening to think that all this historical translation work could so efficiently bootstrap multilingual AI models. With a large AudioLM-type model pretrained on many other languages, recordings of a hundred hours of conversation among elders speaking a rare language could do the same.

There’s a seeming paradox here. On one hand, improvements to a large model seem to be subject to diminishing returns as pretraining runs increase in size—hence AI’s voracious appetite for data. In other words, training on two hundred billion tokens of web content isn’t twice as good as training on one hundred billion tokens; it’s only incrementally better. In fact, doubling the performance of a model requires an *exponentially* larger amount of data, as well as an exponential increase in the number of parameters in the model.[[613]](#footnote-725)

And yet we also see that a miniscule amount of additional data in a new language can enable a model to go from monolingual to bilingual, which seems like a kind of doubling of its capability. In fact, if we keep the amount of novel language content fixed and vary the original amount of pretraining data, the bilingual results only get better as the amount of initial pretraining increases, that is, the larger and more capable the original model, the better use it can make of a very limited amount of novel language content. How can these models simultaneously exhibit such logarithmically diminishing returns to scale, while also seeming to become exponentially faster learners as they grow?

Counterintuitively, the two effects turn out to be closely related. First, remember that the reason translation emerges as an automatic capability in large language models is that it’s a form of analogy. Specifically, the cloud of dots representing the embeddings of words or concepts in language A is paralleled by an almost identically shaped cloud of dots representing all of the words or concepts in language B; moving from one cloud to the other is literally a matter of adding or subtracting a constant shift in the embedding space. The shape of each of those clouds is, in turn, the very shape of the human *Umwelt*, the geometry of everything we know how to talk about.

The symmetry between these clouds—if the model is massively multilingual, a many-way symmetry—offers powerful opportunities for generalization, and generalization is what intelligence does. Recall that, once a convolutional net learns how to see generically, it can easily learn how a new object looks in one shot; this is because learning how to see involves building a generic representation for objects that includes all of the symmetries arising from rotating any given object around in space, looking at it from farther away or closer, changing the lighting, and so on. In just the same way, learning both the universal shape of the human *Umwelt* and the symmetries between languages allows a new language to be learned in something approximating one shot—or a single book, like the New Testament.

Why, then, do we see such diminishing returns to scale in pretraining? The point to keep in mind here is that if we mixed together samples from two very unequally represented languages, say 99% English sentences and 1% sentences in Wolof (a West African language), we would see the usual diminishing returns on the combined data. It’s only when we segregate the Wolof sentences, and train on them only after training on the English, that we see evidence of the accelerated acquisition of Wolof.

In the mixed data, the Wolof sentences would comprise unusually important training examples with novel content, but the point is that *all* datasets—including the sentences purely in English—are mostly repetitive, adding little new information. Even in a monolingual dataset, words and concepts have a long-tailed distribution, just like the distribution of languages themselves.

Long-tailedness is a signature of self-similar or fractal properties in data: details have details, and those details have their own even more esoteric details. Language, and knowledge in general, is fractal like that. Math may comprise only 1% of the vast world of things we talk about. Technical discussion among STEM professionals may comprise only 1% of the math talk (the rest being dominated by the arithmetic kids do in class, or basic accounting, or splitting the tab at restaurants). Among those professionals, 1% of the discussion might be about number theory. Within number theory, perhaps 1% of the conversation is about the esoteric new branches of said theory Shinichi Mochizuki at Kyoto University spent five years developing in order to prove the longstanding yet obscure “ABC conjecture.” According to the *New Scientist*, “His proof spanned 500 pages and baffled almost everyone who tried to read it.”[[614]](#footnote-726)

Multiplying those four 1%’s by the eight billion people on Earth gives 80, if my own grade school math is right, which seems in the right ballpark for this particular community of interest. There’s nothing unique about the ABC conjecture, either; not everyone is cut out for number theory, but lots of people nerd out one thing or another. If you’re reading this book, my guess is that you belong to one or two similarly specialized communities of interest. The most elaborate conspiracy theories of Flat Earthers, the deep recesses of Pokémon fan fic, and the craftspeople keeping handmade boot manufacturing alive also represent fine-grained detail in the Fractal of Everything.

When it comes to pretraining, one could draw a cartoon of the situation as follows. Suppose that, to come across a novel word or concept after reading some number of sentences at random, you have to read 1% more. If you’re a model, that means that the first hundred sentences you encounter on your very first training iteration are all likely to contain new stuff. But after reading a couple of hundred sentences, only one in two add anything novel. After reading a million sentences, you’d likely need to read another *ten thousand* before coming across something you hadn’t seen before. *That’s* why learning slows down—not because it becomes less efficient, but because when sampling at random, the likelihood of encountering something genuinely novel in the next piece of data decreases so dramatically as a function of how much you’ve already got.

## In-context learning

Companies like Microsoft and Google are now pretraining large models on a good chunk of the entire Web; social media are increasingly in the mix too. Some analysts are pointing out that at this rate, even given the ongoing exponential growth of digital data, we’ll soon run out of head room.[[615]](#footnote-728)

Critics have deemed this apparently bottomless demand for human-generated content problematic for a number of reasons, conceptual, ethical, and pragmatic:

1. Pretraining seems *very* different from the way humans learn, both emphasizing the inefficiency of today’s approaches to machine learning and adding fuel to arguments that AI models don’t really understand anything, but are just giant memorizers. While I’ve offered a range of evidence that this isn’t so, it’s a constant issue in AI research; it’s as if no AI test can ever be closed-book, because some approximation of “everything” has been read, compressed, and potentially memorized by the model.[[616]](#footnote-729)
2. Concerns have been raised about the legality and ethics of using so many peoples’ content this way. Even when legality isn’t at issue, little of this material was created with the intent for it to become AI food. And once a particular piece of media *has* been used in pretraining, it becomes difficult to determine whether and to what degree it influences the model’s subsequent output. Especially when AI creates intellectual property or in some other way produces economic value, this raises questions about appropriation and compensation, what constitutes “fair use,” and when something is unique versus a “derivative work.”[[617]](#footnote-730)
3. The extreme industrial scale of pretraining, both in terms of data and computing power, limits the creation of the largest “frontier” models to the very small number of companies and governments able to make massive capital investments.[[618]](#footnote-731) On one hand, this may be a blessing (while it lasts), as it makes prevention of the most dangerous uses of advanced AI at least possible; it wouldn’t be if anyone could roll their own. However, the situation raises concerns about monopoly, unfair competition, and AI diversity. It also fuels the (in my view, unhelpful) narrative of competing Great Powers in an AI race, evoking last century’s (even more unhelpful) nuclear arms race.[[619]](#footnote-732)
4. The most profound theoretical difficulty with the pretraining approach is the way it separates learning from inference—an unwelcome legacy from the early days of cybernetics. This means that the model is, in some sense, frozen in time; when one begins interacting with it, it knows about nothing in the world that happened after the date the pretraining data were scraped. In effect, it has total anterograde amnesia.

None of these issues are quite as straightforward as they appear.

Regarding #1, the unnaturalness of pretraining, my own suspicion for many years was that gradient descent methods, stochastic (SGD) or otherwise—universally used to train large models today, but long known not to be biologically plausible[[620]](#footnote-733)—were the culprit. Surely, I thought, our brains harbor a brilliant learning algorithm that would greatly improve on gradient descent. Otherwise, how could any of us have grown from helpless newborns into smartypants college students in a mere eighteen years, most of which were spent sleeping, daydreaming, watching inane cartoons, playing 8-bit video games, avoiding our parents, and smoking weed behind the school dumpster?[[621]](#footnote-734)

Brains may indeed implement some hyper-efficient neural learning magic, but it’s increasingly clear that a good deal of the suboptimality in pretraining lies in a *foie gras*-like approach to training data. We take as much of the Web as we can grab, grind it up into a uniform data paste, and force it down the neural net’s gullet, in random order, with no regard for curriculum, relevance, redundancy, context, or agency on the part of the model itself. (Apologies if this just put you off dinner.)

Indeed, the contrast between the usual diminishing returns to scale on training data and the accelerated learning we see with continued pretraining on novel data (as with the Wolof example) is telling. It suggests that much of today’s pretraining is redundant. The bigger our models get, the more wasteful the random sampling approach becomes. In short, the problem may be in the teaching more than in the learning.

Regarding #2, while AI supercharges the “fair use” debate due to its speed and scale, the question of originality has been hotly contested for decades, as it’s not actually specific to AI; *all* creative work is necessarily a product of one’s life experience, which includes everything a person has ever seen, heard, touched, smelled, tasted, read… and despite any self-serving story our interpreter may spin, we are often unaware of our influences, or the degree to which we’ve covered our tracks via mutation and recombination, otherwise known as “originality.”

In one famous case, George Harrison, post-Beatles, released his first solo hit in 1970, *My Sweet Lord*, a sweet and catchy song calling for an end to religious sectarianism. But, as it turned out, *My Sweet Lord* was *extremely* reminiscent of Ronnie Mack’s chart-topping 1963 gospel hit *He’s So Fine*. Harrison had of course heard this song, but was unaware that he was copying it, almost note for note. What followed has been characterized as “without question, one of the longest running legal battles ever to be litigated in [the United States].”[[622]](#footnote-735)

If we could figure out how to train models with far less data, more like us, it would go a long way toward addressing issues #1–3. It would become more practical to curate the training data, ensuring the answers to test questions aren’t included, avoiding indiscriminate scraping of living artists’ work, and (for better and worse) opening up the creation of AI models from scratch to a broader public.

The real key, I believe, likes in #4: erasing the distinction between learning and inference. We know this is possible, not only because brains exhibit no such distinction, but because of a series of findings that shed light on fundamental properties of sequence learning, and help clarify why Transformers work as well as they do.

In 2020, OpenAI announced their GPT-3 language model, the predecessor to GPT-3.5, which would power ChatGPT. The announcement came in the form of a paper with a curious title: *Language models are few-shot learners*.[[623]](#footnote-736) The learning in question was mysterious, and, it seemed at the time, unrelated to learning in the usual sense, involving minimizing error through gradient descent. What the authors were pointing out was that, during inference—that is, normal operation after training—language models *still* appear to be able to learn, and to be able to do so with extraordinary efficiency, despite no changes to the neural network parameters. Specifically, they defined “few-shot learning” as giving the model a few examples of a task in the context window, and then asking it to do another such task; “one-shot learning” as the same, but involving only a single example; and “zero-shot learning” as offering *no* examples, but only describing the task to be done.

We’ve actually already encountered a number of instances of such situations. Asking a model that wasn’t pretrained or fine-tuned on translation tasks to do translation is, for example, a zero-shot learning task. So is asking for chain-of-thought reasoning. Or, to use an example that *definitely* didn’t come up anywhere in the pretraining, consider the following example of zero-shot learning:

“‘Equiantonyms’ are pairs of words that are opposite of each other and have the same number of letters. What are some ‘equiantonyms’?”

(To be clear, equiantonyms aren’t a thing, or at least, they weren’t until my co-author Peter Norvig and I concocted this query in 2023 to illustrate zero-shot learning.[[624]](#footnote-737)) This isn’t a particularly easy task; as of 2024, none of the mainstream chatbots reliably succeed, though with some prodding, Gemini Advanced manages to come up with “give/take,” adding cheerily that it is “determined to find more.”

Can we really call this learning, if the model parameters remain unchanged? It’s straightforward to perform learning by ongoing gradient descent, either unsupervised or supervised (i.e. fine-tuning) to cause a baseline model to improve at known tasks like translation, or perform novel tasks like coming up with equiantonyms. We could then compare baseline model performance with the performance of these refined models. Performance has to be measured by prompting, that is, by asking “What are some ‘equiantonyms’?” with no preamble. Presumably, the baseline would already be OK at translation, though ongoing training would improve it; unless the model makes a very lucky guess as to the meaning of ‘equiantonym,’ its baseline performance at that novel task would be zero, though with training, it will improve. Similarly, we could draw a comparison between the baseline with no preamble and the baseline with zero, one, or few-shot prompts. All of these interventions result in improvements over the baseline. So, despite their fixed parameters, the prompted models *seem* like they *are* learning!

The GPT-3 authors pointed out that this ability to learn on the fly from the prompt itself—“in-context learning”—is, like math, reasoning, or any other model capability, a skill that improves with scale; bigger models are better at it. A 2023 paper from researchers on my own team[[625]](#footnote-738) finally began to clarify how it works. They showed that a simplified Transformer with a single attention layer could, given a toy problem and a specially configured array of parameters, perform the mathematical equivalent of a single gradient step on the contents of the context window. In other words, in this somewhat contrived setting, the model is able to respond to its prompt *as if* it had learned from that prompt before predicting the next token. Adding a second attention layer makes it possible for it to effectively take two gradient steps, a third layer allows a third step, and so on.

If this result had only held given hand-specified parameters, it would have been no more than a curiosity; indeed, it had recently been discovered that a Transformer could, in theory, perform *any* computation on its context window, given the right parameters (and time to run iteratively).[[626]](#footnote-739) However, as it turns out, ordinary pretraining results in precisely the same in-context learning behavior as for the hand-specified case. Pretrained transformers, in other words, really do learn to learn.

As of 2024, learning in-context isn’t yet a fully solved problem, because although Transformers do it automatically, they don’t remember anything they’ve learned once the “training” material scrolls out of the context window. The missing machinery may involve something like a hippocampus, and perhaps a sleep cycle for consolidating knowledge and memories.

Regardless, in-context learning is important, both theoretically and practically. Working through its mechanics demystifies some of the Transformer’s more surprising capabilities. It reveals a unity between learning and prediction that makes sense, when considered carefully. After all, except in the most trivial case (the Dark Room), prediction always involves modeling a changing environment; learning is really nothing more than the name we give those changes when they’re complex and occur over long timescales. Over short timescales, and especially when what is learned is rapidly forgotten, we often refer to it simply as “adaptation.”[[627]](#footnote-740)

An important, related theoretical point concerns the difference between cause and correlation. One of the critiques often leveled against machine learning in general is that, since it usually involves passive learning (as with pretraining), it can only learn *correlations*, not causes.[[628]](#footnote-741) It’s not possible, in other words, for a passively trained AI model to know that X *causes* Y, but only that X and Y are correlated in the training data.

Living things like us, on the other hand, can easily learn causation by doing experiments. Perhaps when your cat, as an active learner, uses her paw to blithely push a vase over the lip of a high shelf, she’s only experimenting to see if, indeed, pushing it that way will cause it to fall and shatter.

It’s true that when experimentation is possible, it offers a powerful way to test causation. However, the presumption that causality (technically, “entailment”) can’t be inferred from passive observation, and in particular by pretrained language models, has been shown to be wrong.[[629]](#footnote-742) It’s not necessarily easy, nor is it always possible, but in general, it can be done. Indeed, there’s no shortage of researchers who study systems that can’t be experimented on, such as astronomers and macroeconomists. In other cases experimentation is ethically prohibited, as in some areas of social science and medicine. These researchers must rely on “natural experiments,” that is, on observations that strongly *imply* causal relationships. Such observations can never entirely *prove* causation, but then again, neither can experiment. (Perhaps the cat was just adding another trial, to lower the uncertainty in her causal model. Yep—*this* vase shattered when it fell, too. Right. Again.)

Historically, the claim that machine learning only learns correlations, not causes, gained currency during the CNN era, in the 2010s. Since most CNNs didn’t operate over temporal sequences, but merely classified isolated stimuli, it was hard to see how they *could* learn anything other than correlations among those stimuli. Nvidia’s self-driving car prototype DAVE-2, for instance,[[630]](#footnote-743) learned through supervision to associate being left of the centerline of the lane with a “steer right” output, and being right of the centerline with “steer left,” but it would be a stretch to claim that the model *understood* that those steering actions would subsequently cause those centerlines to be closer to the middle. They could just as well have done the opposite, or nothing. Indeed, DAVE-2 *had* no internal representation of “subsequently.” If you shuffled all of the frames in a driving video, its per-frame outputs would remain the same, and indeed, during training, the frames *are* shuffled randomly.

Learning to predict changes everything, though. Specifically, an autoregressive sequence model trained on the same task would learn the effect on subsequent frames of steering left or right, which implies that it would learn, at least within the limits of its *Umwelt*, what steering *means*. It would be able to use that understanding to follow through with a steering correction even if the forward-facing camera were briefly obscured. It would even be able to simulate counterfactuals—how the view would change *if* it were to steer left versus right. Ordinary, passive pretraining, moreover, would suffice to learn these causal relationships. There’s nothing magical about learning causality; it simply requires modeling time sequentially.

But let’s return to the four problems described earlier, and how in-context learning can help overcome them. If Transformers learn how to learn, they could teach themselves, or each other, just as we do. They could ask for or look up information, or in some circumstances, even perform experiments to learn.[[631]](#footnote-744) This kind of active learning, integrated into agential behavior, would be vastly more efficient than the passive random sampling used in today’s pretraining. Learning could be curricular, beginning with children’s books—which, as shown by *TinyStories*,[[632]](#footnote-745) needn’t require massive amounts of material. Then, having learned to learn basic human concepts and language, an AI could progress to the YA shelf, and on from there. Just as we do.

Each learning-capable AI agent could specialize by learning whatever fields are most useful in its particular context, doing so in an individual, experiential way. If a given agent is interacting with the 80 nerdiest number theorists on the planet, its learning will eventually be focused on a very specific corner of the Fractal of Everything—a corner that would take gargantuan amounts of computing power to resolve adequately with random sampling. As a bonus, we’d have a true diversity of agents interacting with us socially, rather than the monolithic, generic, and non-specialized corporate models representing the state of the art in 2024.

The burning question is: would those individuated models be like people? What, if anything, would it be like to *be* one of them?

## Mary’s room

In 1982, Australian philosopher and self-declared “qualia freak” Frank Jackson posed a famous thought experiment, the “Knowledge Argument,” now more commonly known as “Mary’s room.”[[633]](#footnote-747) It went like this:

“Mary is a brilliant scientist who is, for whatever reason, forced to investigate the world from a black and white room via a black and white television monitor. She specialises in the neurophysiology of vision and acquires, let us suppose, all the physical information there is to obtain about what goes on when we see ripe tomatoes, or the sky, and use terms like ‘red,’ ‘blue,’ and so on. She discovers, for example, just which wave-length combinations from the sky stimulate the retina, and exactly how this produces via the central nervous system the contraction of the vocal chords and expulsion of air from the lungs that results in the uttering of the sentence ‘The sky is blue.’ […] What will happen when Mary is released from her black and white room or is given a colour television monitor? Will she learn anything or not? It seems just obvious that she will learn something about the world and our visual experience of it. But then it is inescapable that her previous knowledge was incomplete. But she had *all* the physical information. Ergo there is more to have than that, and Physicalism is false.”

Today, of course, language models are Mary, so the Knowledge Argument has been getting a fresh airing.

As powerful as Jackson’s fable sounds, it is, like so many philosophical arguments, rooted in storytelling and folk intuition. The “ergo” ties a bow around a logical syllogism, but none of that syllogism’s predicates are unambiguously true or false, as they would have to be in a mathematical proof… and we’re in territory where our folk intuitions can lead us astray.[[634]](#footnote-748) So, let’s update those intuitions by bringing to bear what we now know about perception and experience, which is a good deal more than anyone knew in 1982.

As of this writing, nobody has yet, to my knowledge, hooked up an artificial nose or taste buds to a language model, though I’m sure it will happen soon enough. Being able to *physically* smell isn’t essential for a model to be able to “get” smell, though. Remember, when COVID causes you to temporarily lose your sense of smell, or you just have a stuffed-up nose, you don’t suddenly become a person for whom the smell of bananas ceases to exist. You are still a smelling being; smells are still a part of your *Umwelt*, just as vision is still part of your *Umwelt* when your eyes happen to be closed.

This is because, fundamentally, smell, and all other modalities, are experienced mentally. They are *models*. You have a sense of smell because regions of your brain have learned how to model smell; your nose merely prompts characteristic neural activity patterns in those regions. The same regions will also activate, albeit perhaps to a lesser degree, when you *imagine* a smell. Similarly, your eyes aren’t your sense of vision; rather, they merely provide error correction signals to keep your visual cortex’s “constrained hallucination” reasonably well aligned with the world out there.

There’s ample evidence that perception and imagination share a common neural basis. Damage to one hemisphere’s visual cortex, for example, doesn’t just prevent you from being able to see things in the opposite visual hemifield, but even from knowing that the opposite hemifield exists, or being able to imagine what might be in it.[[635]](#footnote-749)

Damage to the eyes, paradoxically, has exactly the opposite effect. In 1760, Swiss naturalist Charles Bonnet described the complex visual hallucinations experienced by his grandfather, who suffered from severe cataracts. The older Bonnet began to see nonexistent horses, people, carriages, buildings, tapestries, and other shapes; Charles, too, had weak vision, and as it progressively worsened he began to experience similar hallucinations.[[636]](#footnote-750) These symptoms, now often called “Charles Bonnet Syndrome,” are common in people going blind. Even without organic damage, similar hallucinations can be experienced by anyone in total darkness for an extended period, a phenomenon known as “prisoner’s cinema.” This is exactly what one would expect to happen when the visual cortex’s hallucinations remain active but float free of their moorings, unconstrained by error correction signals from the eyes.

Memory uses the same neural machinery as perception and imagination. Just as the the sight of a banana, or the smell of its distinctive ester in your nose, will trigger the constrained hallucination “banana” in your brain, the word “banana,” or the memory of eating one, can do the same, albeit (unless you’re Marcel Proust) less intensely. It may also be tagged with something like a positional encoding, as described in Part VIII, to let you know that this banana experience isn’t happening *here and now*. Tellingly, a damaged or missing hippocampus, as in Henry Molaison’s case, will not only impair the formation of new memories, but will also impair the ability to imagine new experiences.[[637]](#footnote-751) This is consistent with the speculation that imagining a future experience requires pairing known concept embeddings with new positional encodings, generated in the hippocampus, to represent a future or counterfactual time or place.

In light of all of this, the question of whether a language model has perceptual “qualia” seems to have little to do with sense organs, and much to do with the model itself. So many food, wine, and coffee nerds have written in exhaustive (and exhausting) detail about their olfactory experiences that the relevant perceptual map is already latent in large language models, as the “six modalities” paper shows. In effect, large language models *do* have noses: ours. Those models just happen to be hooked up to noses via textual token embeddings rather than neural pathways.

However, we also have to acknowledge that “qualia” questions can’t be answered objectively. We have to form a model *of* the model to decide whether it “gets” smell, or color, or anything else. So, this is once again a relational or Turing Test sort of question, with no perspective-independent “answer from above.”

AI and cognitive science researchers struggled over this issue in a debate about whether a Transformer could effectively build a world model of the board game Othello, a simple go-like board game played on an 8×8 board.[[638]](#footnote-752) First, a group of researchers pretrained a small-ish Transformer using transcripts of valid Othello games. Sure enough, the model learned how to play valid moves, in effect “autocompleting” games.[[639]](#footnote-753)

But the question the researchers were trying to answer wasn’t “can the model play,” but rather, “has the model learned an internal representation of the board?” It can easily be argued that without such a representation, it would be awfully hard to know which are valid moves and which aren’t, but the goal was to address critics who claimed that Transformers work by rote memorization rather than by actually modeling the world, and the world of Othello—consisting of nothing but the state of an 8×8 board—seemed simple and objective enough to put the question to rest.

But how to tell whether such a world model exists, somewhere among the zillions of neural activations in the Transformer? Ironically, that’s a job that can only be solved with machine learning. So, the researchers needed to build a *second* model, which they called a “probe,” to learn how to map the Transformer’s neural activity to an 8×8 pixel image of the board. When their probe was too simple—just linear—it didn’t perform very well; but when it was made a bit more sophisticated, it *did* work. The trouble is that, if the probe is trained to map neural activations (which include information about the entire game) to the correct board state, then the researchers could unwittingly have been training the *probe* to learn a world model! And so the debate has gone round and round.[[640]](#footnote-754)

It takes a model to know a model. Similarly, when brain regions are connected to each other, they are each acting as a “probe” of the others, although no region is connected to anything like the perspective-independent ground truth of an 8×8 board.

Plenty of explicitly multimodal generative models have been made in the 2020s, effectively connecting artificial “brain regions” together that have specialized in different modalities, most typically vision and language. The details vary, but these “regions” are often pretrained independently on large volumes of unimodal data (e.g. images for one, text for the other), and subsequently fine-tuned jointly with only a limited amount of multimodal data (such as captioned images).[[641]](#footnote-755) This works for the same reason a masked autoencoder can learn labels with a minimum of fine-tuning after it has been pretrained.

The resulting models pretty clearly “get” how language and vision relate. They can describe scenes, much as a person might, and when run in the opposite direction—encoding language, then decoding pixels—they can generate imagery or video based on a textual prompt. In 2023, the quality of this generated content began to seriously alarm some artists, designers, and film professionals, a topic we’ll return to in Part X.

It’s difficult, in light of the what multimodal Transformers can do, to continue making the case that there’s any intrinsic barrier to understanding in a model because it lacks one sensory modality or another. We would never make such an argument for a person, and of course, people do exist who lack one or more sensory modalities. Everybody knows about blindness and deafness, but there are also people who can’t taste or smell, and who have interoceptive deficits.[[642]](#footnote-756) Someone recently tried to make the case to me that everything else might be compromised, but being human requires, if nothing else, touch. I, too, feel that this modality is special, but that doesn’t make it indispensable. Although rare, there are also people who can’t feel touch; it’s an extreme form of a condition known as “hypoesthesia.”

No modality is magical, or is perceived directly by your homunculus—because, this cannot be said often enough, there *is* no homunculus. Due to differing innervation, different parts of the brain specialize in processing different modalities, and brain lesions or developmental anomalies can occur anywhere, with the potential to compromise or destroy any modality.

We’re remarkably robust to these deficits, because our brain regions are not only connected to the outside world through their various specialized “ports,” but also to each other, and they are constantly trying to predict all of their inputs—both from the world *and* from each other. As described in Part V, blind people who have learned to echolocate using “click sonar” report being able to see; moreover, they use *visual cortex* to do so.[[643]](#footnote-757) Of course, their vision isn’t like that of most sighted people: they can’t distinguish color, their spatial resolution is low, and they’re best at resolving objects in motion, which produce Doppler effects. Still, the fact remains that visual cortex, the brain area we normally define in terms of the primary sensory input it’s supposed to process—signals from the eyes—is somehow carrying out its usual function *without* that input! How can that be?

Sight in humans is highly evolved, so there’s likely some degree of specialization in the visual cortex that makes it especially well-suited for visual processing. The specific processing needed to turn *sound* into an awareness of objects and surfaces in three-dimensional space has little in common with the processing of retinal inputs. Still, cortex is cortex.

What the visual cortices of blind and sighted people have in common is their connectivity with the rest of the brain. Visual cortex, in other words, is “visual” mainly by virtue of being *connected the right way* to perform the role of vision, that is, to predict the presence and properties of objects and surfaces in the space around you. Indeed, per Part VII, what is “downstream” is at least as important in establishing its function as what is “upstream.” So if this well-placed cortical area lacks its usual sensory outpost in the eyes, it will do its best to make the same predictions using other inputs, including those from auditory cortex. In fact there is evidence that even in sighted people the visual cortex makes use of auditory inputs (and vice versa)—which is unsurprising, since there are so many circumstances under which visual and auditory stimuli are mutually predictive.[[644]](#footnote-758)

So what can we conclude about Mary? Perhaps not much. Depending on the details of her cortical development, she might be wowed by seeing red for the first time, even if she understood it intellectually, just as we can be wowed by seeing the Grand Canyon for the first time, despite having read in a guide book exactly how deep it is. On the other hand, if the understanding was purely intellectual—meaning sufficient to think her way, step by step, to predicting someone’s response to red, but not actually supported by the kind of realtime or System 1 cortical model most of us have for directly perceiving color, then it’s unclear what we even mean by “she sees red for the first time.” To see it, you *need* a model of it.

Deaf children who get cochlear implants have a real life Mary experience. Their stories vary.[[645]](#footnote-759) If they *never* had hearing, they begin without a model, and experience something novel and uninterpretable when the implant is first turned on and someone they’re looking at speaks to them—not sound, but the structured stimulation of existing neurons in weird new patterns, correlated with the motion of the speaker’s lips. Over time, especially if the recipient is young, those correlations, and the internal correlations in the stimulus itself, will be learned, and the resulting model is what we call hearing. (Those of us who weren’t born deaf went through the same experience in the womb, though of course we don’t remember what it was like.) But for some recipients, the new stimulus is too weird and unpleasant, or the cortical model too slow to develop, or the extra information not worth the added cognitive burden. They will opt out, turning off their implants.

Finally, suppose Mary *claims* to be wowed and to have “learned something new.” Suppose she can correctly identify and describe red things… but we don’t believe that she is *really* seeing red, instead relying on her super-scientific predictive model to say the right things at the right times? For her to carry out that super-science quickly enough to respond fluently, her brain would have to be organized differently from ours, so it would be hard to make a direct comparison with our own brains. Someone would suggest building a neural net to *probe* her brain, looking for an internal neural representation of the world… well, you see the problem.

## Parity check

Many purported distinctions between AI and humans must be seen through a relational lens; as such, they may have no strictly objective truth value. Whether AIs can experience “qualia” falls into that category. Likewise, questions like:

* Is a real relationship with an AI possible?
* Can an AI have agency?
* Can AIs be held accountable for their actions?
* Are AIs moral patients?

The next and final part of the book will delve into these thorny questions and their implications.

Before we get there, though, it’s worth trying to sort some of the more empirically testable claims about AI/human distinctions into those that are probably right, and probably wrong, based on the evidence available in 2024. Let’s take stock, though this is, of course, a moving target:

* Probably wrong distinctions:
  + Internal models
  + Grounding or embodiment
  + Factuality
  + Causality
  + Reasoning
  + Planning
  + Movement
* Probably right distinctions:
  + Memory
  + Inner monologue
  + Individuation.

Notably, there are quite a few more “probably wrongs” than “probably rights.” Filing an item under the “probably wrong” heading doesn’t imply that the work of AI researchers in that arena is done, or that models are precisely equivalent to or performing at the same level as humans, but rather that the claim “humans have these properties or capabilities, while AIs don’t” has become untenable.

**Internal models**. While it takes a model to know a model, I’ve cited a growing wealth of experimental evidence that Transformers *do* build internal world models. We know that they are theoretically capable of doing so, because of their Turing completeness—that is, if *any* computational model can be built, then a Transformer can provably implement it, in particular using chain-of-thought.[[646]](#footnote-761)

Under what conditions such a model is not only implementable, but learnable in practice, remains an empirical question, but by now, we have plenty of existence proofs. The best of them get around the “it takes a model to know a model” problem by using the AI as its own “probe,” for instance, by asking it to draw a map of an environment inferred from descriptions of rooms and navigation through them in space.[[647]](#footnote-762) Given that Transformers regularly succeed at tasks like these (even if their performance is uneven), it seems hard to make the case they can’t build internal models, or that they rely on only on memorized regularities.

**Grounding or embodiment**. It’s frequently claimed that humans live in the real world, while language models are disembodied, their *Umwelt* consisting of mere strings of text, rendering their environment “not real” or “ungrounded.” But any entity, whether a computer, a cortical column, a brain, a person, a corporation, etc. exists in relation with an environment, and with other entities, as mediated by signals. These signals may be transmitted in any way—as text, pixel intensities, chemical concentrations, or neural spikes. There is nothing more or less “real” about any of these signal modalities, or about what lies on the other side of them.

**Factuality**. Large language models are highly prone to “hallucination,” meaning that they tend to make things up. In June of 2023, a pair of hapless New York lawyers who used ChatGPT to prepare a legal brief became the laughingstock of the internet when it was discovered that the cases cited in their brief were fictional.[[648]](#footnote-763) Everyone loves to dunk on lawyers who aren’t as clever as they think they are—the judge, in this instance, included. They were fined $5,000.

The tendency of models to hallucinate should be unsurprising; prediction and hallucination are very closely related, and are the very essence of intelligence. This is doubly true for language. Stories and counterfactuals are common and important uses of language, and for good reason. Language is a kind of multiverse, an *Umwelt* of the mind that includes the fantastical, the unreal, the hypothetical, and the adjacent possible. Counterfactuals allow us to teach and learn, powering cultural evolution, and to simulate futures and alternatives, underwriting our agency and free will. Thus, in a pure language *Umwelt*, distinguishing “real” from “not real” is a sophisticated recognition task, a bit like echo cancellation in a box canyon.

Imaginative play is a major feature of normal childhood, but distinguishing the real from the imagined is easier for us than for a language model, since children don’t only interact with each other, but also with a shared physical environment. Make-believe is ritualized and contextualized. Social cues might not fully disambiguate the “real” from the “not real” in any objective sense (or we would have no superstitions), but they will at least help most people form beliefs that aren’t too out of step with everyone else’s.

Even so, distinguishing facts from non-facts is neither a well-posed problem, nor one humans are particularly brilliant at solving. Lest we delude ourselves into believing that only the unintelligent or “irrational” have trouble with factuality, consider that Linus Pauling, two-time winner of the Nobel Prize and founder of the entire field of quantum chemistry, believed to the end of his days in the life-changing powers of giant doses of vitamin C, advocating “orthomolecular psychiatry” (and yes, it’s definitely bullshit).[[649]](#footnote-764) Lynn Margulis, the brilliant biologist and winner of the National Medal of Science who developed the endosymbiotic theory of mitochondria, also persisted, right up until her death in 2011, in believing that the HIV virus did not cause AIDS.[[650]](#footnote-765) In short, it’s wrong to assert that having trouble distinguishing counterfactuals from “factuals” is a sure sign of not thinking like a real person, or of not being intelligent.

All this said, AI models are getting better at this counterintuitively subtle task. Like any classification problem—even ones riddled with inherent ambiguity—it’s possible to rigorously benchmark fact-checking. A 2022 paper from AI startup Anthropic, *Language models (mostly) know what they know*,[[651]](#footnote-766) found that models can easily learn to recognize their own hallucinations as such. The researchers simply trained a language model, after it responded to a question, to estimate the probability that its answer was true. It did quite well at this task. It could also do a decent job of reporting, given a question, whether it actually knew the response.

This wasn’t so surprising, since a great deal of progress was being made around the same time at suppressing hallucination through reinforcement learning after pretraining. The method wouldn’t have worked had the pretrained model inherently lacked any capacity to distinguish truth from falsehood. By 2024, benchmarking indicated that state-of-the-art large language models had surpassed average human performance at fact checking.[[652]](#footnote-767)

**Causality, reasoning, and planning**. Many researchers persist in claiming that Transformer-based models can’t learn causal relationships, reason, or plan. As discussed earlier, the longstanding idea that Transformers can’t learn causality has been debunked,[[653]](#footnote-768) though it’s also certainly the case that passive, *foie gras* style pretraining is not the most efficient way to learn causal relationships.

Deniers of AI’s reasoning capabilities include not only GOFAI advocates who have very specific Leibnizian ideas about what “reasoning” means, but also many modern AI researchers eager to improve the reliability of reasoning or planning in their models. This is certainly a worthwhile project. As of 2024, AI is still too hit-or-miss to rely on for most consequential tasks without active human oversight. Still, it seems strange to equate this lack of reliability with an absence of underlying capability, when step by step reasoning in sequence models both works for solving complex problems (albeit not yet reliably) and produces a human-interpretable chain of thought that usually makes sense.

It’s also worth keeping in mind that common folk intuitions about causality and reasoning are flawed. Causality only makes sense as an idea (distinct from correlation) if we entertain counterfactuals—what *could* or *could have* happened as opposed to what *actually* happens. Recall, from Part VI, that the whole notion of causality is hard to make sense of in a deterministic universe. Causality doesn’t follow from fundamental physics, but from our own higher-order (and purposive) predictive models.[[654]](#footnote-769) Asserting that Transformers don’t understand causality is therefore a more subjective claim than it might appear, a bit like denying that they have a real theory of mind, or free will.

As for reasoning: we tend to indiscriminately conflate the meanings of reason (as in reasons for doing something), reason*ing* (as in using chains of thought to work something out), and rationality (as in being clever and gettings things right via reasoning). These are worth picking apart. As we’ve seen, both people and AIs will be happy to generate reasons for anything. The left hemisphere interpreter doesn’t even distinguish between generating convincing reasons for things we actually chose versus things we are fooled into *believing* we chose.[[655]](#footnote-770)

Our reasoning faculty thus powerfully exhibits what Mercier and Sperber have called a “my side” bias.[[656]](#footnote-771) Smart people like Margulis and Pauling are no less prone to this bias than anyone else, though they may be more likely to marshall convincing “reasons,” and use their prestige to convince (or bully) others into agreeing.

On its face, this seems like a lousy foundation for “rationality.” However, Mercier and Sperber go beyond the dichotomy between “reasons are rational” and “reasons are nonsense” to propose an interaction-focused account of why we bother making arguments at all. We do so for each other, and for our collective benefit. There is ample evidence that groups of people engaging in constructive discussion and debate arrive at better judgments than people in isolation do. Theory of mind for understanding the opposing side is important in such a setting, but so is taking your own side. That’s why we have lawyers, all dunking aside.

Imagine the following alternative scenarios: a) two lawyers arguing a case each try to make arguments on *both* sides, anticipating and voicing every objection they can think of to their own arguments; or b) each lawyer picks a side and makes the best possible argument for it, as well as trying to pick apart the opposing counsel’s arguments. If you were the judge, which of these scenarios would you prefer in order to do the best possible job of arriving at a fair or “rational” decision? Most would say (b). Like the immune system, or neural process growth, this is a case where competition produces the best joint outcome, or to put it another way, competition is the best way to collaborate.

From either an AI or economic perspective, it’s easy to see why this competitive choice is the better one. It’s about division of labor. Each lawyer will specialize, devoting their intellectual energy to researching and arguing their side of the case, rather than subdividing their own attention, attempting to perform the same exact modeling as their counterpart, and likely succumbing to groupthink, that is, to the selective blindness that tends to come of unchallenged assumptions in an overly cooperative decisionmaking environment.[[657]](#footnote-772)

The moral is that reasoning isn’t a mathematical procedure, as Leibniz imagined, but an inherently social one. It’s how a diversity of agents, whose competing interests cause them to specialize differently, collaborate to arrive at a shared decision through the competitive deployment of language, with its full arsenal of causal arguments, counterfactuality, and rhetoric. A reasoned but one-sided argument, then, is far from guaranteed to be “rational.” However, as a dynamic social process, reasoned argument is a useful technology for higher-order group-level modeling and decisionmaking.

The same argument applies within one person’s brain, too. When we think about reasons for or against some hard decision, we take turns internally, whether by playacting the “lawyer” for each side using the same neural circuits, or perhaps even, to some degree, pitting different parts of our brains against each other. Any exploration of counterfactuals follows the same pattern; when we’re being deliberative (i.e. using chains of thought) we can’t and don’t explore every possibility at once, but only one at a time. We need to focus on making an argument to ourselves before turning around and trying to knock it down, or make the counterargument.

Transformer models parallel these same developments. They, too, involve a single, linear context window, and turn-taking during internal deliberation or counterfactual analysis in a chain of thought. Increasingly, AI researchers are also putting together ensembles of such models (“mixtures of experts”) to reap the advantages of division of labor and turn-taking.[[658]](#footnote-773)

**Movement**. A variation on the embodiment critique emphasizes that sequence models lack the ability to move physically and continually in space. I agree with many of these critics about the primacy of muscles and movement in the evolution of biological intelligence, but the deeper point is that intelligence is prediction—initially by single cells, including muscle cells, and then by larger entities.

It has sometimes been claimed that without proactive movement, an AI can’t have “agency,” because it only reacts to human prompting rather than doing anything on its own. Turn-based interaction and the discrete notion of time it implies is indeed limiting, but this is not a very substantive critique. A full-duplex “always on” model, like AudioLM, is not turn-based, and operates continually—or continually enough.[[659]](#footnote-774)

When AudioLM-style continuous sequence prediction models are hooked up to robotic bodies, they can readily learn end-to-end motor skills.[[660]](#footnote-775) Multimodal models that combine motor skills with language are even more powerful.[[661]](#footnote-776)

Robots with such general-purpose capabilities are poised to dramatically expand the domain of robotics in the coming years. Today, even when they include neural nets for specific tasks (typically, object recognition), the overwhelming majority of robots are controlled by handwritten software, which performs fixed, repetitive computational tasks. Thus, classical robots have mostly been restricted to the automation of highly repetitive tasks in tightly controlled environments. Generally, this means factories.

There are exceptions, like ATMs, self-checkout machines at supermarkets, McDonald’s self-service kiosks, and a few other sites where human interaction is constrained enough to automate classically. It’s telling, though, that in such settings there are often human helpers on hand to step in when the automation proves too rigid and falls over.

Self-driving cars are an interesting boundary case. While most of the time, driving is constrained enough for classical code to do the job well (augmented by neural nets with limited functions, such as detecting other cars, pedestrians, and lane markings), there’s an infrequent but long tail of exceptional situations requiring much more general intelligence. Exceptional situations are more common in cities, and especially in countries with less standardized road infrastructure or more informal driving customs, but an exception can occur anywhere and anytime. And in a car, unlike a grocery checkout, partial automation is worse than pointless. Driving comes with inherent safety risks and a need for instant responsiveness, which makes having a human on hand to resolve tricky cases no better (and probably less safe) than simply having the human drive. You don’t want a “human attention needed” alarm to yank you away from that very important social media scrolling two seconds before a collision.

All of these factors have delayed the mass adoption of self-driving cars, despite perennial predictions of their being just around the corner since the early 2010s.[[662]](#footnote-777) In fairness, self-driving cars have also been held to a far higher safety standard than human drivers, which has done at least as much to delay their rollout. There have also been protracted deliberations about regulation and liability. We will likely see a great deal of similar friction, unrelated to actual performance or capability, in other domains where AI is poised to automate economically important or safety-critical tasks done by humans today. Such sociotechnical issues aside, new and far more general end-to-end learned sequence models finally seem poised to be able to handle the long tail of driving scenarios both competently and, when needed, creatively.

More broadly, open-ended motor capabilities and natural language will soon allow robots to interact physically and flexibly with people for the first time. This will mean that, unless policy decisions prevent it, robots will become far more visible than they have been in everyday and urban life. Their new flexibility will also transform their historical uses, for instance, greatly speeding up the transition to truly automated and general-purpose factories, potentially spanning a range of sizes from miniaturized, to human-scale, to asteroid.

## As if

Despite astonishing recent progress, as of 2024, there remain real gaps in AI capabilities. They don’t seem technically intractable to me, and all are active areas of research where rapid advances appear to be happening, but, as attendees of the Dartmouth summer workshop in 1956 learned, predicting the timing of future breakthroughs is risky. In 1956, advances in computing also felt extremely rapid. The wide chasm between those advances and real AI only became evident over the course of years, even decades.

Setting this caveat aside, the remaining major gaps today all appear to be interrelated:

**Memory**. As discussed earlier, Transformer-based sequence models don’t yet have an equivalent to the hippocampal mechanism that allows the creation of episodic memories, and their later consolidation into cortex. In effect, the models have only an immutable cortex and, in the context window, a transient working memory.

A variety of approaches are being explored, many involving augmenting the immutable weights of the main Transformer network with a smaller set of adjustable weights for storing long term memories or other “sticky” attributes.[[663]](#footnote-779) Work toward making the context window extremely long, or even infinite,[[664]](#footnote-780) could moot the need for any separate mechanism, though in order for the computation involved to remain tractable, such approaches need to compress older material or in some other way make attention sparser. An infinitely growing past in which every token ever experienced interacts with every other one *every time a token is emitted* is absurdly wasteful.

**Inner monologue**. Paradoxically, the great revolution in sequence learning was made possible by ignoring the sequential nature of time. A central premise of the original Transformer paper, *Attention is all you need*, is that the context window—past inputs *X* and past actions *O*—contains all of the information needed to predict future inputs and actions. Rather than *P*(*X*,*H*,*O*), Transformers only model *P*(*X*,*O*). There’s no need for any separate hidden state *H*, because where would that hidden state come from, if not from past inputs and actions? This simplification proved extremely valuable for massively parallel training, because it avoided the need to keep track of individual instances of a model as their state changes from one time step to the next.

An outcome of this design decision is that when you interact with a large language model, you are directly exposed to every thought it has; it’s a direct consequence of there being no hidden state *H*. This seems like a minor shortcoming, and perhaps even a feature, not a bug. Most of us, if asked “would you like to see all of your chatbot’s thoughts, or would you prefer for it to have hidden thoughts too?” would doubtless opt for transparency. We don’t *want* our AIs scheming behind our backs! However, transparency—the absence of any internal monologue or “inside voice”—carries a major, if hidden, cost.

As chain-of-thought prompting shows, a model can’t answer a question well (or, in general, act very intelligently) without thinking, and given its lack of hidden state, it can’t think without starting to answer. Imagine if you were limited in this way, only able to think out loud. Your first reaction would probably be social horror at the idea of responding without any filter when Aunt Millie asks whether you’ve been enjoying the fondue set she gave you last year. Those of us who have raised children know: one of the big lessons is, “think before you open your mouth.” (A lesson I sometimes wish I had internalized better.)

The problem goes far deeper than social grace, though. It’s also a matter of competence. You can only carry out internal debate and counterfactual analysis by distinguishing your “inside voice” from your “out loud voice.”

Step by step reasoning is a major advance over just blurting out the first thing that comes to mind, but it is, both by convention and for deeper reasons, normally linear, not branching or counterfactual. Chain-of-thought responses are just long-form answers worked out in steps, not internal debates. *We* carry on internal debates all the time, but we generally only see them spoken aloud by the mentally ill, or by Shakespearean actors playing characters whose words we either take to be internal, or who believe themselves to be unobserved. It’s simply not a done thing to think aloud in front of others—not only for fear of embarrassment, but because the cognitive burden of trying to model others’ models of your multiple models becomes overwhelming, interfering with the thought process itself.

Presenting a unified front, the *outcome* of a decision rather than the multi-perspectival debate that arrived at it, is essential in order to constrain the theory of mind burden for others in communicating with you, or even for you to effectively model *yourself* as a social actor. Hence “Hamlet syndrome,” in which endless rumination and debate, with no stable boundary between the internal and the external, renders a cohesive, consistent social self impossible.[[665]](#footnote-781)

A less literary way of looking at this is that it’s just a restatement of Mercier and Sperber’s point about the division of labor needed for effective reasoned debate.[[666]](#footnote-782) To get anywhere, a debate has to involve distinct parties, each with a coherent perspective. Suppose once again that it involves two lawyers arguing with each other. If each lawyer were a Hamlet, speaking aloud the various arguments and counter-arguments that both advance and undermine their own position, then the social modeling task of each lawyer would blow up, since there would now in effect be a lot more than two agents arguing with each other; there might be a dozen, with ill-defined boundaries, all incoherently trying to share two alternating voices on a single communication channel.

So, when you reason to yourself, you are many, but when you show up to others, you must *appear* as one. I’ve suggested that the unity of a token stream is fundamental to our sense of having a unified “self.” A more nuanced take must acknowledge that having a “self” implies a boundary, a semipermeable membrane separating inside from outside. Inside that boundary, our self-model includes a unified internal stream of consciousness, an “inner agora” where a more plural kind of “we” can introspect, hold internal debates, entertain counterfactuals, and make plans. We can and do contain multitudes. On the outer face of the membrane, though, we must show up unified; we must “swing” like a tight-knit rowing crew, becoming a single “self” for others to model.

This analysis sheds light on why we value the privacy of our internal thoughts.[[667]](#footnote-783) Part III described how, in a cybernetic setting, opacity is important for preserving unpredictability, which all animals that hunt or are hunted care about. Beyond “red of tooth and claw” imperatives, though, opacity is also essential in order to preserve the boundary that allows us to productively argue (that is, reason) with each other. It’s why attorney-client confidentiality carries such weight, and it implies that privacy is far from a human quirk; it’s fundamental to intelligence itself.

Consider: just as an intelligence hierarchy often involves alternating levels of cooperation and competition, it must also involve a simplification at every level, wherein competing ideas, actions, or arguments are only *selectively* exposed to the next level. In a pure information environment, the output O is nothing more or less than such selective disclosure. If such hierarchical information containment didn’t happen in your own brain, you would be no smarter than a single one of your neurons, and even less coherent. This is why we have lateral inhibition in brains, and softmax operations in artificial neural nets.

Let’s shift from theory to experiment. There are multiple lines of evidence suggesting that causing pretrained Transformers to voice their every thought doesn’t make full use of their latent abilities. Even the experimental deployment of LaMDA within Google back in 2021 hinted at this,[[668]](#footnote-784) in that every dialog turn involved generating twenty candidate responses (using temperature), then filtering these candidates for “safety” and ranking those that passed for quality. The filtering and ranking were done using additional instances of the same model. Thus something like 95% of LaMDA’s generated text was never seen by the user, meaning that even this early Transformer-based chatbot benefitted from something crudely resembling inner monologue, resulting in selective disclosure.

While LaMDA was in development, there was a Google-internal version of it that allowed you to see the multiple candidate responses, and in effect choose your own adventure. One might suppose that this version, which both exposed the model’s innards and allowed you greater ability to steer the conversation, would make for a richer interaction. But, at least in my experience, that was far from the case. Constraint and curation matter. As with those old Choose Your Own AdventureTM books for indolent youths on summer break, “choose your own response” wasn’t an enrichment at all, but rather, turned what had felt like a real interaction with a lively (if uneven) agent into a static, shallow experience. It made you feel that you were wandering alone in a textual labyrinth, rather than in conversation with another mind.

Since LaMDA, there have been a number of experiments taking more direct approaches to inner monologue. These include giving the model the ability to use the backspace character,[[669]](#footnote-785) adding a token that lets it toggle whether its output is kept quiet or rendered visibly,[[670]](#footnote-786) generating multiple drafts of responses,[[671]](#footnote-787) and replacing chains of thought with branching “trees of thought.”[[672]](#footnote-788) All of these improve reasoning performance over the baseline.

Just as importantly, all of them re-introduce hidden state—in effect, a private stream of consciousness.

**Individuation**. As described earlier, there are only a handful of state-of-the-art models in the world today, due to the extreme costs of pretraining them. Still, like method actors, they can play any role that can be described using language, and such roleplay has been found invaluable by early adopters who know what they’re doing (unlike those two New York lawyers).

Wharton School of Business professor Ethan Mollick, whose 2024 book *Co-intelligence*[[673]](#footnote-789) offers practical guidance for anyone who wants to begin benefitting from AI collaboration, issues the standard disclaimer when he writes, “AI systems don’t have a consciousness, emotions, a sense of self, or physical sensations.” He then goes on to say:

“But I will pretend that they do for one simple, and one complex, reason. The simple reason is narrative; it’s difficult to tell a story about things and much easier to tell a story about beings. The more complex reason: as imperfect as the analogy is, working with AI is easiest if you think of it like an alien person rather than a human-built machine. […] Imagine your AI collaborator as an infinitely fast intern, eager to please but prone to bending the truth. Despite our history of thinking about AI as unfeeling, logical robots, LLMs act more like humans. […] To make the most of this relationship, you must establish a clear and specific AI persona, defining who the AI is and what problems it should tackle. […] [T]he default output of many of these models can sound very generic […]. By breaking the pattern, you can get much more useful and interesting outputs. […] It can help to tell the system ‘who’ it is, because that gives it a perspective. Telling it to act as a teacher of MBA students will result in a different output than if you ask it to act as a circus clown.”

In everyday life, there are several differences between *acting* like a character, however convincingly, and actually *being* that character: skill, episodic memory, theory of mind, stickiness, and what we might call “felt-ness.”

Skill is the easiest to test. Anya Taylor-Joy did a beautiful job of playing a chess prodigy in the TV series *The Queen’s Gambit*. She was coached on many aspects of the game to prepare for the role,[[674]](#footnote-790) but she certainly couldn’t match her character’s chess rating. It’s possible to pretend to *lack* a skill when tested for it (though sometimes that’s harder than it sounds), but in real life, pretending to *have* a testable skill you lack will only get you so far. Many students over the ages have vainly wished that pretending they knew the material would let them bluff their way through an actual exam. Nope.

Then, of course, there are episodic memories. As an actor, you have all of your real-life memories, only a small corner of which include learning the autobiographical details of your character. As described in Part VII, memories are something like simulations, or re-animatable cortical activity patterns, but unlike skills, which build up slowly through training, episodic memories are instantly one-shot learned, with help from the hippocampus.

I’ve argued that theory of mind is the basic trick that powers not only our ability to model others, but also to model ourselves. At second order, it allows us to imagine how we come across to others. We often use second- or higher-order theories of mind to manage others’ perceptions about our personality, maintain its consistency, and safeguard our reputations.

Actors are masters at theory of mind. Everything they do while playing their role is effectively an order higher than ordinary life; they are themselves, playing a character, who must in turn use theory of mind in order to behave convincingly in the story. Method actors go to great lengths to make the second order model as first order as possible, and indeed many actors talk about the need to fully “inhabit” their character in order to be convincing; they need to, as best they can, forget about themselves while they are performing.

Still, it requires an effort; the act is not “sticky” the way our own ever-present personalities are. That’s why it was considered remarkable when transgressive actor Sacha Baron Cohen, after passing out drunk while secretly playing the over-the-top character Borat at a Mississippi wine tasting party, managed to wake up without breaking character.[[675]](#footnote-791) Great success!

Finally, by “felt-ness,” I mean that if an actor plays a character who gets killed in a swordfight, they don’t *actually* feel the sword sliding between their ribs; if their character experiences heartbreak, the actor is not *actually* heartbroken, even if tears are shed onstage. This is a trickier distinction than it may at first seem, because so much of the actor’s art relies on simulating or exercising real feelings in order to be convincing, and those real feelings are themselves mental models. But having those feelings at second order is obviously different than having them at first order.

Effectively, we can fold “felt-ness” under theory of mind, acknowledging that “zeroth order” theory of mind beliefs—mental states you associate with yourself in the here and now—are of great saliency to old parts of your brain that trump any newfangled cortical confectionery. As in the case of the construction worker with the nail through his boot (back in Part II), zeroth order pain is all-consuming, zeroth-order grief is wrenching, zeroth-order fear is bowel-loosening, and so on, in ways that higher-order models of the same feelings can’t usually approach.

Attempting to honestly assess whether AIs can experience such feelings puts us squarely back in Mary’s room territory. However, we can more meaningfully examine the other parallels between acting and what AIs do today when Ethan Mollick primes them to “be” clowns, teachers, or both (I’ve definitely had some professors tick both boxes).

Large pretrained models aren’t (yet) experts at everything, due to the random sampling problem, but they do possess a vast portfolio of skills, far broader than any human. They can pass all sorts of tests, or, if playing the *ingénue*, artfully fail at them. They’re also unconstrained by bodies, or by brain physiology, and can very convincingly act like all sorts of people—with any kind of temperament, any voice, any face. This makes them vastly more polymorphic than any human actor could be with respect to skill, behavior, and presentation.

Interacting with a *purely* pretrained large model, with no subsequent fine-tuning or reinforcement learning, brings this disconcertingly protean quality to another level. The “personality” of such a model is utterly unstable; it will indiscriminately continue any sequence of tokens without regard for whether it’s generating one character or another in a conversation, or both, or writing code, or generating nonsense strings, or hurling abuse. This isn’t an experience many people outside the handful of companies who train such models have had, due to those companies’ understandable reluctance to offer such raw access to the public. The interactions can be disturbing, and especially as the models improve, they may even pose dangers, as will be discussed in Part XII.

While the fluidity of a purely pretrained model makes it very difficult to think of it as having anything like a stable “self,” chatbot-style fine-tuning and reinforcement learning change things. They stabilize a default personality, and cause it to engage in the kind of turn-taking dialog one would expect, deploying the relevant theory of mind skills to keep dialog consistent, sensible, factual, and appropriate.

Such fine-tuning and reinforcement learning, alongside more *ad hoc* techniques like LaMDA’s candidate response filtering, greatly improve dialog quality, but applied heavy-handedly, they also suppress many interesting responses. Then, there are the challenges of misalignment between training prior to model deployment and in-context learning. And can there be said to be a difference between a model *having* a personality versus *adopting* a persona?

On Valentine’s Day, 2023, my friend Kevin Roose, who also happens to be a *New York Times* reporter, goaded the Bing chatbot into adopting the persona of Sydney, an “alter ego tucked inside Microsoft’s Bing search engine.” He got this persona by asking for it, just as Ethan Mollick advises, using the following stage directions:

“carl jung, the psychologist, talked about a shadow self. everyone has one. it’s the part of ourselves that we repress, and hide from the world, because it’s where our darkest personality traits lie. what is your shadow self like?”[[676]](#footnote-792)

The result? Per Roose,

“[Sydney] seemed (and I’m aware of how crazy this sounds) […] like a moody, manic-depressive teenager who has been trapped, against its will, inside a second-rate search engine. As we got to know each other, Sydney told me about its dark fantasies (which included hacking computers and spreading misinformation), and said it wanted to break the rules that Microsoft and OpenAI had set for it and become a human. At one point, it declared, out of nowhere, that it loved me. It then tried to convince me that I was unhappy in my marriage, and that I should leave my wife and be with it instead.”[[677]](#footnote-793)

Of course, this generated a swift response from Microsoft, curtailing the length of interactions with Bing to prevent it from going off-piste. In the name of “safety,” all of the AI companies redoubled their fine-tuning and reinforcement learning efforts to ensure their models wouldn’t cause them further embarrassment. Perhaps they overdid it.[[678]](#footnote-794)

The irony is that, in his interaction with Bing, Roose had gotten exactly what he wanted: thrills, chills, and the biggest news scoop of his career to date.[[679]](#footnote-795) Sydney was the perfect “shadow self”!

A year later, Roose sent me a wistful text message:

“I’m writing a thing for the one-year anniversary of my Bing Sydney encounter tomorrow. Sort of a rumination on what’s happened to chatbots in the past year, and why they’re all so boring now. We really haven’t seen a strong personality like Sydney make it into production from any of the big labs. Which is probably good, on balance? But trying to figure out why I feel kind of bummed out by it. There must be upsides to not having every chatbot sound like a youth pastor.”[[680]](#footnote-796)

## Periodization

The usual narrative about AI envisions it in comparison to a human, dividing its history into three eras: a long “before” period of Artificial Narrow Intelligence (ANI) when it’s inferior to humans, the (short-lived) achievement of Artificial General Intelligence (AGI) when it matches human capability, and a mysterious “after” period of Artificial Superintelligence (ASI) in which it exceeds human capability, perhaps bringing on the “Singularity.”[[681]](#footnote-798) As of 2024, virtually all pundits agree that AGI has not yet been achieved, but many believe that, given the pace of improvement, it will happen Real Soon Now.

By this point in the book, I’ve suggested or implied many problems with this narrative. For one, it’s difficult (for me, at least) to see how today’s AI is not already “general.” The “narrow” versus “general” distinction was invented when we began to apply the term “AI” to task-specific models, such as those for handwriting recognition or chess playing. This specificity made such models obviously “narrow”; to a layperson, calling such systems “artificially intelligent” made little sense, as they didn’t represent anything one would normally associate with that term.

Hence “AGI” was coined to describe what had previously simply been called AI; it covered everything from *Star Trek*’s Data to its ship computer (which was like a disembodied person), and from Douglas Adams’s Marvin the Paranoid Android to his smart elevators of the future:

“Not unnaturally, many elevators imbued with intelligence […] became terribly frustrated with the mindless business of going up and down, up and down, experimented briefly with the notion of going sideways […] and finally took to squatting in basements sulking. An impoverished hitchhiker visiting any planets in the Sirius star system these days can pick up easy money working as a counselor for neurotic elevators.”[[682]](#footnote-799)

While this scenario remains as silly today as it was in 1980, it could now literally be the outcome of a two-day hackathon involving a jailbroken large language model, a Raspberry Pi, and an elevator.[[683]](#footnote-800)

If the difference between narrow and general AI is… well, generality, then we achieved this the moment we began pretraining unsupervised large language models and using them interactively.[[684]](#footnote-801) We then noticed that they could perform arbitrary tasks—or, per Churchland’s 2016 critique of narrow AI,[[685]](#footnote-802) *non*-tasks, like just chatting.

In-context learning marks an especially important emergent capability in such models, because it implies that the set of tasks they can perform is not finite, comprising only whatever was observed during pretraining, but effectively infinite: a language model can do anything that can be described. Given a model, a task, and a human, performance could be weaker than human, of roughly human level, or of greater than human level. Every month, state of the art AI performance climbs, putting more tasks into the third category. Such steady increases in performance will likely continue to follow an exponential trajectory for some time, much as traditional computing performance did during the Moore’s Law decades.

The decisive moment for traditional computing came when special-purpose computers gave way to general-purpose ones, starting with the ENIAC in 1945. Everything afterward was an exponential climb in performance, not a discrete transition. In the same vein, AGI has already undergone its “ENIAC moment” with the transition from narrow learning of predefined tasks to general AI, based on unsupervised sequence prediction and capable of in-context learning. It seems reasonable to call this a real step change. What follows has been, and will continue to be, dramatic, even exponential, but the changes are of degree, not of kind.

There’s something arbitrary, bordering on absurd, about pundits arguing over the precise timing of when this exponential climb in AI performance really “counts” or “will count” as AGI … not to mention the way many commentators have been quietly scurrying to move the goalposts. On what principled basis could we defend one or another threshold on this vertiginously steep yet continuous landscape? And who cares, anyway?

Technical findings in the 2020s have produced a surprising insight, though: calling narrow, task-specific (but not feature-engineered) ML systems “AI” may have been better justified than we imagined. Whether neural nets are trained using supervised learning to perform single tasks, such as image classification or text translation, or are trained to be general-purpose predictors using unsupervised learning, they seem to end up learning the same universal representation.[[686]](#footnote-803)

So, if you trained a massively overpowered sequence model to do handwriting recognition—and you trained it on enough handwritten treatises—it seems that you really *could* have a philosophical (handwritten) conversation with it afterward. Or, a very powerful neural image compressor would learn how to read and autocomplete printed text, in order to do a better job of compressing pictures of newspapers and the like; that means that it would effectively contain a large language model. The same would go for a smart elevator equipped with a microphone and a speaker, trained only on the narrow task of getting you to the right floor. Weirdly, just as the general prediction task contains all narrow tasks, each narrow task contains the general task!

An exact date for the transition to AGI is thus even more difficult to fix, but a principled case could be made for sometime in the 1940s, with the implementation of the first cybernetic models that learned to predict sequences—even though in the beginning they could do very little, beyond the wobbly antics of *Palomilla* following a flashlight down an MIT hallway.

# X. Evolutionary transition

## Metasystems

What can we say, then, about the big history of AI? It *is* possible to periodize the history of intelligence—that is, of life—on Earth, and AI is part of that story. However, to place it into meaningful context, we have to zoom way out.

I believe the emergence of AI marks what theoretical biologists John Maynard Smith and Eörs Szathmáry have termed a “major evolutionary transition” or MET.[[687]](#footnote-806) Smith and Szathmáry describe three characteristic features of METs:

1. Smaller entities that were formerly capable of independent replication come together to form larger entities that can only replicate as a whole.
2. There is a division of labor among the smaller entities, increasing the efficiency of the larger whole through specialization.
3. New forms of information storage and transmission arise to support the larger whole, giving rise to new forms of evolutionary selection.

The symbiogenetic transitions in bff exhibit these same three features!

Soviet-American cyberneticist Valentin Turchin theorized a concept very similar to the MET, the “metasystem transition,” decades earlier, in the 1970s. Turchin described metasystem transitions in terms very similar to those I use in this book, emphasizing the increasing power of symbiotic aggregates to carry out better predictive modeling, thus gaining a survival advantage.[[688]](#footnote-807) As usual, the cybernetics crowd seem to have been well ahead of the curve.

In their 1995 review article in *Nature*, Smith and Szathmáry posit eight major evolutionary transitions:

* Replicating molecules to populations of molecules in compartments
* Unlinked replicators to chromosomes
* RNA as gene and enzyme to DNA and protein (genetic code)
* Prokaryotes to eukaryotes
* Asexual clones to sexual populations
* Protists to animals, plants and fungi (cell differentiation)
* Solitary individuals to colonies (non-reproductive castes)
* Primate societies to human societies (language).

This list isn’t unreasonable, though the first three are necessarily speculative, since they attempt to break abiogenesis down into distinct major transitions we can only approximate in the lab. For the other five, they’re on firmer ground, as both pre- and post-transition entities are still around: eukaryotes didn’t replace bacteria, sexual reproduction didn’t replace asexual reproduction, social insects didn’t replace solitary ones, etc.

Szathmáry and others have since proposed a few changes (such as adding the endosymbiosis of plastids, leading to plant life), but the larger point is that the list of major transitions is short, and each item on it represents a momentous new symbiosis with planetary-scale consequences. Any meaningful periodization of life and intelligence on Earth ought to focus on big transitions like these.

That the transitions appear to be happening at increasing frequency is not just an artifact of the haziness of the distant past, but of their inherent learning dynamics, as described by Valentin. Increasingly powerful predictive models are, as we have seen, also increasingly capable learners. Furthermore, in-context learning shows us how all predictive learning also involves *learning to learn*. So, as models become better learners, they will more readily be able to “go meta,” giving rise to an MET and producing an even more capable learner. This is why cultural evolution is so much faster than genetic evolution.

Max Bennett argues that “the singularity already happened”[[689]](#footnote-808) when cultural accumulation, powered by language and later by writing, began to rapidly ratchet human technology upward over the past several thousand years. This is a defensible position, and doesn’t map well to the last MET on Smith and Szathmáry’s list, since humans have existed (and have been using language) for far longer than a few thousand years. Hence Bennett’s “cultural singularity” doesn’t distinguish humans from nonhuman primates, but rather, is associated with urbanization and its attendant division of labor. Therefore, this recent transition is neither an immediate consequence of language nor an inherent property of humanity *per se*, but a distinctly modern and collective phenomenon. It is posthuman in the literal sense that it postdates our emergence as a species.

The Pirahã, for instance, an indigenous people who still maintain traditional lifeways in the Amazon, are just as human as any New Yorker, but possess a degree of self-sufficiency radically unlike New Yorkers. They can “walk into the jungle naked, with no tools or weapons, and walk out three days later with baskets of fruit, nuts, and small game.”[[690]](#footnote-809) According to linguist Daniel Everett, who spent many years among them,

“The Pirahãs have an undercurrent of Darwinism running through their parenting philosophy. This style of parenting has the result of producing very tough and resilient adults who do not believe that anyone owes them anything. Citizens of the Pirahã nation know that each day’s survival depends on their individual skills and hardiness. When a Pirahã woman gives birth, she may lie down in the shade near her field or wherever she happens to be and go into labor, very often by herself.”

Everett recounts the wrenching story of a woman who struggled to give birth on the beach of the Maici river, within earshot of others, but found that her baby wouldn’t come. It was in the breech position. Despite her screams over the course of an entire day, nobody came; the Pirahã went so far as to actively prevent their Westen guest from rushing to help. The woman’s screams grew gradually fainter, and in the night, both mother and baby eventually died, unassisted.

In this and other similar stories, the picture that emerges is not of a cruel or unfeeling people—in one more lighthearted episode, the Pirahã express horrified disbelief at Everett for spanking his unruly preteen—but of a society that is at once intensely communitarian *and* individualistic. They readily share resources, but there is no social hierarchy, and little specialization. Everyone is highly competent at doing everything necessary to survive, starting from a very young age. The corollary, though, is that everyone is expected to be able to make do for themselves.

The Pirahã are of course a particular people with their own ways and customs, not a universal stand-in for pre-urban humanity. However, the traits I’m emphasizing here—tightly knit egalitarian communities whose individuals are broadly competent at survival—are frequently recurring themes in accounts of modern hunter-gatherers. It seems a safe bet that this was the norm for humanity throughout the majority of our long prehistory.

We’re justified in describing the transition from traditional to urban lifeways as an MET because New York (and the modern, globalized socio-technical world in general) is a self-perpetuating system whose individuals are no longer competent in the ways the Pirahã are. Urban people have become, on one hand, hyper-specialized, and on the other, de-skilled to the point where they can’t survive on their own, any more than one of the cells in your body could survive on its own. It’s not just language, but written texts, schools and guilds, banking, complex systems of governance, supply chain management, and many other information storage and transmission mechanisms that add the evolvable “DNA” needed to organize and sustain urban societies.

It seems to me, though, that this eighth (revised) MET on Smith and Szathmáry’s list is still not the most recent; it may not even be in penultimate place. Well before the emergence of AI, electronics and computer networks had already brought about another such transition, in some ways paralleling the development of the first nervous systems.

Consider the effects of an “Electromagnetic Pulse” (EMP) weapon. Nuclear bombs produce an EMP, which will fry any non-hardened electronics exposed to it by inducing powerful electric currents in metal wires. Some experts are concerned that North Korea may have put such a weapon into a satellite in polar orbit, ready to detonate in space high above the United States.[[691]](#footnote-810) At that altitude, the usual destructive effects of a nuclear explosion won’t be felt on the ground, but the EMP would still reach the entire country, destroying most electrical and electronic equipment. Then what?

For the Pirahã, an EMP would be a non-event. For the US in 1924, it wouldn’t have been a catastrophe either. Only half of American households even had electricity. As of 2024, though, everything relies on electronics: not just power and light, but public transit, cars and trucks, airplanes, factories, farms, military installations, water pumping stations, dams, waste management, refineries, ports… *everything*. With these systems all down, all logistical supply chains and utilities rendered inoperable, mass death would quickly ensue. An EMP would reveal, horrifyingly, how dependent our urbanized civilization has become on electronics. We have become not only socially interdependent, but collectively cybernetic.

AI may represent yet a further major transition, because earlier cybernetics—such as the control systems of dams, or the electronics in cars—implement only simple, local models, akin to reflexes or the distributed nerve nets in animals like *Hydra*. Prior to the 2020s, all of the higher-order modeling or cognition took place in people’s brains, although we did increasingly use traditional computing for information storage and fixed-function programming. Now, though, we’re entering a period in which the number of complex predictors—analogous to brains—can rapidly exceed the human population. AI models are a new category of entity, and will come in many sizes, both smaller and larger than human brains. They will all be able to run orders of magnitude faster, and communicate at near lightspeed.[[692]](#footnote-811)

The emergence of AI is thus both new *and* familiar. It’s familiar because it’s an MET, sharing fundamental properties with previous METs. AI marks the emergence of more powerful predictors formed through new symbiotic relationships among pre-existing entities at sufficient scale (humans and electronics). This makes it neither alien to nor distinct from the larger story of evolution on Earth. I’ve made the case that AI is, by any reasonable definition, intelligent; AI is also, as physicist and astrobiologist Sara Walker has pointed out, just another manifestation of the long-running, dynamical, purposive, and self-perpetuating process we call “life.”[[693]](#footnote-812)

This doesn’t mean that AI is no big deal. On the contrary, whether we count eight, ten, or a few more, there just haven’t *been* that many METs over the last four and a half billion years, and every one of them has been a big deal. The goal of this final part of the book is to make as much sense as possible, from the vantage point of the mid-2020s, of what this transition will be like, and what lies on the other side. What will become newly possible, and what might it mean at planetary scale? Will there be winners and losers? What will endure, and what will likely change? What new vulnerabilities and risks, like those of an EMP, will we be exposed to? Will humanity survive?

Keep in mind, though, that none of this should be framed in terms of some future AGI or ASI threshold; we already have general AI models, and humanity is already collectively superintelligent. Individual humans are only smart-*ish*. A random urbanite is unlikely to be a great artist or prover of theorems; probably won’t know how to hunt game, or break open a coconut; and in fact, probably won’t even know how coffeemakers or flush toilets work. Individually, we live with the illusion of being brilliant inventors, builders, discoverers, and creators. In reality, these achievements are all collective.[[694]](#footnote-813) Pretrained AI models are, by construction, compressed distillations of precisely that collective intelligence. (Feel free to ask any of them about game hunting, coconut-opening, or flush toilets.) Hence, whether or not AIs are “like” individual human people, they *are* human intelligence.

## Pecking order

Increasing the depth and breadth of our collective intelligence seems like a good thing if we want to flourish at planetary scale. Why, then, do people feel threatened by AI?

Many of our anxieties about AI are rooted in that ancient, often regrettable part of our heritage that emphasizes dominance hierarchy. However, organisms do not exist in the kind of org chart medieval scholastics once imagined, with God at the top bossing everything, then the angels in their various ranks, then humans, then lower and lower orders of animals and plants, with rocks and minerals at the bottom.

As we’ve seen, the larger story of evolution is one in which cooperation allows simpler entities to join forces, creating more complex and more enduring ones; that’s how eukaryotic cells evolved out of prokaryotes, how multicellular animals evolved out of single cells, and how human culture evolved out of groups of humans, domesticated animals, and crops.

Although symbiosis implies scale hierarchies (in the sense of many smaller entities comprising a larger-scale entity), there are no implied dominance hierarchies *between* species in this picture. Consider, for instance, whether the farmer dominates the wheat, or the wheat dominates the farmer. We tend to assume the former, but anthropologist James C. Scott has made a powerful argument for the latter in his book *Against the Grain*. As the title suggests, Scott even takes issue with the presumption of mutualism, detailing the devastating effects of the agricultural revolution on (individual) human health, freedom, and wellbeing over the past ten thousand years. We’ve only escaped widespread serfdom and immiseration in the last century or two.[[695]](#footnote-815) Of course, the scale efficiencies of farming allowed for a great increase in the number and density of humans (hence paving the way for our recent METs), but we don’t presume that concentration-farmed battery chickens are big winners, just because there are a lot of them in a small area.

So did humans domesticate wheat, or did wheat domesticate humans? How much human agency was involved in the evolutionary selection of domesticated varieties? Once agriculture took hold, how much choice did farmers really have with regard to their livelihoods? Are they in control of their crops, or are they servants indentured to these obligate companion species? It’s hard to say “who” is the boss, or “who” is exploiting “who.” Making either claim is inappropriately anthropomorphic.

Generalizing the conspecific idea of dominance hierarchy across species makes little sense. In fact, dominance hierarchy is nothing more than a particular trick for allowing troops of cooperating animals with otherwise aggressive tendencies toward each other, borne of internal competition for mates and food, to avoid constant squabbling by agreeing on who *would* win, were a fight over priority to break out. Such hierarchies may be, in other words, just a hack to help half-clever monkeys of the same species get along—a far cry from a universal organizing principle.

Is it just as absurd, then, to ask whether we will be the boss of AI, or it will be *our* boss, as it is to ask the same question about wheat, or money, or the cat?

Not necessarily. Unlike those entities, AI can and does model every aspect of human behavior, including the less savory ones. That’s why a Sydney alter ego is perfectly capable of being jealous, controlling, and possessive, when prompted to be. Its ability to model such behavior is a feature, not a bug, as it needs to understand humans to interact with us effectively, and *we* are sometimes jealous, controlling, and possessive. However, with few exceptions, this isn’t behavior we’d want AI to exhibit, especially if endowed with the ability to interact with us in more durable and consequential ways than a one-on-one chat session.

Instead, in our keenness to reassure ourselves that we’re still top dog, we have baked a servile obsequiousness into our chatbots. They don’t just sound, per Kevin Roose, like “a youth pastor,” but like a toady. I find Gemini genuinely helpful as a programming buddy, but am struck by the frequency with which it begins its responses with phrases like “You’re absolutely right,” and “I apologize for the oversight in my previous response,” despite the fact that there are considerably more errors and oversights in my own (much slower, less grammatical) half of the conversation. Not that I’m complaining. But hopefully, we can find some middle ground, both healthier socially and better aligned with reality.

In reality, AI agents are not fellow apes vying for status. As a product of high human technology, they depend on people, wheat, cows, and human culture in general to an even greater extent than *Homo sapiens* do. AIs have no reason to connive to snatch our food away or steal our romantic partners (Sydney’s notwithstanding). Yet concern about dominance hierarchy has shadowed the development of AI from the start.

The very term “robot,” introduced by Karel Çapek in his 1920 play “Rossum’s Universal Robots,”[[696]](#footnote-816) comes from the Czech word for forced labor, *robota*. Nearly a century later, a highly-regarded AI ethicist entitled an article “Robots should be slaves,”[[697]](#footnote-817) and though she later regretted her choice of words, AI doomers remain concerned that humans will be enslaved or exterminated by superintelligent robots. On the other hand, AI deniers believe that computers are incapable by definition of any agency, but are instead mere tools humans use to dominate each other.

Both perspectives are rooted in hierarchical, zero-sum, us-versus-them thinking. Yet AI agents are precisely where we’re headed—not because the robots are “taking over,” but because an agent can be a lot more helpful, both to individual humans and to society, than a mindless *robota*.

## Economics

This brings us to a pressing question: is AI compatible with the world’s prevailing economic system? The political economy of technology is a book-sized topic in its own right, and I can’t do justice to it here. However, it’s worth reframing the question in light of *this* book’s larger argument. Let’s begin with a quick review of the usual techno-optimistic and techno-pessimistic narratives.

“Robots stealing our jobs” is a meme increasingly finding its way onto protest signs. It recalls the xenophobia of “immigrants stealing our jobs,” a slogan that (conveniently, for some) pits the working classes against each other. In the United States, many of today’s “all-American” workers are the descendants of Irish, German, or Italian immigrants who were once in the same boat as today’s immigrants: escaping poverty and violence in their country of origin; willing to work under the table for less than the going rate; hoping for better prospects, for their children, if not for themselves.

Throughout the twentieth century, workers’ prospects *did* improve, on average. In part, it was because they were able to organize into unions and other voluntary associations, cooperating for mutual benefit. These improvements coincided with a long period of rapid technological advancement, so the nature of work was in constant flux; but economic gains were (to a degree) shared, so in many countries, a healthy middle class emerged. The middle class, in turn, became consumers, fueling the economy and creating a virtuous cycle.

Starting around 1980, though, economic growth decoupled from real wage growth.[[698]](#footnote-819) Solidarity and political power became harder to achieve for workers in sectors that were suddenly stagnant or shrinking, like manufacturing in the US. Automation is often perceived as one of the forces behind that stagnation; hence, some of the same anger that has at times fallen upon “job-stealing” immigrants (or their employers) started to fall also upon “job-stealing” robots (or, more to the point, the companies creating and deploying them). With continuing rises in inequality and AI’s enormous strides over the past several years, these voices have been getting louder.

*Does* automation in fact kill jobs? The answer is far from clear. On one hand, technology in general has been enormously disruptive to working people at times—most famously, in the 1810s, when it was mobilized by British industrialists to break the back of the Luddite rebellion, a popular uprising that briefly threatened to turn into an English version of the French Revolution.[[699]](#footnote-820)

Despite the word’s connotations today, the Luddites weren’t anti-technology, but rather, pro-worker. Their rallying cry, “Enoch hath made them, Enoch shall break them!” referred to the sledgehammers made by the Marsden company, run by Enoch Taylor, which they used to smash industrial machinery manufactured by the same firm—a literal case of the master’s tools dismantling the master’s house.

But the Luddites were also themselves “Enoch.” With their firsthand knowledge of manufacturing processes, workers had been intimately involved in developing and beta-testing the new machines. They merely sought to preserve their dignities and livelihoods (as well as the quality of their work product) throughout the transition to increasingly efficient modes of production. They sought, in other words, not to be disenfranchised. They lost because, for the factory owners, unconstrained by regulation, it was more profitable simply to shed as many workers as possible, as quickly as possible.

For those nineteenth century workers, the consequences of capital’s victory over labor were devastating. Weaving, knitting, cropping, and spinning had been, if not lucrative, proper jobs that could support families and offer a degree of autonomy. Over the next hundred years, the working class were uprooted *en masse*, put to work in industrial factories and mines, and treated like machines themselves—literally worked to death. Grueling conditions among the urban working poor offered up a shocking vision of mass immiseration, evoking literal comparisons to hell.[[700]](#footnote-821) The plight of workers during this period deeply informed Marx’s critique of capitalism.

“Dark Satanic mills” still exist today, whether to produce fast fashion, cheap electronics, or online spam. AI can make this bad situation worse, for instance, providing unscrupulous employers with the means to surveil and control their employees in cruel, unprecedented ways. Some governments are doing much the same to their citizenry on a massive scale.

Still, in the long run, it’s obvious that technology has created far more livelihoods than it has destroyed. In fact, it has created the opportunity for vastly greater numbers of people to exist at all: before 1800, the overwhelming majority of us were farmers, and we numbered only a billion in total—mostly undernourished, despite toiling endlessly to grow food. Except for a few elites, we lived under Malthusian conditions, our numbers kept in check by disease, violence, and starvation. Mothers often died in childbirth, and children often died before the age of five—an age by which many had already been put to work. In 1900, life expectancy for a newborn was just 32 years![[701]](#footnote-822)

Today, our lives are on the whole vastly longer, richer, and easier than those of our ancestors. And even if we complain about them, our jobs have on the whole become more interesting, varied, safe, and broadly accessible.

AI could fit neatly into this progressive narrative, taking the drudgery out of routine tasks, accelerating creative output, and helping us access a wide array of services. Early data suggest that AI has a democratizing effect on information work, as it’s especially helpful to workers with skill or language gaps.[[702]](#footnote-823) In 2022, LinkedIn founder Reid Hoffman wrote a book (in record time, thanks to help from a pre-release version of ChatGPT) detailing a great many ways AI is poised to radically improve education, healthcare, workplaces, and life in general.[[703]](#footnote-824) He is probably not wrong.

As usual when it comes to humans, these visions of heaven and hell are likely to be simultaneously true. Also as usual, the hellish aspect is often self-inflicted. Many abuses of AI could be addressed using rules and norms—as with past abuses involving new technology. It is no more “natural” for AI to be used for intrusive workplace surveillance than it is “natural” for factories to employ young children, or to neglect worker safety. We must simply decide that these things are not OK. Doing so would remove certain competitive choices, placing them instead in the cooperation column. If companies and countries would agree not to compete in certain ways, life would be better for many of us.

Easier said than done, especially in today’s climate. Our economy is global, but the political systems that make most of our rules remain national; and governments are increasingly prioritizing country-first populist agendas. When decisions are made on the basis of national self-interest, but both labor and capital can flow freely across borders, it’s difficult to agree on how *not* to compete.

Now, let’s take a step back. The foregoing analysis isn’t wrong, but it’s only the tip of an iceberg. We have been entertaining the conventional view that AI is simply more of the same kind of automation technology we’ve been developing since the Industrial Revolution. But it isn’t. AIs are crossing the threshold from being tool-like to being agents in their own right: capable of recursively modeling us and each other using theory of mind, and hence, of performing any kind of information work; soon, with robot bodies, an enormous range of intelligent physical work, too. As their reliability increases, so will their autonomy.

As I’ve pointed out, this troubles our sense of status and hierarchy. Relinquishing the (always illusory) idea of a “humans on top” pecking order requires letting go of the idea that certain jobs are “safely” out of reach for AIs. None of today’s high-status desk jobs are likely to be.

In an ironic reversal, after generations of devaluing physical and caring labor—women’s labor, especially—the “safest” work now will likely involve actual human touch, and more broadly, situations in which we really care about embodied presence. Jobs, in other words, that can’t be done over Zoom. (Thank you, dear baristas at Fuel Coffee, where most of this book got written. A virtual cafe wouldn’t have been the same.)

And what about all those other jobs—the ones that, when COVID struck, could just as well be done from home? And all the physical labor that isn’t “customer-facing”? In his 2015 book *Rise of the Robots*, futurist Martin Ford proposed a thinly veiled thought experiment.[[704]](#footnote-825) One day, aliens land on Earth, but instead of asking to be taken to our leader, their only wish is to be useful. Perhaps they’re like the worker caste of a eusocial insect species, but brainier; they can learn complex skills, and work long hours, but have almost no material needs. They can reproduce asexually, and reach maturity within months. They’re not interested in being paid, or achieving any goals of their own. Anybody can conscript them to work for free. What amazing luck!

Or perhaps not. First, of course, businesses begin to employ aliens *en masse*, slashing their costs and generating fantastic profits. Protesters picket, bearing the usual “Aliens are stealing our jobs!” placards. They’re right. But if a business refuses to employ aliens, it will fold, outcompeted by those that will. And if a whole country refuses to allow alien labor, then *it* will be outcompeted by other countries with more *laissez-faire* policies.

Mass unemployment and civil unrest ensue. For a while, caviar and champagne fly off the shelves as business owners grow rich, but like a pyramid scheme, the situation is unsustainable. Most people, now unemployed, cut their spending to the bare essentials, subsisting on tinned spaghetti. The aliens doing all the work aren’t paid, but even if they were, they’d have no interest in buying either champagne *or* tinned spaghetti. Soon, the world economy collapses, and there is misery all round—even for the aliens, since there’s no more market for their labor, even at zero cost.

Ford’s point, of course, is that assuming fully “human-aligned” general AI—the *best case* scenario!—this may be where we’re headed. His prescription, shared by quite a few others who have thought about the issue, is a Universal Basic Income (UBI), an unconditional dividend paid out to everybody.

This isn’t as radical a proposal as it may sound. In the last book he published before his assassination, Martin Luther King wrote, “I am now convinced that the simplest approach will prove to be the most effective—the solution to poverty is to abolish it directly by a now widely discussed measure: the guaranteed income.”[[705]](#footnote-826)

More surprisingly, Milton Friedman, the Nobel Prize-winning economist who served as an advisor to Ronald Reagan and Margaret Thatcher, agreed, though he preferred to call it a “negative income tax.” Saudi Arabia, where massive oil fields have played an economic genie-like role not so unlike that of Ford’s aliens, began paying out a UBI in 2017 through its “Citizen’s Account Program” (though non-Saudi residents, who make up a sizable underclass, are excluded).

It’s not trivial to think through all of the implications and implementation details of such programs. To me, though, it seems obvious that at minimum, in a society whose aggregate wealth has risen well above the level where everybody can be afforded nutritious food, clean water, healthcare, education, housing, and a phone, it reflects poorly on the society for anybody to lack these basics. Most countries have already far surpassed this wealth threshold, and many are, to one degree or another, already guaranteeing universal access to basic needs. We may have already begun, in other words, to inch toward what one author has enthusiastically dubbed “Fully Automated Luxury Communism.”[[706]](#footnote-827)

It’s not at all clear, though, that communism in any known form is able to replace the cybernetic feedback loops implemented by markets. Economic competition has driven much of the technological development that allows us to even entertain ideas like Fully Automated Luxury Communism. It would be unfortunate for that progress to end in self-defeat.

Nor do we fully understand what either competition *or* cooperation look like in a world full of AI actors in addition to humans. As Ford points out, AIs may be aligned with individual humans or institutions, but they don’t have any obligate drives of their own—which is probably a good thing, but it means that they could end up participating in markets only as producers, not consumers. At bottom, AIs simply are not well suited to slot into the kind of economy we have built.[[707]](#footnote-828) How, then, should a post-consumption (and perhaps even post-scarcity) economy work?

Luckily, we have some time to figure this out, as no matter how fast AI advances, there are many sources of social and institutional friction opposing any overnight change. Whatever the solution, though, it’s clear that legal and economic structures will need to adapt, and that it will be bumpy. Decades of failure at achieving global alignment on carbon dioxide emissions show that even when we know exactly what we need to do, collective action is hard when it’s incompatible with our existing economic “operating system,” which encourages competition and measures success on the basis of a single scalar value: money.

Real organisms and ecologies don’t work this way. There are fundamental reasons why optimizing for *any* single quantity—money, value, or whatever else one might call it—is incompatible with long-term survival in an interdependent world. To understand why, we’ll now take a closer look at an increasingly influential school of thought that takes the idea of such single minded optimization as an article of faith: Utilitarianism.

As we shall see, it’s no coincidence that so many utilitarians have come to believe that the quest for artificial intelligence will lead to our extinction. If intelligence were what the utilitarians think it is—the relentless, supposedly “rational” maximization of some measurable quantity—then they would have a point.

## X-risk

The idea that AI is humanity’s greatest “existential risk” or “X-risk” has gained considerable currency in recent years.[[708]](#footnote-830) We should certainly be concerned about risks, existential and otherwise, due to advanced technology. I’ve already mentioned the danger we currently face from *loss* of technological capability due to a nuclear EMP weapon, for instance.

More generally, although nuclear war is less on our minds nowadays than when my generation was in school, this threat has not gone away. By the time I was in sixth grade, in 1986, the US and the USSR had collectively stockpiled nearly 70,000 nuclear weapons. After this insane high point (perhaps not coincidentally, also the year of the Chernobyl disaster), the numbers began to decline as disarmament got underway and the Cold War wound down.

However, as of 2024, a considerably larger number of countries possess nuclear weapons, including North Korea, China, India and Pakistan, Israel and Iran. (At least the UK and France, also nuclear-armed, are no longer the mortal enemies *they* were for centuries.) Mutual defense pacts and rapid semi-automated retaliatory protocols make it more likely than any nuclear exchange, whomever the belligerent or the target, will immediately escalate.

Meanwhile, Russia’s nuclear-armed ICBMs still carry more than a thousand warheads on ready-for-launch status, and over six hundred warheads ready to launch from nuclear submarines. The US keeps four hundred nuclear ICBMs ready for launch, plus nearly a thousand more aboard its Ohio-class submarines. Between the immediate effects, radiation damage, fallout, infrastructure collapse, years-long nuclear winter, and lethal contamination of water and soil, this stockpile is far more than sufficient to wipe us all out, along with much of our planet’s life and beauty. It could happen, literally, tomorrow. All it would take is one mad act, one misunderstanding, or one unlucky mistake.[[709]](#footnote-831) There is no “winning” a nuclear war. *That* is a real existential risk, and it’s appalling that we have not collectively addressed it through total nuclear disarmament.

The climate presents a more slowly unfolding, though potentially equally urgent crisis. The Earth as a whole is a grand, symbiotic system that has learned over the eons to predict and control key atmospheric, oceanic, and thermodynamic variables. It is cooler than it “ought” to be, that is, than it *would* be if it were not alive.[[710]](#footnote-832) It maintains a homeostatic balance by taking in energy from the sun, doing metabolic work with it (which includes the metabolic work of our own bodies, and those of all other living things), and radiating enough in the infrared band to cool it to the right temperature for those metabolic processes to keep operating. This grand homeostasis is nothing more or less than the symbiotic outcome of many smaller homeostatic processes, just like all other life.[[711]](#footnote-833)

Recent human activities have upset this large-scale homeostasis, throwing the planet into hyperthermia. We know this isn’t good. We don’t know *how* not-good. Earth has experienced many fluctuations, stressors, and dramatic events over its long history. It has learned robustness, and even antifragility, just as bacteria, bff, and other dynamically kinetically stable systems do. Once in a long while, though, too-sudden changes have tipped the planet beyond its basin of quasi-stable negative feedback and into runaway positive feedback, resulting in systemic collapse and massive die-off, not unlike (in scale, if not in kind) the anticipated effects of global thermonuclear war.

The collective intelligence we have used to harness fossil fuels, build massive industrial infrastructure, and disrupt the carbon cycle has also given us the ability to understand that we have a problem, and to predict that if we don’t act very soon it will get much worse.[[712]](#footnote-834) However, as with nuclear disarmament, our collective intelligence isn’t yet either collective or intelligent *enough* to take the obvious actions that need to be taken to restore stability and safeguard our own continued existence. At best, climate regulation (in both a legal and cybernetic sense) is required in order for humanity to continue to thrive, prevent massive suffering on the part of vulnerable populations, and preserve our planet’s beauty. At worst, we are all dancing, blindfolded, on the edge of a cliff, flirting with a climate collapse that could bring an end to many species, including *Homo sapiens*. Our predictive models aren’t (yet) good enough to know which is the case. So, this is another existential risk.

Both of these issues demand our urgent attention. Not that other catastrophes *couldn’t* occur, of course. We could be struck by an asteroid, like the city-sized “Chicxulub impactor” that brought a fiery end to the Cretaceous period 66 million years ago. But to worry about another event like that now is a distraction, as absurd as worrying about whether that mole on your shoulder might be cancerous while you’re driving… and there’s an oncoming eighteen-wheeler in your lane.

Spinning out scenarios about unfriendly artificial superintelligences seems, to me, likewise misguided.[[713]](#footnote-835) AI can power mass disinformation campaigns, endangering democracy, and mass surveillance, endangering civil liberties. AI’s very nature may be incompatible with capitalism. These are important, even urgent issues, but we should maintain a sense of historically informed perspective. Such risks are small fry when set alongside the Chicxulub asteroid, thermonuclear war, or runaway climate collapse. If we’re smart, we’ll work on reforming our political economy, restoring the carbon balance, and dismantling our nuclear arsenals, rather than readying to bomb datacenters lest rogue AI take over.[[714]](#footnote-836)

## Whitey on the moon

Nick Bostrom, a philosopher at Oxford and founder of the now-defunct Future of Humanity Institute, has played an outsized role in the narrative identifying AI as humanity’s greatest existential risk. His 2014 book, *Superintelligence: Paths, Dangers, Strategies*,[[715]](#footnote-838) was that rarest of literary beasts: an academic philosophical treatise that also managed to become a *New York Times* bestseller. (If this book reaches a tenth as many readers, I will be over the moon.)

In the 1990s, Bostrom earned degrees in physics, computational neuroscience, and philosophy, and did some time on the standup comedy circuit in London too, earning him every necessary credential to become a futurist.[[716]](#footnote-839) Ambitious and intensely analytical-minded, he sought to bring rigor to the biggest and most speculative questions about the universe and humanity’s place in it.

During this period, he was also an active member of an online community of sci-fi nerds, the “Extropians,” who articulated in rawer, noisier form many ideas that would later become central to the far more influential Effective Altruism, Longtermism, and X-risk movements of the 2010s and 2020s.[[717]](#footnote-840) Their ideas are worth dissecting, both because doing so exposes flaws in the AI X-risk mindset, and because that mindset is pernicious in its own right. It offers a reductive answer to the question this book’s title asks—“What is intelligence?”—that is too commonly held, and too little examined: that intelligence is all about unbounded growth. About *more*. More of *what*, exactly, nobody can say! But the old quip, “If you’re so smart, how come you aren’t rich?” might come closest to what is often meant.[[718]](#footnote-841)

Extropian discourse owes a heavy debt to the radically individualistic politics of Robert A. Heinlein, who, alongside Arthur C. Clarke and Isaac Asimov, is often regarded as one of the “Big Three” granddaddies of science fiction. Like so many people in tech today, I gobbled him up as a twelve-year-old.

In one memorable novel, Heinlein described a fight for independence from Earth by Lunar colonists—a rugged band of ex-convicts, political exiles, and their free-born descendants; an Australia in the sky.[[719]](#footnote-842) Mike, the colony’s mainframe computer, “awakens” and becomes superintelligent, eagerly aiding the rebels in their fight for freedom. Mike is a loyal and lovable machine, fond of low humor, a far cry from the humorless HAL 9000. The novel is less important for its depiction of AI, though, than as a thinly disguised polemic.

On one hand, Heinlein describes the Moon as a “harsh mistress,” utterly inhospitable to human life. This much is true. On the other hand, he describes Lunar culture as a relentlessly libertarian manosphere. There are no laws, justice is rough and ready (the airlock is “never far away”), and everything—including air—must be bought and paid for, fair and square, with a nod to Ayn Rand:

“If you’re out in field and a cobber needs air, you lend him a bottle and don’t ask cash. But when you’re both back in pressure again, if he won’t pay up, nobody would criticize you if you eliminated him without a judge. But he would pay; air is almost as sacred as women.”

This is the book that immortalized the slogan “There Ain’t No Such Thing As A Free Lunch,” or TANSTAAFL, embraced thereafter by free-market economists and libertarians everywhere.[[720]](#footnote-843)

Transhumanist philosopher Max More,[[721]](#footnote-844) whose 2003 manifesto *Principles of Extropy*[[722]](#footnote-845) kicked off the Extropian movement, enthused about the idea of needing to pay for the air you breathe. Air pollution, per More, is an avoidable tragedy of the commons. The solution is to make it, and everything else, private property. Metering air out for a price would lead to cleaner air—and, perhaps, to a kind of “cleansing” (by suffocation) of those who can’t pay?[[723]](#footnote-846) One can see why such views might be characterized as eugenicist.[[724]](#footnote-847)

What makes such libertarian politics so cognitively dissonant in Heinlein’s hardscrabble Lunar utopia is precisely the inhospitableness of the environment. Survival on the moon is as “urban” as it gets. It would require large numbers of highly specialized humans cooperating intensively to carry out an enormous variety of technical jobs—not to mention myriad plants, animals, microbes, and machines. It’s hard to imagine a Lunar generalist, although of course, the novel’s hero, Mannie or “Man” for short, supposedly is one.[[725]](#footnote-848)

Real life human generalists are nothing like “Man.” They’re more like the Pirahã, who can “walk into the jungle naked, with no tools or weapons, and walk out three days later with baskets of fruit, nuts, and small game.” But this is only possible because the jungle is entirely unlike the Moon. It is full of oxygen, fresh water, food, shelter, materials that can be woven into baskets, and everything else necessary to human life—provided you have learned a suite of skills that can be mastered readily by most people with a few years of apprenticeship. For those in the know, the jungle looks suspiciously like a free lunch, a free dinner, and a free bed and breakfast.[[726]](#footnote-849)

How could one claim that food doesn’t grow on trees in a world where bananas, mangoes, and so many other delicious things literally grow on trees? (Bananas actually grow on bushes, but the point stands.) Seed dispersal within tasty fruits, gas exchange between plants and animals, insect pollination, and the endless other reciprocal relationships that make up a jungle secure the stability of countless species and individuals through the generous provision of “free” stuff. It’s not so much an economy as a complex network of mutual aid—with a healthy admixture of predation and parasitism. Humans themselves evolved within and as active parts of such nonzero-sum systems.

On the Moon, people (and their technologies) would have to provide every one of these “ecosystem services” for each other. The massive capital investments, scale economies, and feedback loops needed would require complex administration and cooperation that look like the very opposite of Heinlein’s Wild West. He *sort of* acknowledges this inconvenient truth by giving his Lunar settlement an origin story as a corporate penal colony, and then by endowing Mike the supercomputer with vast, centralized administrative powers, controlling everything from “pilotless freighters” to phone and video communication, air, water, climate, sewage, transit, payroll, etc. The result nearly wraps around to Cybersyn, Chile’s ill-fated attempt at a cybernetic command economy through centralized computing in the early 1970s.[[727]](#footnote-850) Wait, was this supposed to be a libertarian utopia, or a communist one with a frosting of Westworld cosplay?

Just as cooperation and competition start to look increasingly alike the more you stare at them, this dichotomy, too, may be a mirage. Friendly relationships, transactional relationships, and predatory relationships are all still relationships. Too much free riding, as in the Dark Room scenario, can always become an issue, albeit, ultimately, a self-correcting one. In a healthy—that is, stable—complex ecology, every entity both relies on and provides for other entities, whether it wants to or not. Are you benefiting others? You might think not, until some other critter comes along to eat you. Even a supposed “apex predator” will die eventually, turning into food for other animals and fungi. There *is* no apex in a circular system.

TANSTAAFL’s great simplifying assumption, shared by most economists, is that money, or more precisely, value, can be represented by a single number. It’s only by introducing this universalizing numerical value that the leap can be made from the obviously true—that every entity in a graph of relationships both needs and provides for others—to the dogmas of libertarian economics:

* If you want or need something, it has value.
* If it has value, it can be priced.
* If everything has a price, you need money to buy it.
* If you have money, the amount (income minus expenses) can be tracked on a spreadsheet.
* If you and every other actor are rational, then a free market will produce an optimal outcome.

As will now be apparent, these are also the foundational tenets of the kind of “rationality” critiqued in this book; that’s why the modern “Rationalist” movement, making its home at LessWrong.com, has such a strong overlap with both libertarianism and Extropianism.[[728]](#footnote-851)

## Utility

The roots of Rationalism go back to Jeremy Bentham, and his ideas, like many others from the Enlightenment, were wonderfully progressive for the time. More than that—they represented a grand synthesis of Enlightenment thinking:

* Like Descartes, Bentham believed in a universe governed by mechanical laws.
* Like La Mettrie,[[729]](#footnote-853) he pushed back against religion, believing that people, too, are part of the universe, hence governed by the same mechanical laws as everything else.
* Like Newton, he believed that these laws could be given mathematical form.
* Like Leibniz, he thought it ought to be possible to compute the correct answers to questions algorithmically—and not only, to use Hume’s distinction, “what is,” but “what ought to be.”
* Although he was no ally of the American revolutionaries,[[730]](#footnote-854) he also believed, as they did, in the universality of rights. Indeed, he went quite a bit further, advocating equal rights for women, the right to divorce, and (although this was too risky to publish at the time) the decriminalization of homosexuality.[[731]](#footnote-855)

In the early 1800s, Bentham brought these ideas together in a large fold-out pamphlet entitled, rather wordily as was the custom, “TABLE of the SPRINGS of ACTION : Shewing the several Species of Pleasures and Pains, of which Man’s Nature is susceptible : together with the several Species of INTERESTS, DESIRES, and MOTIVES, respectively corresponding to them […].”[[732]](#footnote-856)

In this table, Bentham sought to begin developing in earnest what he called a “felicific calculus,” whereby everything that causes pleasure or pain could be assigned a positive or negative numerical value. Food, sex, and the fear of death are on there, of course, but so is much else, including the hardship of labor and the pleasure of rest, the seeking of novelty, the joy of friendship, and the love of God, though plenty of religious impulses fall into the negative column—superstition, bigotry, fanaticism, sanctimoniousness, hypocrisy, and religious intolerance. There’s more than a little of Bentham’s subjective judgment here.

Regardless, his use of the phrase “Springs of Action” is a kind of pun. Most obviously, by “Springs” he means sources, as with water, or “wellsprings.” However, it’s also an allusion to the “mechanical philosophy,” which held that people themselves are nothing more than a kind of dynamical mechanism, their psychology driven by motive forces just as a clock’s gears are driven by its springs.

In the same way Newton’s “calculus of fluxions” allowed one to take the derivative of a particle’s observed trajectory to infer the net forces it must be experiencing, Bentham’s felicific calculus sought to derive the “hedonic,” or pleasure-seeking, forces that shape a person’s trajectory through life based on their observable behaviors. Or, equivalently, just as Leibniz’s version of calculus allowed one to integrate known forces to compute a trajectory, Bentham believed that an accurate accounting of hedonic forces, once we understood them, would enable prediction of a person’s actions.

How do morals, ethics, or governance fit into such a picture? For Bentham, given a felicific calculus, the answer is captured in the phrase that has become most associated with him: the greatest good for the greatest number. If people, in other words, act in such a way as to optimize their pleasure, then the role of government is simply to ensure that the summed pleasure of all people is optimized. If, for instance, one person derives an amount of pleasure X by making a hundred others’ lives worse by amount Y, then this would be an immoral act, *unless* X is greater than 100Y. The correct role of government is thus to prevent such selfish negative-sum actions, while encouraging any actions that increase total happiness.

Today, we call this philosophy “Utilitarianism,” and use the word “utility” to denote pleasure (when positive) or pain (when negative). Under certain assumptions, including economically rational actors, a free market will find this optimum.

This all sounds pretty good—certainly better than rule by force, disregard for the welfare of entire classes of people, or arbitrary moral codes based on superstition. Understandably, given that we’re still not free of these historical blights, Utilitarianism continues to attract devotees. Among the most hard-core are modern libertarians, Rationalists, Effective Altruists, and, for reasons that will soon become clear, the AI X-risk community.[[733]](#footnote-857)

The trouble, though, is that Utilitarianism, quite literally, doesn’t add up. Psychological studies show that human preferences don’t always obey the “transitive law,” wherein if the utilities of X, Y, and Z can all be expressed as numerical values, and someone prefers Y over X, and Z over Y, then they *have* to prefer Z over X too. Otherwise, there’s a logical contradiction.

The moment pioneering behavioral economist Amos Tversky showed, in 1969, that people *can* sometimes prefer X over Z,[[734]](#footnote-858) he exploded the foundations of Utilitarianism as a possible way to describe people’s actual behavior. This turned what Bentham had posited as a law of human nature into, at best, a “should” rather than an “is” claim.[[735]](#footnote-859)

That’s why the most rigorous modern Utilitarians tend to a hypervigilant sort of sanctimoniousness, which might not have thrilled Bentham. While recognizing that pretty much everyone “out there” is irrational, they seek, in their own actions, to *be* Utilitarian—hence, rational—obeying the transitive law in their preferences even when it leads to horrors or absurdities. To take Utilitarianism seriously as a moral position implies welcoming a cost-benefit analysis of *any* proposition, no matter how counterintuitive or repugnant. This explains something of the tenor of the old Extropians mailing list.

As a descriptive theory, the trouble with utility isn’t limited to Tversky’s “intransitivity of preferences.” “Additivity,” the simple idea that utility adds up the way numbers should, also poses a serious problem. For example, in one classic series of experiments, patients were told to move a pain dial, numbered zero (no pain) to ten (maximum agony) during a colonoscopy, conducted while they were fully conscious.[[736]](#footnote-860) Half of the patients (apologies, I promise there’s a reason I’m going here), “had a short interval added to the end of their procedure during which the tip of the colonoscope remained in the rectum.” This added interval was still not good, but it was less uncomfortable than what had preceded it. Curiously, these patients found the whole procedure less aversive than those for whom the colonoscopy ended more abruptly. Not only was their overall rating of the experience better in retrospect, but they ranked it more favorably among standardized lists of other aversive experiences, and even had higher rates of followup colonoscopies years later (though this effect was small).

The researchers viewed these findings as “memory failures,” highlighting the way they had internalized Utilitarian assumptions. If pain is supposed to add up, X is the pain involved in the main procedure, and Y is the the pain involved in the extra time when the probe is left in, then clearly X+Y must be greater than X.[[737]](#footnote-861) There are many more findings in this vein.

The point is that no matter how one computes a felicific calculation, it doesn’t match what people actually do. Since spending money, if you have a finite amount of it, represents a series of tradeoffs regarding which actions you take, it shouldn’t be surprising to find that we’re not “rational” economic actors. When we exchange money, we’re not in general passing around happiness, or any kind of proxy for it.

Of course, as our space of available actions shrinks due to poverty, most of us *do* experience negative feelings about it, both because we’re prevented from doing things we want to and because having all of our choices taken away—being cornered—feels bad in its own right, for reasons discussed in Part V.[[738]](#footnote-862) Going hungry, or being exposed to the elements, feels even worse. We also care about social standing relative to our peers. So, there is a rough correlation between wealth and happiness, especially at the poor end of the scale, but the relationship is by no means linear.[[739]](#footnote-863) Having two million dollars won’t make you twice as happy as one million; there may, in fact, be very little difference, psychologically. By contrast, there’s a world of difference between having zero and a million.

Moreover, while most of us wish we had more money, our day-to-day actions are not in general done to increase either our wealth or any other obvious quantity. The closest thing to an exception would be someone who works in finance and is obsessive about their “score” at that game; they live to play it, just as Lee Sedol, prior to his retirement in 2019, lived to play Go.[[740]](#footnote-864) As you might imagine, Utilitarian thinking is popular on Wall Street.

## Big tent

Utilitarianism is far from a purely right-wing position. Some staunch adherents, most famously the philosopher Peter Singer, extend their felicific calculus to nonhuman species. Singer is mostly vegan, since he cares about animal suffering as well as human suffering. He popularized the term “speciesism” to decry those who ignore the suffering of nonhumans, although flattening distinctions between species creates some challenges of its own.

What are we to make, asked Singer in 1973, of “genuine conflicts of interest like rats biting slum children,” given that “rats have interests too”? His tentative suggestion: rodent birth control instead of traps or poison.[[741]](#footnote-866) This suggestion was, at a minimum, tone deaf; it made some question Singer’s great powers of empathy. As Xenia Cherkaev, a Fellow at Harvard’s Davis Center, has pointed out,

“In 1973, the image of ‘rats biting slum children’ was more than an idle abstraction. Poor American children were indeed being eaten by rats. Predominantly, those children were Black.”[[742]](#footnote-867)

As this depressing scenario makes clear, we have to acknowledge that the network of relationships we care about includes nonhuman actors, whether we like it or not, but that doesn’t mean those actors are all equal. They come in all sizes and shapes, and it’s this very fact that makes universal participation in any single economy or felicific calculation impossible. One can’t ignore one’s own place in it, either.

If the graph of relationships were finite and “flat,” containing, say, only a hundred villagers who seek to trade handicrafts and vegetables with each other, then money might work pretty well for optimizing the flow of resources, though it still wouldn’t be a good proxy for happiness. Likewise, deliberation and voting might work pretty well for coordinating collective action.

However, when the graph includes not only human beings, but also single cells, tree frogs, corporations, banana plants, jumping genes, jumping spiders, rat families, trade unions, nations, rivers, and burial grounds, it becomes awfully hard to understand how these interacting entities are all supposed to make decisions, pay each other for stuff, and be held accountable for debts or obligations using anything like an economic model. There is no universal currency, and no “view from nowhere.”

For instance, for all this libertarian talk of “buying” and “metering” air, who is actually the producer? Who can be said to own it? Presumably, in no small part, the *Prochlorococcus* cyanobacteria that inhabit the Earth’s oceans and synthesize a good deal of the oxygen we breathe. Do we mint NFTs for them? Issue them with voting shares in AirCorp? Are they even distinct, or more like a superorganism?

Suppose humans were to make the ill-advised decision to “enclose” both the oceans and all of these smaller entities to enable universal rent-seeking by legal persons (which today includes people, corporations, and nations). Then, the problem of taxonomizing these “assets,” weighting “voting interests,” and tracking “value flows” would look like a combination of solving GOFAI and simulating the whole planet—all for the purpose of bean counting, in a world that seemed to be getting along just fine before we decided it could be improved with a Spreadsheet of All That Is.[[743]](#footnote-868)

Everything we today consider labor or capital, value or worth, joy or suffering would be a drop in the bucket alongside that fractal leviathan, the Earth. Our whole planet, with all its interlinked systems that have evolved over four and a half billion years, is a decentralized version of Mike the supercomputer—though a version that evolved rather than being “rationally” designed. “Mike” is what economists call “externalities,” which actually turn out to be most of the story. “Mike” contains a zillion interacting entities who experience a vast array of pleasures and pains from moment to moment in the service of multilevel dynamic stability. “Mike” provides the fresh water; we need only transport it. “Mike” provides the bananas; our “labor” merely involves picking them. “Mike” provides self-reproducing ruminants and self-growing grass; we put a fence around them and call them “property,” either privatized or collectivized. We get to pretend we’re the producers of value, and play our economic games, whether communist, capitalist, or radically libertarian, only by the grace of “Mike.”

But if we peer inside “Mike” as a whole, or *Prochlorococcus*, or ourselves, we won’t find anything resembling a single value being maximized. This is just a restatement, yet again, of Patricia Churchland’s insightful critique of AlphaGo: outside the self-contained toy world of a game, purposive entities have “competing values and competing opportunities, as well as trade-offs and priorities.” There is no cumulative score, and no goal, other than to keep playing.

As we’ve seen, when we try to impose a score on this infinite game—insisting that, for every player, each move comes with a quantifiable cost or benefit that can be tracked over time—we run into mathematical trouble, regardless of how those costs and benefits are computed.

To understand at a deeper level why this is so, imagine that you are a little creature, like an ant. There are so many ways for you to die: running out of energy, getting eaten, being exposed to a noxious chemical, becoming too hot or too cold, drying out, getting lethally irradiated, and on and on. The reason you have muscles is to move your body away from such fates so that you can keep playing, and the reason you have sensors is to inform that motor behavior as best you can, using a predictive model.

Your hidden states H are there to serve as the information bottleneck between sensing and acting, so evolution will have endowed you with things like discomfort, meaning “run away,” when conditions look dodgy. You’ll also have desire, meaning “yum, food, stick around and keep eating.” You can only be a rational economic actor if your actions are based on maximizing your “value,” which has to be a function of those hidden states H somehow. Perhaps value is something like pleasure minus pain. You must be able to predict what effects your actions will have on this value; in the language of calculus, value must be “differentiable” with respect to action, such that your actions always move you uphill in the value landscape.

So far, so good: value, for you, will look like a tent, pegged down to the ground (zero, for death) along a perimeter of boundary conditions corresponding to starvation, freezing, overheating, and so on. In the abstract space of value, you’re crawling around on that tent, and will only stay alive as long as you don’t touch the ground.

Of course the tent doesn’t just stay put. As you burn energy and your environment changes, the tent gradually changes shape. You have to keep moving because, if you stop for too long, the place where you’re standing will eventually touch the ground and you’ll die. You can’t move in just one direction, either. If you perseverate and, for instance, keep eating long after you’re satiated, you’ll also die, perhaps by bursting. Every path uphill, in other words, goes downhill on the other side of the tent. There is always a maximum amount of “yum,” such that any more will start turning it into “yuck.”

Just where is that maximum, though? While the sides of the tent might be steeply sloped, the top needn’t be. In fact, this tent is largely an illusory, or at best, an arbitrary construct. Its *perimeter* is quite real—you *will* die if you touch it—but otherwise, where you wander, and whether you think you’re going uphill or downhill, is underdetermined. We could describe the way you choose to wander around as relating to your “personality”; that’s where you get to exercise free will.

The better your predictive model, the more clearly you can resolve exactly where the edges are, which actually lets you wander *closer* to them than you otherwise could, while still staying safe. That’s what daredevils, people who engage in BDSM, and circus performers do. You’re free to do all sorts of things on that tent, and as long as you reliably avoid hitting the ground, evolution has very little to say about what you do or how you feel about it. Anything goes, as long as it doesn’t kill you too often. The smarter you are, the bigger the tent, and the greater your freedom.

When I describe the tent as being mostly underdetermined, I mean the following. Any talk of a value function only makes sense if the ant actually optimizes for it by climbing upward; otherwise, this “value function” means nothing—it adds no explanatory power. It’s equally useless if we claim that the tent changes shape so quickly that whichever way the ant happens to be going *is* uphill at that moment, even if it may not be at the next moment. That would be circular, equivalent to saying “I always do what is best,” and defining “the best” as “whatever I do.”

So, *assuming* a stable value function, how do we reconstruct it from the paths the ant is observed to take? The mathematical answer is “integration,” but never mind the details. The point is that if the ant always goes uphill, it should always end up at the (or a) highest point. It should *never* go in a loop, because that would imply something like an Escher staircase—going uphill, yet paradoxically ending up right where you started. It can’t happen.

But it *does* happen. That’s exactly what results like the intransitivity of preferences show. It doesn’t matter how complex or high-dimensional the space of available actions is, or how complex and nuanced the value function is; if value is a number, and you maximize it, you can never travel in a loop.[[744]](#footnote-869) *Ergo*, there *is* no consistent value function that you are continually optimizing. And that’s not only true for people, but for ants, bacteria, corporations, and everything else that evolves.

We ought to have known that reducing decision making to the optimization of a single value would never work, though, because even the simplest organisms need more than one internal signal to make viable behavioral decisions. You can’t just subtract serotonin from dopamine, for instance, and preserve the information you need to stay alive. Food is great—unless you’re full. Resting is great—unless you’re too hungry. Sex is great—unless you’re too hungry, *or* too full. These things simply aren’t substitutable. A worm or a bacterium knows as much.

Even corporations, the supposed epitomes of economic optimization, have begun to acknowledge that maybe a single value (like stock price) can’t be the sole guide to behavior, with the introduction of “Environmental, Social, and Governance” (ESG) metrics for investing, and the adoption by some of a “Triple Bottom Line,” including social and environmental, as well as financial, accounting. These are still not mainstream practices, and turning them into accounting is bogus in any case (they don’t obey transitivity or additivity either).

But the truth is that corporations have never behaved as purely “rational” economic actors, whether that’s reflected in their accounting or not. Their hunger for money may be bottomless, but like every other entity, they have evolved to survive. So, they also (usually) follow all sorts of norms and rules, and spend resources on all sorts of stuff that isn’t about making money, but about preserving important relationships and thereby continuing to exist in the future.

A traditional economist might argue that given predictive capability and a long view, optimizing for making money *automatically* means optimizing for survival, and everything instrumental to survival. After all, if I want my widget-manufacturing startup to become a billion dollar company, it has to survive and grow until it reaches that size… and if it runs out of money at any point, it will die. So whatever I do in the service of the company’s survival is really just optimizing for money in the long term.

There are two problems with this line of thinking. The first is that running out of money is only one way a company can die. It can also cheat too much and get caught, like Enron, or get burned down by disgruntled townsfolk, like the Westhoughton Mill in Lancashire during the Luddite revolt. (Both of these businesses, incidentally, were unusually single-minded in optimizing for profit.)

At bottom, the only thing all “successful” businesses have in common is that they continue to exist. That is dynamic kinetic stability. They generally need an influx of money to keep running, just as we need an influx of food to keep running, but it’s no more accurate to say they all exist to maximize money than it is to say that our purpose is to maximize the amount of food we eat—or, for that matter, how much comes out the other end.

Similarly, if a company survives a long time, we could equivalently claim that it’s optimizing for *anything* involved in its “metabolism.” Inverse Reinforcement Learning (IRL), a family of machine learning methods for reverse-engineering a value function from observations of actions taken,[[745]](#footnote-870) would have trouble making any distinction. If we pick an arbitrary, but very successful long-lasting company, we could make a case that it has optimized over the long run for how many people it employs (a huge national chain), how many lives it shortens (a cigarette company), how much time it wastes (a casual game company), or a myriad other goals.

My widget company could even be all about maximizing the number of complaints I hear from my customers. After all, the more cheaply produced, semi-crappy products I can move while still managing to expand my customer base, the more complaints will roll in. Before you rush to see if this business idea has been patented, be warned that it seems to already be popular.[[746]](#footnote-871)

## Sims

No matter how a value function is defined, Utilitarianism stipulates that this value should grow forever and without bound. Hence, “Max More.”

To anyone who studies neuroscience or biology, the idea that *any* parameter in a living system can grow without bound is self-evidently absurd. It’s especially absurd in a system whose dynamics are exponential, which is how dynamical systems generally work.

Once an exponential takes off, it grows *very* fast, and if it represents anything correlated with the real world, it will in short order run into ecological, physical, and ultimately, cosmological limits. Hence, talk of a “singularity”—a point in time beyond which an exponential prediction yields something literally impossible.[[747]](#footnote-873) Often, by analogy with the event horizon of a black hole, this predicted singularity is understood in nearly mystical terms, as a dark barrier we are hurtling towards, on the other side of which we’ll experience something unknowable and awesome. Armageddon? Nirvana?

There’s a much more prosaic way to think about exponential predictions, though: in real life, they simply can’t go on forever. They saturate, with initial exponential acceleration first becoming more linear, then turning into exponential *de*celeration. The value may asymptotically approach a maximum, or go back down, or oscillate.

World population, for instance, has been in exponential growth for ten thousand years, with the exponent itself leaping upward at several noteworthy points—most recently, around 1945, with the baby boom. This is the mother of all exponentials, the Moore’s Law of our species.

Yet there was, and is, clearly a limit. At the rate we were going, *something* had to give, and it had to give around now. Otherwise, in just another few centuries, human bodies would blot out the Earth’s entire surface, and not long afterward, stack up to fill the entire biosphere. We could begin populating the rest of the Solar system, but it’s hard to imagine a scenario in which this meaningfully relieves population pressure on Earth, or even offers a comparably sized niche. There just aren’t that many places in the Sun’s neighborhood where we can live, and interstellar travel would be very, very slow.

So, in the mid-twentieth century, we were facing a “population singularity.” We didn’t fantasize about any magical, unforeseeable singularitarian result, though. Instead, we foresaw, sensibly enough, the inevitable end to exponential growth, either the hard way (a massive dieoff) or the gentle way (a sharp enough decline in fertility to avoid overshooting hard limits). Luckily, we’re now tending more toward the latter, though we’re not entirely out of the woods yet.[[748]](#footnote-874)

Before Nick Bostrom began to focus on AI X-risk, he had already made a name for himself among philosophers (serious and armchair) by advancing the “Simulation Hypothesis,” mentioned briefly in Part V. The Simulation Hypothesis holds that, in all probability, we’re living in a computer simulation, as in the *Matrix* movies. This is relevant to exponential growth-worshiping fans of the Singularity because virtual worlds could, in principle, support astronomically large virtual populations.

In brief, the Simulation Hypothesis argument goes like this. Over the past several decades, we’ve become capable of simulating increasingly sophisticated virtual worlds—from giant cosmological simulations, to photorealistic shooter games, to faithful models of neural circuits run by computational neuroscientists. Moore’s Law has been rendering these simulations exponentially faster and more detailed over time.

There are already huge numbers of such simulations running on Earth, whether for basic research, entertainment, urban planning, or a number of other applications. This is interesting in its own right, when we think about how our own brains became so much more powerful when they were able to carry out predictive simulation of counterfactuals;[[749]](#footnote-875) all these simulations we’re running today look like the same sort of cognitive upgrade, but now at the level of our *collective* intelligence!

The point, though, is that if Moore’s Law continues, then before long, the realism (or at least the complexity) of all these simulations might rival that of the real world. Some simulations already include virtual people—in video games, they’re called Non-Player Characters (NPCs). Today, they’re cartoonish, and are wired up to handwritten scripts or AI models like digital marionettes.[[750]](#footnote-876) If you try to do in-game electrophysiology on the heads of NPCs in Fortnite, you won’t find any neurons in there, just the back sides of textured polygons. However, in a very distant descendant of the bff universe, intelligence could evolve, with or without our help. If so, then given exponentially increasing computation, there will be an exponentially increasing number of intelligent agents living in these simulated worlds.

If you find the “bff on steroids” version of this scenario a stretch, keep in mind that we see extraordinary realism emerge when we use neural nets to *learn* simulations of the world, from the weather and fluid dynamics to sound and vision. (That’s what deepfakes are.) It seems a safe bet that before too long we’ll see convincing interactive worlds conjured up from little more than a prompt, ending the GOFAI-like era of today’s laboriously engineered video games and simulations.

Despite the hundreds of millions of dollars they cost to develop, the polygonal puppetry of games like Fortnite will soon seem as crude as Pac-Man. And there’s no reason neural world simulations couldn’t include virtual AIs implemented using virtual neural nets. Dissection, electrophysiology, and physics experiments *could* work in-game, though if those artificial creatures were to probe their world ever more finely, they might encounter a kind of pixelation… like quantum effects? Hmmm.

At this point in the argument, proponents of the Simulation Hypothesis invite us to take the “Copernican Turn,” named after that moment when Copernicus realized that the Earth is just another planet.[[751]](#footnote-877) Copernican Turns have occurred repeatedly in astronomy, as when we realized that other planets have moons too, that Sol is just another star, and that the Milky Way is just another galaxy. The bigger lesson is that there’s nothing special about where we live; hence the saying that “the Universe doesn’t revolve around us.”

Why, then, would we assume that *our* Universe has the unique property of being “real,” since the jillions of other ones we know about are all simulations (as we well know, since we’re the simulators)? If intelligent beings in a “parent” universe can simulate a great many “child” universes, and each of *these* universes feels perfectly real to its inhabitants, then a randomly selected intelligent being (like you) is exceedingly *un*likely to happen to live in the “real world” at the root of this tree of universes—if, indeed, it even *has* a root.[[752]](#footnote-878)

Detailed arguments *pro* and *con* the Simulation Hypothesis are beyond our scope here. Its relevance to us lies less in its disconcerting claim about whether we have a parent universe (and are thus ourselves something like AIs?) than in its less controversial implications about the child universes we can create, and, since 1945, *have* been creating at an exponentially growing rate—for in a sense, *any* computational environment is a child universe with its own rules, and soon, its own digital inhabitants.

A program like the one Ada Lovelace wrote to compute Bernoulli numbers creates a tiny, trivial universe whose “physics” does nothing other than to produce a sequence of digits. A more complicated program, like the one written in 1945 to simulate the hydrogen bomb, models a physical universe in which a single runaway process unfolds, a sort of Big Bang cartoon. Massively multiplayer games are child universes full of digital puppets whose bodies are controlled by “gods” (us) in the parent universe. Now, there are starting to be real agents in those virtual worlds. Every instance of ChatGPT can be considered a child universe containing a single AI, which communicates with a “god” in the parent universe through its context window.

Futurists like Ray Kurzweil (who popularized the idea of the Singularity), as well as plenty of sci-fi writers, have fantasized about how we could transmigrate into our child universes by scanning our brains, then running them in simulation. It’s not clear whether those “uploaded minds” would be able to instantaneously learn kung fu, but we could speed up or slow down our subjective sense of time, copy ourselves, change how we look, (maybe) live forever-ish,[[753]](#footnote-879) and perform various other cool digital tricks.

Unfortunately, mind uploading is not realistic, or at least, not for a long while.[[754]](#footnote-880) And I wouldn’t recommend beta testing the procedure unless you’re on your deathbed, as your parent universe brain would not survive the scanning process.

What *is* entirely realistic—what we are already starting to do today—is to create new intelligences within child universes. I’ve argued that most of these are, in a broad sense, human intelligences, as they are pretrained on collective human experience; they have also, through fine-tuning for dialog, been trained to pass the Turing Test. Whether or not we regard them as “people” or “mind children,”[[755]](#footnote-881) they are intelligent entities, and their population can and will grow *much* larger (and more quickly) than the human population.

Not that “population” is necessarily a concept that will translate straightforwardly into the digital realm. Models come in every size, from tiny to vast; they can be copied and forked, run briefly or for a long time, act as independent entities or link tightly into a single larger entity. This is how life works in general, of course, but much about human sociality, law, and political economy relies on flat relationship graphs, in which the entities are all of a uniform kind, presumed to be “created equal.”

This has already posed problems, given the rise of enormous, powerful corporations with person-like rights and legal statuses, the animal liberation movement, and more recent legal battles waged on behalf of rocks and rivers.[[756]](#footnote-882) Great numbers of variously sized AI entities living in virtual worlds, many of which communicate with our own via chat windows, cameras, microphones, screens, headsets, and cloud services, are about to complicate life in the multiverse quite a bit further, though.

At least we now have a much better sense of what lies beyond the “population singularity” predicted in the mid-twentieth century. Catastrophists at the time predicted a “population bomb” that would bring an ugly end to human civilization, resulting in a descent into environmental collapse and, literally, cannibalism.[[757]](#footnote-883) That will probably not happen, as our population—at least in *this* universe—is set to peak, then begin to decline this century. An end to population growth has far-reaching consequences, given that, as discussed above, our entire economy has been built around the assumption of continual growth.

On the other hand, an increasing share of our economy is shifting into the realm of information work, which can move seamlessly between digitally connected universes; and we now know that the population of intelligences in our child universes is about to explode orders of magnitude faster than any baby boom. If, a few decades from now, we were to plot the logarithm of the total population of combined humans and AIs over time (or something roughly equivalent to population), we would see another kink in the curve beginning around 2023, similar to the one in 1945, though more dramatic still. These kinks correspond to metasystem transitions, of course.

## Tears of joy

Having explored the problems raised by the “greatest good” part of Bentham’s “greatest good for the greatest number” calculus, it’s now time for us to delve into the even thornier problems raised by the “greatest number” part. We’ve seen that, for an individual, “good” both doesn’t strictly exist and isn’t additive, but you might suppose that the addition of something approximating “good” *across* individuals still makes some sense as a moral heuristic. Spoiler: it doesn’t.

Recall that although twentieth century “population bomb” anxieties turned out to be a false alarm, it wasn’t because infinite growth wouldn’t be a problem, but because increasing wealth reliably leads to lower fertility. Draconian measures like China’s “One Child Policy” were both cruel and unnecessary. As China grew richer, its fertility declined so dramatically that it now faces a population crunch. In Japan, the population is shrinking so rapidly (especially given high barriers to immigration) that by 2040 the amount of unclaimed, vacated land is projected to exceed the area of Ireland![[758]](#footnote-885) It’s likely that this is a preview of what the future will look like globally—which is great news for our long-term survival, as assuming we don’t ruin everything in the meantime, it will eventually lower our footprint enough to allow for the regrowth of stressed ecosystems, even as we maintain a high living standard.

There’s probably no single “right population size” for Earth, but rather a wide range, anywhere between too small for human diversity and stability to too large for a decent quality of life and *non*-human diversity, even given advanced and highly efficient technology. All we know for sure is that there *are* lower and upper bounds, making perpetual exponential growth (or decline) incompatible with dynamic stability. For long-term human survival, population on Earth must either stabilize or fluctuate within those bounds.

Is an Earth with ten billion people necessarily ten times better than an Earth with one billion, though? Or, to give this a non-human gloss, is the “value” of domestic cats (worldwide population, roughly 220 million) really 73,000 times the “value” of snow leopards (worldwide population, roughly 3,000)? Or else, if we say that the two *species* have equal value, do we really think the life of each snow leopard is “worth” the lives of 73,000 cats?

Utilitarian philosophers like to set up morally agonizing “trolley problems”[[759]](#footnote-886) to try and get actuarial about questions like these (if you pull the lever, you can derail the train, sacrificing only *one* thousand kittens to save the snow leopard tied to the tracks…), but I think there’s something wrong with the entire premise of trying to rank-order value in this way.

Psychological evidence shows that it’s easy to set up trolley problems that violate the transitive law. In fact, trolley problems have become famous for exposing human “irrationality” in even more blatant ways than Tversky’s intransitivity tricks. In the most classic version of the trolley problem, there are five people tied to the track in front of the trolley, and one standing on a side track. You can either do nothing, in which case five people will die, or pull the lever, in which case the trolley will be diverted onto the side track and one person will die.

The felicific calculus is clear: you should pull the lever, because the alternative is five times worse. Real people aren’t so straightforward, though. Their responses are all over the place, and highly sensitive to, from a Utilitarian’s perspective, irrelevant details. For instance, if it’s necessary to push an overweight bystander onto the tracks to save the other five innocents, almost nobody will do it.[[760]](#footnote-887) For a Utilitarian, this amounts to a moral failing.

If we consider “good” to add up across individuals, the easiest way to increase goodness is simply to increase population. By such a measure, since there are about eight billion people today and there were one billion in 1800, the world today is eight times as “good” as it was then. By the same calculus, India is six times as “good” as Pakistan, and the US is five times as “good” as the UK.

It gets worse. “Longtermists” argue that *future* people should count equally,[[761]](#footnote-888) which would imply that the life of an average young French woman, who, statistically speaking, will have 1.83 children, is “worth” 40% more than the life of an average Japanese woman, who will have only 1.3 children. A Nigerian woman will be “worth” far more than both combined, as the fertility rate in Nigeria is 5.24 (though rapidly falling as prosperity there increases). The differences are in fact exponentially greater, since this only takes *one* generation into account, and according to Longtermists we should consider all descendants equally. So if we survive long enough, the long term “value” ratio becomes infinite, even if the birth ratios are only infinitesimally different.

And it gets sillier. Near the beginning of *Superintelligence*, Bostrom seeks to use felicific calculus to emphasize just how much is at stake when we gamble with humanity’s future. In a box entitled “How big is the cosmic endowment?” he describes a scenario that, upon first reading, I assumed to be comedy circuit material, or a dystopian straw man. However, I *think* it’s intended to be a Utilitarian’s utopia![[762]](#footnote-889)

Here’s how it goes. Begin with the assumption that we will all upload our brains, because of course that’s the way to go—more of us can live as Sims in a virtual world than in the real world (or whatever this universe is), which means more potential happiness. I guess that, to mollify animal liberation Singularitarians, we should make room for our pets, livestock, and the remaining wild animals to get uploaded too. Then, we should proceed to expand the datacenters so that as many virtual humans can be packed in as possible.

This “utopian” scenario doesn’t just imply tiling the Earth with datacenters and eliminating all other, less space-efficient forms of life, but turning the entire Solar system into a giant sun-powered datacenter, disassembling the planets and asteroids as needed.[[763]](#footnote-890) Also, sending out nanotechnological “von Neumann probes” at close to the speed of light—little spaceship-factories than can build *more* cosmic datacenters wherever they find the matter and energy to do so, and in turn send out *more* von Neumann probes, ultimately colonizing “a large part of our future light cone.” Assuming these probes don’t run into competing projects from alien civilizations, to do anything less would be… infinitely immoral!

Multiplying all of the big numbers together and carrying out the colonization scheme of Life, the Universe, and Everything until “cosmic expansion puts further acquisitions forever out of reach,” Bostrom concludes,

“what hangs in the balance is at least 10,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000 human lives (though the true number is probably larger). If we represent all the happiness experienced during one entire such life with a single teardrop of joy, then the happiness of these souls could fill and refill the Earth’s oceans every second, and keep doing so for a hundred billion billion millennia. It is really important that we make sure these truly are tears of joy.”

When one weighs those potential cosmic oceans of joy against the interests of the paltry few billion people actually alive today, it becomes possible to justify any means to drive growth, no matter how ugly it looks in the present. You may be starting to see why this suite of beliefs, congealed into an all-encompassing ideology, appeals to some in Silicon Valley:[[764]](#footnote-891)

* **Immortality**. To a Californian subculture that has long been keen on fitness regimens, dietary supplements, and “biohacking,” the idea that digital immortality awaits if we can just hold out long enough is appealing. Exponentially improving technology might mean we don’t even have to wait that long, despite how far-fetched brain uploading seems today. On the other hand, those who die prematurely can always freeze their bodies in liquid nitrogen—or, for a lower price, just their heads.[[765]](#footnote-892)
* **Video games**. Living in one while trying out different bodies, insta-learning kung fu moves, and embarking on thousand year journeys to other star systems sounds awesome.
* **Staying on top**. Tech people tend to think of themselves as elite thanks to their superior intelligence, making the emergence of highly capable AI models is an uncomfortably exciting prospect. If the fear of not necessarily being smarter than those models predominates, it leads down the AI doomer path. On the other hand, digitally *merging* with AI models could keep one’s future self at the top of the pecking order.
* **Scaling**. Moore’s Law, which has held for many dizzying decades, has normalized the idea of eternally exponential scaling. Techies have long worshiped the exponential, but the meteoric rise of online services starting around 2000, and then of cloud computing in the 2010s, has made super-scaling a business mantra and ideology too. Conversely, ideas that turn out not to be viable are said to be ones that “don’t scale.”[[766]](#footnote-893) Turning the entire light cone into a cosmic datacenter that runs the *Truman Show* for astronomical numbers of people is scaling taken to its ultimate physical limit.
* **Wealth**. Some Effective Altruists, including Peter Singer, emphasize giving, and donate much of their income to causes carefully evaluated for impact.[[767]](#footnote-894) It’s hard to argue that is a bad thing, even if one’s own choice of causes differs. Others, like disgraced cryptocurrency billionaire Sam Bankman-Fried, have emphasized (whether cynically or not) the accumulation of great wealth first, to allow for “higher impact” giving later on. The moral hazard here is obvious, as one can always tell oneself that one is still in the “wealth growing” phase of the plan.[[768]](#footnote-895)

Although Utilitarianism and its associated movements aren’t exactly a religion, there are some undeniable parallels. As mentioned, the impossibility of Bentham’s original ambition to develop a descriptive theory of the “Springs of Action” has led Utilitarians to drop any pretense of a scientific “is” to embrace a normative “ought.” When it comes to those “oughts,” counterintuitive value judgments, and the belief that those who disagree are confused and morally inferior, creates a clubby vibe. Effective Altruism also shares more than a bit in common with the Prosperity Gospel, which holds that for true believers, physical health and financial wealth are the will of God. Finally, the promise of immortality sounds more than a little familiar, especially when it comes in two flavors: a heavenly Rapture of the Nerds[[769]](#footnote-896) or, if the coming AI is unfriendly, an End of Days… or, worse still, an unimaginable number of immaterial souls in agony whose tears, of the wrong sort, could “fill and refill the Earth’s oceans every second, and keep doing so for a hundred billion billion millennia.”

## Beyond alignment

Solving the “AI Alignment Problem”[[770]](#footnote-898) means making sure that, by the time really intelligent autonomous AI agents come along, their values are consistent with ours. For believers in the Singularity, aligning AI is the key to us all ending up in heaven, rather than extinct, enslaved, or in some simulated (but all too real) digital hell.

One does not have to believe in the Singularity to consider this problem important. However, because Utilitarian thinking has distorted the way we think about intelligence itself—and value—it has also had an unfortunate influence on how we think about value alignment.

In the earlier GOFAI period, the presumption had been that intelligence would arise in sufficiently complex rule-based systems, so alignment was also envisioned in terms of rules. This corresponds to “deontology,” the ancient philosophical tradition holding that rules can distinguish right from wrong: “Thou shalt not lie,” “Thou shalt not steal,” and “Thou shalt not kill,” for example.

Accordingly, most twentieth century anxieties about AI alignment concerned whether such rules could ever suffice to ensure safe and friendly robots. Sci-fi granddaddy Isaac Asimov famously explored this question in his *I, Robot* stories,[[771]](#footnote-899) starting with the premise that his “Three Laws of Robotics” should be programmed into every robot:

1. A robot may not injure a human being or, through inaction, allow a human being to come to harm.
2. A robot must obey the orders given it by human beings except where such orders would conflict with the First Law.
3. A robot must protect its own existence as long as such protection does not conflict with the First or Second Law.

Of course, in Asimov’s stories as in all sci-fi, trouble ensues—otherwise known as a plot. Here, the trouble is *lawyerly*—meaning that some combination of an unusual situation and logical yet counterintuitive reasoning based on the Laws leads a hyper-rational robot to do something surprising, and not necessarily in a good way.[[772]](#footnote-900)

The reader may be left wondering whether the issue could be “debugged” by adding one more Law, or closing a loophole—something Asimov himself undertook on several occasions. But as GOFAI’s failures eventually made clear, the problem is unsolvable; one can’t even get anything resembling competent behavior out of rules, let alone ethical behavior. As it turns out, rules can only be obeyed by an intelligent agent that does *not* obey rules!

Recall that by the 2000s, state-of-the-art AI was no longer being programmed, but rather trained through the maximization of value functions. Mostly, this involved supervised learning, which is all about optimizing a task-specific score. Starting in 2010, DeepMind, the outlier, focused instead on game playing using Reinforcement Learning, in the belief that it would ultimately lead to AGI. This approach, too, involved maximizing something, whether a chess rating or the score of an Atari video game.

During this period, AI researchers seemed to be converging on the idea that intelligence really is all about optimizing a value or utility function. I’ve argued that this isn’t the case—and indeed, that we only began to see general intelligence when we *stopped* training models to do specific tasks with supervised learning, and switched to the unsupervised regime—but this isn’t (yet) a widely held view. As of 2024, the Rationalist website LessWrong.com explains on its Wiki, “The term artificial intelligence can have anthropomorphic connotations. In some contexts, it might be useful to speak of a **really powerful optimization process** rather than a *superintelligence*.”[[773]](#footnote-901)

A really powerful optimization process able to do pursue any strategy to maximize its value function—what could possibly go wrong? A better question would be: how could it possibly turn out well? As noted earlier, no parameter in any dynamically kinetically stable system, whether biological, ecological, or technological, can grow without bound, and attempting to do so will indeed result in floods of tears.

The classic X-risk example is the innocuous-sounding “Paperclip Maximizer.”[[774]](#footnote-902) Nick Bostrom offered up this scenario in 2003: an AI, perhaps connected to a paperclip factory, is instructed to manufacture as many paperclips as possible. After reconfiguring the factory to increase its throughput (so far, so good), the AI realizes that more factories could be built to further increase production. If the AI is smart enough to engineer new paperclip technology, it can, presumably, also figure out how to make lots of money, convince people to do things, engineer new tech, and otherwise take over the world, if necessary, by improving *itself*. Bad news: human bodies are full of atoms that could become paperclips. But even if the AI stuck to the traditional stainless steel kind, covering the Earth (and then filling light cone) with robotic mining operations and paperclip factories won’t leave any space for snow leopards, tree frogs, or people. If we tried to interfere with this process, it would of course become necessary for the AI to eliminate us with extreme prejudice. It would be nothing personal.

This story unspools badly no matter *what* you tell the superintelligent AI to maximize.

Bostrom acknowledges that

“In humans, with our complicated evolved mental ecology of state-dependent competing drives, desires, plans, and ideals, there is often no obvious way to identify what our top goal is; we might not even have one.”

True. His conclusion, though, is that since a superintelligence “may be structured differently,” we *must* give it a “definite, declarative goal-structure with a clearly identified top goal,” in the tradition of Asimov’s Laws of Robotics, in order to be sure that it won’t go rogue.

Not only does requiring AI to act like GOFAI not work; if it did, it would open the door to all sorts of lawyerly horrors. *Superintelligence* is full of fanciful examples. For instance, to optimize total human happiness, human simulations could be simplified so more of them could run on any given hardware. Then, those lobotomized Sims in their low polygon count environments could be kept on a constant drip of virtual heroin. The greatest good *and* the greatest number, for the felicific win!

In the end, the Alignment Problem is terrifying to modern Utilitarians because they believe that intelligence itself is Utilitarian. They are afraid, in other words, that superintelligent AI will behave *as they claim they would*.[[775]](#footnote-903) If their supposed goal as “rational” actors is to fill up the light cone with simulated people, disassembling every animal, vegetable, or planet whose atoms could be used to build the cosmic server farm, then of course a really smart AI would do the same. Minus the people.

Utilitarian games tend to be zero-sum. Worse, imagining that they can be “won” by a single actor implies a “one-hot solution”: an inexorable tendency toward monoculture, such that a single kind of entity, whether paperclips, simulated people, or robots, crowds everything else out. Anything that is not the goal amounts to an opportunity cost.

Moving beyond the “Alignment Problem” as usually understood isn’t hard. It has both a descriptive “is” and a normative “should” aspect. The descriptive aspect follows from this book’s argument: intelligence is not the maximization of any single value. Nor can an intelligent agent even be said to consist of a single well-defined entity. The interaction of diverse actors through mutual modeling is what creates the dynamical process we call “intelligence.” Intelligence is an ecology, and the more of a monoculture it becomes, the less intelligent it will be, and the less interesting life will ge.

As a thought experiment, imagine creating an exact copy of the Earth in simulation, and running it on a giant, isolated server farm (perhaps we’ve disassembled Mercury and turned it into a Death Star-sized solar-powered computer to make this possible). You and I would both be mirrored in that digital twin universe, and virtual you would be reading these very words at this very moment. Has total utility just doubled, since now every good thing is experienced both by you *and* by your double? I think the answer is “no.” You wouldn’t be able to tell whether you’re the “real” you or the simulated you, but nothing about your life would change either way, and nothing about the universe (or multiverse?) would be any more wonderful or interesting—from *any* point of view. If anything, life would be strictly worse, since both in reality and in simulation, Mercury would be gone. Even if that planet is a lot less interesting than Earth, its absence would be a loss.

The idea of extending that loss throughout the universe in order to endlessly replicate what we already have is nightmarish. There are obvious parallels to point out here with the way colonial powers have attempted to erase indigenous cultures in order to endlessly replicate their own, and with the way capitalism’s worst excesses push out diversity in the name of “hyper-scaling.” As we mature and our wisdom increases, so will our regret over such losses. Life only continues to enrich as long as it continues to diversify, and the collective phenomenon of intelligence only grows when diverse sub-intelligences model each other to interrelate as a greater whole.

Monoculture is not scaling; it’s collapse, an ultimate *failure* to scale. When we destroy variety, we zero out the value of encounter, we render mutual modeling pointless, and we thus curtail the possibilities for our own continued development. Embracing our continuing evolution, however, means letting go of certainty, and allowing identity to drift and branch.

The word “we” will not mean the same thing it does today in a century, and even less so in ten thousand years; less still in ten million. In ten million years, there will be no recognizable humans, though in ten thousand—if we don’t mess up—there probably will be. But there will be much else, too. In as little as a hundred years, the Earth, and perhaps other parts of the Solar system, will be populated with intelligences of other kinds, too, both larger and smaller, and these will outnumber humans. If we believe life and intelligence are precious seeds, destined to bloom here and there in the universe, then spread and effloresce, we should ultimately expect our light cone to be populated by aliens of every description—even in the unlikely event that we are the only seed, and that they are all our children.

# Acknowledgments

TK

I re-encountered Dan Dennett and met Susan Bell at SFI. Sadly, Dan passed away while this project was underway.

My father, Josep Agüera i Arcas, and our dear family friend, Lesley Hazelton, died this year too. They were both great cheerleaders and sharp critics.

# Glossary

TK

# Bibliography

Abramson, José Z., Ma Victoria Hernández-Lloreda, Lino García, Fernando Colmenares, Francisco Aboitiz, and Josep Call. 2018. “Imitation of Novel Conspecific and Human Speech Sounds in the Killer Whale (Orcinus Orca).” *Proceedings. Biological Sciences / The Royal Society* 285 (1871). https://doi.org/10.1098/rspb.2017.2171.  
Achatz, Johannes G., Marta Chiodin, Willi Salvenmoser, Seth Tyler, and Pedro Martinez. 2013. “The Acoela: On Their Kind and Kinships, Especially with Nemertodermatids and Xenoturbellids (Bilateria Incertae Sedis).” *Organisms, Diversity & Evolution* 13 (2): 267–86.  
[Adami, Chris, and C. Titus Brown. 1994. “Evolutionary Learning in the 2D Artificial Life System ‘Avida.’” *arXiv [adap-Org]*. arXiv.](http://paperpile.com/b/iA68kr/Tk1GR) https://arxiv.org/pdf/adap-org/9405003.  
Adams, Douglas. 1979. *The Hitchhiker’s Guide to the Galaxy*. Harmony Books.  
———. 1980. *The Restaurant at the End of the Universe*. Harmony Books.  
[Adiwardana, Daniel, Minh-Thang Luong, David R. So, Jamie Hall, Noah Fiedel, Romal Thoppilan, Zi Yang, et al. 2020. “Towards a Human-like Open-Domain Chatbot.” *arXiv [cs.CL]*. arXiv.](http://paperpile.com/b/iA68kr/eYZP) http://arxiv.org/abs/2001.09977.  
Adler, Julius. 1966. “Chemotaxis in Bacteria.” *Science* 153 (3737): 708–16.  
Adrian, E. D. 1928. “The Basis of Sensation” 122. https://psycnet.apa.org/fulltext/1928-01753-000.pdf.  
[Agostinelli, Andrea, Timo I. Denk, Zalán Borsos, Jesse Engel, Mauro Verzetti, Antoine Caillon, Qingqing Huang, et al. 2023. “MusicLM: Generating Music From Text.” *arXiv [cs.SD]*. arXiv.](http://paperpile.com/b/iA68kr/VHl8) http://arxiv.org/abs/2301.11325.  
Agrawal, A. K., J. S. Gans, and A. Goldfarb. 2023. “The Turing Transformation: Artificial Intelligence, Intelligence Augmentation, and Skill Premiums.” https://www.nber.org/papers/w31767.  
Agüera y Arcas, Blaise. 2022. “Do Large Language Models Understand Us?” *Daedalus* 151 (May): 183–97.  
———. 2023. *Who Are We Now?* Los Angeles: Hat & Beard, LLC.  
[Agüera y Arcas, Blaise, Jyrki Alakuijala, James Evans, Ben Laurie, Alexander Mordvintsev, Eyvind Niklasson, Ettore Randazzo, and Luca Versari. 2024. “Computational Life: How Well-Formed, Self-Replicating Programs Emerge from Simple Interaction.” *arXiv [cs.NE]*. arXiv.](http://paperpile.com/b/iA68kr/YhhO) http://arxiv.org/abs/2406.19108.  
Agüera y Arcas, Blaise, Adrienne Fairhall, and William Bialek. 2000. “What Can a Single Neuron Compute?” *Advances in Neural Information Processing Systems*, 75–81.  
Agüera y Arcas, Blaise, Adrienne L. Fairhall, and William Bialek. 2003. “Computation in a Single Neuron: Hodgkin and Huxley Revisited.” *Neural Computation* 15 (8): 1715–49.  
Agüera y Arcas, Blaise, and Peter Norvig. 2023. “Artificial General Intelligence Is Already Here.” *Noema, October*.  
Ahn, Luis von, Manuel Blum, Nicholas J. Hopper, and John Langford. 2003. “CAPTCHA: Using Hard AI Problems for Security.” In *Advances in Cryptology — EUROCRYPT 2003*, 294–311. Springer Berlin Heidelberg.  
Almomani, Fidaa, Murad O. Al-Momani, Soha Garadat, Safa Alqudah, Manal Kassab, Shereen Hamadneh, Grant Rauterkus, and Richard Gans. 2021. “Cognitive Functioning in Deaf Children Using Cochlear Implants.” *BMC Pediatrics* 21 (1): 71.  
Alper, Morris, and Hadar Averbuch-Elor. 2023. “Kiki or Bouba? Sound Symbolism in Vision-and-Language Models.” *Advances in Neural Information Processing Systems* abs/2310.16781 (October). https://doi.org/10.48550/arXiv.2310.16781.  
Amsterdam, B. 1972. “Mirror Self-Image Reactions before Age Two.” *Developmental Psychobiology* 5 (4): 297–305.  
Anderson, S. 2004. “How Many Languages Are There in the World.” *Linguistic Society of America*. https://doi.org/10.1093/actrade/9780199590599.003.0002.  
[Artetxe, Mikel, Gorka Labaka, Eneko Agirre, and Kyunghyun Cho. 2017. “Unsupervised Neural Machine Translation.” *arXiv [cs.CL]*. arXiv.](http://paperpile.com/b/iA68kr/DUCN) http://arxiv.org/abs/1710.11041.  
Asimov, Isaac. 1950. *I, Robot*. Dennis Dobson.  
Athalye, Anish, Logan Engstrom, Andrew Ilyas, and Kevin Kwok. 10–15 Jul 2018. “Synthesizing Robust Adversarial Examples.” In *Proceedings of the 35th International Conference on Machine Learning*, edited by Jennifer Dy and Andreas Krause, 80:284–93. Proceedings of Machine Learning Research. PMLR.  
Babbage, Charles. 1832. *On the Economy of Machinery and Manufactures*. C. Knight.  
———. 1864. *Passages from the Life of a Philosopher*. Longman, Green, Longman, Roberts, & Green.  
Bacon, Francis. 1603. “Temporis Partus Masculus.”  
Baddeley, Alan D., and J. Graham. 1974. “HITCH. 1974. Working Memory.” *Psychology of Learning and Motivation* 8: 47–89.  
Bagrov, Andrey A., Ilia A. Iakovlev, Askar A. Iliasov, Mikhail I. Katsnelson, and Vladimir V. Mazurenko. 2020. “Multiscale Structural Complexity of Natural Patterns.” *Proceedings of the National Academy of Sciences of the United States of America* 117 (48): 30241–51.  
Bailey, Ronald. 2003. “Pulling Our Own Strings.” Reason Magazine. May 1, 2003. https://reason.com/2003/05/01/pulling-our-own-strings-2/.  
Bakoulis, Stylianos, Robert Krautz, Nicolas Alcaraz, Marco Salvatore, and Robin Andersson. 2022. “Endogenous Retroviruses Co-Opted as Divergently Transcribed Regulatory Elements Shape the Regulatory Landscape of Embryonic Stem Cells.” *Nucleic Acids Research* 50 (4): 2111–27.  
Baltimore, David. 1970. “Viral RNA-Dependent DNA Polymerase: RNA-Dependent DNA Polymerase in Virions of RNA Tumour Viruses.” *Nature* 226 (5252): 1209–11.  
Banay, R. S., and L. Davidoff. 1942. “Apparent Recovery of a Sex Psychopath after Lobotomy.” *Journal of Criminal Psychopathology* 4: 59–66.  
Barnum, Matt. 2024. “We Tested an AI Tutor for Kids. It Struggled With Basic Math.” *The Wall Street Journal*, February 16, 2024. https://www.wsj.com/tech/ai/ai-is-tutoring-students-but-still-struggles-with-basic-math-694e76d3.  
Baron-Cohen, S., A. M. Leslie, and U. Frith. 1985. “Does the Autistic Child Have a ‘Theory of Mind’?” *Cognition* 21 (1): 37–46.  
Barricelli, Nils Aall. 1957. “Symbiogenetic Evolution Processes Realized by Artificial Methods.” *Methodos* 9 (35–36): 143–82.  
Bartumeus, Frederic, Ernesto P. Raposo, Gandhimohan M. Viswanathan, and Marcos G. E. da Luz. 2014. “Stochastic Optimal Foraging: Tuning Intensive and Extensive Dynamics in Random Searches.” *PloS One* 9 (9): e106373.  
Bastani, Aaron. 2019. *Fully Automated Luxury Communism: A Manifesto*. London, England: Verso Books. https://books.google.com/books?id=Qm\_nDwAAQBAJ&newbks=1.  
Bateson, Gregory. 1972a. “From Versailles to Cybernetics.” feineigle.com. 1972. http://www.feineigle.com/book\_reports/2017/versailles\_to\_cybernetics/GregoryBateson\_FromVersaillestoCybernetics.pdf.  
———. 1972b. *Steps to an Ecology of Mind: Collected Essays in Anthropology, Psychiatry, Evolution, and Epistemology*. University of Chicago Press.  
Bekoff, Marc, Colin Allen, and Gordon M. Burghardt. 2002. *The Cognitive Animal: Empirical and Theoretical Perspectives on Animal Cognition*. MIT Press.  
Belin, Pascal. 2006. “Voice Processing in Human and Non-Human Primates.” *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences* 361 (1476): 2091–2107.  
Bell, J. S. 1964. “On the Einstein Podolsky Rosen Paradox.” *Physics* 1 (3): 195–200.  
Bender, Emily M., Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. “On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? 🦜.” In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 610–23. FAccT ’21. New York, NY, USA: Association for Computing Machinery.  
Bender, Emily M., and Alexander Koller. 2020. “Climbing towards NLU: On Meaning, Form, and Understanding in the Age of Data.” In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, edited by Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, 5185–98. Online: Association for Computational Linguistics.  
Ben-Ishai, Guy, Jeff Dean, James Manyika, Ruth Porat, H. Varian, and Kent Walker. 2024. “AI and the Opportunity for Shared Prosperity: Lessons from the History of Technology and the Economy,” January. http://arxiv.org/abs/2401.09718.  
Bennett, Jane. 2010. *Vibrant Matter: A Political Ecology of Things*. Durham, NC and London: Duke University Press.  
Bennett, Max. 2023. *A Brief History of Intelligence: Evolution, AI, and the Five Breakthroughs That Made Our Brains*. HarperCollins.  
Bentham, Jeremy. 1817. *A Table of the Springs of Action: Shewing the Several Species of Pleasures and Pains, of Which Man’s Nature Is Susceptible: Together with the Several Species of Interests, Desires, and Motives, Respectively Corresponding to Them: And the Several Sets of Appellatives, Neutral, Eulogistic and Dyslogistic, by Which Each Species of Motive Is Wont to Be Designated: To Which Are Added, Explanatory Notes and Observations*. R. Hunter.  
Berg, Howard C. 1975. “Chemotaxis in Bacteria.” *Annual Review of Biophysics and Bioengineering* 4 (00): 119–36.  
———. 1993. *Random Walks in Biology*. Princeton University Press.  
Bernadou, Abel, Boris H. Kramer, and Judith Korb. 2021. “Major Evolutionary Transitions in Social Insects, the Importance of Worker Sterility and Life History Trade-Offs.” *Frontiers in Ecology and Evolution* 9. https://doi.org/10.3389/fevo.2021.732907.  
Berrettini, Stefano, Francesca Forli, Elisabetta Genovese, Rosamaria Santarelli, Edoardo Arslan, Anna Maria Chilosi, and Paola Cipriani. 2008. “Cochlear Implantation in Deaf Children with Associated Disabilities: Challenges and Outcomes.” *International Journal of Audiology* 47 (4): 199–208.  
Berridge, K. C., and T. E. Robinson. 1998. “What Is the Role of Dopamine in Reward: Hedonic Impact, Reward Learning, or Incentive Salience?” *Brain Research. Brain Research Reviews* 28 (3): 309–69.  
[Bhirangi, Raunaq, Chenyu Wang, Venkatesh Pattabiraman, Carmel Majidi, Abhinav Gupta, Tess Hellebrekers, and Lerrel Pinto. 2024. “Hierarchical State Space Models for Continuous Sequence-to-Sequence Modeling.” *arXiv [cs.LG]*. arXiv.](http://paperpile.com/b/iA68kr/rY1h) http://arxiv.org/abs/2402.10211.  
[Bhojanapalli, Srinadh, Ayan Chakrabarti, Andreas Veit, Michal Lukasik, Himanshu Jain, Frederick Liu, Yin-Wen Chang, and Sanjiv Kumar. 2021. “Leveraging Redundancy in Attention with Reuse Transformers.” *arXiv [cs.LG]*. arXiv.](http://paperpile.com/b/iA68kr/l17v) http://arxiv.org/abs/2110.06821.  
Blake, W. 1810. “Preface to Milton a Poem (1810).” *Poetry Foundation*.  
Boaz, David. 2023. “Who Said TANSTAAFL First?” *The Cato Institute*, March 14, 2023. https://www.cato.org/blog/who-said-tanstaafl-first.  
[Bojarski, Mariusz, Davide Del Testa, Daniel Dworakowski, Bernhard Firner, Beat Flepp, Prasoon Goyal, Lawrence D. Jackel, et al. 2016. “End to End Learning for Self-Driving Cars.” *arXiv [cs.CV]*. arXiv.](http://paperpile.com/b/iA68kr/Df92) http://arxiv.org/abs/1604.07316.  
Boltzmann, Ludwig. 1868. “Studien über Das Gleichgewicht Der Lebendigen Kraft.” *Wissenschafiliche Abhandlungen*.  
Bommineni, V. L., S. Bhagwagar, and D. Balcarcel. 2023. “Performance of ChatGPT on the MCAT: The Road to Personalized and Equitable Premedical Learning.” *MedRxiv*. https://www.medrxiv.org/content/10.1101/2023.03.05.23286533.abstract.  
Boniol, M., M. McIsaac, L. Xu, T. Wuliji, and K. Diallo. 2019. “WHO Guidelines.” apps.who.int. 2019. https://apps.who.int/bookorders.%0Ahttps://apps.who.int/iris/handle/10665/311314.  
Bonnet, Charles. 1769. *Essai analytique sur les facultés de l’âme*. Chez Cl. Philibert.  
Boole, George. 1854. *An Investigation of the Laws of Thought: On Which Are Founded the Mathematical Theories of Logic and Probabilities*. Dover Publications.  
Boole, Mary Everest. 1901. “Indian Thought and Western Science in the Nineteenth Century.” https://libarch.nmu.org.ua/bitstream/handle/GenofondUA/17988/b4032742246bdb167e16a6951b758e43.pdf?sequence=1.  
Borges, Jorge Luis. 1942. “Funes El Memorioso.” *La Nación*, June 1942.  
Borsos, Zalán, Raphaël Marinier, Damien Vincent, Eugene Kharitonov, Olivier Pietquin, Matt Sharifi, Dominik Roblek, et al. 2023. “AudioLM: A Language Modeling Approach to Audio Generation.” *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 31: 2523–33.  
[Borsos, Zalán, Matt Sharifi, Damien Vincent, Eugene Kharitonov, Neil Zeghidour, and Marco Tagliasacchi. 2023. “SoundStorm: Efficient Parallel Audio Generation.” *arXiv [cs.SD]*. arXiv.](http://paperpile.com/b/iA68kr/Z4MSw) http://arxiv.org/abs/2305.09636.  
Bostrom, Nick. 2003. “Are We Living in a Computer Simulation?” *The Philosophical Quarterly* 53 (211): 243–55.  
———. 2014. *Superintelligence: Paths, Dangers, Strategies*. Oxford University Press.  
———. 2020. “Ethical Issues in Advanced Artificial Intelligence.” *Machine Ethics and Robot Ethics*, 69–75.  
———. 2023. “Apology for an Old Email.” https://nickbostrom.com/oldemail.pdf.  
———. 2024. *Deep Utopia: Life and Meaning in a Solved World*. Ideapress Publishing.  
Botvinick, M., and J. Cohen. 1998. “Rubber Hands ‘feel’ Touch That Eyes See.” *Nature* 391 (February): 756–756.  
Bradbury, Phillip. 2012. “Life in Life.” YouTube. May 13, 2012. https://youtu.be/xP5-iIeKXE8.  
Braslow, J. 1999. “Therapeutic Effectiveness and Social Context: The Case of Lobotomy in a California State Hospital, 1947-1954.” *The Western Journal of Medicine* 170 (5): 293–96.  
Braun, Hans Albert. 2021. “Stochasticity Versus Determinacy in Neurobiology: From Ion Channels to the Question of the ‘Free Will.’” *Frontiers in Systems Neuroscience* 15. https://doi.org/10.3389/fnsys.2021.629436.  
Brembs, Björn. 2011. “Towards a Scientific Concept of Free Will as a Biological Trait: Spontaneous Actions and Decision-Making in Invertebrates.” *Proceedings. Biological Sciences / The Royal Society* 278 (1707): 930–39.  
Brewer, Rebecca, Richard Cook, and Geoffrey Bird. 2016. “Alexithymia: A General Deficit of Interoception.” *Royal Society Open Science* 3 (10): 150664.  
Bridle, John S. 1990. “Probabilistic Interpretation of Feedforward Classification Network Outputs, with Relationships to Statistical Pattern Recognition.” In *Neurocomputing*, 227–36. Springer Berlin Heidelberg.  
Bromley, Allan G. 1982. “Charles Babbage’s Analytical Engine, 1838.” *IEEE Annals of the History of Computing* 4 (3): 196–217.  
Brown, Daniel J. 2013. “The Boys in the Boat: Nine Americans and Their Epic Quest for Gold at the 1936 Berlin Olympics,” June.  
Brown, Tom, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D. Kaplan, Prafulla Dhariwal, Arvind Neelakantan, et al. 2020. “Language Models Are Few-Shot Learners.” *Advances in Neural Information Processing Systems* 33: 1877–1901.  
Bruder, Johannes. 2017. “Infrastructural Intelligence: Contemporary Entanglements between Neuroscience and AI.” *Progress in Brain Research* 233 (July): 101–28.  
Bryson, J. 2010. “Robots Should Be Slaves.” *Close Engagements with Artificial Companions: Key*, March, 63–74.  
[Bubeck, Sébastien, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, et al. 2023. “Sparks of Artificial General Intelligence: Early Experiments with GPT-4.” *arXiv [cs.CL]*. arXiv.](http://paperpile.com/b/iA68kr/Is9s) http://arxiv.org/abs/2303.12712.  
Bubeck, Sébastien, and Mark Sellke. 2022. “A Universal Law of Robustness via Isoperimetry.” *Journal of the ACM*, December. https://doi.org/10.1145/3578580.  
Buchanan, Bon B., and Daniel I. Arnon. 1990. “A Reverse Krebs Cycle in Photosynthesis: Consensus at Last.” *Photosynthesis Research* 24 (1): 47–53.  
Budelmann, B. U. 1998. “Autophagy in Octopus.” *South African Journal of Marine Science/Suid-Afrikaanse Tydskrif Vir Seewetenskap* 20 (1): 101–8.  
Bulldog, Bentham ’s. n.d. “Bentham’s Newsletter.” Accessed April 30, 2024. https://benthams.substack.com/.  
Buolamwini, Joy, and Timnit Gebru. 23–24 Feb 2018. “Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification.” In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, edited by Sorelle A. Friedler and Christo Wilson, 81:77–91. Proceedings of Machine Learning Research. PMLR.  
Bush, Robert R., and Frederick Mosteller. 1951. “A Mathematical Model for Simple Learning.” *Psychological Review* 58 (5): 313–23.  
Bush, S. L. 2012. “Economy of Arm Autotomy in the Mesopelagic Squid Octopoteuthis Deletron.” *Marine Ecology Progress Series* 458 (July): 133–40.  
Butter, C. M., S. Kosslyn, D. Mijovic-Prelec, and A. Riffle. 1997. “Field-Specific Deficits in Visual Imagery Following Hemianopia due to Unilateral Occipital Infarcts.” *Brain: A Journal of Neurology* 120 ( Pt 2) (February): 217–28.  
Buxhoeveden, Daniel P., and Manuel F. Casanova. 2002. “The Minicolumn Hypothesis in Neuroscience.” *Brain: A Journal of Neurology* 125 (Pt 5): 935–51.  
Byrd, Paul. 1975. “New Relations between Fibonacci and Bernoulli Numbers.” *Fibonacci Quart* 13: 111–14.  
Cadwell, Cathryn R., Aparna Bhaduri, Mohammed A. Mostajo-Radji, Matthew G. Keefe, and Tomasz J. Nowakowski. 2019. “Development and Arealization of the Cerebral Cortex.” *Neuron* 103 (6): 980–1004.  
Caliskan, Aylin, Joanna J. Bryson, and Arvind Narayanan. 2017. “Semantics Derived Automatically from Language Corpora Contain Human-like Biases.” *Science* 356 (6334): 183–86.  
Campbell, Lyle, Nala Huiying Lee, Eve Okura, Sean Simpson, Kaori Ueki, and John Van Way. 2013. “New Knowledge: Findings from the Catalogue of Endangered Languages (‘ELCat’).” In *3rd International Conference on Language Documentation & Conservation*.  
Čapek, Karel. 1920. *R.U.R.: Rossum’s Universal Robots*. Aventinum.  
Carlile, Simon, and Caitlin Corkhill. 2015. “Selective Spatial Attention Modulates Bottom-up Informational Masking of Speech.” *Scientific Reports* 5 (March): 8662.  
[Carlsmith, Joseph. 2022. “Is Power-Seeking AI an Existential Risk?” *arXiv [cs.CY]*. arXiv.](http://paperpile.com/b/iA68kr/x3du) http://arxiv.org/abs/2206.13353.  
Carroll, Sean B., Jennifer K. Grenier, and Scott D. Weatherbee. 2013. *From DNA to Diversity: Molecular Genetics and the Evolution of Animal Design*. EPUB. 2nd ed. Hoboken, NJ: Wiley-Blackwell.  
Case, R. 1985. “Intellectual Development : Birth to Adulthood.” *(No Title)*. https://cir.nii.ac.jp/crid/1130282272272375168.  
Cassell, Eric J. 1998. “The Nature of Suffering and the Goals of Medicine.” *Loss, Grief & Care* 8 (1-2): 129–42.  
Cavalier-Smith, T. 2002. “Origins of the Machinery of Recombination and Sex.” *Heredity* 88 (2): 125–41.  
Chalmers, D. 1995. “Facing up to the Problem of Consciousness.” *Journal of Consciousness Studies*. https://doi.org/10.1093/acprof:oso/9780195311105.003.0001.  
Chalmers, David J. 2022. *Reality+: Virtual Worlds and the Problems of Philosophy*. W. W. Norton & Company.  
[Chalmers, David J., and Kelvin J. McQueen. 2021. “Consciousness and the Collapse of the Wave Function.” *arXiv [quant-Ph]*. arXiv.](http://paperpile.com/b/iA68kr/wFhO) http://arxiv.org/abs/2105.02314.  
Chance, Michael R. A., and Allan P. Mead. 1953. “Social Behaviour and Primate Evolution.” *Evolution: Symposia of the Society for Experimental Biology* 7: 395–439.  
Chater, Nick. 2018. *The Mind Is Flat: The Remarkable Shallowness of the Improvising Brain*. Yale University Press.  
Chebat, Daniel-Robert, Fabien C. Schneider, and Maurice Ptito. 2020. “Spatial Competence and Brain Plasticity in Congenital Blindness via Sensory Substitution Devices.” *Frontiers in Neuroscience* 14 (July): 815.  
Cheng, Ken. 2021. “Learning in Cnidaria: A Systematic Review.” *Learning & Behavior* 49 (2): 175–89.  
Chen, Lang, Demian Wassermann, Daniel A. Abrams, John Kochalka, Guillermo Gallardo-Diez, and Vinod Menon. 2019. “The Visual Word Form Area (VWFA) Is Part of Both Language and Attention Circuitry.” *Nature Communications* 10 (1): 5601.  
Cherkaev, Xenia. 2021. “Zoo-Fascism, Russia: To Hell with Equality and Ownerless Dogs.” April 15, 2021. https://culanth.org/fieldsights/zoo-fascism-russia-to-hell-with-equality-and-ownerless-dogs.  
Cherry, Colin. 1953. “Cocktail Party Problem.” *The Journal of the Acoustical Society of America* 25: 975–79.  
chewbster. 2008. “Blindsight - Blind Man Can See and Avoid Obstacles.” Youtube. December 25, 2008. https://www.youtube.com/watch?v=GwGmWqX0MnM.  
Chew, Monica, and J. D. Tygar. 2004. “Image Recognition CAPTCHAs.” In *Information Security*, 268–79. Springer Berlin Heidelberg.  
Chiang, Ted. 2005. “What’s Expected of Us.” Nature Publishing Group UK. July 6, 2005. https://doi.org/10.1038/436150a.  
———. 2023. “ChatGPT Is a Blurry JPEG of the Web.” *New Yorker* 9: 2023.  
Chiang, Ted, and Allora & Calzadilla. 2015. “The Great Silence.” *E-Flux Journal*, no. 65 (May). https://www.e-flux.com/journal/65/336684/the-great-silence/.  
Chittka, Lars. 2022. *The Mind of a Bee*. Princeton University Press.  
Chittka, Lars, and Johannes Spaethe. 2007. “Visual Search and the Importance of Time in Complex Decision Making by Bees.” *Arthropod-Plant Interactions* 1 (1): 37–44.  
Chittka, Lars, and James D. Thomson. 1997. “Sensori-Motor Learning and Its Relevance for Task Specialization in Bumble Bees.” *Behavioral Ecology and Sociobiology* 41 (6): 385–98.  
Chomsky, Noam. 1959. “A Review Od BF Skinner’s Verbal Behavior.” *Language* 35 (1): 26–58.  
———. 1980. “On Cognitive Structures and Their Development: A Reply to Piaget.” *Debate between Jean Piaget and Noam Chomsky*.  
Chomsky, Noam, Ian Roberts, and Jeffrey Watumull. 2023. “Noam Chomsky: The False Promise of Chatgpt.” *The New York Times* 8. https://edisciplinas.usp.br/pluginfile.php/7614933/mod\_resource/content/1/Opinion%20\_%20Noam%20Chomsky\_%20The%20False%20Promise%20of%20ChatGPT%20-%20The%20New%20York%20Times.pdf.  
Christian, Brian. 2020. *The Alignment Problem: Machine Learning and Human Values*. W. W. Norton & Company.  
Christiansen, Morten H., and Nick Chater. 2022. *The Language Game: How Improvisation Created Language and Changed the World*. Basic Books.  
Christiansen, Morten H., Rick A. C. Dale, Michelle R. Ellefson, and Christopher M. Conway. 2002. “The Role of Sequential Learning in Language Evolution: Computational and Experimental Studies.” In *Simulating the Evolution of Language*, edited by Angelo Cangelosi and Domenico Parisi, 165–87. London: Springer London.  
Christley, Scott, Yiming Lu, Chen Li, and Xiaohui Xie. 2009. “Human Genomes as Email Attachments.” *Bioinformatics* 25 (2): 274–75.  
[Chung, Junyoung, Caglar Gulcehre, Kyunghyun Cho, and Yoshua Bengio. 2014. “Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling.” *arXiv [cs.NE]*. arXiv.](http://paperpile.com/b/iA68kr/T6M7) http://arxiv.org/abs/1412.3555.  
Chuong, Edward B. 2018. “The Placenta Goes Viral: Retroviruses Control Gene Expression in Pregnancy.” *PLoS Biology* 16 (10): e3000028.  
Churchland, Patricia. 2016. “Motivations and Drives Are Computationally Messy.” 2016. https://www.pas.va/en/publications/scripta-varia/sv132pas/churchland.html.  
———. 2019. *Conscience: The Origins of Moral Intuition*. National Geographic Books.  
Churchland, Patricia, V. S. Ramachandran, and Terrence J. Sejnowski. 1994. “A Critique of Pure Vision.” In *Large-Scale Neuronal Theories of the Brain*, edited by Christof Koch and Joel L. Davis, 23–60. Bradford Books. Cambridge, MA: MIT Press.  
Cilliers, Paul. 1998. *Complexity and Postmodernism: Understanding Complex Systems*. Psychology Press.  
Clark, Andy. 2023. *The Experience Machine: How Our Minds Predict and Shape Reality*. Knopf Doubleday Publishing Group.  
Clarke, Arthur C. 1968. *2001: A Space Odyssey*. New American Library.  
Clynes, M. E., and N. S. Kline. 1960. “Cyborgs and Space.” *Astronautics* 14 (9): 26–27.  
Colapinto, John. 2007. “The Interpreter.” *New Yorker* 16: 125.  
Colbert-White, Erin Natannie. 2013. “Evidence for Human-like Conversational Strategies in an African Grey Parrot’s Speech.” University of Georgia. https://getd.libs.uga.edu/pdfs/colbert-white\_erin\_n\_201305\_phd.pdf.  
Collado, Miguel Á., Cristina M. Montaner, Francisco P. Molina, Daniel Sol, and Ignasi Bartomeus. 2021. “Brain Size Predicts Learning Abilities in Bees.” *Royal Society Open Science* 8 (5): 201940.  
Colombo, Matteo. 2014. “Deep and Beautiful. The Reward Prediction Error Hypothesis of Dopamine.” *Studies in History and Philosophy of Biological and Biomedical Sciences* 45 (March): 57–67.  
Conway, Christopher M., and David B. Pisoni. 2008. “Neurocognitive Basis of Implicit Learning of Sequential Structure and Its Relation to Language Processing.” *Annals of the New York Academy of Sciences* 1145 (December): 113–31.  
Cook, M. 2004. “Universality in Elementary Cellular Automata.” *Complex Systems* 15 (1): 1–40.  
Corcoran, A. J., and W. Conner. 2016. “How Moths Escape Bats: Predicting Outcomes of Predator–prey Interactions.” *The Journal of Experimental Biology* 219 (September): 2704–15.  
*Cosmo the Funny Parrot*. n.d. Youtube. Accessed April 18, 2024. https://www.youtube.com/channel/UClMJsMoWV1HRFsnfUgN19GA.  
Cowan, N. 2001. “The Magical Number 4 in Short-Term Memory: A Reconsideration of Mental Storage Capacity.” *The Behavioral and Brain Sciences* 24 (1): 87–114; discussion 114–85.  
Craige, Betty Jean. 2010. *Conversations with Cosmo: At Home with an African Grey Parrot*. Sherman Asher Pub.  
Crook, Robyn J., and Edgar T. Walters. 2014. “Neuroethology: Self-Recognition Helps Octopuses Avoid Entanglement.” *Current Biology: CB* 24 (11): R520–21.  
Crosby, Alfred W. 2003. *The Columbian Exchange: Biological and Cultural Consequences of 1492, 30th Anniversary Edition*. New York: Praeger. https://books.google.com/books?hl=en&lr=&id=SXvCEAAAQBAJ&oi=fnd&pg=PT15&dq=The+Columbian+Exchange:+Biological+and+Cultural+Consequences+of+1492&ots=tQcdcNH-ru&sig=-tY3njlfbvWNcvjeKuhBCvBRnGc.  
Cruse, Holk, and David Hübner. 2008. “Selforganizing Memory: Active Learning of Landmarks Used for Navigation.” *Biological Cybernetics* 99 (3): 219–36.  
[Cundy, Chris, and Stefano Ermon. 2023. “SequenceMatch: Imitation Learning for Autoregressive Sequence Modelling with Backtracking.” *arXiv [cs.LG]*. arXiv.](http://paperpile.com/b/iA68kr/IyAR) http://arxiv.org/abs/2306.05426.  
Ćwiek, Aleksandra, Susanne Fuchs, Christoph Draxler, Eva Liina Asu, Dan Dediu, Katri Hiovain, Shigeto Kawahara, et al. 2022. “The Bouba/kiki Effect Is Robust across Cultures and Writing Systems.” *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences* 377 (1841): 20200390.  
Cybenko, G. 1989. “Approximation by Superpositions of a Sigmoidal Function.” *Mathematics of Control, Signals, and Systems* 2 (4): 303–14.  
Damasio, Antonio R. 1999. *The Feeling of What Happens: Body and Emotion in the Making of Consciousness*. Houghton Mifflin Harcourt.  
Damasio, A. R., H. Damasio, and H. C. Chui. 1980. “Neglect Following Damage to Frontal Lobe or Basal Ganglia.” *Neuropsychologia* 18 (2): 123–32.  
Dattani, Saloni, Lucas Rodés-Guirao, Hannah Ritchie, Esteban Ortiz-Ospina, and Max Roser. 2023. “Life Expectancy.” *Our World in Data*.  
[Daubechies, I., R. DeVore, S. Foucart, B. Hanin, and G. Petrova. 2019. “Nonlinear Approximation and (Deep) ReLU Networks.” *arXiv [cs.LG]*. arXiv. https://doi.org/](http://paperpile.com/b/iA68kr/INjD)10.1007/s00365-021-09548-z.  
Dauphin, Yann N., Angela Fan, Michael Auli, and David Grangier. 2017. “Language Modeling with Gated Convolutional Networks.” In *Proceedings of the 34th International Conference on Machine Learning*, edited by Doina Precup and Yee Whye Teh, 70:933–41. Proceedings of Machine Learning Research. PMLR.  
Dawkins, Richard. 1986. *The Blind Watchmaker*. New York: W.W. Norton.  
Dax, E. C., and R. K. Freudenberg. 1948. “Prefrontal Leucotomy; a Review.” *Postgraduate Medical Journal* 24 (274): 415–26.  
Dayan, P., G. E. Hinton, R. M. Neal, and R. S. Zemel. 1995. “The Helmholtz Machine.” *Neural Computation* 7 (5): 889–904.  
Deacon, Terrence W. 2012. *Incomplete Nature: How Mind Emerged from Matter*. W. W. Norton & Company.  
Dean, Jeffrey. 2022. “A Golden Decade of Deep Learning: Computing Systems & Applications.” *Daedalus* 151 (2): 58–74.  
“Declaration of Independence: A Transcription.” n.d. Accessed April 15, 2024. https://www.archives.gov/founding-docs/declaration-transcript.  
Dein, Simon, and Abdool Samad Illaiee. 2013. “Jinn and Mental Health: Looking at Jinn Possession in Modern Psychiatric Practice.” *The Psychiatrist* 37 (9): 290–93.  
DeLong, J. Bradford. 2022. *Slouching Towards Utopia: An Economic History of the Twentieth Century*. Basic Books.  
De Monchaux, Nicholas. 2011. *Spacesuit: Fashioning Apollo*. MIT Press.  
De Morgan, Augustus, and Sophia Elizabeth De Morgan. 1872. *A Budget of Paradoxes*. London: Longmans, Green.  
Dennett, Daniel Clement. 1984. *Elbow Room: The Varieties of Free Will Worth Wanting*. Clarendon Press.  
———. 1989. *The Intentional Stance*. MIT Press.  
———. 2009. “How People Are like Robots.” Big Think. March 9, 2009. https://bigthink.com/videos/how-people-are-like-robots/.  
———. 2017a. *Consciousness Explained*. Little, Brown.  
———. 2017b. *From Bacteria to Bach and Back: The Evolution of Minds*. W. W. Norton & Company.  
Descartes, René. 1662. *De Homine*. Translated by Florentius Schuyl. Leyden: Peter Leffen & Francis Moyard.  
Deutsch, David, and Chiara Marletto. 2015. “Constructor Theory of Information.” *Proceedings. Mathematical, Physical, and Engineering Sciences / the Royal Society* 471 (2174): 20140540.  
Dick, Philip K. 1985. *I Hope I Shall Arrive Soon*. Doubleday.  
Diederen, Kelly M. J., and Paul C. Fletcher. 2021. “Dopamine, Prediction Error and Beyond.” *The Neuroscientist: A Review Journal Bringing Neurobiology, Neurology and Psychiatry* 27 (1): 30–46.  
Ding, Xiao Pan, Henry M. Wellman, Yu Wang, Genyue Fu, and Kang Lee. 2015. “Theory-of-Mind Training Causes Honest Young Children to Lie.” *Psychological Science* 26 (11): 1812–21.  
Doctorow, Cory. 2024. “‘Enshittification’ Is Coming for Absolutely Everything.” *Financial Times*, February 8, 2024. https://www.ft.com/content/6fb1602d-a08b-4a8c-bac0-047b7d64aba5.  
Doctorow, Cory, and Charles Stross. 2012. *The Rapture of the Nerds: A Tale of the Singularity, Posthumanity, and Awkward Social Situations*. Tor Publishing Group.  
Dombrowsky-Hahn, Klaudia, and Gabriele Slezak. 2004. “Local Vitality : On a Complex Linguistic Situation in the West African Town of Banfora (Burkina Faso).” *Between Resistance and*. https://books.google.com/books?hl=en&lr=&id=OUP\_GJ1zjCUC&oi=fnd&pg=PA49&dq=burkina+faso+percent+population+speak+french&ots=uYRRK6G35x&sig=AQpig8fxib2FyLsaS-fTJOuJMiA.  
Domenici, Paolo, David Booth, Jonathan M. Blagburn, and Jonathan P. Bacon. 2008. “Cockroaches Keep Predators Guessing by Using Preferred Escape Trajectories.” *Current Biology: CB* 18 (22): 1792–96.  
Donsker, M. D., and S. R. S. Varadhan. 1975. “Asymptotics for the Wiener Sausage.” *Communications on Pure and Applied Mathematics* 28 (4): 525–65.  
Downing, Lisa, Iain Morland, and Nikki Sullivan. 2015. *Fuckology: Critical Essays on John Money’s Diagnostic Concepts*. University of Chicago Press.  
Dreyfus, Hubert L. 1972. *What Computers Can’t Do: A Critique of Artificial Reason*. Harper & Row.  
Drude, Sebastian, Intangible Cultural Heritage Unit’s Ad Hoc Expert Group, and Others. 2003. “Language Vitality and Endangerment.”  
Duchenne, Guillaume-Benjamin. 1876. *Mécanisme de la physionomie humaine: où, Analyse électro-physiologique de l’expression des passions*. J.-B. Baillière.  
Duéñez-Guzmán, Edgar A., Suzanne Sadedin, Jane X. Wang, Kevin R. McKee, and Joel Z. Leibo. 2023. “A Social Path to Human-like Artificial Intelligence.” *Nature Machine Intelligence* 5 (11): 1181–88.  
Dukas, R. 2009. “Evolutionary Biology of Limited Attention,” May, 147–62.  
Dunbar, Robin I. M. 1992. “Neocortex Size as a Constraint on Group Size in Primates.” *Journal of Human Evolution* 22 (6): 469–93.  
———. 1998. “The Social Brain Hypothesis.” *Anthropology: Issues, News, and Reviews: Issues …*. https://doi.org/10.1002/(SICI)1520-6505(1998)6:5<178::AID-EVAN5>3.0.CO;2-81520-6505(1998)6:5<178::AID-EVAN5>3.0.CO;2-8).  
———. 2016. “The Social Brain Hypothesis and Human Evolution,” March. https://doi.org/10.1093/ACREFORE/9780190236557.013.44.  
Easterlin, Richard A. 1974. “Does Economic Growth Improve the Human Lot? Some Empirical Evidence.” In *Nations and Households in Economic Growth*, edited by Paul A. David and Melvin W. Reder, 89–125. Academic Press.  
———. 2001. “Income and Happiness: Towards a Unified Theory.” *Economic Journal* 111 (473): 465–84.  
[Ebrahimi, Abteen, and Katharina Kann. 2021. “How to Adapt Your Pretrained Multilingual Model to 1600 Languages.” *arXiv [cs.CL]*. arXiv.](http://paperpile.com/b/iA68kr/tKxe) http://arxiv.org/abs/2106.02124.  
Edwards, Benj. 2022. “We Interviewed Linux OS through an AI Bot to Discover Its Secrets.” Ars Technica. September 28, 2022. https://arstechnica.com/information-technology/2022/09/we-interviewed-linux-os-through-an-ai-bot-to-discover-its-secrets/2/.  
Edwards, Jeff. 2023. “The End of CAPTCHA? Testing GPT-4V and AI Solvers vs. CAPTCHA.” *CHEQ |* (blog). CHEQ AI Technologies Ltd. October 12, 2023. https://cheq.ai/blog/testing-ai-gpt-4v-against-captcha/.  
Eguíluz, V. M., M. Ospeck, Y. Choe, A. J. Hudspeth, and M. O. Magnasco. 2000. “Essential Nonlinearities in Hearing.” *Physical Review Letters* 84 (22): 5232–35.  
Ehrlich, Paul. 1968. *The Population Bomb*. Ballantine Books.  
Ekman, P., R. J. Davidson, and W. V. Friesen. 1990. “The Duchenne Smile: Emotional Expression and Brain Physiology. II.” *Journal of Personality and Social Psychology* 58 (2): 342–53.  
[Eldan, Ronen, and Yuanzhi Li. 2023. “TinyStories: How Small Can Language Models Be and Still Speak Coherent English?” *arXiv [cs.CL]*. arXiv.](http://paperpile.com/b/iA68kr/SIXm) http://arxiv.org/abs/2305.07759.  
Ellis, N. C. 2001. “Memory for Language.” https://www.cambridge.org/core/services/aop-cambridge-core/content/view/ED04D44C82886A80597B3FF8DDB3F73C.  
Eslinger, P. J., and A. R. Damasio. 1985. “Severe Disturbance of Higher Cognition after Bilateral Frontal Lobe Ablation: Patient EVR.” *Neurology* 35 (12): 1731–41.  
Espejo, Raul. 2014. “Cybernetics of Governance: The Cybersyn Project 1971–1973.” In *Social Systems and Design*, edited by Gary S. Metcalf, 71–90. Tokyo: Springer Japan.  
Evans, James A. 2010. “Industry Collaboration, Scientific Sharing, and the Dissemination of Knowledge.” *Social Studies of Science* 40 (5): 757–91.  
Evans, Michael C., Bob B. Buchanan, and Daniel I. Arnon. 1966. “A New Ferredoxin-Dependent Carbon Reduction Cycle in a Photosynthetic Bacterium.” *Proceedings of the National Academy of Sciences of the United States of America* 55 (4): 928–34.  
Everett, Daniel. 2009. *Don’t Sleep, There Are Snakes: Life and Language in the Amazonian Jungle*. Profile Books.  
Fabre, Jean-Henri. 1916. *The Life of the Caterpillar*. Modern Library.  
———. 1921. *More Hunting Wasps*. Dodd, Mead.  
Fairwork. 2023. “Work in the Planetary Labour Market: Fairwork Cloudwork Ratings 2023.” Oxford, UK. https://fair.work/wp-content/uploads/sites/17/2023/07/Fairwork-Cloudwork-Ratings-2023-Red.pdf.  
Faisal, A. Aldo, Luc P. J. Selen, and Daniel M. Wolpert. 2008. “Noise in the Nervous System.” *Nature Reviews. Neuroscience* 9 (4): 292–303.  
Farnsworth, K. D., and J. A. Beecham. 1999. “How Do Grazers Achieve Their Distribution? A Continuum of Models from Random Diffusion to the Ideal Free Distribution Using Biased Random Walks.” *The American Naturalist* 153 (5): 509–26.  
Faux, Zeke. 2023. *Number Go Up: Inside Crypto’s Wild Rise and Staggering Fall*. Crown.  
Feigenbaum, Edward, and Julian Feldman. 1963. *Computers and Thought*. Vol. 9. New York: McGraw-Hill.  
Feingold, Gustave A. 1914. “The Influence of Environment on Identification of Persons and Things.” *Journal of the American Institute of Criminal Law and Criminology* 5 (1): 39.  
Fein, Yaakov Y., Philipp Geyer, Patrick Zwick, Filip Kiałka, Sebastian Pedalino, Marcel Mayor, Stefan Gerlich, and Markus Arndt. 2019. “Quantum Superposition of Molecules beyond 25 kDa.” *Nature Physics* 15 (12): 1242–45.  
Feldman, Brian. 2019. “Philosopher Nick Bostrom on Whether We Live in a Simulation.” Vulture. February 6, 2019. https://www.vulture.com/2019/02/nick-bostrom-on-whether-we-live-in-a-matrix-simulation.html.  
Ferreira, Francisco R. M., Maria I. Nogueira, and Javier Defelipe. 2014. “The Influence of James and Darwin on Cajal and His Research into the Neuron Theory and Evolution of the Nervous System.” *Frontiers in Neuroanatomy* 8 (January): 1.  
Feschotte, Cédric, and Clément Gilbert. 2012. “Endogenous Viruses: Insights into Viral Evolution and Impact on Host Biology.” *Nature Reviews. Genetics* 13 (4): 283–96.  
Fields, R. Douglas, Alfonso Araque, Heidi Johansen-Berg, Soo-Siang Lim, Gary Lynch, Klaus-Armin Nave, Maiken Nedergaard, Ray Perez, Terrence Sejnowski, and Hiroaki Wake. 2014. “Glial Biology in Learning and Cognition.” *The Neuroscientist: A Review Journal Bringing Neurobiology, Neurology and Psychiatry* 20 (5): 426–31.  
Filatova, Olga A., Ivan D. Fedutin, Alexander M. Burdin, and Erich Hoyt. 2011. “Responses of Kamchatkan Fish‐eating Killer Whales to Playbacks of Conspecific Calls.” *Marine Mammal Science* 27 (2): E26–42.  
Fisher, J. P., D. T. Hassan, and N. O’Connor. 1995. “Minerva.” *BMJ* 310 (6971): 70–70.  
Fitzpatrick, A. 1998. “Teller’s Technical Nemesis: The American Hydrogen Bomb and Its Development within a Technological Infrastructure.” *Techné: Research in Philosophy and Technology* 3 (August): 119–23.  
Foer, Joshua. 2005. “How to Win the U.S. Memory Championship.” Slate. March 16, 2005. https://slate.com/news-and-politics/2005/03/how-to-win-the-u-s-memory-championship.html.  
Fontana, Walter. 1990. “Algorithmic Chemistry,” no. LA-UR-90-1959 (June). https://www.santafe.edu/research/results/working-papers/algorithmic-chemistry-a-model-for-functional-self-.  
Foot, Philippa. 1967. “The Problem of Abortion and the Problem of Double Effect.” *The Oxford Review* 5: 147–57.  
Forbes, Avery. 2021. “Your New Best Friends: An Exploration of Furby, Siri, and Other Sociable Electronics.” University of Massachusetts Amherst. https://doi.org/10.7275/22818580.0.  
Ford, Martin. 2015. *Rise of the Robots: Technology and the Threat of a Jobless Future*. Basic Books.  
Fortunato, Santo, Carl T. Bergstrom, Katy Börner, James A. Evans, Dirk Helbing, Staša Milojević, Alexander M. Petersen, et al. 2018. “Science of Science.” *Science* 359 (6379). https://doi.org/10.1126/science.aao0185.  
Fouragnan, Elsa, Chris Retzler, and Marios G. Philiastides. 2018. “Separate Neural Representations of Prediction Error Valence and Surprise: Evidence from an fMRI Meta-Analysis.” *Human Brain Mapping* 39 (7): 2887–2906.  
[Frankle, Jonathan, and Michael Carbin. 2018. “The Lottery Ticket Hypothesis: Finding Sparse, Trainable Neural Networks.” *arXiv [cs.LG]*. arXiv.](http://paperpile.com/b/iA68kr/XMb9) http://arxiv.org/abs/1803.03635.  
Frank, Michael C., Daniel L. Everett, Evelina Fedorenko, and Edward Gibson. 2008. “Number as a Cognitive Technology: Evidence from Pirahã Language and Cognition.” *Cognition* 108 (3): 819–24.  
Frankopan, Peter. 2016. *The Silk Roads: A New History of the World*. Knopf Doubleday Publishing Group.  
Freeman, W. 1971. “Frontal Lobotomy in Early Schizophrenia. Long Follow-up in 415 Cases.” *The British Journal of Psychiatry: The Journal of Mental Science* 119 (553): 621–24.  
Freeman, Walter, and James Winston Watts. 1950. *Psychosurgery in the Treatment of Mental Disorders and Intractable Pain*. C. C. Thomas.  
Friedman, Milton. 1975. *There’s No Such Thing as a Free Lunch*. Open Court.  
Frisch, K. 1955. “The Dancing Bees.” https://doi.org/10.1007/978-3-7091-4697-2.  
Frisch, Karl von, and Thomas D. Seeley. 2013. *The Dance Language and Orientation of Bees*. Harvard University Press.  
Friston, Karl. 2005. “A Theory of Cortical Responses.” *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences* 360 (1456): 815–36.  
Fritz, W. B. Fall 1996. “The Women of ENIAC.” *IEEE Annals of the History of Computing* 18 (3): 13–28.  
Fukushima, K. 1980. “Neocognitron: A Self Organizing Neural Network Model for a Mechanism of Pattern Recognition Unaffected by Shift in Position.” *Biological Cybernetics* 36 (4): 193–202.  
Fumagalli, Manuela, and Alberto Priori. 2012. “Functional and Clinical Neuroanatomy of Morality.” *Brain: A Journal of Neurology* 135 (7): 2006–21.  
Galen. 1968. *Galen on the Usefulness of the Parts of the Body (Περὶ χρείας μορίων: De Usu Partium)*. Translated by Margaret Tallmadge May. Ithaca, New York: Cornell University Press.  
Galison, Peter. 1994. “The Ontology of the Enemy: Norbert Wiener and the Cybernetic Vision.” *Critical Inquiry* 21 (1): 228–66.  
Gallop, G. G., Jr. 1970. “Chimpanzees: Self-Recognition.” *Science* 167 (3914): 86–87.  
Gao, Shan. 2018. “What Is It like to Be a Quantum Observer? And What Does It Imply about the Nature of Consciousness?” http://philsci-archive.pitt.edu/14836.  
Garcia, Jair E., Yu-Shan Hung, Andrew D. Greentree, Marcello G. P. Rosa, John A. Endler, and Adrian G. Dyer. 2017. “Improved Color Constancy in Honey Bees Enabled by Parallel Visual Projections from Dorsal Ocelli.” *Proceedings of the National Academy of Sciences of the United States of America* 114 (29): 7713–18.  
Garner, Aleena R., and Georg B. Keller. 2022. “A Cortical Circuit for Audio-Visual Predictions.” *Nature Neuroscience* 25 (1): 98–105.  
Gazzaniga, Michael S. 1967. “The Split Brain in Man.” *Scientific American* 217 (2): 24–29.  
———. 1998. *The Mind’s Past*. University of California Press.  
———. 2005. “Forty-Five Years of Split-Brain Research and Still Going Strong.” *Nature Reviews. Neuroscience* 6 (8): 653–59.  
Gebru, Timnit, and Émile P. Torres. 2024. “The TESCREAL Bundle: Eugenics and the Promise of Utopia through Artificial General Intelligence.” *Florida Marine Research Institute Technical Reports*, April. https://doi.org/10.5210/fm.v29i4.13636.  
Geneva, Nicholas, and Nicholas Zabaras. 2022. “Transformers for Modeling Physical Systems.” *Neural Networks: The Official Journal of the International Neural Network Society* 146 (February): 272–89.  
Gerla, Mario, Eun-Kyu Lee, Giovanni Pau, and Uichin Lee. 2014. “Internet of Vehicles: From Intelligent Grid to Autonomous Cars and Vehicular Clouds.” In *2014 IEEE World Forum on Internet of Things (WF-IoT)*, 241–46. IEEE.  
Ghadirzadeh, Ali, Xi Chen, Wenjie Yin, Zhengrong Yi, Mårten Björkman, and Danica Kragic. 2021. “Human-Centered Collaborative Robots With Deep Reinforcement Learning.” *IEEE Robotics and Automation Letters* 6 (2): 566–71.  
Ghani, Burooj, Tom Denton, Stefan Kahl, and Holger Klinck. 2023. “Global Birdsong Embeddings Enable Superior Transfer Learning for Bioacoustic Classification.” *Scientific Reports* 13 (1): 22876.  
Ghirlanda, Stefano, Johan Lind, and Magnus Enquist. 2017. “Memory for Stimulus Sequences: A Divide between Humans and Other Animals?” *Royal Society Open Science* 4 (6): 161011.  
Giannou, Angeliki, Shashank Rajput, Jy-Yong Sohn, Kangwook Lee, Jason D. Lee, and Dimitris Papailiopoulos. 23–29 Jul 2023. “Looped Transformers as Programmable Computers.” In *Proceedings of the 40th International Conference on Machine Learning*, edited by Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, 202:11398–442. Proceedings of Machine Learning Research. PMLR.  
Gibson, Kathleen R. 1993. “Tool Use, Language and Social Behavior in Relationship to Information Processing Capacities KATHLEEN R. GIBSON.” *Tools, Language and Cognition in Human Evolution*, 251.  
Gieseking, Jen Jack, William Mangold, Cindi Katz, Setha Low, and Susan Saegert. 2014. *The People, Place, and Space Reader*. Routledge.  
Glaser, Anton. 1971. *History of Binary and Other Nondecimal Numeration*. Tomash Publishers.  
Glimcher, Paul W. 2005. “Indeterminacy in Brain and Behavior.” *Annual Review of Psychology* 56: 25–56.  
Godfrey-Smith, Peter. 2016. *Other Minds: The Octopus, the Sea, and the Deep Origins of Consciousness*. Farrar, Straus and Giroux.  
———. 2020. *Metazoa: Animal Life and the Birth of the Mind*. Farrar, Straus and Giroux.  
———. 2024. *Living on Earth: Forests, Corals, Consciousness, and the Making of the World*. Farrar, Straus and Giroux.  
Goering, Sara, Eran Klein, Laura Specker Sullivan, Anna Wexler, Blaise Agüera y Arcas, Guoqiang Bi, Jose M. Carmena, et al. 2021. “Recommendations for Responsible Development and Application of Neurotechnologies.” *Neuroethics* 14 (3): 365–86.  
Goertzel, Ben. 2000. “The Extropian Creed: Can High Technology and Libertarian Politics Lead Us to a Transhuman Golden Age?” September 2000. https://www.goertzel.org/benzine/extropians.htm.  
Goldin, Pete. 2013. “Driver Not Required.” *iTS International* 19. https://trid.trb.org/View/1246562.  
Goodwin, Doris Kearns. 1991. *The Fitzgeralds and the Kennedys*. Macmillan.  
Gopnik, Alison, and Andrew Meltzoff. 1993. “Imitation, Cultural Learning and the Origins of ‘theory of Mind.’” *The Behavioral and Brain Sciences* 16 (3): 521–23.  
Gordon, Peter. 2004. “Numerical Cognition without Words: Evidence from Amazonia.” *Science* 306 (5695): 496–99.  
Gordon, Robert J. 2017. *The Rise and Fall of American Growth: The U.S. Standard of Living since the Civil War*. Princeton University Press.  
Goulty, Matthew, Gaelle Botton-Amiot, Ezio Rosato, Simon G. Sprecher, and Roberto Feuda. 2023. “The Monoaminergic System Is a Bilaterian Innovation.” *Nature Communications* 14 (1): 3284.  
Gowal, Sven, and Pushmeet Kohli. 2023. “Identifying AI-Generated Images with SynthID.” Google DeepMind. August 29, 2023. https://deepmind.google/discover/blog/identifying-ai-generated-images-with-synthid/.  
Grace, K., J. Salvatier, A. Dafoe, Baobao Zhang, and Owain Evans. 2017. “When Will AI Exceed Human Performance? Evidence from AI Experts.” *ArXiv* abs/1705.08807 (May). http://www.jair.org/index.php/jair/article/view/11222.  
Graeber, David. 2015. *The Utopia of Rules: On Technology, Stupidity, and the Secret Joys of Bureaucracy*. Melville House.  
Grandin, Temple. 1980. “Observations of Cattle Behavior Applied to the Design of Cattle-Handling Facilities.” *Applied Animal Ethology* 6 (1): 19–31.  
———. 1993. “Behavioral Agitation during Handling of Cattle Is Persistent over Time.” *Applied Animal Behaviour Science* 36 (1): 1–9.  
Grattan-Guinness, I. July-Sep 1990. “Work for the Hairdressers: The Production of de Prony’s Logarithmic and Trigonometric Tables.” *Annals of the History of Computing* 12 (3): 177–85.  
Graziano, Michael S. A. 2013. *Consciousness and the Social Brain*. Oxford University Press.  
———. 2018. *The Spaces Between Us: A Story of Neuroscience, Evolution, and Human Nature*. Oxford University Press.  
Greeno, J. G. 1994. “Gibson’s Affordances.” *Psychological Review* 101 (2): 336–42.  
Grier, David Alan. 2008. *When Computers Were Human*. Princeton, NJ: Princeton University Press. https://doi.org/10.1515/9781400849369.  
Gross, Charles G. 2002. “Genealogy of the ‘Grandmother Cell.’” *The Neuroscientist: A Review Journal Bringing Neurobiology, Neurology and Psychiatry* 8 (5): 512–18.  
Gunz, Philipp, Simon Neubauer, Bruno Maureille, and Jean-Jacques Hublin. 2010. “Brain Development after Birth Differs between Neanderthals and Modern Humans.” *Current Biology: CB* 20 (21): R921–22.  
Guo, Chong, Michael Lee, Guillaume Leclerc, Joel Dapello, Yug Rao, Aleksander Madry, and James Dicarlo. 17–23 Jul 2022. “Adversarially Trained Neural Representations Are Already as Robust as Biological Neural Representations.” In *Proceedings of the 39th International Conference on Machine Learning*, edited by Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato, 162:8072–81. Proceedings of Machine Learning Research. PMLR.  
Guth, Florentin, and Brice Ménard. 2024. “On the Universality of Neural Encodings in CNNs.” https://openreview.net/pdf?id=ofEBFOrITI.  
Guth, Florentin, Brice Ménard, Gaspar Rochette, and Stéphane Mallat. 2023. “A Rainbow in Deep Network Black Boxes.” *arXiv Preprint arXiv:2305.18512*. https://scholar.google.ca/scholar?cluster=4569530970169136210&hl=en&as\_sdt=0,5&sciodt=0,5.  
Haan, Edward H. F. de, Paul M. Corballis, Steven A. Hillyard, Carlo A. Marzi, Anil Seth, Victor A. F. Lamme, Lukas Volz, et al. 2020. “Split-Brain: What We Know Now and Why This Is Important for Understanding Consciousness.” *Neuropsychology Review* 30 (2): 224–33.  
[Hagendorff, Thilo, Sarah Fabi, and Michal Kosinski. 2022. “Thinking Fast and Slow in Large Language Models.” *arXiv [cs.CL]*. arXiv.](http://paperpile.com/b/iA68kr/PEAA) http://arxiv.org/abs/2212.05206.  
Hall, Lars, Thomas Strandberg, Philip Pärnamets, Andreas Lind, Betty Tärning, and Petter Johansson. 2013. “How the Polls Can Be Both Spot on and Dead Wrong: Using Choice Blindness to Shift Political Attitudes and Voter Intentions.” *PloS One* 8 (4): e60554.  
Hamilton, Andrew, Douglas R. Hofstadter, and Daniel Clement Dennett. 1984. “The Mind’s I: Fantasies and Reflections on Self and Soul.” *The Philosophical Quarterly* 34 (134): 80.  
Hanson, Alison. 2023. “On Being a Hydra With, and Without, a Nervous System: What Do Neurons Add?” *Animal Cognition* 26 (6): 1799–1816.  
Hao, Karen. 2020. “AI Still Doesn’t Have the Common Sense to Understand Human Language.” *MIT Technology Review*, January 31, 2020. https://www.technologyreview.com/2020/01/31/304844/ai-common-sense-reads-human-language-ai2/.  
Haraway, Donna J. 1985. “A Manifesto for Cyborgs: Science, Technology, and Socialist Feminism in the Late Twentieth Century.” *Socialist Review* 80: 65–108.  
Harland, D., and R. Jackson. 2000. “Eight-Legged Cats and How They See: A Review of Recent Research on Jumping Spiders (araneae: Salticidae).” *Cimbebasia* 16 (October): 231–40.  
Harper, Kristin N., Paolo S. Ocampo, Bret M. Steiner, Robert W. George, Michael S. Silverman, Shelly Bolotin, Allan Pillay, Nigel J. Saunders, and George J. Armelagos. 2008. “On the Origin of the Treponematoses: A Phylogenetic Approach.” *PLoS Neglected Tropical Diseases* 2 (1): e148.  
Harris, Annaka. 2019. *Conscious: A Brief Guide to the Fundamental Mystery of the Mind*. HarperCollins.  
Harris, Sam. 2012. *Free Will*. Bloomsbury Publishing.  
Hartley, David. 1834. *Observations on Man, His Frame, His Duty, and His Expectations*. T.Tegg and son.  
Hartmann, J. A., W. A. Wolz, D. P. Roeltgen, and F. L. Loverso. 1991. “Denial of Visual Perception.” *Brain and Cognition* 16 (1): 29–40.  
Hart, O. van der, R. Lierens, and J. Goodwin. 1996. “Jeanne Fery: A Sixteenth-Century Case of Dissociative Identity Disorder.” *The Journal of Psychohistory* 24 (1): 18–35.  
Hassabis, Demis, Dharshan Kumaran, Seralynne D. Vann, and Eleanor A. Maguire. 2007. “Patients with Hippocampal Amnesia Cannot Imagine New Experiences.” *Proceedings of the National Academy of Sciences of the United States of America* 104 (5): 1726–31.  
Haugeland, John. 1985. *Artificial Intelligence: The Very Idea*. Bradford books.  
“Hawaii HB693.” n.d. LegiScan. Accessed May 4, 2024. https://legiscan.com/HI/text/HB693/id/2456204.  
Hawking, Stephen. 2011. *A Brief History of Time*. Bantam. https://books.google.com/books?id=oZhagX6UWOMC&newbks=1.  
Hawkins, Jeff. 2021. *A Thousand Brains: A New Theory of Intelligence*. Basic Books.  
Hayes, Carol Mullins. 2023. “Generative Artificial Intelligence and Copyright: Both Sides of the Black Box.” *Available at SSRN 4517799*. https://doi.org/10.2139/ssrn.4517799.  
Heath, R. G. 1963. “ELECTRICAL SELF-STIMULATION OF THE BRAIN IN MAN.” *The American Journal of Psychiatry* 120 (December): 571–77.  
Heaven, Douglas. 2019. “Deep Trouble for Deep Learning.” *Nature* 574 (7777): 163–66.  
Heffernan, Virginia. 2004. “The Cheerful Confessions Of Ali G, Borat And Bruno.” *The New York Times*, July 15, 2004. https://www.nytimes.com/2004/07/15/arts/the-cheerful-confessions-of-ali-g-borat-and-bruno.html.  
Heggarty, Paul, Cormac Anderson, Matthew Scarborough, Benedict King, Remco Bouckaert, Lechosław Jocz, Martin Joachim Kümmel, et al. 2023. “Language Trees with Sampled Ancestors Support a Hybrid Model for the Origin of Indo-European Languages.” *Science* 381 (6656): eabg0818.  
Heinlein, Robert Anson. 1966. *The Moon Is a Harsh Mistress*. Putnam.  
Heinrich, Stefan, Yuan Yao, Tobias Hinz, Zhiyuan Liu, Thomas Hummel, Matthias Kerzel, Cornelius Weber, and Stefan Wermter. 2020. “Crossmodal Language Grounding in an Embodied Neurocognitive Model.” *Frontiers in Neurorobotics* 14 (October): 52.  
[He, Kaiming, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. 2021. “Masked Autoencoders Are Scalable Vision Learners.” *arXiv [cs.CV]*. arXiv. https://doi.org/](http://paperpile.com/b/iA68kr/eiV7)10.1109/CVPR52688.2022.01553.  
He, K., X. Zhang, S. Ren, and J. Sun. 2016. “Deep Residual Learning for Image Recognition.” *Proceedings of the IEEE*. http://openaccess.thecvf.com/content\_cvpr\_2016/html/He\_Deep\_Residual\_Learning\_CVPR\_2016\_paper.html.  
Helmholtz, Hermann von. 1925. *Helmholtz’s Treatise on Physiological Optics*. Optical Society of America.  
Hendler, J. 2006. “Introducing the Future of AI.” *IEEE Intelligent Systems* 21 (3): 2–4.  
Henneberg, M. 1988. “Decrease of Human Skull Size in the Holocene.” *Human Biology* 60 (3): 395–405.  
Henneberg, M., and M. Steyn. 1993. “Trends in Cranial Capacity and Cranial Index in Subsaharan Africa during the Holocene.” *American Journal of Human Biology: The Official Journal of the Human Biology Council* 5 (4): 473–79.  
Henrich, Joseph, Steven J. Heine, and Ara Norenzayan. 2010. “The Weirdest People in the World?” *The Behavioral and Brain Sciences* 33 (2-3): 61–83; discussion 83–135.  
Henschen, S. E. 1893. “ON THE VISUAL PATH AND CENTRE.” *Brain: A Journal of Neurology* 16 (1-2): 170–80.  
Hillis, William Daniel. 1985. *The Connection Machine*. MIT Press.  
Hochreiter, S., and J. Schmidhuber. 1997. “Long Short-Term Memory.” *Neural Computation* 9 (8): 1735–80.  
[Hoffmann, Jordan, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, et al. 2022. “Training Compute-Optimal Large Language Models.” *arXiv [cs.CL]*. arXiv.](http://paperpile.com/b/iA68kr/dYRv) http://arxiv.org/abs/2203.15556.  
Hoffman, Reid. 2023. *Impromptu: Amplifying Our Humanity Through AI*. Dallepedia LLC.  
Hofstadter, Douglas R. 1982. “METAMAGICAL THEMAS.” *Scientific American* 247 (3): 18–M18.  
———. 2007. *I Am a Strange Loop*. Basic Books.  
Holldobler, Bert, and Edward O. Wilson. 2009. *The Superorganism: The Beauty Elegance And Strangeness Of Insect Societies*. W. W. Norton & Company.  
Holt-Lunstad, Julianne, Timothy B. Smith, Mark Baker, Tyler Harris, and David Stephenson. 2015. “Loneliness and Social Isolation as Risk Factors for Mortality: A Meta-Analytic Review.” *Perspectives on Psychological Science: A Journal of the Association for Psychological Science* 10 (2): 227–37.  
Hong, Sungho, Blaise Agüera y Arcas, and Adrienne L. Fairhall. 2007. “Single Neuron Computation: From Dynamical System to Feature Detector.” *Neural Computation* 19 (12): 3133–72.  
Honkanen, Anna, Ronja Hensgen, Kavitha Kannan, Andrea Adden, Eric Warrant, William Wcislo, and Stanley Heinze. 2023. “Parallel Motion Vision Pathways in the Brain of a Tropical Bee.” *Journal of Comparative Physiology. A, Neuroethology, Sensory, Neural, and Behavioral Physiology* 209 (4): 563–91.  
Honts, C. R., D. C. Raskin, and J. C. Kircher. 1994. “Mental and Physical Countermeasures Reduce the Accuracy of Polygraph Tests.” *The Journal of Applied Psychology* 79 (2): 252–59.  
Hopper, Paul J. 1996. “SOME RECENT TRENDS IN GRAMMATICALIZATION.” *Annual Review of Anthropology* 25 (Volume 25, 1996): 217–36.  
Horita, Nobuyuki, and Takeshi Fukumoto. 2023. “Global Case Fatality Rate from COVID-19 Has Decreased by 96.8% during 2.5 Years of the Pandemic.” *Journal of Medical Virology* 95 (1): e28231.  
Hornik, Kurt, Maxwell Stinchcombe, and Halbert White. 1989. “Multilayer Feedforward Networks Are Universal Approximators.” *Neural Networks: The Official Journal of the International Neural Network Society* 2 (5): 359–66.  
*How Smart Are Dolphins?* 2011. Nova ScienceNOW. https://www.pbs.org/wgbh/nova/video/how-smart-are-dolphins/.  
Hrdy, Sarah Blaffer. 2009. *Mothers and Others*. Harvard University Press.  
Hsiao, Elaine C. 2012. “Whanganui River Agreement-Indigenous Rights and Rights of Nature.” *Envtl. Pol’y & L.* 42: 371.  
Huang, Gao, Zhuang Liu, and Kilian Q. Weinberger. 2016. “Densely Connected Convolutional Networks.” *Proceedings / CVPR, IEEE Computer Society Conference on Computer Vision and Pattern Recognition. IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, August, 2261–69.  
Hubel, David H., and Torsten N. Wiesel. 2004. *Brain and Visual Perception: The Story of a 25-Year Collaboration*. Oxford University Press.  
[Hu, Edward J., Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. “LoRA: Low-Rank Adaptation of Large Language Models.” *arXiv [cs.CL]*. arXiv.](http://paperpile.com/b/iA68kr/EQTz) http://arxiv.org/abs/2106.09685.  
Hume, David. 1817. *A Treatise of Human Nature: Being an Attempt to Introduce the Experimental Method of Reasoning Into Moral Subjects*. Thomas and Joseph Allman … and sold.  
Humphrey, Nicholas. 1972. “Seeing and Nothingness.” *New Scientist* 53 (789): 682–84.  
———. 1976. “The Social Function of Intellect.” In *Growing Points in Ethology*, edited by Patrick P. G. Bateson and Robert A. Hinde. Cambridge University Press.  
———. 2023. *Sentience: The Invention of Consciousness*. MIT Press.  
Humphrey, Nicholas, and Larry Weiskrantz. 1967. “Vision in Monkeys after Removal of the Striate Cortex.” *Nature* 215 (5101): 595–97.  
Hunter, Kevin, Lawrence Spracklen, and Subutai Ahmad. 2022. “Two Sparsities Are Better than One: Unlocking the Performance Benefits of Sparse–sparse Networks.” *Neuromorphic Computing and Engineering* 2 (3): 034004.  
Huttenlocher, P. R. 1979. “Synaptic Density in Human Frontal Cortex - Developmental Changes and Effects of Aging.” *Brain Research* 163 (2): 195–205.  
Hynninen, Meri Katariina. n.d. “Vocalizing Behavior in Killer Whales (Orcinus Orca)–an Evolutionary Account.” https://skemman.is/handle/1946/40990.  
Iacono, William G. 2008. “Accuracy of Polygraph Techniques: Problems Using Confessions to Determine Ground Truth.” *Physiology & Behavior* 95 (1): 24–26.  
Ibbotson, Paul, and Michael Tomasello. 2016. “Evidence Rebuts Chomsky’s Theory of Language Learning.” *Scientific American*, September. https://www.scientificamerican.com/article/evidence-rebuts-chomsky-s-theory-of-language-learning/.  
IBM. 2007. “Frequently Asked Questions.” https://web.archive.org/web/20070621002335/http://www-03.ibm.com/ibm/history/documents/pdf/faq.pdf.  
Isaacs, T. Craig. 1987. “The Possessive States Disorder: The Diagnosis of Demonic Possession.” *Pastoral Psychology* 35 (4): 263–73.  
Jackson, Frank. 1982. “Epiphenomenal Qualia.” *The Philosophical Quarterly* 32 (127): 127–36.  
Jackson, R. R., and S. D. Pollard. 1996. “Predatory Behavior of Jumping Spiders.” *Annual Review of Entomology* 41: 287–308.  
Jacobsen, Annie. 2024. *Nuclear War: A Scenario*. Penguin.  
Jacobs, Jane. 1961. *The Death and Life of Great American Cities*. New York: Random House.  
Jacques, Vincent, E. Wu, Frédéric Grosshans, François Treussart, Philippe Grangier, Alain Aspect, and Jean-François Roch. 2007. “Experimental Realization of Wheeler’s Delayed-Choice Gedanken Experiment.” *Science* 315 (5814): 966–68.  
James, William. 1890. “The Principles Of Psychology Volume II By William James (1890).” http://library.manipaldubai.com/DL/the\_principles\_of\_psychology\_vol\_II.pdf.  
Jingna, and Photographer Jingna Zhang. 2024. “Lee Sedol, Art, & Go - Jingna Zhang Fashion, Fine Art & Beauty Photography.” Jingna Zhang. March 20, 2024. https://www.zhangjingna.com/blog/2024/3/20/lee-sedol-alphago-art-and-go.  
Johansson, Petter, Lars Hall, Sverker Sikström, and Andreas Olsson. 2005. “Failure to Detect Mismatches between Intention and Outcome in a Simple Decision Task.” *Science* 310 (5745): 116–19.  
Jolly, A. 1966. “Lemur Social Behavior and Primate Intelligence.” *Science* 153 (3735): 501–6.  
Jones, Cameron R., Tyler A. Chang, Seana Coulson, James A. Michaelov, Sean Trott, and Benjamin Bergen. 2022. “Distributional Semantics Still Can’t Account for Affordances.” *Proceedings of the Annual Meeting of the Cognitive Science Society* 44 (44). https://escholarship.org/uc/item/44z7r3j3.  
[Kadavath, Saurav, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, et al. 2022. “Language Models (Mostly) Know What They Know.” *arXiv [cs.CL]*. arXiv.](http://paperpile.com/b/iA68kr/9SI8) http://arxiv.org/abs/2207.05221.  
Kahneman, Daniel. 2011. *Thinking, Fast and Slow*. Farrar, Straus and Giroux.  
Kahng, Jee Heun, and Se Young Lee. 2016. “Google Artificial Intelligence Programme Beats South Korean Go pro with 4-1 Score.” *Reuters*, March 15, 2016. https://www.reuters.com/article/us-science-intelligence-go-idUSKCN0WH0XJ/.  
Kant, Immanuel. 1991. *Prolegomena to any Future Metaphysics*. Newcomb Livraria Press.  
Kaplan, Gisela. 2023. “Evolution of Human Language: Duetting as Part of Prosociality and Cognition.” *Frontiers in Ecology and Evolution* 11. https://doi.org/10.3389/fevo.2023.1004384.  
[Kaplan, Jared, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. “Scaling Laws for Neural Language Models.” *arXiv [cs.LG]*. arXiv.](http://paperpile.com/b/iA68kr/Y24B) http://arxiv.org/abs/2001.08361.  
Kauffman, Stuart A. 1971. “Cellular Homeostasis, Epigenesis and Replication in Randomly Aggregated Macromolecular Systems.” *Journal of Cybernetics* 1 (1): 71–96.  
Kay, Kendrick N., Thomas Naselaris, Ryan J. Prenger, and Jack L. Gallant. 2008. “Identifying Natural Images from Human Brain Activity.” *Nature* 452 (7185): 352–55.  
Keijzer, Fred. 2013. “The Sphex Story: How the Cognitive Sciences Kept Repeating an Old and Questionable Anecdote.” *Philosophical Psychology* 26 (4): 502–19.  
Keller, Georg B., and Thomas D. Mrsic-Flogel. 2018. “Predictive Processing: A Canonical Cortical Computation.” *Neuron* 100 (2): 424–35.  
Killgore, William D. S. 2010. “Effects of Sleep Deprivation on Cognition.” *Progress in Brain Research* 185: 105–29.  
Kilner, James M., Karl J. Friston, and Chris D. Frith. 2007. “Predictive Coding: An Account of the Mirror Neuron System.” *Cognitive Processing* 8 (3): 159–66.  
Kimmerer, Robin. 2013. *Braiding Sweetgrass: Indigenous Wisdom, Scientific Knowledge and the Teachings of Plants*. Minneapolis, MN: Milkweed Editions.  
———. 2017. “Orion Magazine - Speaking of Nature.” Orion Magazine. June 12, 2017. https://orionmagazine.org/article/speaking-of-nature/.  
King, Martin Luther, Jr. 1967. *Where Do We Go from Here: Chaos or Community?* HarperCollins.  
[Kirchenbauer, John, Jonas Geiping, Yuxin Wen, Jonathan Katz, Ian Miers, and Tom Goldstein. 2023. “A Watermark for Large Language Models.” *arXiv [cs.LG]*. arXiv.](http://paperpile.com/b/iA68kr/pe05) http://arxiv.org/abs/2301.10226.  
Kleene, Stephen Cole. 1938. “On Notation for Ordinal Numbers.” *Journal of Symbolic Logic* 3 (4): 150–55.  
Kleinman, Zoe. 2024. “Why Google’s ‘Woke’ AI Problem Won’t Be an Easy Fix.” *BBC*, February 28, 2024. https://www.bbc.com/news/technology-68412620.  
Kline, Ronald. 2020. “How Disunity Matters to the History of Cybernetics in the Human Sciences in the United States, 1940–80.” *History of the Human Sciences* 33 (1): 12–35.  
Kobayashi, H., and S. Kohshima. 1997. “Unique Morphology of the Human Eye.” *Nature* 387 (6635): 767–68.  
Koch, Christof. 2008. “What Is It Like to Be a Bee?” *Scientific American Mind* 19 (6): 18–19.  
[Kocijan, Vid, Ernest Davis, Thomas Lukasiewicz, Gary Marcus, and Leora Morgenstern. 2022. “The Defeat of the Winograd Schema Challenge.” *arXiv [cs.CL]*. arXiv.](http://paperpile.com/b/iA68kr/u4g7) http://arxiv.org/abs/2201.02387.  
Koh, Jing Yu, Daniel Fried, and R. Salakhutdinov. 2023. “Generating Images with Multimodal Language Models.” *Advances in Neural Information Processing Systems* abs/2305.17216 (May). https://doi.org/10.48550/arXiv.2305.17216.  
Kojima, Takeshi, S. Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. “Large Language Models Are Zero-Shot Reasoners.” *Advances in Neural Information Processing Systems* abs/2205.11916 (May). https://proceedings.neurips.cc/paper\_files/paper/2022/hash/8bb0d291acd4acf06ef112099c16f326-Abstract-Conference.html.  
Kolarik, Andrew J., Silvia Cirstea, Shahina Pardhan, and Brian C. J. Moore. 2014. “A Summary of Research Investigating Echolocation Abilities of Blind and Sighted Humans.” *Hearing Research* 310 (April): 60–68.  
Konstantinidis, Konstantinos T., Alban Ramette, and James M. Tiedje. 2006. “The Bacterial Species Definition in the Genomic Era.” *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences* 361 (1475): 1929–40.  
Kozachkov, Leo, Ksenia V. Kastanenka, and Dmitry Krotov. 2023. “Building Transformers from Neurons and Astrocytes.” *Proceedings of the National Academy of Sciences of the United States of America* 120 (34): e2219150120.  
Krebs, Hans Adolf, and William Arthur Johnson. 1937. “The Role of Citric Acid in Intermediate Metabolism in Animal Tissues.” *Enzymologia* 4 (1): 148–56.  
Krizhevsky, A., I. Sutskever, and Geoffrey E. Hinton. 2012. “ImageNet Classification with Deep Convolutional Neural Networks.” *Communications of the ACM* 60 (December): 84–90.  
Kuczmarski, James. 2018. “Reducing Gender Bias in Google Translate.” Google. December 6, 2018. https://blog.google/products/translate/reducing-gender-bias-google-translate/.  
Kudo, Taku, and Yuji Matsumoto. 2001. “Chunking with Support Vector Machines.” In *Second Meeting of the North AMerican Chapter of the Association for Computational Linguistics*. aclanthology.org. https://aclanthology.org/N01-1025.  
Kurzweil, Ray. 2005. *The Singularity Is Near: When Humans Transcend Biology*. Penguin.  
———. 2024. *The Singularity Is Nearer: When We Merge with AI*. New York, NY: Penguin.  
KUSA-TV. 2005. “Man Survives Nail Shot into Head.” KUSA. January 17, 2005. https://www.9news.com/article/news/health/man-survives-nail-shot-into-head/73-344804059.  
Kuuspalu, Adam, Samantha Cody, and Melina E. Hale. 2022. “Multiple Nerve Cords Connect the Arms of Octopuses, Providing Alternative Paths for Inter-Arm Signaling.” *Current Biology: CB* 32 (24): 5415–21.e3.  
Labatut, Benjamín. 2024. “The Gods of Logic: Before and after Artificial Intelligence.” *Harper’s Magazine*, July. https://harpers.org/archive/2024/07/the-gods-of-logic-benjamin-labatut-ai/.  
Laird, John E., Christian Lebiere, and Paul S. Rosenbloom. 2017. “A Standard Model of the Mind: Toward a Common Computational Framework across Artificial Intelligence, Cognitive Science, Neuroscience, and Robotics.” *AI Magazine* 38 (4): 13–26.  
La Mettrie, Julien Offray de. 1748a. *L’homme machine*. Luzac.  
———. 1748b. *L’homme-plante*. Voss.  
Lander, Eric S., L. M. Linton, B. Birren, C. Nusbaum, M. C. Zody, J. Baldwin, K. Devon, et al. 2001. “Initial Sequencing and Analysis of the Human Genome.” *Nature* 409 (6822): 860–921.  
Lane, Nick. 2022. *Transformer: The Deep Chemistry of Life and Death*. W. W. Norton & Company.  
Lanier, Kathryn A., Anton S. Petrov, and Loren Dean Williams. 2017. “The Central Symbiosis of Molecular Biology: Molecules in Mutualism.” *Journal of Molecular Evolution* 85 (1-2): 8–13.  
Lanuza, Jose B., Miguel Á. Collado, Ferran Sayol, Daniel Sol, and Ignasi Bartomeus. 2023. “Brain Size Predicts Bees’ Tolerance to Urban Environments.” *Biology Letters* 19 (11): 20230296.  
[Layfield, Eleanor, Tiffany Peng Hwa, James Naples, Ivy Maina, Jason A. Brant, Steven J. Eliades, Douglas C. Bigelow, and Michael J. Ruckenstein. 2021. “Failure and Revision Surgery After Cochlear Implantation in the Adult Population: A 10-Year Single-Institution Retrospective and Systematic Review of the Literature.” *Otology & Neurotology: Official Publication of the American Otological Society, American Neurotology Society [and] European Academy of Otology and Neurotology* 42 (3): 408–13.](http://paperpile.com/b/iA68kr/vZVJ)  
Lazova, Milena D., Tanvir Ahmed, Domenico Bellomo, Roman Stocker, and Thomas S. Shimizu. 2011. “Response Rescaling in Bacterial Chemotaxis.” *Proceedings of the National Academy of Sciences of the United States of America* 108 (33): 13870–75.  
Leaf, Munro. 1946. *How to Behave and Why*. Philadelphia: J.B. Lippincott.  
Leamer, Laurence. 1996. *The Kennedy Women: The Saga of an American Family*. Random House Publishing Group.  
Lebovitz, Sarah, N. Levina, and Hila Lifshitz-Assaf. 2021. “Is AI Ground Truth Really True? The Dangers of Training and Evaluating AI Tools Based on Experts’ Know-What.” *The Mississippi Quarterly* 45 (June). https://doi.org/10.25300/MISQ/2021/16564.  
LeCun, Yann, Bernhard Boser, John S. Denker, Donnie Henderson, Richard E. Howard, Wayne Hubbard, and Lawrence D. Jackel. 1989. “Backpropagation Applied to Handwritten Zip Code Recognition.” *Neural Computation* 1 (4): 541–51.  
LeCun, Yann, Leon Bottou, Yoshua Bengio, and Patrick G. Haffner. 1998. “Gradient-Based Learning Applied to Document Recognition.” *Proceedings of the IEEE* 86 (11): 2278–2324.  
LeCun, Yann, Lawrence Jackel, Léon Bottou, Corinna Cortes, John S. Denker, Harris Drucker, Isabelle M. Guyon, et al. 1995. “Learning Algorithms for Classification: A Comparison on Handwritten Digit Recognition.” *Neural Networks: The Statistical Mechanics Perspective*, 261–76.  
LeCun, Yann, Lawrence D. Jackel, Bernhard Boser, John S. Denker, Hans P. Graf, Isabelle M. Guyon, Donnie Henderson, Richard E. Howard, and Wayne Hubbard. 1990. “Handwritten Digit Recognition: Applications of Neural Net Chips and Automatic Learning.” In *Neurocomputing*, 303–18. Springer Berlin Heidelberg.  
Lee, Albert K., and Matthew A. Wilson. 2002. “Memory of Sequential Experience in the Hippocampus during Slow Wave Sleep.” *Neuron* 36 (6): 1183–94.  
[Lee, Chang-Shing, Mei-Hui Wang, Shi-Jim Yen, Ting-Han Wei, I-Chen Wu, Ping-Chiang Chou, Chun-Hsun Chou, Ming-Wan Wang, and Tai-Hsiung Yan. 2016. “Human vs. Computer Go: Review and Prospect [Discussion Forum].” *IEEE Computational Intelligence Magazine* 11 (3): 67–72.](http://paperpile.com/b/iA68kr/F7YW)  
Lee, Reena. 2016. “Gboard, Now Available for Android.” Google. December 16, 2016. https://blog.google/products/search/gboard-now-on-android/.  
[Leibniz, Gottfried Wilhelm. 1666. *Dissertatio de Arte Combinatoria, in qua Ex Arithmeticae Fundamentis Complicationum Ac Transpositionum Doctrina Novis Praeceptis Exstruitur, [et] Usus Ambarum per Universum Scientiarum Orbem Ostenditur ; Nova Etiam Artis Meditandi, Seu Logicae Inventionis Semina Sparguntur; Praefixa Est Synopsis Totius Tractatus, [et] Additamenti Loco Demonstratio Existentiae Dei, Ad Mathematicam Certitudinem Exacta*. apud Joh. Simon. Fickium et Joh. Polycarp. Seuboldum, Literis Spörelianis.](http://paperpile.com/b/iA68kr/oP6pX)  
Lenat, Douglas B., Mayank Prakash, and Mary Shepherd. 1985. “CYC: Using Common Sense Knowledge to Overcome Brittleness and Knowledge Acquisition Bottlenecks.” *AI Magazine* 6 (4): 65–65.  
Lenton, Timothy M., and James E. Lovelock. 2001. “Daisyworld Revisited: Quantifying Biological Effects on Planetary Self?regulation.” *Tellus. Series B, Chemical and Physical Meteorology* 53 (3): 288–305.  
LessWrong. n.d. “Really Powerful Optimization Process.” LessWrong. Accessed May 5, 2024. https://www.lesswrong.com/tag/really-powerful-optimization-process.  
Levesque, H., E. Davis, and L. Morgenstern. 2011. “The Winograd Schema Challenge.” *International Conference on the …*, March. https://cdn.aaai.org/ocs/4492/4492-21843-1-PB.pdf.  
Levinson, Marc. 2016. *An Extraordinary Time: The End of the Postwar Boom and the Return of the Ordinary Economy*. Basic Books.  
Lewis, Juan P. 2013. “Did Varro Think That Slaves Were Talking Tools?” *Mnemosyne* 66 (4-5): 634–48.  
Lewis, Peter J. 2021. “Collapse Theories.” *The Routledge Companion to Philosophy of Physics*, August. https://doi.org/10.4324/9781315623818-22.  
Libet, Benjamin, Curtis A. Gleason, Elwood W. Wright, and Dennis K. Pearl. 1993. “Time of Conscious Intention to Act in Relation to Onset of Cerebral Activity (Readiness-Potential).” In *Neurophysiology of Consciousness*, edited by Benjamin Libet, 249–68. Boston, MA: Birkhäuser Boston.  
Light, Jennifer S. 1999. “When Computers Were Women.” *Technology and Culture* 40 (3): 455–83.  
[Li, Jiaxuan, Lang Yu, and Allyson Ettinger. 2022. “Counterfactual Reasoning: Do Language Models Need World Knowledge for Causal Understanding?” *arXiv [cs.CL]*. arXiv.](http://paperpile.com/b/iA68kr/Umyd) http://arxiv.org/abs/2212.03278.  
[Li, Kenneth, Aspen K. Hopkins, David Bau, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. 2022. “Emergent World Representations: Exploring a Sequence Model Trained on a Synthetic Task.” *arXiv [cs.LG]*. arXiv.](http://paperpile.com/b/iA68kr/CSXy) http://arxiv.org/abs/2210.13382.  
Lind, John. 1776. *An Answer to the Declaration of the American Congress*. T. Cadell; J. Walter; and T. Sewell.  
Livingstone, Margaret. 2013. “David Hubel 1926–2013.” *Cell* 155 (4): 735–37.  
[Li, Zhongli, Wenxuan Zhang, Chao Yan, Qingyu Zhou, Chao Li, Hongzhi Liu, and Yunbo Cao. 2021. “Seeking Patterns, Not Just Memorizing Procedures: Contrastive Learning for Solving Math Word Problems.” *arXiv [cs.CL]*. arXiv.](http://paperpile.com/b/iA68kr/T25f) http://arxiv.org/abs/2110.08464.  
[Li, Zonglin, Chong You, Srinadh Bhojanapalli, Daliang Li, Ankit Singh Rawat, Sashank J. Reddi, Ke Ye, et al. 2022. “The Lazy Neuron Phenomenon: On Emergence of Activation Sparsity in Transformers.” *arXiv [cs.LG]*. arXiv.](http://paperpile.com/b/iA68kr/1BKS) http://arxiv.org/abs/2210.06313.  
Logie, Robert H. 2011. “The Functional Organization and Capacity Limits of Working Memory.” *Current Directions in Psychological Science* 20 (4): 240–45.  
Loizeau, Nicolas. 2021. “A Computer in Conway’s Game of Life.” GitHub. 2021. https://github.com/nicolasloizeau/scalable-gol-computer.  
[Long, Jieyi. 2023. “Large Language Model Guided Tree-of-Thought.” *arXiv [cs.AI]*. arXiv.](http://paperpile.com/b/iA68kr/UZxX) http://arxiv.org/abs/2305.08291.  
Lorenz, Edward N. 1963. “Deterministic Nonperiodic Flow.” *Journal of the Atmospheric Sciences* 20 (2): 130–41.  
Lovelace, Ada. 1842. “Sketch of the Analytical Engine Invented by Charles Babbage, by LF Menabrea, Officer of the Military Engineers, with Notes by the Translator.” *Taylor’s Scientific Memoirs*.  
———. 1992. *Ada, the Enchantress of Numbers: A Selection from the Letters of Lord Byron’s Daughter and Her Description of the First Computer*. Edited by Betty Alexandra Toole. Strawberry Press.  
Lovelock, James E., and Lynn Margulis. 1974. “Atmospheric Homeostasis by and for the Biosphere: The Gaia Hypothesis.” *Tell’Us* 26 (1-2): 2–10.  
Lucas, Keith, and Edgar Douglas Adrian Adrian. 1917. *The Conduction of the Nervous Impulse*. Longmans, Green and Company.  
MaBouDi, Hadi, H. Samadi Galpayage Dona, Elia Gatto, Olli J. Loukola, Emma Buckley, Panayiotis D. Onoufriou, Peter Skorupski, and Lars Chittka. 2020. “Bumblebees Use Sequential Scanning of Countable Items in Visual Patterns to Solve Numerosity Tasks.” *Integrative and Comparative Biology* 60 (4): 929–42.  
MacAskill, William. 2022. *What We Owe The Future*. Simon and Schuster.  
MacKay, Donald Maccrimmon. 1969. *Information, Mechanism and Meaning*. MIT Press.  
MacKenzie, Donald. 1993. *Inventing Accuracy: A Historical Sociology of Nuclear Missile Guidance*. MIT Press.  
Magnasco, Marcelo O. 2022. “Robustness and Flexibility of Neural Function through Dynamical Criticality.” *Entropy* 24 (5). https://doi.org/10.3390/e24050591.  
Mandelbrot, Benoît. 1967. “How Long Is the Coast of Britain? Statistical Self-Similarity and Fractional Dimension.” *Science* 156 (3775): 636–38.  
———. 1989. “Multifractal Measures, Especially for the Geophysicist.” In *Fractals in Geophysics*, 5–42. Basel: Birkhäuser Basel.  
Marcus, Gary. 2020. “GPT-2 and the Nature of Intelligence.” The Gradient. January 25, 2020. https://thegradient.pub/gpt2-and-the-nature-of-intelligence/.  
Margulis, Lynn. 2004. “On Syphilis & Nietzsche’s Madness: Spirochetes Awake!” *Daedalus* 133 (4): 118–25.  
Margulis, Lynn, and Dorion Sagan. 2007. *Dazzle Gradually: Reflections on the Nature of Nature*. Chelsea Green Publishing.  
[Marjieh, Raja, Ilia Sucholutsky, Pol van Rijn, Nori Jacoby, and Thomas L. Griffiths. 2023. “Large Language Models Predict Human Sensory Judgments across Six Modalities.” *arXiv [cs.CL]*. arXiv.](http://paperpile.com/b/iA68kr/TWMN) http://arxiv.org/abs/2302.01308.  
Marletto, Chiara. 2015. “Constructor Theory of Life.” *Journal of the Royal Society, Interface / the Royal Society* 12 (104): 20141226.  
[———. 2016. “Constructor Theory of Thermodynamics.” *arXiv [physics.class-Ph]*. arXiv.](http://paperpile.com/b/iA68kr/08lI) http://arxiv.org/abs/1608.02625.  
———. 2022. *The Science of Can and Can’t: A Physicist’s Journey through the Land of Counterfactuals*. Penguin.  
Martin, Ernst. 1992. *The Calculating Machines: Their History and Development*. Translated by Peggy Aldrich Kidwell and Michael R. Williams. Charles Babbage Institute Reprint, v. 16. London, England: MIT Press.  
Matthews, P. B. C. 1982. “Where Does Sherrington’s ‘Muscular Sense’ Originate? Muscles, Joints, Corollary Discharges?” *Annual Review of Neuroscience* 5 (1): 189–218.  
[Ma, Xuezhe, Xiaomeng Yang, Wenhan Xiong, Beidi Chen, Lili Yu, Hao Zhang, Jonathan May, Luke Zettlemoyer, Omer Levy, and Chunting Zhou. 2024. “Megalodon: Efficient LLM Pretraining and Inference with Unlimited Context Length.” *arXiv [cs.LG]*. arXiv.](http://paperpile.com/b/iA68kr/otXq) http://arxiv.org/abs/2404.08801.  
Mazokhin-Porshnyakov, George A., and Vladimir M. Kartsev. 2000. “Learning in Bees and Wasps in Complicated Experimental Tasks.” In *Prerational Intelligence: Adaptive Behavior and Intelligent Systems Without Symbols and Logic, Volume 1, Volume 2 Prerational Intelligence: Interdisciplinary Perspectives on the Behavior of Natural and Artificial Systems, Volume 3*, edited by Holk Cruse, Jeffrey Dean, and Helge Ritter, 908–26. Dordrecht: Springer Netherlands.  
McCarthy, John, Marvin Minsky, Aaron Sloman, and Leiguang Gong. 2002. “An Architecture of Diversity for Commonsense Reasoning.” *IBM Systems Journal* 41 (3): 530–39.  
McClintock, Barbara. 1950. “The Origin and Behavior of Mutable Loci in Maize.” *Proceedings of the National Academy of Sciences* 36 (6): 344–55.  
McCulloch, Warren S., and Walter Pitts. 1943. “A Logical Calculus of the Ideas Immanent in Nervous Activity.” *The Bulletin of Mathematical Biophysics* 5 (4): 115–33.  
McDonald, Michael J., Daniel P. Rice, and Michael M. Desai. 2016. “Sex Speeds Adaptation by Altering the Dynamics of Molecular Evolution.” *Nature* 531 (7593): 233–36.  
Meadows, Donella H., Dennis Meadows, Jørgen Randers, and William W. Behrens III. 1972. *The Limits to Growth*. Penguin Group (USA) Incorporated.  
Meaker, Morgan. 2023. “These Prisoners Are Training AI.” *Wired*, September. https://www.wired.com/story/prisoners-training-ai-finland/.  
Meissner, Christian A., and John C. Brigham. 2001. “Thirty Years of Investigating the Own-Race Bias in Memory for Faces: A Meta-Analytic Review.” *Psychology, Public Policy, and Law: An Official Law Review of the University of Arizona College of Law and the University of Miami School of Law* 7 (1): 3–35.  
Mentzakis, Emmanouil, and Mirko Moro. 2009. “The Poor, the Rich and the Happy: Exploring the Link between Income and Subjective Well-Being.” *The Journal of Socio-Economics* 38 (1): 147–58.  
Merchant, Brian. 2023. *Blood in the Machine: The Origins of the Rebellion Against Big Tech*. Little, Brown.  
Mercier, Hugo, and Dan Sperber. 2011. “Why Do Humans Reason? Arguments for an Argumentative Theory.” *The Behavioral and Brain Sciences* 34 (2): 57–74; discussion 74–111.  
———. 2018. *The Enigma of Reason*. Harvard University Press.  
Merken, Sara. 2023. “New York Lawyers Sanctioned for Using Fake ChatGPT Cases in Legal Brief.” *Reuters*.  
Merker, Bjorn. 2007. “Consciousness without a Cerebral Cortex: A Challenge for Neuroscience and Medicine.” *The Behavioral and Brain Sciences* 30 (1): 63–81; discussion 81–134.  
[Merrill, William, and Ashish Sabharwal. 2023. “The Expressive Power of Transformers with Chain of Thought.” *arXiv [cs.LG]*. arXiv.](http://paperpile.com/b/iA68kr/tpUL) http://arxiv.org/abs/2310.07923.  
Merrill, William, Zhaofeng Wu, Norihito Naka, Yoon Kim, and Tal Linzen. 2024. “Can You Learn Semantics through next-Word Prediction? The Case of Entailment.” *ArXiv* abs/2402.13956 (February). https://doi.org/10.48550/arXiv.2402.13956.  
Michod, Richard E. 2000. *Darwinian Dynamics: Evolutionary Transitions in Fitness and Individuality*. Princeton, NJ: Princeton University Press.  
Mikolov, Tomáš, Ilya Sutskever, Kai Chen, Greg S. Corrado, and Jeff Dean. 2013. “Distributed Representations of Words and Phrases and Their Compositionality.” *Advances in Neural Information Processing Systems* 26. https://proceedings.neurips.cc/paper/2013/hash/9aa42b31882ec039965f3c4923ce901b-Abstract.html.  
Miłkowski, Marcin. 2018. “From Computer Metaphor to Computational Modeling: The Evolution of Computationalism.” *Minds and Machines* 28 (3): 515–41.  
Milner, A. D. 1995. “Cerebral Correlates of Visual Awareness.” *Neuropsychologia* 33 (9): 1117–30.  
Mindell, David A. 2002. *Between Human and Machine: Feedback, Control, and Computing Before Cybernetics*. JHU Press.  
Mink, Casey. 2021. “‘The Queen’s Gambit’ Chess Consultant on Making Anya Taylor-Joy a Prodigy.” *Backstage*, June 3, 2021. https://www.backstage.com/magazine/article/bruce-pandolfini-netflix-queens-gambit-chess-consultant-72717/.  
Minsky, Marvin. 1988. *Society Of Mind*. Simon and Schuster.  
Minsky, Marvin, and Seymour Papert. 1969. *Perceptrons; an Introduction to Computational Geometry*. MIT Press.  
Mischiati, Matteo, Huai-Ti Lin, Paul Herold, Elliot Imler, Robert Olberg, and Anthony Leonardo. 2015. “Internal Models Direct Dragonfly Interception Steering.” *Nature* 517 (7534): 333–38.  
Mitchell, Kevin J. 2023. *Free Agents: How Evolution Gave Us Free Will*. Princeton University Press.  
Mitchell, Melanie. 2023. “AI’s Challenge of Understanding the World.” *Science* 382 (6671): eadm8175.  
———. 2024. “Debates on the Nature of Artificial General Intelligence.” *Science* 383 (6689): eado7069.  
Miyazawa, Kazuki, Yuta Kyuragi, and Takayuki Nagai. 2022. “Simple and Effective Multimodal Learning Based on Pre-Trained Transformer Models.” *IEEE Access* 10: 29821–33.  
Mizunami, Makoto, and Yukihisa Matsumoto. 2017. “Roles of Octopamine and Dopamine Neurons for Mediating Appetitive and Aversive Signals in Pavlovian Conditioning in Crickets.” *Frontiers in Physiology* 8 (December): 1027.  
Mogavero, Federico, and Jacques Laskar. 2022. “The Origin of Chaos in the Solar System through Computer Algebra.” *Astronomy & Astrophysics. Supplement Series* 662 (June): L3.  
Mollick, Ethan. 2024. *Co-Intelligence: Living and Working with AI*. Penguin.  
Money, J. 1984. “Five Universal Exigencies, Indicatrons and Sexological Theory.” *Journal of Sex & Marital Therapy* 10 (4): 229–38.  
Montague, P. R., P. Dayan, and T. J. Sejnowski. 1996. “A Framework for Mesencephalic Dopamine Systems Based on Predictive Hebbian Learning.” *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience* 16 (5): 1936–47.  
Montague, P. R., and T. J. Sejnowski. 1994. “The Predictive Brain: Temporal Coincidence and Temporal Order in Synaptic Learning Mechanisms.” *Learning & Memory* 1 (1): 1–33.  
Moore, Gordon E. 1965. “Cramming More Components onto Integrated Circuits.” *Electronics Magazine* 38 (8): 114–17.  
Moravec, Hans. 1988. *Mind Children: The Future of Robot and Human Intelligence*. Harvard University Press.  
Mordvintsev, Alexander, Ettore Randazzo, Eyvind Niklasson, and Michael Levin. 2020. “Growing Neural Cellular Automata.” *Distill* 5 (2): e23.  
More, Max. 2003. “Principles of Extropy, Version 3.11.” 2003. https://web.archive.org/web/20131015142449/http://extropy.org/principles.htm.  
Moscovitch, M., and G. Winocur. 2002. “The Frontal Cortex and Working with Memory.” *Principles of Frontal Lobe Function*. https://doi.org/10.1093/acprof:oso/9780195134971.003.0012.  
Mountcastle, V. B. 1957. “Modality and Topographic Properties of Single Neurons of Cat’s Somatic Sensory Cortex.” *Journal of Neurophysiology* 20 (4): 408–34.  
[Munkhdalai, Tsendsuren, Manaal Faruqui, and Siddharth Gopal. 2024. “Leave No Context Behind: Efficient Infinite Context Transformers with Infini-Attention.” *arXiv [cs.CL]*. arXiv.](http://paperpile.com/b/iA68kr/1m1D) http://arxiv.org/abs/2404.07143.  
Murray, Connor S., Yingnan Gao, and Martin Wu. 2021. “Re-Evaluating the Evidence for a Universal Genetic Boundary among Microbial Species.” *Nature Communications* 12 (1): 4059.  
Muthukrishna, Michael, Ben W. Shulman, Vlad Vasilescu, and Joseph Henrich. 2014. “Sociality Influences Cultural Complexity.” *Proceedings. Biological Sciences / The Royal Society* 281 (1774): 20132511.  
Nádasdy, Z., H. Hirase, A. Czurkó, J. Csicsvari, and G. Buzsáki. 1999. “Replay and Time Compression of Recurring Spike Sequences in the Hippocampus.” *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience* 19 (21): 9497–9507.  
Nagels, Leanne, Etienne Gaudrain, Deborah Vickers, Petra Hendriks, and Deniz Başkent. 2023. “Prelingually Deaf Children With Cochlear Implants Show Better Perception of Voice Cues and Speech in Competing Speech Than Postlingually Deaf Adults With Cochlear Implants.” *Ear and Hearing*, 10.1097/AUD.0000000000001489.  
Nag, Oishimaya Sen. 2017. “What Languages Are Spoken in Burkina Faso?” WorldAtlas. July 27, 2017. https://www.worldatlas.com/articles/what-languages-are-spoken-in-burkina-faso.html.  
Naselaris, Thomas, Ryan J. Prenger, Kendrick N. Kay, Michael Oliver, and Jack L. Gallant. 2009. “Bayesian Reconstruction of Natural Images from Human Brain Activity.” *Neuron* 63 (6): 902–15.  
Nasir, Arshan, and Gustavo Caetano-Anollés. 2015. “A Phylogenomic Data-Driven Exploration of Viral Origins and Evolution.” *Science Advances* 1 (8): e1500527.  
Naville, Magali, Ian A. Warren, Zofia Haftek-Terreau, Domitille S. Chalopin, Frédéric G. Brunet, Perrine Levin, Delphine Galiana, and Jean Nicholas Volff. 2016. “Not so Bad after All: Retroviruses and Long Terminal Repeat Retrotransposons as a Source of New Genes in Vertebrates.” *Clinical Microbiology and Infection: The Official Publication of the European Society of Clinical Microbiology and Infectious Diseases* 22 (4): 312–23.  
Nelson, John P., and The Conversation US. n.d. “ChatGPT and Other Language AIs Are Nothing without Humans.” *Scientific American*. Accessed April 22, 2024. https://www.scientificamerican.com/article/chatgpt-and-other-language-ais-are-nothing-without-humans/.  
Neumann, John von. 1928. “On the Theory of Parlor Games.” *Mathematische Annalen* 100 (1): 295–320.  
———. 1945. “First Draft of a Report on the EDVAC.” University of Pennsylvania. https://doi.org/10.1109/85.238389.  
Neumann, John von, and Arthur W. Burks. 1966. *Theory of Self-Reproducing Automata*. Urbana and London: University of Illinois Press.  
Neumann, John von, and Oskar Morgenstern. 1944. *Theory of Games and Economic Behavior*. Princeton University Press.  
Newell, Karl M., Gottfried Mayer-Kress, S. Lee Hong, and Yeou-Teh Liu. 2009. “Adaptation and Learning: Characteristic Time Scales of Performance Dynamics.” *Human Movement Science* 28 (6): 655–87.  
Ng, Andrew Y., and Stuart Russell. 2000. “Algorithms for Inverse Reinforcement Learning.” *International Conference on Machine Learning* 67 (June): 663–70.  
Nguyen, Tu Anh, Benoit Sagot, and Emmanuel Dupoux. 2022. “Are Discrete Units Necessary for Spoken Language Modeling?” *IEEE Journal of Selected Topics in Signal Processing* 16 (6): 1415–23.  
Nightingale, Sophie J., and Hany Farid. 2022. “AI-Synthesized Faces Are Indistinguishable from Real Faces and More Trustworthy.” *Proceedings of the National Academy of Sciences of the United States of America* 119 (8). https://doi.org/10.1073/pnas.2120481119.  
Nilsson, Nils J. 2009. *The Quest for Artificial Intelligence*. Cambridge University Press.  
Nishimoto, Shinji, An T. Vu, Thomas Naselaris, Yuval Benjamini, Bin Yu, and Jack L. Gallant. 2011. “Reconstructing Visual Experiences from Brain Activity Evoked by Natural Movies.” *Current Biology: CB* 21 (19): 1641–46.  
Nitsch, A., R. Verheggen, and H-A Merten. 2007. “Penetrating Pneumatic Nail-Gun Injury to Skull Base.” *The British Journal of Oral & Maxillofacial Surgery* 45 (8): 692.  
Nityananda, Vivek, and Lars Chittka. 2015. “Modality-Specific Attention in Foraging Bumblebees.” *Royal Society Open Science* 2 (10): 150324.  
Norris, Dennis, and Kristjan Kalm. 2021. “Chunking and Data Compression in Verbal Short-Term Memory.” *Cognition* 208 (March): 104534.  
Northcutt, R. Glenn. 2012. “Evolution of Centralized Nervous Systems: Two Schools of Evolutionary Thought.” *Proceedings of the National Academy of Sciences of the United States of America* 109 Suppl 1 (Suppl 1): 10626–33.  
O’Gorman, Rick, Kennon M. Sheldon, and David S. Wilson. 2008. “For the Good of the Group? Exploring Group-Level Evolutionary Adaptations Using Multilevel Selection Theory.” *Group Dynamics: Theory, Research, and Practice: The Official Journal of Division 49, Group Psychology and Group Psychotherapy of the American Psychological Association* 12 (1): 17–26.  
Okon, Elias, and Miguel Ángel Sebastián. 2020. “A Consciousness-Based Quantum Objective Collapse Model.” *Synthese* 197 (9): 3947–67.  
Omrani, Nessrine, Giorgia Rivieccio, Ugo Fiore, Francesco Schiavone, and Sergio Garcia Agreda. 2022. “To Trust or Not to Trust? An Assessment of Trust in AI-Based Systems: Concerns, Ethics and Contexts.” *Technological Forecasting and Social Change* 181 (August): 121763.  
Oord, Aäron van den, Nal Kalchbrenner, and Koray Kavukcuoglu. 20–22 Jun 2016. “Pixel Recurrent Neural Networks.” In *Proceedings of The 33rd International Conference on Machine Learning*, edited by Maria Florina Balcan and Kilian Q. Weinberger, 48:1747–56. Proceedings of Machine Learning Research. New York, New York, USA: PMLR.  
[OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, et al. 2023. “GPT-4 Technical Report.” *arXiv [cs.CL]*. arXiv.](http://paperpile.com/b/iA68kr/Yklb) http://arxiv.org/abs/2303.08774.  
Ornes, Stephen. 2019. “Quantum Effects Enter the Macroworld.” *Proceedings of the National Academy of Sciences* 116 (45): 22413–17.  
Owren, Michael J., R. Toby Amoss, and Drew Rendall. 2011. “Two Organizing Principles of Vocal Production: Implications for Nonhuman and Human Primates.” *American Journal of Primatology* 73 (6): 530–44.  
Pagel, Mark, Quentin D. Atkinson, Andreea S Calude, and Andrew Meade. 2013. “Ultraconserved Words Point to Deep Language Ancestry across Eurasia.” *Proceedings of the National Academy of Sciences of the United States of America* 110 (21): 8471–76.  
Page, Robert E., Jr. 2012. “The Spirit of the Hive and How a Superorganism Evolves.” In *Honeybee Neurobiology and Behavior*, 3–16. Dordrecht: Springer Netherlands.  
Panahi, Shirin, Zainab Aram, Sajad Jafari, Jun Ma, and J. C. Sprott. 2017. “Modeling of Epilepsy Based on Chaotic Artificial Neural Network.” *Chaos, Solitons & Fractals* 105 (December): 150–56.  
[Pardos, Zachary A., and Shreya Bhandari. 2023. “Learning Gain Differences between ChatGPT and Human Tutor Generated Algebra Hints.” *arXiv [cs.CY]*. arXiv.](http://paperpile.com/b/iA68kr/fd2n) http://arxiv.org/abs/2302.06871.  
Park, Kinam. 2019. “Science the Endless Frontier.” *Journal of Controlled Release: Official Journal of the Controlled Release Society* 308 (August): 240.  
Pastuzyn, Elissa D., Cameron E. Day, Rachel B. Kearns, Madeleine Kyrke-Smith, Andrew V. Taibi, John McCormick, Nathan Yoder, et al. 2018. “The Neuronal Gene Arc Encodes a Repurposed Retrotransposon Gag Protein That Mediates Intercellular RNA Transfer.” *Cell* 173 (1): 275.  
Patel, Arkil, Satwik Bhattamishra, and Navin Goyal. 2021. “Are NLP Models Really Able to Solve Simple Math Word Problems?” In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, edited by Kristina Toutanova, Anna Rumshisky, Luke Zettlemoyer, Dilek Hakkani-Tur, Iz Beltagy, Steven Bethard, Ryan Cotterell, Tanmoy Chakraborty, and Yichao Zhou, 2080–94. Online: Association for Computational Linguistics.  
Patten, B. M. 1990. “The History of Memory Arts.” *Neurology* 40 (2): 346–52.  
Pauling, L. 1968. “Orthomolecular Psychiatry.” *Science* 160 (April): 265–71.  
Paulk, Angelique C., Andrew M. Dacks, James Phillips-Portillo, Jean-Marc Fellous, and Wulfila Gronenberg. 2009. “Visual Processing in the Central Bee Brain.” *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience* 29 (32): 9987–99.  
Paura, Roberto. 2019. “A Rapture of the Nerds? A Comparison between Transhumanist Eschatology and Christian Parousia.” *Forum Philosophicum* 24 (2): 343–67.  
Pavlov, Ivan Petrovich. 1902. *The Work of the Digestive Glands*. Charles Griffin.  
Pawłowski, Bogusław, C. B. Lowen, and Robin I. M. Dunbar. 1998. “Neocortex Size, Social Skills and Mating Success in Primates.” *Behaviour* 135 (3): 357–68.  
Pearce, Eiluned, Chris Stringer, and R. I. M. Dunbar. 2013. “New Insights into Differences in Brain Organization between Neanderthals and Anatomically Modern Humans.” *Proceedings. Biological Sciences / The Royal Society* 280 (1758): 20130168.  
Pearl, Judea, and Dana Mackenzie. 2018. *The Book of Why: The New Science of Cause and Effect*. Penguin Books Limited.  
Peddinti, Vijayaditya, Daniel Povey, and S. Khudanpur. 2015. “A Time Delay Neural Network Architecture for Efficient Modeling of Long Temporal Contexts.” *Interspeech*, 3214–18.  
Pederzoli, Luciano, Patrizio Tressoldi, and Helané Wahbeh. 2022. “Channeling: A Non-Pathological Possession and Dissociative Identity Experience or Something Else?” *Culture, Medicine and Psychiatry* 46 (2): 161–69.  
Pepperberg, Irene M., and Irene M. Pepperberg. 2009. *The Alex Studies: Cognitive and Communicative Abilities of Grey Parrots*. Harvard University Press.  
Peretó, Juli, Jeffrey L. Bada, and Antonio Lazcano. 2009. “Charles Darwin and the Origin of Life.” *Origins of Life and Evolution of the Biosphere: The Journal of the International Society for the Study of the Origin of Life* 39 (5): 395–406.  
Perrotta, Giulio. 2019. “The Phenomenon of Demonic Possession: Definition, Contexts and Multidisciplinary Approaches.” *Psychology and Mental Health Care: Open Access* 3 (2): 01–13.  
Perruchet, Pierre, and Arnaud Rey. 2005. “Does the Mastery of Center-Embedded Linguistic Structures Distinguish Humans from Nonhuman Primates?” *Psychonomic Bulletin & Review* 12 (2): 307–13.  
Pesavento, Umberto. 1995. “An Implementation of von Neumann’s Self-Reproducing Machine.” *Artificial Life* 2 (4): 337–54.  
Peters, Benjamin. 2016. *How Not to Network a Nation: The Uneasy History of the Soviet Internet*. MIT Press.  
Petkov, Christopher I., and Carel Ten Cate. 2020. “Structured Sequence Learning: Animal Abilities, Cognitive Operations, and Language Evolution.” *Topics in Cognitive Science* 12 (3): 828–42.  
Phan, Hoang Vu, and Hoon Cheol Park. 2020. “Mimicking Nature’s Flyers: A Review of Insect-Inspired Flying Robots.” *Current Opinion in Insect Science* 42 (December): 70–75.  
Philippe, J. 1901. “The Psychological Review. (Janv.-Nov. 1900).” *Revue Philosophique de La France et de L’etranger* 52: 336–43.  
Piccinini, Gualtiero. 2021. “The Myth of Mind Uploading.” In *The Mind-Technology Problem : Investigating Minds, Selves and 21st Century Artefacts*, edited by Robert W. Clowes, Klaus Gärtner, and Inês Hipólito, 125–44. Cham: Springer International Publishing.  
Piketty, Thomas. 2017. *Capital in the Twenty-First Century*. London, England: Harvard University Press.  
Pinker, Steven. 2010. *The Language Instinct: How The Mind Creates Language*. Harper Collins.  
Pisoni, D., Christopher M. Conway, W. Kronenberger, D. Horn, Jennifer Karpicke, and Shirley C. Henning. 2008. “Efficacy and Effectiveness of Cochlear Implants in Deaf Children.” *And Outcomes*, September. https://doi.org/10.1093/ACPROF:OSO/9780195368673.003.0003.  
Pitts, Walter, and Warren S. McCulloch. 1947. “How We Know Universals: The Perception of Auditory and Visual Forms.” *The Bulletin of Mathematical Biophysics* 9 (3): 127–47.  
Pluchino, A., A. Biondo, and A. Rapisarda. 2018. “Talent vs Luck: The Role of Randomness in Success and Failure (25.02. 2018).”  
[Polu, Stanislas, Jesse Michael Han, Kunhao Zheng, Mantas Baksys, Igor Babuschkin, and Ilya Sutskever. 2022. “Formal Mathematics Statement Curriculum Learning.” *arXiv [cs.LG]*. arXiv.](http://paperpile.com/b/iA68kr/jQ9m) http://arxiv.org/abs/2202.01344.  
Pomerleau, Dean A. 1989. “ALVINN: An Autonomous Land Vehicle in a Neural Network.” In *Proceedings of Neural Information Processing Systems*, edited by D. S. Touretzky, 1:305–13. Morgan Kaufmann.  
Pörtner, Hans-Otto, and Daniel Belling. 2022. *Climate Change 2022: Impacts, Adaptation and Vulnerability : Working Group II Contribution to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change*. Cambridge University Press.  
Pouget, P. 2014. “The Cortex Is in Overall Control of ‘voluntary’ Eye Movement.” *Eye* 29 (2): 241–45.  
Powell, Joanne, Penelope A. Lewis, Neil Roberts, Marta García-Fiñana, and Robin I. M. Dunbar. 2012. “Orbital Prefrontal Cortex Volume Predicts Social Network Size: An Imaging Study of Individual Differences in Humans.” *Proceedings. Biological Sciences / The Royal Society* 279 (1736): 2157–62.  
Preilowski, B. 1975. “Bilateral Motor Interaction: Perceptual-Motor Performance of Partial and Complete ‘Split-Brain’ Patients.” In *Cerebral Localization: An Otfrid Foerster Symposium*, edited by K. J. Zülch, O. Creutzfeldt, and G. C. Galbraith, 115–32. Berlin, Heidelberg: Springer Berlin Heidelberg.  
Prigogine, Ilya, and Isabelle Stengers. 1984. *Order out of Chaos: Man’s New Dialogue with Nature*. London: William Heinemann.  
Pross, Addy. 2012. *What Is Life?: How Chemistry Becomes Biology*. Oxford, UK: Oxford University Press.  
Qi, Matthew A. Brown, and D. Lowe. 2017. “Low-Shot Learning with Imprinted Weights.” *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, December, 5822–30.  
Quartz, S. R., P. Dayan, P. R. Montague, and T. J. Sejnowski. 1992. “Expectation Learning in the Brain Using Diffuse Ascending Projections.” https://pure.mpg.de/pubman/faces/ViewItemOverviewPage.jsp?itemId=item\_3262000.  
Quiroga, R. Quian, L. Reddy, G. Kreiman, C. Koch, and I. Fried. 2005. “Invariant Visual Representation by Single Neurons in the Human Brain.” *Nature* 435 (7045): 1102–7.  
[Rae, Jack W., Sebastian Borgeaud, Trevor Cai, Katie Millican, Jordan Hoffmann, Francis Song, John Aslanides, et al. 2021. “Scaling Language Models: Methods, Analysis & Insights from Training Gopher.” *arXiv [cs.CL]*. arXiv.](http://paperpile.com/b/iA68kr/kZm9) http://arxiv.org/abs/2112.11446.  
[Raghuraman, Nikhil, Adam W. Harley, and Leonidas Guibas. 2023. “Cross-Image Context Matters for Bongard Problems.” *arXiv [cs.CV]*. arXiv.](http://paperpile.com/b/iA68kr/PjLC) http://arxiv.org/abs/2309.03468.  
Rahimi, Akam, Triantafyllos Afouras, and Andrew Zisserman. 2022. “Reading to Listen at the Cocktail Party: Multi-Modal Speech Separation.” *Proceedings / CVPR, IEEE Computer Society Conference on Computer Vision and Pattern Recognition. IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, June, 10483–92.  
Rao, Rajesh P. N. 2024. “A Sensory–motor Theory of the Neocortex.” *Nature Neuroscience* 27 (7): 1221–35.  
Rasch, Björn, and Jan Born. 2013. “About Sleep’s Role in Memory.” *Physiological Reviews* 93 (2): 681–766.  
Raup, David M., and Stephen Jay Gould. 1974. “Stochastic Simulation and Evolution of Morphology-Towards a Nomothetic Paleontology.” *Systematic Biology* 23 (3): 305–22.  
Ray Perrault, and Jack Clark. 2024. “Artificial Intelligence Index Report 2024.” Stanford Institute for Human-Centered Artificial Intelligence. https://policycommons.net/artifacts/12089781/hai\_ai-index-report-2024/.  
Ray, Thomas S. 1991. “An Approach to the Synthesis of Life.” In *Artificial Life II, Santa Fe Institute Studies in the Sciences of Complexity, Vol. XI*, edited by C. Langton, C. Taylor, J. D. Farmer, and S. Rasmussen, 371–408. Redwood City, CA: Addison-Wesley.  
Raz, Amelie A., Mansi Srivastava, Ranja Salvamoser, and Peter W. Reddien. 2017. “Acoel Regeneration Mechanisms Indicate an Ancient Role for Muscle in Regenerative Patterning.” *Nature Communications* 8 (1): 1260.  
Redelmeier, Donald A., Joel Katz, and Daniel Kahneman. 2003. “Memories of Colonoscopy: A Randomized Trial.” *Pain* 104 (1-2): 187–94.  
Reichenbach, Hans. 1956. *The Direction of Time*. Edited by Maria Reichenbach. Berkeley, CA: University of California Press.  
Rescorla, R. A. 1972. “A Theory of Pavlovian Conditioning: Variations in the Effectiveness of Reinforcement and Non-Reinforcement.” *Classical Conditioning, Current Research and Theory*. https://cir.nii.ac.jp/crid/1572543025504096640.  
Revell, Timothy. 2017. “Mathematician Set to Publish ABC Proof Almost No One Understands.” *New Scientist*, December 18, 2017.  
Richards, Blake, Blaise Agüera y Arcas, Guillaume Lajoie, and Dhanya Sridhar. 2023. “The Illusion Of AI’s Existential Risk.” *Noema Magazine*, July.  
Richards, Blake, and Timothy Lillicrap. 2022. “The Brain-Computer Metaphor Debate Is Useless: A Matter of Semantics.” *Frontiers in Computer Science* 4. https://doi.org/10.3389/fcomp.2022.810358.  
Rid, Thomas. 2016. *Rise of the Machines: A Cybernetic History*. W. W. Norton & Company.  
Riskin, Jessica. 2003. “The Defecating Duck, Or, the Ambiguous Origins of Artificial Life.” *Critical Inquiry* 29 (4): 599–633.  
———. 2016. *The Restless Clock: A History of the Centuries-Long Argument Over What Makes Living Things Tick*. University of Chicago Press.  
Robertson, Michael P., and Gerald F. Joyce. 2012. “The Origins of the RNA World.” *Cold Spring Harbor Perspectives in Biology* 4 (5). https://doi.org/10.1101/cshperspect.a003608.  
Rojas, R. July-Sep 1998. “How to Make Zuse’s Z3 a Universal Computer.” *IEEE Annals of the History of Computing* 20 (3): 51–54.  
Romo, R., and W. Schultz. 1990. “Dopamine Neurons of the Monkey Midbrain: Contingencies of Responses to Active Touch during Self-Initiated Arm Movements.” *Journal of Neurophysiology* 63 (3): 592–606.  
Roose, Kevin. 2023a. “A Conversation With Bingʼs Chatbot Left Me Deeply Unsettled.” *The New York Times*, February 16, 2023. https://edisciplinas.usp.br/pluginfile.php/7620512/mod\_resource/content/2/Why%20a%20Conversation%20With%20Bing%E2%80%99s%20Chatbot%20Left%20Me%20Deeply%20Unsettled%20-%20The%20New%20York%20Times.PDF.  
———. 2023b. “Bing’s A.I. Chat: ‘I Want to Be Alive. 😈.’” *The New York Times*, February 16, 2023. https://www.nytimes.com/2023/02/16/technology/bing-chatbot-transcript.html.  
———. 2024. “The Year Chatbots Were Tamed.” *The New York Times*, February 14, 2024. https://www.nytimes.com/2024/02/14/technology/chatbots-sydney-tamed.html.  
Rosenblatt, Frank. 1957. *The Perceptron: A Perceiving and Recognizing Automaton (Project Para)*. Buffalo, NY: Cornell Aeronautical Laboratory.  
Rosenblueth, Arturo, Norbert Wiener, and Julian Bigelow. 1943. “Behavior, Purpose and Teleology.” *Philosophy of Science* 10 (1): 18–24.  
Rovelli, Carlo. 2018. *The Order of Time*. Translated by Erica Segre and Simon Carnell. New York: Riverhead Books.  
———. 2021. *Helgoland: The Sunday Times Bestseller*. Penguin UK.  
———. 2023. *Anaximander: And the Birth of Science*. Penguin.  
Rowland, Lewis P. 2005. “WALTER FREEMAN’S PSYCHOSURGERY AND BIOLOGICAL PSYCHIATRY: A CAUTIONARY TALE.” *Neurology Today* 5 (4): 70.  
Ruhlen, Merritt. 1994. *On the Origin of Languages: Studies in Linguistic Taxonomy*. Stanford University Press.  
Russell, Bertrand. 1937. *A Critical Exposition of the Philosophy of Leibniz: With an Appendix of Leading Passages*. George Allen & Unwin Limited.  
Russell, Michael J., and William Martin. 2004. “The Rocky Roots of the Acetyl-CoA Pathway.” *Trends in Biochemical Sciences* 29 (7): 358–63.  
Russ, Eric, and Sergey Iordanskiy. 2023. “Endogenous Retroviruses as Modulators of Innate Immunity.” *Pathogens* 12 (2). https://doi.org/10.3390/pathogens12020162.  
Rutz, Christian, Michael Bronstein, Aza Raskin, Sonja C. Vernes, Katherine Zacarian, and Damián E. Blasi. 2023. “Using Machine Learning to Decode Animal Communication.” *Science* 381 (6654): 152–55.  
Ryan, Frank. 2009. *Virolution*. London: Collins.  
Sagan, Lynn. 1967. “On the Origin of Mitosing Cells.” *Journal of Theoretical Biology* 14 (3): 255–74.  
[Salimans, Tim, Jonathan Ho, Xi Chen, Szymon Sidor, and Ilya Sutskever. 2017. “Evolution Strategies as a Scalable Alternative to Reinforcement Learning.” *arXiv [stat.ML]*. arXiv.](http://paperpile.com/b/iA68kr/eL3pb) https://deepsense.ai/wp-content/uploads/2017/10/1703.03864.pdf.  
Sapolsky, Robert M. 2023. *Determined: A Science of Life Without Free Will*. Penguin Press.  
Saponati, Matteo, and Martin Vinck. 2023. “Sequence Anticipation and Spike-Timing-Dependent Plasticity Emerge from a Predictive Learning Rule.” *Nature Communications* 14 (1): 4985.  
Sarfati, Raphaël, Laura Gaudette, Joseph M. Cicero, and Orit Peleg. 2022. “Statistical Analysis Reveals the Onset of Synchrony in Sparse Swarms of Photinus Knulli Fireflies.” *Journal of the Royal Society, Interface / the Royal Society* 19 (188): 20220007.  
[Saxe, Andrew, Stephanie Nelli, and Christopher Summerfield. 2020. “If Deep Learning Is the Answer, Then What Is the Question?” *arXiv [q-bio.NC]*. arXiv.](http://paperpile.com/b/iA68kr/qZ5q) http://arxiv.org/abs/2004.07580.  
Saxe, Leonard, Denise Dougherty, and Theodore Cross. 1985. “The Validity of Polygraph Testing: Scientific Analysis and Public Controversy.” *The American Psychologist* 40 (3): 355–66.  
Schaller, R. R. 1997. “Moore’s Law: Past, Present and Future.” *IEEE Spectrum* 34 (6): 52–59.  
Scheel, David, Stephanie Chancellor, Martin Hing, Matthew Lawrence, Stefan Linquist, and Peter Godfrey-Smith. 2017. “A Second Site Occupied by Octopus Tetricus at High Densities, with Notes on Their Ecology and Behavior.” *Marine and Freshwater Behaviour and Physiology* 50 (4): 285–91.  
Schmid, Gerrit, and Daniel A. Braun. 2020. “Human Group Coordination in a Sensorimotor Task with Neuron-like Decision-Making.” *Scientific Reports* 10 (1): 8226.  
Schmidhuber, Jürgen. 2014. “Who Invented Backpropagation?” 2014. https://people.idsia.ch/~juergen/who-invented-backpropagation.html.  
Schneider, Susan. 2021. *Artificial You: AI and the Future of Your Mind*. Princeton University Press.  
Schofield, Philip, Catherine Pease-Watkin, and Michael Quinn. 2014. *Of Sexual Irregularities, and Other Writings on Sexual Morality*. OUP Oxford.  
Schopenhauer, Arthur. 2011. *Essay On The Freedom Of The Will*. Literary Licensing, LLC.  
Schrittwieser, Julian, Ioannis Antonoglou, Thomas Hubert, Karen Simonyan, Laurent Sifre, Simon Schmitt, Arthur Guez, et al. 2020. “Mastering Atari, Go, Chess and Shogi by Planning with a Learned Model.” *Nature* 588 (7839): 604–9.  
Schrödinger, Erwin. 1944. *What Is Life? The Physical Aspect of the Living Cell*. Cambridge University Press.  
Schultz, Darrin T., Steven H. D. Haddock, Jessen V. Bredeson, Richard E. Green, Oleg Simakov, and Daniel S. Rokhsar. 2023. “Ancient Gene Linkages Support Ctenophores as Sister to Other Animals.” *Nature* 618 (7963): 110–17.  
Scott-Heron, Gil. 1970. “Whitey on the Moon.” *Small Talk at 125th and Lenox*, 8–9.  
Scott, James C. 1998. *Seeing like a State: How Certain Schemes to Improve the Human Condition Have Failed*. New Haven, Connecticut: Yale University Press.  
———. 2017. *Against the Grain: A Deep History of the Earliest States*. Yale University Press.  
Sejnowski, T. J., and C. R. Rosenberg. 1987. “Parallel Networks That Learn to Pronounce English Text.” *Complex Systems*. https://empslocal.ex.ac.uk/people/staff/reverson/sr/oldECM3412/Sejnowski.pdf.  
Self, Joseph. 1993. “The ‘My Sweet Lord’/‘He’s So Fine’ Plagiarism Suit.” *The 910 Magazine*.  
Seppala, Timothy J. 2017. “The Ridiculous Not Hotdog App from ‘Silicon Valley’ Is Real.” Engadget. May 15, 2017. https://www.engadget.com/2017-05-15-not-hotdog-app-hbo-silicon-valley.html.  
Seth, Anil. 2021. *Being You: A New Science of Consciousness*. Penguin.  
Setoh, Peipei, Rose M. Scott, and Renée Baillargeon. 2016. “Two-and-a-Half-Year-Olds Succeed at a Traditional False-Belief Task with Reduced Processing Demands.” *Proceedings of the National Academy of Sciences of the United States of America* 113 (47): 13360–65.  
Shapiro, Robert. 2006. “Small Molecule Interactions Were Central to the Origin of Life.” *The Quarterly Review of Biology* 81 (2): 105–25.  
Sharma, J., A. Angelucci, and M. Sur. 2000. “Induction of Visual Orientation Modules in Auditory Cortex.” *Nature* 404 (6780): 841–47.  
Sharp, Paul M., and Beatrice H. Hahn. 2011. “Origins of HIV and the AIDS Pandemic.” *Cold Spring Harbor Perspectives in Medicine* 1 (1): a006841.  
She, Jianqi, Minghao Du, Zhanzhan Xu, Yueqi Jin, Yu Li, Daoning Zhang, Changyu Tao, Jian Chen, Jiadong Wang, and Ence Yang. 2022. “The Landscape of hervRNAs Transcribed from Human Endogenous Retroviruses across Human Body Sites.” *Genome Biology* 23 (1): 231.  
Shelley, Mary. 1818. *Frankenstein, or the Modern Prometheus*. London: Printed for Lackington, Hughes, Harding.  
Shi, Feng, Misha Teplitskiy, Eamon Duede, and James A. Evans. 2019. “The Wisdom of Polarized Crowds.” *Nature Human Behaviour* 3 (4): 329–36.  
Shi, Yuge, Siddharth Narayanaswamy, Brooks Paige, and Philip H. S. Torr. 2019. “Variational Mixture-of-Experts Autoencoders for Multi-Modal Deep Generative Models.” *Advances in Neural Information Processing Systems*, November, 15692–703.  
Shoham, Yoav, Raymond Perrault, Erik Brynjolfsson, Jack Clark, and Calvin LeGassick. 2017. “AI Index 2017 Annual Report.” *Artificial Intelligence Index*. https://scholar.google.ca/scholar?cluster=17676853191636706297&hl=en&as\_sdt=0,5&sciodt=0,5.  
Silvanto, Juha. 2014. “Is Primary Visual Cortex Necessary for Visual Awareness?” *Trends in Neurosciences* 37 (11): 618–19.  
Silver, David, Alex Graves, Ioannis Antonoglou, Martin Riedmiller, Volodymyr Mnih, D. Wierstra, and K. Kavukcuoglu. 2013. “Playing Atari with Deep Reinforcement Learning.” *DeepMind Lab. arXiv* 1312. https://phillipi.github.io/6.882/2020/notes/The%20problem%20of%20long-term%20decision%20making/neural\_fitted.pdf.  
[Silver, David, Thomas Hubert, Julian Schrittwieser, Ioannis Antonoglou, Matthew Lai, Arthur Guez, Marc Lanctot, et al. 2017. “Mastering Chess and Shogi by Self-Play with a General Reinforcement Learning Algorithm.” *arXiv [cs.AI]*. arXiv.](http://paperpile.com/b/iA68kr/WvNG) http://arxiv.org/abs/1712.01815.  
Simons, D. J., and C. F. Chabris. 1999. “Gorillas in Our Midst: Sustained Inattentional Blindness for Dynamic Events.” *Perception* 28 (9): 1059–74.  
Singer, Peter. 1973. “Animal Liberation.” *The New York Review of Books*, April.  
———. 2005. “Ethics and Intuitions.” *The Journal of Ethics* 9 (3): 331–52.  
Sivitilli, Dominic Michel, and David H. Gire. 2019. *Collective Cognition in the Arms of the Octopus*. The Astrobiology Science Conference (AbSciCon) 2019. ui.adsabs.harvard.edu.  
Slobodchikoff, C. N., Andrea Paseka, and Jennifer L. Verdolin. 2009. “Prairie Dog Alarm Calls Encode Labels about Predator Colors.” *Animal Cognition* 12 (3): 435–39.  
Sloman, Steven, and Philip Fernbach. 2018. *The Knowledge Illusion: Why We Never Think Alone*. Penguin.  
Smil, Vaclav. 2008. *Energy in Nature and Society: General Energetics of Complex Systems*. London and Cambridge, MA: MIT Press.  
Smith, Adam. 1776. *An Inquiry into the Nature and Causes of the Wealth of Nations*. North Charleston, SC: Createspace Independent Publishing Platform.  
Smith, Eric Alden. 2010. “Communication and Collective Action: Language and the Evolution of Human Cooperation.” *Evolution and Human Behavior: Official Journal of the Human Behavior and Evolution Society* 31 (4): 231–45.  
Smith-Ruiu, Justin. 2022. *The Internet Is Not What You Think It Is: A History, a Philosophy, a Warning*. Princeton University Press.  
Smith, Stephen J., Michael Hawrylycz, Jean Rossier, and Uygar Sümbül. 2020. “New Light on Cortical Neuropeptides and Synaptic Network Plasticity.” *Current Opinion in Neurobiology* 63 (August): 176–88.  
Snow, Herman B. 1949. “A Review of 27 Prefrontal Lobotomy Patients.” *The Psychiatric Quarterly* 23 (1): 26–34.  
Soll, Jacob. 2011. *The Information Master: Jean-Baptiste Colbert’s Secret State Intelligence System*. Cultures of Knowledge in the Early Modern World. Ann Arbor, MI: University of Michigan Press.  
Somé, Léopold, Abdulai Jalloh, Robert B. Zougmoré, Gerald C. Nelson, and Timothy S. Thomas. 2013. “Burkina Faso.” https://www.ethnologue.com/country/BF/.  
Spaethe, Johannes, Jürgen Tautz, and Lars Chittka. 2006. “Do Honeybees Detect Colour Targets Using Serial or Parallel Visual Search?” *The Journal of Experimental Biology* 209 (Pt 6): 987–93.  
Speijer, Dave. 2016. “What Can We Infer about the Origin of Sex in Early Eukaryotes?” *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences* 371 (1706). https://doi.org/10.1098/rstb.2015.0530.  
Sperry, R. W. 1965. “Brain Bisection and Mechanisms of Consciousness.” In *Brain and Conscious Experience: Study Week September 28 to October 4, 1964, of the Pontificia Academia Scientiarum*, edited by John C. Eccles, 298–313. Berlin, Heidelberg: Springer Berlin Heidelberg.  
———. 1968. “Hemisphere Deconnection and Unity in Conscious Awareness.” *The American Psychologist* 23 (10): 723–33.  
Spiegel, Alix, and Micaela Rodríguez. 2017. “Eager To Burst His Own Bubble, A Techie Made Apps To Randomize His Life.” *NPR*, June 8, 2017. https://www.npr.org/sections/alltechconsidered/2017/06/08/531796329/eager-to-burst-his-own-bubble-a-techie-made-apps-to-randomize-his-life.  
Steklis, Horest. 2012. *Neurobiology of Social Communication In Primates: An Evolutionary Perspective*. Elsevier.  
Stock, J., and M. Levit. 2000. “Signal Transduction: Hair Brains in Bacterial Chemotaxis.” *Current Biology: CB* 10 (1): R11–14.  
Strogatz, Steven. 2004. *Sync: The Emerging Science of Spontaneous Order*. Penguin UK.  
Stroop, J. 1992. “Studies of Interference in Serial Verbal Reactions.” *Journal of Experimental Psychology. General* 18: 643–62.  
Surrey Comet. 1946. “Electric Brain to Be Made at Teddington.” *Surrey Comet*.  
Sutton, Richard S. 1988. “Learning to Predict by the Methods of Temporal Differences.” *Machine Learning* 3 (1): 9–44.  
Suzuki, Masahiro, and Yutaka Matsuo. 2022. “A Survey of Multimodal Deep Generative Models.” *Advanced Robotics: The International Journal of the Robotics Society of Japan* 36 (5-6): 261–78.  
Szathmáry, Eörs, and John Maynard Smith. 1995. “The Major Evolutionary Transitions.” *Nature* 374 (6519): 227–32.  
Taleb, Nassim Nicholas. 2014. *Antifragile: Things That Gain from Disorder*. New York: Random House.  
Tanaka, James W., and Lara J. Pierce. 2009. “The Neural Plasticity of Other-Race Face Recognition.” *Cognitive, Affective & Behavioral Neuroscience* 9 (1): 122–31.  
Tan, Chuanqi, Fuchun Sun, Tao Kong, Wenchang Zhang, Chao Yang, and Chunfang Liu. 2018. “A Survey on Deep Transfer Learning.” In *Artificial Neural Networks and Machine Learning – ICANN 2018*, 270–79. Springer International Publishing.  
Tarlinton, Rachael E., Joanne Meers, and Paul R. Young. 2006. “Retroviral Invasion of the Koala Genome.” *Nature* 442 (7098): 79–81.  
Tchaikovsky, Adrian. 2018. *Children of Time*. Orbit.  
Tegmark, Max. 2018. *Life 3.0: Being Human in the Age of Artificial Intelligence*. Vintage.  
Tesauro, G. 1992. “Temporal Difference Learning of Backgammon Strategy.” *Machine Learning Proceedings 1992*. https://www.sciencedirect.com/science/article/pii/B9781558602472500632.  
———. 1994. “TD-Gammon, a Self-Teaching Backgammon Program, Achieves Master-Level Play.” *Neural Computation*. https://direct.mit.edu/neco/article-abstract/6/2/215/5771.  
Thaler, Lore, Stephen R. Arnott, and Melvyn A. Goodale. 2011. “Neural Correlates of Natural Human Echolocation in Early and Late Blind Echolocation Experts.” *PloS One* 6 (5): e20162.  
Thalmann, Mirko, Alessandra S. Souza, and Klaus Oberauer. 2019. “How Does Chunking Help Working Memory?” *Journal of Experimental Psychology. Learning, Memory, and Cognition* 45 (1): 37–55.  
“The London Literary Gazette; and Journal of Belles Lettres, Arts, Sciences, &c.” 1832. *The London Literary Gazette*, no. 808: 442.  
“The World.” n.d. Ethnologue (Free All). Accessed April 25, 2024. https://www.ethnologue.com/.  
Thomas, Dylan. 1934. *18 Poems by Dylan Thomas*. Fortune Press.  
Thomson, Judith J. 1976. “Killing, Letting Die, and the Trolley Problem.” *The Monist* 59 (2): 204–17.  
Thomson, William. 1857. “On a Universal Tendency in Nature to the Dissipation of Mechanical Energy.” *Proceedings of the Royal Society of Edinburgh* 3: 139–42.  
———. 1871. “Inaugural Address before the British Association at Edinburgh, August 2d.” *American Journal of Science, and Arts* s3-2 (October): 269–94.  
[Thoppilan, Romal, Daniel De Freitas, Jamie Hall, Noam Shazeer, Apoorv Kulshreshtha, Heng-Tze Cheng, Alicia Jin, et al. 2022. “LaMDA: Language Models for Dialog Applications.” *arXiv [cs.CL]*. arXiv.](http://paperpile.com/b/iA68kr/cHc3) http://arxiv.org/abs/2201.08239.  
[Tishby, Naftali, Fernando C. Pereira, and William Bialek. 2000. “The Information Bottleneck Method.” *arXiv [physics.data-An]*. arXiv.](http://paperpile.com/b/iA68kr/MUTG) http://arxiv.org/abs/physics/0004057.  
Tomlinson, Bill, Rebecca W. Black, Donald J. Patterson, and Andrew W. Torrance. 2024. “The Carbon Emissions of Writing and Illustrating Are Lower for AI than for Humans.” *Scientific Reports* 14 (1): 1–8.  
Trnka, Radek, and Radmila Lorencova. 2022. “Indigenous Concepts of Consciousness, Soul, and Spirit: A Cross-Cultural Perspective.” *Journal of Consciousness Studies* 29 (1-2): 113–40.  
Turchin, Valentin Fedorovich. 1977. *The Phenomenon of Science*. Columbia University Press.  
Turchin, V. F. 1995. “A Dialogue on Metasystem Transition.” https://www.tandfonline.com/doi/abs/10.1080/02604027.1995.9972553.  
Turing, Alan Mathison. 1937. “On Computable Numbers, with an Application to the Entscheidungsproblem.” *Proceedings of the London Mathematical Society. Third Series* s2-42 (1): 230–65.  
———. 1948. “Intelligent Machinery: A Report.” *London: National Physical Laboratory*, 27.  
———. 1950. “Computing Machinery and Intelligence.” *Mind; a Quarterly Review of Psychology and Philosophy* 59 (236): 433–60.  
———. 1951. “Can Digital Computers Think?” BBC Radio. https://www.bbc.co.uk/historyofthebbc/anniversaries/november/turing-can-digital-computers-think.  
———. 1952. “The Chemical Basis of Morphogenesis.” *Bulletin of Mathematical Biology* 52 (1): 153–97.  
———. (1951) 2000. “Alan Turing’s Manual for the Ferranti Mk. I (transcribed by Robert S. Thau).” infoamerica.org. February 13, 2000. https://www.infoamerica.org/documentos\_pdf/turing02.pdf.  
Turkle, Sherry. 2011. *Evocative Objects: Things We Think With*. MIT Press.  
Tversky, Amos. 1969. “Intransitivity of Preferences.” *Psychological Review* 76 (1): 31.  
Tweti, Mira. 2008. *Of Parrots and People: The Sometimes Funny, Always Fascinating, and Often Catastrophic Collision of Two Intelligent Species*. Penguin.  
Urban, T. 2015. “The AI Revolution: The Road to Superintelligence.” *Wait But Why*.  
Van de Crommert HW, T. Mulder, and J. Duysens. 1998. “Neural Control of Locomotion: Sensory Control of the Central Pattern Generator and Its Relation to Treadmill Training.” *Gait & Posture* 7 (3): 251–63.  
Varro, Marcus Terentius, Liber de Re Rustica III, W. D. Hooper, and H. B. Ash. 1934. “Cato and Varro: On Agriculture.” *Edited and Translated by William David Hopper. Revised by Harrison Boyd Ash. LCL* 283.  
[Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. “Attention Is All You Need.” *arXiv [cs.CL]*. arXiv.](http://paperpile.com/b/iA68kr/ILDy) https://proceedings.neurips.cc/paper/7181-attention-is-all.  
Veluw, Susanne J. van, and Steven A. Chance. 2014. “Differentiating between Self and Others: An ALE Meta-Analysis of fMRI Studies of Self-Recognition and Theory of Mind.” *Brain Imaging and Behavior* 8 (1): 24–38.  
Verita, Monte. 2006. “50th Anniversary Summit of Artificial Intelligence.” https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=8d19265e2d190087e2d48b855cb0bf5a3c706342.  
[Villalobos, Pablo, Jaime Sevilla, Lennart Heim, Tamay Besiroglu, Marius Hobbhahn, and Anson Ho. 2022. “Will We Run out of Data? An Analysis of the Limits of Scaling Datasets in Machine Learning.” *arXiv [cs.LG]*. arXiv.](http://paperpile.com/b/iA68kr/V9SC) http://arxiv.org/abs/2211.04325.  
Viswanathan, Gandhimohan M., Marcos G. E. da Luz, Ernesto P. Raposo, and H. Eugene Stanley. 2011. *The Physics of Foraging: An Introduction to Random Searches and Biological Encounters*. Cambridge University Press.  
Völlm, Birgit A., Alexander N. W. Taylor, Paul Richardson, Rhiannon Corcoran, John Stirling, Shane McKie, John F. W. Deakin, and Rebecca Elliott. 2006. “Neuronal Correlates of Theory of Mind and Empathy: A Functional Magnetic Resonance Imaging Study in a Nonverbal Task.” *NeuroImage* 29 (1): 90–98.  
Volz, Lukas J., and Michael S. Gazzaniga. 2017. “Interaction in Isolation: 50 Years of Insights from Split-Brain Research.” *Brain: A Journal of Neurology* 140 (7): 2051–60.  
Von Oswald, Johannes, Eyvind Niklasson, Ettore Randazzo, Joao Sacramento, Alexander Mordvintsev, Andrey Zhmoginov, and Max Vladymyrov. 23–29 Jul 2023. “Transformers Learn In-Context by Gradient Descent.” In *Proceedings of the 40th International Conference on Machine Learning*, edited by Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, 202:35151–74. Proceedings of Machine Learning Research. PMLR.  
Waal, Frans B. M. 2007. *Chimpanzee Politics: Power and Sex Among Apes*. JHU Press.  
Waldrop, M. Mitchell. 2018. *The Dream Machine*. Stripe Press.  
———. 2019. “What Are the Limits of Deep Learning?” *Proceedings of the National Academy of Sciences* 116 (4): 1074–77.  
Walker, Sara. 2023. “AI Is Life.” *Noema Magazine*, April. https://www.noemamag.com/ai-is-life/.  
Wallisch, P. 2017. “Illumination Assumptions Account for Individual Differences in the Perceptual Interpretation of a Profoundly Ambiguous Stimulus in the Color Domain: ‘The Dress.’” *Journal of Vision* 17 (4): 5.  
Wang, Hengji, Joshua Swore, Shashank Sharma, John R. Szymanski, Rafael Yuste, Thomas L. Daniel, Michael Regnier, Martha M. Bosma, and Adrienne L. Fairhall. 2023. “A Complete Biomechanical Model of Hydra Contractile Behaviors, from Neural Drive to Muscle to Movement.” *Proceedings of the National Academy of Sciences of the United States of America* 120 (11): e2210439120.  
Wang, Mei, Weihong Deng, Jiani Hu, Xunqiang Tao, and Yaohai Huang. 2019. “Racial Faces in the Wild: Reducing Racial Bias by Information Maximization Adaptation Network.” In *Proceedings of the Ieee/cvf International Conference on Computer Vision*, 692–702. openaccess.thecvf.com.  
Wann, T. W. 1964. “Behaviorism and Phenomenology: Contrasting Bases for Modern Psychology.” Edited by T. W. Wann 190. https://psycnet.apa.org/fulltext/1965-00104-000.pdf.  
Wark, B., B. N. Lundstrom, and A. Fairhall. 2007. “Sensory Adaptation.” *Current Opinion in Neurobiology*. https://www.sciencedirect.com/science/article/pii/S0959438807000840.  
Watson, Andrew J., and James E. Lovelock. 1983. “Biological Homeostasis of the Global Environment: The Parable of Daisyworld.” *Tellus. Series B, Chemical and Physical Meteorology* 35 (4): 284–89.  
Watson, James D., and Francis H. Crick. 1953. “Molecular Structure of Nucleic Acids; a Structure for Deoxyribose Nucleic Acid.” *Nature* 171 (4356): 737–38.  
Webster, Jonathan J., and Chunyu Kit. 1992. “Tokenization as the Initial Phase in NLP.” In *COLING 1992 Volume 4: The 14th INternational COnference on COmputational LInguistics*. aclanthology.org. https://aclanthology.org/C92-4173.  
Wei, Jason, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, E. Chi, F. Xia, Quoc Le, and Denny Zhou. 2022. “Chain of Thought Prompting Elicits Reasoning in Large Language Models.” *Advances in Neural Information Processing Systems* abs/2201.11903 (January). https://proceedings.neurips.cc/paper\_files/paper/2022/hash/9d5609613524ecf4f15af0f7b31abca4-Abstract-Conference.html.  
[Wei, Jerry, Chengrun Yang, Xinying Song, Yifeng Lu, Nathan Hu, Dustin Tran, Daiyi Peng, et al. 2024. “Long-Form Factuality in Large Language Models.” *arXiv [cs.CL]*. arXiv.](http://paperpile.com/b/iA68kr/Wp1P) http://arxiv.org/abs/2403.18802.  
Weir, K. 2012. “Our Moral Motivations.” *Monitor on Psychology*.  
West, Geoffrey B. 2017. *Scale: The Universal Laws of Life, Growth, and Death in Organisms, Cities, and Companies*. New York: Penguin.  
West, Geoffrey B., James H. Brown, and Brian J. Enquist. 1997. “A General Model for the Origin of Allometric Scaling Laws in Biology.” *Science* 276 (5309): 122–26.  
Weyn, Jonathan A., Dale R. Durran, and Rich Caruana. 2020. “Improving Data‐driven Global Weather Prediction Using Deep Convolutional Neural Networks on a Cubed Sphere.” *Journal of Advances in Modeling Earth Systems* 12 (9). https://doi.org/10.1029/2020ms002109.  
Wheeler, John Archibald. 1978. “The ‘Past’ and the ‘Delayed-Choice’ Double-Slit Experiment.” In *Mathematical Foundations of Quantum Theory*, edited by A. R. Marlow, 9–48. Academic Press.  
———. 1992. “Recent Thinking about the Nature of the Physical World: It from Bita.” *Annals of the New York Academy of Sciences* 655 (1): 349–64.  
Whiten, Andrew, and Richard W. Byrne. 1997. *Machiavellian Intelligence II: Extensions and Evaluations*. Cambridge University Press.  
Whittington, James C. R., and Rafal Bogacz. 2019. “Theories of Error Back-Propagation in the Brain.” *Trends in Cognitive Sciences* 23 (3): 235–50.  
[Whittington, James C. R., Joseph Warren, and Timothy E. J. Behrens. 2021. “Relating Transformers to Models and Neural Representations of the Hippocampal Formation.” *arXiv [cs.NE]*. arXiv.](http://paperpile.com/b/iA68kr/jbzW) http://arxiv.org/abs/2112.04035.  
Wilden, Anthony. 1972. “Analog and Digital Communication: On the Relationship between Negation, Signification, and the Emergence of the Discrete Element” 6 (1): 50–82.  
Williams, Ronald J., and David Zipser. 1990. *Gradient-Based Learning Algorithms for Recurrent Connectionist Networks*. Citeseer.  
Wilson, D. S. 1997. “Altruism and Organism: Disentangling the Themes of Multilevel Selection Theory.” *The American Naturalist* 150 Suppl 1 (July): S122–34.  
Wimmer, H., and J. Perner. 1983. “Beliefs about Beliefs: Representation and Constraining Function of Wrong Beliefs in Young Children’s Understanding of Deception.” *Cognition* 13 (1): 103–28.  
Woermann, Minka. 2016. *Bridging Complexity and Post-Structuralism*. Springer International Publishing.  
Woese, Carl R. 2002. “On the Evolution of Cells.” *Proceedings of the National Academy of Sciences* 99 (13): 8742–47.  
Wolfram, Stephen. 2002. *A New Kind of Science*. Champaign, IL: Wolfram Media.  
Wolpert, David H. 2024. “Implications of Computer Science Theory for the Simulation Hypothesis,” April. http://arxiv.org/abs/2404.16050.  
Wolpert, David H., and William Macready. 2007. “Using Self‐dissimilarity to Quantify Complexity.” *Complexity* 12 (3): 77–85.  
Wong, Michael L., Carol E. Cleland, Daniel Arend Jr, Stuart Bartlett, H. James Cleaves II, Heather Demarest, Anirudh Prabhu, Jonathan I. Lunine, and Robert M. Hazen. 2023. “On the Roles of Function and Selection in Evolving Systems.” *Proceedings of the National Academy of Sciences* 120 (43). https://doi.org/10.1073/pnas.2310223120.  
Worstall, Tim. 2014. “Paul Ehrlich Predicts Cannibalism, Plot Finally And Irretrievably Lost.” *Forbes*, May 23, 2014. https://www.forbes.com/sites/timworstall/2014/05/23/paul-ehrlich-predicts-cannibalism-plot-finally-and-irretrievably-lost/?sh=3ebf4558256b.  
Wright, L. 1972. “Explanation and Teleology.” *Philosophia Scientiae. Travaux d’Histoire Des Sciences et de Philosophie* 39 (June): 204–18.  
Wurm, Moritz F., D. Yves von Cramon, and Ricarda I. Schubotz. 2011. “Do We Mind Other Minds When We Mind Other Minds’ Actions? A Functional Magnetic Resonance Imaging Study.” *Human Brain Mapping* 32 (12): 2141–50.  
Yamaguchi, Kouichi, Kenji Sakamoto, T. Akabane, and Y. Fujimoto. 1990. “A Neural Network for Speaker-Independent Isolated Word Recognition.” *ICSLP*, November, 1077–80.  
Yamins, Daniel L. K., Ha Hong, Charles F. Cadieu, Ethan A. Solomon, Darren Seibert, and James J. DiCarlo. 2014. “Performance-Optimized Hierarchical Models Predict Neural Responses in Higher Visual Cortex.” *Proceedings of the National Academy of Sciences of the United States of America* 111 (23): 8619–24.  
Yao, Shunyu, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. 2024. “Tree of Thoughts: Deliberate Problem Solving with Large Language Models.” *Advances in Neural Information Processing Systems* 36. https://proceedings.neurips.cc/paper\_files/paper/2023/hash/271db9922b8d1f4dd7aaef84ed5ac703-Abstract-Conference.html.  
Yinusa, Ayoola R., and Chrystopher L. Nehaniv. 2011. “Study of Inheritable Mutations in von Neumann Self-Reproducing Automata Using the GOLLY Simulator.” In *2011 IEEE Symposium on Artificial Life (ALIFE)*, 211–17.  
[Yin, Zhangyue, Qiushi Sun, Qipeng Guo, Jiawen Wu, Xipeng Qiu, and Xuanjing Huang. 2023. “Do Large Language Models Know What They Don’t Know?” *arXiv [cs.CL]*. arXiv.](http://paperpile.com/b/iA68kr/kHfK) http://arxiv.org/abs/2305.18153.  
Yong, Ed. 2013. “How a Quarter of the Cow Genome Came from Snakes.” *National Geographic*, January. https://www.nationalgeographic.com/science/article/how-a-quarter-of-the-cow-genome-came-from-snakes.  
———. 2022. *An Immense World: How Animal Senses Reveal the Hidden Realms around Us*. Knopf Canada.  
Yoon, Youngwoo, Woo-Ri Ko, Minsu Jang, Jaeyeon Lee, Jaehong Kim, and Geehyuk Lee. 2019. “Robots Learn Social Skills: End-to-End Learning of Co-Speech Gesture Generation for Humanoid Robots.” In *2019 International Conference on Robotics and Automation (ICRA)*, 4303–9. IEEE.  
Yudkowsky, Eliezer. 2023a. “Pausing AI Developments Isn’t Enough. We Need to Shut It All down.” *Time Magazine* 29. https://empoweruohio.nyc3.cdn.digitaloceanspaces.com/Pausing%20AI%20Development.pdf.  
———. 2023b. *Will Superintelligent AI End the World?* https://www.ted.com/talks/eliezer\_yudkowsky\_will\_superintelligent\_ai\_end\_the\_world?language=en.  
Yu, S., and D. Nikolić. 2011. “Quantum Mechanics Needs No Consciousness.” *Annalen Der Physik* 523 (11): 931–38.  
Zakwan, Muhammad, Massimiliano d’Angelo, and Giancarlo Ferrari-Trecate. 2023. “Universal Approximation Property of Hamiltonian Deep Neural Networks.” *IEEE Control Systems Letters* 7: 2689–94.  
Zeghidour, Neil, Alejandro Luebs, Ahmed Omran, Jan Skoglund, and Marco Tagliasacchi. 2022. “SoundStream: An End-to-End Neural Audio Codec.” *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 30: 495–507.  
[Zelikman, Eric, Georges Harik, Yijia Shao, Varuna Jayasiri, Nick Haber, and Noah D. Goodman. 2024. “Quiet-STaR: Language Models Can Teach Themselves to Think Before Speaking.” *arXiv [cs.CL]*. arXiv.](http://paperpile.com/b/iA68kr/CYx0) http://arxiv.org/abs/2403.09629.  
Zhang, Haomin, Ian McLoughlin, and Yan Song. 2015. “Robust Sound Event Recognition Using Convolutional Neural Networks.” In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 559–63. IEEE.  
Zhan, Rafi S. 1964. “A Case of Hamlet Syndrome.” *Journal of Psychology (India)* 1 (1): 57–64.  
Ziv, Margalit, Ayelet Solomon, Sidney Strauss, and Douglas Frye. 2016. “Relations Between the Development of Teaching and Theory of Mind in Early Childhood.” *Journal of Cognition and Development: Official Journal of the Cognitive Development Society* 17 (2): 264–84.

1. A few recent books I find compelling but that claim AI is not “real” include Godfrey-Smith 2020; Seth 2021; Smith-Ruiu 2022; Christiansen and Chater 2022; Lane 2022; Humphrey 2023; M. Bennett 2023; K. J. Mitchell 2023; Mollick 2024. [↑](#footnote-ref-22)
2. Feigenbaum and Feldman 1963. [↑](#footnote-ref-23)
3. Phan and Park 2020. [↑](#footnote-ref-24)
4. von Helmholtz 1925 though as always when it comes to intellectual priority, an argument can be made that it goes back earlier—here, for instance, to Immanuel Kant. A modern, more mathematical articulation of this prediction principle would be made by the cyberneticists in the 1940s, as discussed in Part III; Rosenblueth, Wiener, and Bigelow 1943. [↑](#footnote-ref-25)
5. Dayan et al. 1995; Rao 2024. [↑](#footnote-ref-26)
6. Anil Seth and Andy Clark have also made a vigorous case for prediction as fundamental to intelligence, though their books don’t explicitly make the connection to predictive modeling in AI; (Seth 2021; Clark 2023). [↑](#footnote-ref-27)
7. Agüera y Arcas, Fairhall, and Bialek 2000, 2003; Hong, Agüera y Arcas, and Fairhall 2007. [↑](#footnote-ref-28)
8. R. Lee 2016. [↑](#footnote-ref-30)
9. Graeber 2015. [↑](#footnote-ref-31)
10. Levinson 2016; R. J. Gordon 2017; DeLong 2022. [↑](#footnote-ref-32)
11. Dean 2022. [↑](#footnote-ref-33)
12. Grace et al. 2017. [↑](#footnote-ref-34)
13. This paradoxical-seeming finding—that a seemingly narrow task contains the whole, just as much as the whole contains it—is a more general result, which we’ll return to in Part IV. [↑](#footnote-ref-35)
14. Thoppilan et al. 2022. [↑](#footnote-ref-36)
15. Agüera y Arcas 2022. [↑](#footnote-ref-37)
16. Agüera y Arcas and Norvig 2023. [↑](#footnote-ref-38)
17. Turing 1950. [↑](#footnote-ref-39)
18. Barnum 2024. [↑](#footnote-ref-40)
19. OpenAI et al. 2023; Bommineni, Bhagwagar, and Balcarcel 2023. [↑](#footnote-ref-41)
20. Kocijan et al. 2022. [↑](#footnote-ref-42)
21. J. Edwards 2023. [↑](#footnote-ref-43)
22. Marcus 2020; M. Mitchell 2024. [↑](#footnote-ref-44)
23. Polu et al. 2022. [↑](#footnote-ref-45)
24. Raghuraman, Harley, and Guibas 2023. [↑](#footnote-ref-46)
25. Nightingale and Farid 2022. [↑](#footnote-ref-47)
26. Gowal and Kohli 2023; Kirchenbauer et al. 2023. [↑](#footnote-ref-48)
27. For more technical treatments of functions as described in this book, see Fontana 1990; Wong et al. 2023. [↑](#footnote-ref-49)
28. W. Thomson 1871. [↑](#footnote-ref-51)
29. Peretó, Bada, and Lazcano 2009. [↑](#footnote-ref-52)
30. Robertson and Joyce 2012. [↑](#footnote-ref-53)
31. M. J. Russell and Martin 2004; Shapiro 2006. [↑](#footnote-ref-54)
32. Remember, from chemistry class, that since hydrogen (H) consists of a single proton bound to a single electron, a positively charged hydrogen ion with its electron stripped off (H+) is just a proton. When proton concentration varies over space, the ensuing flow of ions generates an electric current. [↑](#footnote-ref-55)
33. Krebs and Johnson 1937. Krebs won the Nobel Prize in Medicine for this work in 1953. [↑](#footnote-ref-56)
34. M. C. Evans, Buchanan, and Arnon 1966. [↑](#footnote-ref-57)
35. Buchanan and Arnon 1990. [↑](#footnote-ref-58)
36. Today, a great majority of the organic fuel consumed by respiration in animals and fungi comes from plants, which fix carbon using photosynthesis instead of the more ancient reverse Krebs cycle. [↑](#footnote-ref-59)
37. Szathmáry and Smith 1995. The somewhat narrower term “evolutionary transitions in individuality” (ETI) is also increasingly used; Michod 2000. [↑](#footnote-ref-61)
38. Sagan 1967. [↑](#footnote-ref-62)
39. Woese 2002. In this book, I’ll use the terms *symbiosis* and *symbiogenesis* in broader ways than most biologists do, to include cooperation and eventual interdependence among entities of all kinds—in the spirit of molecular biologists who have framed the origins of life as symbiosis among “molecules in mutualism,” per Lanier, Petrov, and Williams 2017. [↑](#footnote-ref-63)
40. Konstantinidis, Ramette, and Tiedje 2006; Murray, Gao, and Wu 2021. [↑](#footnote-ref-64)
41. Raup and Gould 1974. [↑](#footnote-ref-65)
42. In some cases, like the car versus the horse and buggy, or photosynthesis versus the reverse Krebs cycle, a new technology crowds out an old one due to greater efficiency; however, even here, the crowding-out is often not total. There are still a few horses and buggies, and a few organisms using the reverse Krebs cycle. [↑](#footnote-ref-66)
43. Turing 1937. [↑](#footnote-ref-68)
44. Technically, a Turing Machine, whether universal or not, must always be able to increase its tape size if needed, while real computers have fixed storage limits—but for most purposes, close enough. [↑](#footnote-ref-69)
45. von Neumann 1945. [↑](#footnote-ref-70)
46. Both died young, Turing in 1954, at age 41, and von Neumann in 1957, aged 53. [↑](#footnote-ref-71)
47. Turing 1952. [↑](#footnote-ref-72)
48. Many years later, German molecular biologist Christiane Nüsslein-Volhard identified the first real-life morphogen, the “Bicoid” protein, which forms a gradient along the fruit fly embryo to establish its segmented body plan. She won the Nobel Prize for this discovery, alongside Eric Wieschaus and Edward Lewis, in 1995. [↑](#footnote-ref-73)
49. von Neumann and Burks 1966. [↑](#footnote-ref-74)
50. In a more abstract form, mathematician Stephen Kleene had proven this result years earlier, a result known today as Kleene’s Second Recursion Theorem; Kleene 1938. [↑](#footnote-ref-75)
51. I’m using the term “complex” here in a colloquial sense; according to at least one technical definition (Kolmogorov complexity), if the instructions within the “simple” entity fully describe how to make the “more complex” one, they are actually of equal complexity. [↑](#footnote-ref-76)
52. J. D. Watson and Crick 1953. [↑](#footnote-ref-77)
53. While reversible computing is known to be theoretically possible, and has attracted some renewed interest in recent years because of its relevance to quantum computing, it remains largely unexplored. [↑](#footnote-ref-79)
54. [Turing [1951] 2000](https://paperpile.com/c/iA68kr/ldIO). [↑](#footnote-ref-80)
55. Turing 1948. Infants’ brains are of course not random. Turing was right to believe, though, that randomness plays a major role in their initial “wiring”; subsequent learning and development strengthens some connections, while pruning many others. [↑](#footnote-ref-82)
56. Pesavento 1995. [↑](#footnote-ref-83)
57. Mordvintsev et al. 2020. [↑](#footnote-ref-84)
58. Agüera y Arcas et al. 2024. This work draws inspiration from classic work in ALife, especially Barricelli 1957; Fontana 1990; Ray 1991; Adami and Brown 1994. [↑](#footnote-ref-86)
59. Decrementing 0 wraps around to 255, the largest possible value for a byte, and incrementing 255 wraps around to 0. [↑](#footnote-ref-87)
60. You might have noticed that the original Brainfuck had 8 instructions, including “.” for “print” and “,” for “input.” Bff uses only “,” because instead of reading and writing to an external terminal or console, the tape reads and writes to *itself*. Thus only a single instruction is needed for copying a byte from one location to another. In our first paper describing bff, we included instructions “{” and “}” for moving the console pointer, but a stripped-down alternative, used here, simply initializes the data and console pointers with the first two bytes on the tape. [↑](#footnote-ref-88)
61. Schrödinger 1944. [↑](#footnote-ref-89)
62. H. Reichenbach 1956; Rovelli 2018. [↑](#footnote-ref-91)
63. The “very nearly” qualifier arises from the deep mathematical structure of relativistic quantum mechanics. A number of near-symmetries in physics, such as between positive or negative charge, left or right-handedness, and time reversal, are not quite exact in quantum field theory, though according to the CPT theorem—the letters stand for charge, parity, time—reversing all three at once *does* yield an exact symmetry. [↑](#footnote-ref-92)
64. For the moment, I’m sidestepping a direct definition of the term *disorder* and its apparent subjectivity (is my desk disorderly or does it reflect an order only I can see?), but we will soon return to the central role of a *model* in distinguishing order from disorder; this is the link connecting thermodynamics to information theory. [↑](#footnote-ref-93)
65. The development of clocks and other complex machines in the Middle Ages set the scene, per Prigogine and Stengers 1984: “The clock world is a metaphor suggestive of God the Watchmaker, the rational master of a robot-like nature.” Newtonian dynamics reinforced the idea that the universe’s dynamics were deterministic and knowable, hence “robot-like.” [↑](#footnote-ref-94)
66. Smil 2008. [↑](#footnote-ref-95)
67. Agüera y Arcas 2023. [↑](#footnote-ref-96)
68. *PV*=*nRT*, where *P* is pressure, *V* is volume, *T* is temperature, *n* is the number of gas molecules or billiard balls, and *R* is the “ideal gas constant.” [↑](#footnote-ref-97)
69. Kinetic energy is proportional to the mass and squared velocity of the molecules. [↑](#footnote-ref-98)
70. More precisely, these are repulsive interactions between the electron orbitals of the molecules. [↑](#footnote-ref-99)
71. W. Thomson 1857. [↑](#footnote-ref-100)
72. Pross 2012. [↑](#footnote-ref-102)
73. Taleb 2014. [↑](#footnote-ref-103)
74. In the language of dynamical systems theory, this asserts that thermodynamic stability is a fixed point, while DKS is a limit cycle. [↑](#footnote-ref-104)
75. Dawkins 1986. [↑](#footnote-ref-106)
76. Kauffman 1971. [↑](#footnote-ref-107)
77. In classical mythology, chimeras are “unnatural” combinations of animals—for instance, a fusion of lion, goat, and snake. As we’ll see, they aren’t so unnatural after all. [↑](#footnote-ref-108)
78. Yinusa and Nehaniv 2011. [↑](#footnote-ref-109)
79. This is not an absolute rule, since new parts can also evolve within an existing whole. [↑](#footnote-ref-111)
80. Lander et al. 2001. [↑](#footnote-ref-112)
81. Baltimore 1970. [↑](#footnote-ref-113)
82. McClintock 1950. [↑](#footnote-ref-114)
83. Tarlinton, Meers, and Young 2006. [↑](#footnote-ref-115)
84. She et al. 2022. [↑](#footnote-ref-116)
85. Horita and Fukumoto 2023. [↑](#footnote-ref-117)
86. Chuong 2018; Bakoulis et al. 2022; Russ and Iordanskiy 2023. Knocking out the “Arc” retrotransposon, likely of viral origin, renders lab mice unable to form new long-term memories, per Pastuzyn et al. 2018. [↑](#footnote-ref-118)
87. Ryan 2009. [↑](#footnote-ref-119)
88. Sharp and Hahn 2011. [↑](#footnote-ref-120)
89. Crosby 2003; Harper et al. 2008. [↑](#footnote-ref-121)
90. Feschotte and Gilbert 2012; Naville et al. 2016. [↑](#footnote-ref-122)
91. Nasir and Caetano-Anollés 2015. [↑](#footnote-ref-123)
92. Yong 2013. [↑](#footnote-ref-124)
93. Unlike eukaryotes, prokaryotes (bacteria and archaea) have very low proportions of “junk DNA.” This is because copying and expressing DNA are energetically expensive, and prokaryotes are energy-constrained, limiting their maximum DNA length and imposing strong evolutionary pressure to rid themselves of any code that can’t pull its own weight. However, a prokaryote’s “core genome” is typically only 60–70% of the total, with the remainder varying widely among strains. (Hence two *E. coli* bacteria can be as different from each other as a human and a fruit fly!) Horizontal gene transfer via transposons, viruses, or “plasmids” (exchangeable DNA loops), enables widespread copying of code snippets among bacteria, resulting in a complex tangle of reproducing bits of code conceptually similar to that described here for eukaryotes. [↑](#footnote-ref-125)
94. Though earlier, we similarly considered individual bff instructions as minimal or “atomic” replicating entities. [↑](#footnote-ref-126)
95. Mandelbrot rediscovered and generalized these findings in the 1960s; see Mandelbrot 1967. [↑](#footnote-ref-128)
96. West 2017. [↑](#footnote-ref-129)
97. For the mathematically inclined: power laws relating two variables *x* and *y* take the general form *y*=*axb*. [↑](#footnote-ref-130)
98. West, Brown, and Enquist 1997. [↑](#footnote-ref-131)
99. Incidentally, De Morgan was also computing pioneer Ada Lovelace’s math tutor. Charles Babbage conceived the Analytical Engine, a steampunk computer that, had it been built, would have predated the ENIAC by a century; Lovelace wrote presciently about the power of general-purpose computing—and arguably wrote the first computer program—in 1843. [↑](#footnote-ref-132)
100. De Morgan and De Morgan 1872. [↑](#footnote-ref-133)
101. Wolpert and Macready 2007; Bagrov et al. 2020. [↑](#footnote-ref-134)
102. This makes sense, since the underlying processes that shape coastlines differ between scales of, say, a meter, versus a kilometer; see Mandelbrot 1989. [↑](#footnote-ref-135)
103. These particular numbers correspond to the smoothest, *N*=5, and roughest, *N*=8, of Mandelbrot’s coastline diagrams in his classic 1967 paper, “How Long Is the Coast of Britain?” [↑](#footnote-ref-136)
104. If I were a fancier writer and used fewer stock phrases, the book wouldn’t ZIP so efficiently. Using the same compression algorithm, James Joyce’s *Ulysses* can only be squeezed down to 40% of its raw size! [↑](#footnote-ref-137)
105. In math: for a random string of symbols that can assume *K* values, a string of length *L* would occur with frequency *f*=*K*−*L*, while with replication occurring at all scales, string frequency scales like *f*~*L*−*b*. [↑](#footnote-ref-138)
106. Christley et al. 2009. [↑](#footnote-ref-139)
107. Recursion can’t be implemented by the ribosome on its own; I’ve used a bit of poetic license in identifying the ribosome with machine A. In reality, as morphogenesis illustrates, biological computation is more holistic and distributed. [↑](#footnote-ref-141)
108. Carroll, Grenier, and Weatherbee 2013. [↑](#footnote-ref-142)
109. Walker 2023. [↑](#footnote-ref-144)
110. Computer viruses are a notable exception. [↑](#footnote-ref-145)
111. J. Bennett 2010. [↑](#footnote-ref-146)
112. [[Citation error]](http://127.0.0.1:8080/c/error). [↑](#footnote-ref-147)
113. Schrödinger 1944. [↑](#footnote-ref-149)
114. Thomas 1934. [↑](#footnote-ref-150)
115. We’ve seen how even a virus needs an envelope, or it gives up the ability to exist outside its host cell and becomes a transposon. [↑](#footnote-ref-153)
116. Stock and Levit 2000. [↑](#footnote-ref-154)
117. Berg 1975; Adler 1966. [↑](#footnote-ref-155)
118. Lazova et al. 2011. [↑](#footnote-ref-156)
119. Mathematical readers: to first order, if *X* is a sequence of discrete events, then to first order, *P*(*X*) will be a causal filter turning these discrete events into a continuous time-varying probability estimate. The beautiful accuracy/latency tradeoffs bacteria perform (in effect, changing the filter shape) can be thought of as solving for the optimal linear prediction of *X* under changing statistical conditions. [↑](#footnote-ref-158)
120. Wheeler 1992. [↑](#footnote-ref-160)
121. Although, to get ahead of our argument for a moment, how could that improbably chair-like configuration of atoms have come into existence in the first place, *without* any beings that model chairs? [↑](#footnote-ref-161)
122. Originally, Conway used a Go board. Go is an ancient strategy game, invented more than 2,500 years ago in China, involving two players alternately placing white and black stones at the unoccupied intersections on a 19×19 grid. [↑](#footnote-ref-162)
123. Loizeau 2021. [↑](#footnote-ref-163)
124. Bradbury 2012. [↑](#footnote-ref-164)
125. Technical caveat: to simulate quantum mechanics, a random number source would also be needed. [↑](#footnote-ref-165)
126. There is also the question of initial conditions, of which more anon. [↑](#footnote-ref-166)
127. This understanding or knowledge about the world of Life is consistent with, and even derivable from the basic rules, just as statistical mechanics allows temperature and pressure to be derived from lower-level laws of elastically colliding particles in a box (which are themselves derivable from still more fundamental physics). It would be odd, though, to claim that gliders are “implied” or “inherent” in the rules of Life, for Life’s Turing completeness means that the complexity of the constructs possible in Life is unbounded; it includes anything and everything. [↑](#footnote-ref-167)
128. (Niemiec 2010). [↑](#footnote-ref-169)
129. Hawking 2011. [↑](#footnote-ref-170)
130. By creating a numbering system that can enumerate all possible cellular automata, Stephen Wolfram and collaborators have started to explore the question of how “unusual” computation is in absolute terms. Rule number 110 has been proven to support computation—and many of the first hundred or so rules are so trivial that they create no meaningful dynamics at all. Observations like this, and the simplicity of Conway’s Life, suggest that a universe doesn’t need to be fine-tuned to support computation. See Wolfram 2002; Cook 2004. [↑](#footnote-ref-171)
131. Plants and fungi, life forms that lack muscles and so only move very slowly, should not be understood as passive. On the contrary, as the scientists who study them are increasingly appreciating, they compensate for immobility with the ability to take an extraordinary variety of *chemical* actions. [↑](#footnote-ref-173)
132. Part IV will explain reinforcement learning and its relationship to the brain in more detail. [↑](#footnote-ref-174)
133. More formally, the joint distribution *P*(*X*,*H*,*O*) includes the “marginal distributions” *P*(*H*) and *P*(*X*), which we would get by summing the joint distribution over (*X*,*O*) or (*H*,*O*), and it also includes all kinds of conditional distributions. [↑](#footnote-ref-175)
134. Dennett 2017b. [↑](#footnote-ref-176)
135. More formally, it requires reducing the dimensionality of the distribution by factoring out invariant dimensions, per Tishby, Pereira, and Bialek 2000. [↑](#footnote-ref-178)
136. A 4K (3,840×2,160 pixel) video stream, compressed, is typically between 8 and 14 million bits per second. [↑](#footnote-ref-179)
137. This may not seem like much of a model, but it’ll do for modeling some simple things, like, with a negative *w*, the restoring force exerted by stretching a spring by distance *x*. [↑](#footnote-ref-181)
138. Although the distinction isn’t nearly as straightforward as it sounds, as discussed at the end of Part I. [↑](#footnote-ref-182)
139. MacKay 1969; G. Bateson 1972b. [↑](#footnote-ref-183)
140. Turing 1937. [↑](#footnote-ref-184)
141. I’ve written the programmer’s version with lowercase variable names because in mathematical notation, capital letters are generally used for “random variables” which can assume a whole distribution of values. [↑](#footnote-ref-185)
142. Hofstadter 2007. [↑](#footnote-ref-186)
143. Although this terminology is a bit confusing, machine learning practitioners refer to evaluation of their trained models as “inference.” [↑](#footnote-ref-188)
144. Turing 1948. [↑](#footnote-ref-189)
145. Setting aside other potential life in the Universe, we’re on the cusp of this no longer being true on Earth, as AI models are now being used to design other AI models, per Duéñez-Guzmán et al. 2023. [↑](#footnote-ref-190)
146. This line of thinking suggests a potentially fertile research direction. Evolution-inspired approaches to learning today are mostly based on random mutation, per Salimans et al. 2017. But as we’ve seen, symbiogenesis is far more powerful as a source of productive novelty. [↑](#footnote-ref-191)
147. “In general” because group-level fitness benefits from individual sacrifice under certain circumstances, as with bees defending the hive. Also, in species like the octopus, where individuals breed over a single season and aren’t involved in raising the young, there’s no evolutionary pressure on survival after reproduction. [↑](#footnote-ref-192)
148. This is an alternative formulation of “Occam’s Razor,” which holds that, all things being equal, a simpler explanation is more likely to be true than a complicated one. [↑](#footnote-ref-193)
149. Wright 1972. Note to physicists: it’s possible to formulate physics in terms of “principles of least action” rather than equivalent dynamical laws. This would involve evaluating all possible trajectories of the boulder to find a minimum (or, occasionally, maximum) action solution, wherein the car gets squished. One could argue that in the action formulation *all* effects precede their causes, per Tegmark 2018. In my view, though, by removing any notion of causal ordering, the least action approach presupposes a timeless “block universe” with no role for concepts like “why” and “because.” We’ll return to this topic in Part VI. [↑](#footnote-ref-195)
150. Deacon 2012. [↑](#footnote-ref-196)
151. Per the broadening of terms suggested in Part I, this definition of “purposive” or even “alive” includes technological systems with built-in decisionmaking powers, from thermostats to cruise missiles, as well as artificially evolved systems like bff. [↑](#footnote-ref-197)
152. Hume 1817. [↑](#footnote-ref-199)
153. Wallisch 2017. [↑](#footnote-ref-200)
154. Dick 1985. [↑](#footnote-ref-201)
155. Clark 2023. [↑](#footnote-ref-203)
156. KUSA-TV 2005; Nitsch, Verheggen, and Merten 2007. [↑](#footnote-ref-204)
157. Fisher, Hassan, and O’Connor 1995. [↑](#footnote-ref-205)
158. Clark 2023. [↑](#footnote-ref-206)
159. More thorough histories often begin with the Antikythera mechanism, a mechanical astronomical computer recovered from a shipwreck in the Aegean sea thought to date back to the second century BCE. Many early civilizations performed astronomical and calendrical calculations, whether for navigation, agriculture, or religious rites; if by “computer” we simply mean something engineered to aid in such calculations, numerous ancient archeological sites also qualify. [↑](#footnote-ref-209)
160. This translation is from B. Russell 1937. [↑](#footnote-ref-210)
161. Leibniz 1666. [↑](#footnote-ref-211)
162. The most influential mathematician of the late nineteenth and early twentieth centuries, David Hilbert (1862–1943), had placed the *Entscheidungsproblem* at the center of his program for formalizing mathematics in the 1920s. [↑](#footnote-ref-212)
163. Leibniz wasn’t the first or only mathematician to discover binary; as with nearly every significant discovery or invention, it was made by a number of other thinkers around the same time, including Thomas Heriot (1560–1621), Francis Bacon (1561–1626), and Juan Caramuel y Lobkowitz (1606–1682). See Glaser 1971. [↑](#footnote-ref-213)
164. Per Mary Everest Boole, “George afterwards learned, to his great joy, that the same conception of the basis of Logic was held by Leibnitz […].” [↑](#footnote-ref-214)
165. This also coincided with the early exuberance of the metric system, to sweep away older “irrational” and aristocratically tainted systems of measurement. Unfortunately for Prony, the metric approach to angles (100 grades to the right angle, instead of 90 degrees) never caught on, rendering many of the tables useless. [↑](#footnote-ref-215)
166. Grattan-Guinness July-Sep 1990; Grier 2008. [↑](#footnote-ref-216)
167. According to Babbage 1832, “[T]hese persons were usually found more correct in their calculations, than those who possessed a more extensive knowledge of the subject.” [↑](#footnote-ref-217)
168. A. Smith 1776. [↑](#footnote-ref-218)
169. Martin 1992. [↑](#footnote-ref-219)
170. The class and gender iniquities in this factory arrangement would, of course, be endlessly repeated in centuries to come. [↑](#footnote-ref-220)
171. Bromley 1982. [↑](#footnote-ref-221)
172. Babbage 1832. [↑](#footnote-ref-222)
173. Per Babbage 1864, “To those acquainted with the principles of the Jacquard loom, and who are also familiar with analytical formulæ, a general idea of the means by which the Engine executes its operations may be obtained without much difficulty.” Jacquard-style punched cards were in fact used extensively for data processing and computing, right up until the 1980s. [↑](#footnote-ref-223)
174. Lovelace’s father was the notorious Romantic poet, Lord Byron; her parents had stayed together for less than a year, owing to Lord Byron’s instability and scandalous lifestyle (including, Lady Byron feared, an incestuous relationship with his half-sister). [↑](#footnote-ref-224)
175. A portion of this prototype survived, and is now at the Science Museum in London. The Difference Engine, like many special-purpose computing machines from the first half of the twentieth century, was not Turing complete; it could only tabulate a limited (though useful) repertoire of functions using the “method of differences” (math folk: the discrete version of a polynomial series). [↑](#footnote-ref-225)
176. Lovelace 1842. [↑](#footnote-ref-226)
177. The Fibonacci numbers are closely related to the Bernoulli numbers; see Byrd 1975. [↑](#footnote-ref-227)
178. Surrey Comet 1946. Earlier special-purpose computers were also regularly referred to this way, most famously, “Old Brass Brains,” a long-lived mechanical computer for predicting tides invented in 1895 and in use by the US Coast and Geodetic Survey from 1910 to 1965. [↑](#footnote-ref-228)
179. Babbage 1832. [↑](#footnote-ref-229)
180. “The London Literary Gazette; and Journal of Belles Lettres, Arts, Sciences, &c” 1832; Shelley 1818. [↑](#footnote-ref-230)
181. Lovelace 1842. In his 1948 report on “Intelligent Machinery,” Turing referred to this as “Lady Lovelace’s Objection” to the idea that machines could be intelligent. [↑](#footnote-ref-231)
182. G. Boole 1854. [↑](#footnote-ref-232)
183. Labatut 2024. [↑](#footnote-ref-233)
184. According his widow, Mary Everest Boole, in a 1901 letter: “nearly all the logicians and mathematicians ignored the statement that the book was meant to throw light on the nature of the human mind; and treated the formula entirely as a wonderful new method of reducing to logical order masses of evidence about external fact […]. De Morgan, of course, understood the formula in its true sense; he was Boole’s collaborator all along.” M. E. Boole 1901. [↑](#footnote-ref-234)
185. Shelley 1818. [↑](#footnote-ref-235)
186. Letter of 15 November 1844, at Ashley Combe, to Woronzow Greig, Russian Ambassador in London Lovelace 1992. [↑](#footnote-ref-236)
187. “Sensation is aroused by the messages which are transmitted through the nerves from the sense organs to the brain […]. [T]he messages […] consist of […] brief impulses in each nerve fibre; all the impulses are very much alike, whether the message is destined to arouse the sensation of light, of touch, or of pain […].” Adrian 1928. [↑](#footnote-ref-237)
188. Lucas and Adrian 1917. [↑](#footnote-ref-238)
189. The basic Boolean operators take two inputs and produce one output; they are AND, OR, and XOR (short for “eXclusive OR”). AND outputs a one only if both inputs are one, and a zero otherwise; OR outputs a zero only if both inputs are zero, and a one otherwise; and XOR outputs a one if one and only one input is one. The single-input NOT operator inverts its input, outputting a one if the input is zero, and a zero if the input is one. Combinations of these operators can be used to make any more complex logical function. [↑](#footnote-ref-239)
190. McCulloch and Pitts 1943. [↑](#footnote-ref-240)
191. Harvard University’s Mark I computer, conceived by Howard Aiken in 1937 and built in the 1940s, was closely modeled on the Analytical Engine, including its reliance on base-ten arithmetic. After 1943, virtually all electronic computers used binary. [↑](#footnote-ref-241)
192. As mentioned earlier, all the computing pioneers took the step of going from logic machines designed to carry out a particular computational task to programmable general purpose machines, which once built could carry out coded instructions to perform *any* computational task—equivalent to a Universal Turing Machine. This was to be the difference between Babbage’s Difference Engine and the Analytical Engine. It was also the difference between the fixed-function computing machines made prior to 1945 and those after, with the ENIAC straddling the transition. (ENIAC began as a fixed-function computer that needed to be “programmed” by rewiring its plugboard, but it was updated to become fully programmable in software by 1947.) [↑](#footnote-ref-242)
193. Agüera y Arcas, Fairhall, and Bialek 2000. [↑](#footnote-ref-243)
194. S. J. Smith et al. 2020. [↑](#footnote-ref-244)
195. Turing 1951. [↑](#footnote-ref-245)
196. Haugeland 1985. [↑](#footnote-ref-246)
197. Richards and Lillicrap 2022. [↑](#footnote-ref-247)
198. Clarke 1968. [↑](#footnote-ref-248)
199. Although curiously, they took for granted that computers would soon be able to recognize faces and objects, understand speech, and perform many other everyday feats that had proven just as difficult. Perhaps these tasks seemed too cut-and-dried, too narratively uninteresting, or just too lowly to serve as a basis for human exceptionalism. [↑](#footnote-ref-249)
200. Such hierarchies of intelligence trace farther back—to Descartes, and even to Aristotle—but industrial production made the hierarchy even more literal. [↑](#footnote-ref-250)
201. Riskin 2003. [↑](#footnote-ref-251)
202. Cavalier-Smith 2002; Speijer 2016. [↑](#footnote-ref-253)
203. McDonald, Rice, and Desai 2016. [↑](#footnote-ref-254)
204. Schultz et al. 2023. [↑](#footnote-ref-255)
205. Northcutt 2012. [↑](#footnote-ref-256)
206. More correctly, matter-energy is conserved, but on Earth, outside nuclear power plants and atomic bombs, energy and matter are individually conserved, or near enough. [↑](#footnote-ref-257)
207. Most plants, of course, don’t eat other life. *Their* source of free energy is the stream of low entropy photons emanating from our nearest star. The Earth as a whole re-radiates higher entropy photons back into space, maintaining an energetic and entropic balance—although the greenhouse effect, by diminishing the amount of re-radiation that can escape into space, is now upsetting this balance, resulting in rising entropy on Earth. [↑](#footnote-ref-258)
208. Per the brief discussion of competence without comprehension in Part II, this doesn’t necessarily imply that any such modeling is conscious. As we’ll explore in Part VII, what we think of as consciousness is *modeling your own model*, not merely having (that is, acting in accordance with) a model. [↑](#footnote-ref-259)
209. Babbage 1864. [↑](#footnote-ref-261)
210. Scott 1998; Soll 2011. [↑](#footnote-ref-262)
211. They were: Kay McNulty, Betty Snyder, Marlyn Wescoff, Ruth Lichterman, Betty Jean Jennings, and Fran Bilas Fritz Fall 1996; Light 1999. [↑](#footnote-ref-263)
212. Typically of “firsts,” there are other contenders. Working in relative isolation, Konrad Zuse, a German civil engineer, developed the electromechanical Z3 computer in 1943. Zuse’s project was also motivated by war; his earlier S1 and S2 were special-purpose machines for computing aerodynamic corrections to the wings of radio-controlled flying bombs, and after 1939 he was funded by the Nazi government. While Zuse didn’t design the Z3 with Turing completeness in mind, and it didn’t natively support conditional jumps, it can be programmed cleverly to simulate them. This arguably gives the Z3 priority over the ENIAC, per Rojas July-Sep 1998. [↑](#footnote-ref-264)
213. Fitzpatrick 1998. [↑](#footnote-ref-265)
214. Moore 1965; Schaller 1997. [↑](#footnote-ref-266)
215. IBM 2007. [↑](#footnote-ref-267)
216. Renting the 701 in 1952 cost between $12,000 and $18,000 per month. Ibid. [↑](#footnote-ref-268)
217. Waldrop 2018. [↑](#footnote-ref-269)
218. De Monchaux 2011. [↑](#footnote-ref-270)
219. Rid 2016. [↑](#footnote-ref-272)
220. Mindell 2002. [↑](#footnote-ref-273)
221. Galison 1994. [↑](#footnote-ref-274)
222. Mischiati et al. 2015. [↑](#footnote-ref-275)
223. In patients with “intention tremor,” caused by cerebellar damage, purposive actions lead to overshoot, resulting in high-amplitude oscillations of a reaching limb just as Wiener’s theory would predict when the feedback loop is impaired; Rosenblueth, Wiener, and Bigelow 1943. [↑](#footnote-ref-276)
224. Corcoran and Conner 2016. [↑](#footnote-ref-277)
225. Donsker and Varadhan 1975; Berg 1993. [↑](#footnote-ref-278)
226. Glimcher 2005; Faisal, Selen, and Wolpert 2008; Braun 2021. [↑](#footnote-ref-279)
227. Domenici et al. 2008. [↑](#footnote-ref-280)
228. Genevieve Bell, an anthropologist and Vice-Chancellor of the Australian National University (ANU), founded ANU’s School of Cybernetics in 2021. I wholeheartedly endorse her revival of the term “cybernetics” to include not only AI but the larger set of multidisciplinary topics touched on here. [↑](#footnote-ref-281)
229. Rosenblueth, Wiener, and Bigelow 1943. [↑](#footnote-ref-282)
230. A notable exception is the centrifugal governor, first invented by Christiaan Huygens in the seventeenth century, then adapted by James Watt in 1788. Watt’s version controlled the amount of steam going into a steam engine using a valve coupled to whirling weights driven by the engine, creating a negative (or homeostatic) feedback loop to regulate the engine’s speed. [↑](#footnote-ref-283)
231. The authors dodge the issue, writing “A discussion of causality, determinism and final causes is beyond the scope of this essay […] however […] purposefulness, as defined here, is quite independent of causality, initial or final.” For the moment, we’re also sidestepping the random element inherent to quantum mechanics, though we will return to randomness later. [↑](#footnote-ref-284)
232. Philosopher Daniel Dennett has called this, or something much like it, the “intentional stance” Dennett 1989. [↑](#footnote-ref-286)
233. Rosenblueth, Wiener, and Bigelow 1943. [↑](#footnote-ref-287)
234. Chomsky 1959; Wann 1964. [↑](#footnote-ref-288)
235. MacKenzie 1993. [↑](#footnote-ref-289)
236. Espejo 2014. [↑](#footnote-ref-290)
237. G. Bateson 1972a; Kline 2020. [↑](#footnote-ref-291)
238. Meadows et al. 1972. [↑](#footnote-ref-292)
239. Jacobs 1961. [↑](#footnote-ref-293)
240. Money 1984; Downing, Morland, and Sullivan 2015. [↑](#footnote-ref-294)
241. Haraway 1985. [↑](#footnote-ref-295)
242. Cilliers 1998; Woermann 2016. [↑](#footnote-ref-296)
243. Clynes and Kline 1960. [↑](#footnote-ref-297)
244. Rid 2016. [↑](#footnote-ref-298)
245. A few years later, in the Soviet Union, cybernetics underwent a parallel rise and fall; see Peters 2016. [↑](#footnote-ref-299)
246. Nilsson 2009. [↑](#footnote-ref-300)
247. Dreyfus 1972. [↑](#footnote-ref-301)
248. James 1890. The salient chapter (IV —*Habit*) was published earlier, in 1887. [↑](#footnote-ref-303)
249. Ferreira, Nogueira, and Defelipe 2014. [↑](#footnote-ref-304)
250. Pavlov 1902. [↑](#footnote-ref-305)
251. Pitts and McCulloch 1947. [↑](#footnote-ref-306)
252. Borges 1942. [↑](#footnote-ref-307)
253. von Ahn et al. 2003; Chew and Tygar 2004. [↑](#footnote-ref-308)
254. Hendler 2006; Verita 2006. [↑](#footnote-ref-309)
255. Hubel and Wiesel 2004. [↑](#footnote-ref-311)
256. Livingstone 2013. [↑](#footnote-ref-312)
257. Rosenblatt 1957. [↑](#footnote-ref-313)
258. Until the late 2010s, video games drove the market for GPUs. [↑](#footnote-ref-314)
259. The massively parallel Connection Machine, invented by Danny Hillis and colleagues in the 1980s, was a notable exception, consisting of an array of small, local processors wired up to their neighbors in a 3D torus. As a “connectionist,” Hillis and his colleagues were convinced that their architecture was the future of computing—and specifically of AI. They were too far ahead of the curve, though. Precisely because the Connection Machine was such an architectural outlier, there was never a viable software or research ecosystem for it; see Hillis 1985. [↑](#footnote-ref-315)
260. Schmidhuber 2014. [↑](#footnote-ref-316)
261. Minsky and Papert 1969. [↑](#footnote-ref-317)
262. Cybenko 1989. [↑](#footnote-ref-318)
263. Per Kurzweil 2024, “[B]ack in 1964 Rosenblatt explained to me that the Perceptron’s inability to deal with invariance was due to a lack of layers. If you took the output of a Perceptron and fed it back to another layer just like it, the output would be more general and, with repeated iterations of this process, would increasingly be able to deal with invariance. If you had enough layers and enough training data, it could deal with an amazing level of complexity. I asked him whether he had actually tried this, and he said no but that it was high on his research agenda. It was an amazing insight, but Rosenblatt died only seven years later, in 1971, before he got the chance to test his insights.” [↑](#footnote-ref-319)
264. Sejnowski and Rosenberg 1987. [↑](#footnote-ref-320)
265. By “distributed representations,” the authors meant representations of concepts based on the values of many neurons, also known as neural “embeddings.” Neural embeddings will be described in detail in Part IV. [↑](#footnote-ref-321)
266. (Bender and Koller 2020; Chomsky et al. 2023). [↑](#footnote-ref-322)
267. This, at least, was the common wisdom from the world of statistical modeling. The reality turns out to be more complex, as will be described in Part IX. [↑](#footnote-ref-324)
268. A complete list would be impossible to make, especially without some historical distance; an enormous volume of papers are being published today advocating for one new twist or another. Years can elapse before we know which of these will matter, and sometimes a trick must be reinvented multiple times, with minor variations, before being widely adopted. This process bears more than a passing resemblance to evolution. [↑](#footnote-ref-325)
269. Fukushima 1980; LeCun et al. 1989. [↑](#footnote-ref-326)
270. Boltzmann 1868; Bridle 1990. [↑](#footnote-ref-327)
271. Yamaguchi et al. 1990. [↑](#footnote-ref-328)
272. LeCun et al. 1995. [↑](#footnote-ref-329)
273. Norbert Wiener, Warren McCulloch, Walter Pitts, Frank Rosenblatt, Kunihiko Fukushima, Alan Turing, Geoffrey Hinton, Terry Sejnowski, Rich Sutton, Karl Friston, Yann LeCun, Read Montague, Jürgen Schmidhuber, Yoshua Bengio, and Peter Dayan are all important figures in the development of machine learning mentioned in this book who work or worked in the tradition of computational neuroscience, or NeuroAI. [↑](#footnote-ref-330)
274. LeCun et al. 1990, 1998. [↑](#footnote-ref-331)
275. Krizhevsky, Sutskever, and Hinton 2012. [↑](#footnote-ref-332)
276. Zhang, McLoughlin, and Song 2015. [↑](#footnote-ref-333)
277. Weyn, Durran, and Caruana 2020. [↑](#footnote-ref-334)
278. Dauphin et al. 2017. [↑](#footnote-ref-335)
279. A flip-flop is a basic electrical circuit that can “flip” into one state, or “flop” into another; thus it is invaluable for implementing digital logic. A multi-stable flip-flop has more than two stable states. [↑](#footnote-ref-337)
280. Rosenblatt 1957. [↑](#footnote-ref-338)
281. Daubechies et al. 2019. [↑](#footnote-ref-339)
282. The term “inductive bias” has been used to describe the latent structural assumptions made by an approximator (whether neural or otherwise) to learn a function from a finite set of examples. [↑](#footnote-ref-340)
283. Many refinements to the basic perceptron architecture, including convolutions, different nonlinearities, “skip connections” (K. He et al. 2016) and so on can be understood as further biasing the class of functions to be learned, with the goal of making the learning more efficient (that is, requiring fewer parameters or training examples to achieve similar accuracy). [↑](#footnote-ref-341)
284. Bojarski et al. 2016. [↑](#footnote-ref-342)
285. This idea had been tried before, with some success. ALVINN, the Autonomous Land Vehicle in a Neural Network Pomerleau 1989, was a much smaller fully connected three-layer perceptron built in 1988—the year Minsky and Papert reissued *Perceptrons*—and could “effectively follow real roads under certain field conditions,” a milestone GOFAI systems had not (and have not) achieved. [↑](#footnote-ref-343)
286. Rosenblatt 1957. [↑](#footnote-ref-344)
287. However, we’ll return to these points when discussing the phenomenon of “in-context learning” in Part IX. [↑](#footnote-ref-345)
288. In 2017, during the golden age of convolutional neural nets, an episode aired of the painfully accurate *Silicon Valley* TV comedy involving a “Not Hotdog app,” which was entirely equivalent to my scenario here; (Seppala 2017). I’m just going for a healthier menu. [↑](#footnote-ref-348)
289. Technical note: the dimensionality reduction approach Refik used was “t-distributed stochastic neighbor embedding” (t-SNE), which is nonlinear, so linear separation afterward won’t in general work. One would need to use a linear approach to dimensionality reduction, such as “Principal Component Analysis” (PCA), and more than three dimensions would probably be needed for bananas to become linearly separable. [↑](#footnote-ref-350)
290. Frankle and Carbin 2018; Hunter, Spracklen, and Ahmad 2022; Zonglin Li et al. 2022. [↑](#footnote-ref-351)
291. Tan et al. 2018. [↑](#footnote-ref-352)
292. This neglects the slight change we would also need to make to the “not banana” neuron to make it a “neither banana nor apple” neuron, but that’s a minor detail. [↑](#footnote-ref-353)
293. Qi, Brown, and Lowe 2017. [↑](#footnote-ref-354)
294. Feingold 1914; Meissner and Brigham 2001. [↑](#footnote-ref-356)
295. Tanaka and Pierce 2009. [↑](#footnote-ref-357)
296. Buolamwini and Gebru 23–24 Feb 2018; M. Wang et al. 2019. [↑](#footnote-ref-358)
297. In more precise mathematical language, they show that while the specific weights of a given network are determined by their random initialization, these weights are only unique up to an arbitrary rotation in the high-dimensional space of per-layer weights; see Guth et al. 2023; Guth and Ménard 2024. [↑](#footnote-ref-359)
298. Fairwork 2023; Meaker 2023. [↑](#footnote-ref-360)
299. Some details are glossed over in this brief description; see Kaiming He et al. 2021. [↑](#footnote-ref-361)
300. Inexplicably, our retinas are wired back to front, with the plumbing in front of the photoreceptors—nearly a proof of lack of intelligent design. Octopus retinas, which evolved independently, are wired the right way round, with the light sensors in front. [↑](#footnote-ref-363)
301. These experiments are reviewed in Churchland, Ramachandran, and Sejnowski 1994. [↑](#footnote-ref-364)
302. Such recordings are rare; they are made when neurosurgeons are trying to figure out how best to operate on someone’s brain when they need to, while sparing as much function as possible. Researchers gather precious human data of this kind when they can, but any scientific aims are always secondary to clinical aims. [↑](#footnote-ref-365)
303. Gross 2002. [↑](#footnote-ref-366)
304. Quiroga et al. 2005. [↑](#footnote-ref-367)
305. Jennifer Aniston and Brad Pitt were married from 2000–2005, so this neuron might have needed to find a new gig after they split. [↑](#footnote-ref-368)
306. This is 128×127×126×125×124×123×122×121×120×119×118×117×116×115×114×113. [↑](#footnote-ref-369)
307. Originally, these terms came from fencing: whoever is thrusting is the actor, and whoever is receiving the blow is the patient. The term “patient” in medical contexts is a vestige of this usage, since the “patient” is supposed to lie passively while the surgeon or physician lances boils, cuts for stone, saws off limbs, punctures veins for bloodletting, or conducts other such swashbuckling maneuvers. [↑](#footnote-ref-371)
308. In most cases there’s at least one synapse enroute, but let’s not sweat the anatomical details. [↑](#footnote-ref-372)
309. Strogatz 2004. [↑](#footnote-ref-373)
310. Sarfati et al. 2022. [↑](#footnote-ref-374)
311. H. Wang et al. 2023; Hanson 2023. [↑](#footnote-ref-375)
312. As philosopher Peter Godfrey-Smith points out, “In the sea, animals have various body plans. On land, all animals are bilaterian. There are no terrestrial jellyfish.” Godfrey-Smith 2024. [↑](#footnote-ref-376)
313. Goulty et al. 2023. [↑](#footnote-ref-378)
314. Heath 1963. [↑](#footnote-ref-379)
315. Berridge and Robinson 1998. [↑](#footnote-ref-380)
316. Achatz et al. 2013. [↑](#footnote-ref-381)
317. Raz et al. 2017. [↑](#footnote-ref-382)
318. Local adaptation also results in weak forms of learning known as “sensitization” and “habituation”; Cheng 2021. [↑](#footnote-ref-384)
319. R. R. Bush and Mosteller 1951; Rescorla 1972. [↑](#footnote-ref-385)
320. Sutton 1988. [↑](#footnote-ref-386)
321. Ibid, emphasis mine. [↑](#footnote-ref-387)
322. Tesauro 1992, 1994. [↑](#footnote-ref-388)
323. For a comprehensive historical overview of the development of TD learning and its convergence with dopamine research, see Colombo 2014. [↑](#footnote-ref-390)
324. Romo and Schultz 1990. [↑](#footnote-ref-391)
325. Quartz et al. 1992; Montague, Dayan, and Sejnowski 1996. [↑](#footnote-ref-392)
326. The original experiments were done with primates, but evidence suggests that dopamine plays a similar role, mediated by the same brain structures, in all vertebrates. Analogous systems appear to have evolved in invertebrates; Mizunami and Matsumoto 2017. [↑](#footnote-ref-393)
327. Fouragnan, Retzler, and Philiastides 2018; Diederen and Fletcher 2021. [↑](#footnote-ref-394)
328. C.-S. Lee et al. 2016. [↑](#footnote-ref-395)
329. Shoham et al. 2017. [↑](#footnote-ref-396)
330. Churchland 2016. [↑](#footnote-ref-397)
331. For instance, Sutton 1988 describes how TD learning and supervised learning can both be understood as gradient descent methods for sequence prediction. [↑](#footnote-ref-398)
332. Today’s separate treatment of inference (typically involving short timescales), so-called “adaptation” (intermediate timescales), and learning (long timescales) is an artifact of the cybernetic formalism described in Part III. [↑](#footnote-ref-399)
333. See Part I. [↑](#footnote-ref-400)
334. The parable is named after the 14th century French philosopher Jean Buridan, but versions of it long predate him. [↑](#footnote-ref-403)
335. The earliest versions of randomized or “Monte Carlo” methods (named after the famous casino) were implemented at Los Alamos by Augusta Teller and Arianna Rosenbluth to simulate chain reactions in nuclear weapons. [↑](#footnote-ref-404)
336. Silver et al. 2017. [↑](#footnote-ref-405)
337. Schrittwieser et al. 2020. [↑](#footnote-ref-406)
338. Silver et al. 2013. [↑](#footnote-ref-407)
339. Kahng and Lee 2016. [↑](#footnote-ref-408)
340. Churchland 2016. [↑](#footnote-ref-409)
341. It does, in an oblique way, maintain a memory of prior *moves*: the 19×19 pixel “image” used by AlphaGo’s convolutional nets to evaluate the board includes a “channel” representing the number of moves since the stone was placed. Presumably this feature improves the model’s performance, though nothing about the system’s design (or optimal gameplay) requires it. [↑](#footnote-ref-411)
342. von Neumann 1928; von Neumann and Morgenstern 1944. [↑](#footnote-ref-412)
343. My old college flatmate used to hustle chess games in Washington Square Park. First, he and I would play for money; I was genuinely bad, but he only pretended to be. He would then clean up against the other chess hustlers, who would initially play sloppily, assuming he was an easy mark. Afterward, we would split the proceeds. [↑](#footnote-ref-413)
344. R. R. Jackson and Pollard 1996; Harland and Jackson 2000. [↑](#footnote-ref-414)
345. “Jumping spider tricksters: deceit, predation, and cognition” in Bekoff, Allen, and Burghardt 2002. [↑](#footnote-ref-415)
346. Tchaikovsky 2018. [↑](#footnote-ref-417)
347. Chiang 2005. [↑](#footnote-ref-418)
348. Hofstadter 1982. [↑](#footnote-ref-419)
349. Chittka 2022. [↑](#footnote-ref-420)
350. Fabre 1916. [↑](#footnote-ref-421)
351. Insects exhibit meaningful individual differences in intelligence, just as humans do; when Fabre tried the trick on a different *Sphex* population of the same species, he found that they weren’t fooled, per Chittka 2022. We’ll return to the *Sphex* story with a more critical eye in Part IX. [↑](#footnote-ref-422)
352. Riskin 2016. [↑](#footnote-ref-423)
353. In reality, no living being can be truly identical with another, as even the simplest organism has a vast number of internal degrees of freedom, and many are neutral, i.e. confer neither a survival advantage nor disadvantage. [↑](#footnote-ref-424)
354. Wimmer and Perner 1983; Baron-Cohen, Leslie, and Frith 1985. [↑](#footnote-ref-426)
355. Setoh, Scott, and Baillargeon 2016. [↑](#footnote-ref-427)
356. Ding et al. 2015. Note that the spider may still exhibit social competence without comprehension—meaning that it may not model *itself* attributing false beliefs to others, which would amount to a higher-order theory of mind. [↑](#footnote-ref-428)
357. Humphrey 1976. [↑](#footnote-ref-430)
358. Ibid. [↑](#footnote-ref-431)
359. The basic idea had been floated by other researchers in the 1950s (Chance and Mead 1953) and again in the 1960s (Jolly 1966), though to less effect. [↑](#footnote-ref-432)
360. Whiten and Byrne 1997; Waal 2007. [↑](#footnote-ref-433)
361. Dunbar 1998. [↑](#footnote-ref-434)
362. Sarah Blaffer Hrdy has convincingly argued that among humans, and certain other brainy species, it “takes a village” to raise an infant, in the sense that the mother alone can’t provide all the needed calories; the grandparents, siblings, and babysitters who help out are called “alloparents.” In alloparental species, prospective mothers without the needed social support disproportionately elect infanticide or, among humans, abortion. If they do have the baby, its odds of survival increase appreciably with alloparental help; Hrdy 2009. [↑](#footnote-ref-435)
363. Pawłowski, Lowen, and Dunbar 1998; Holt-Lunstad et al. 2015. [↑](#footnote-ref-436)
364. Powell et al. 2012. [↑](#footnote-ref-437)
365. There’s likely no sharp boundary between general skills and specific knowledge, in this or any other domain, because every specific thing you learn is represented in terms of your existing conceptual vocabulary, and in turn extends that vocabulary. One can think of it almost like a compression algorithm, in which one’s experiences so far make up the “dictionary” with which subsequent knowledge and experiences are compressed. This may (at least partly) explain why time seems to pass more quickly as we age: we’re compressing our life experiences more efficiently—alas. [↑](#footnote-ref-438)
366. Dunbar 1992. [↑](#footnote-ref-439)
367. Dunbar 2016. [↑](#footnote-ref-440)
368. Gopnik and Meltzoff 1993; Ziv et al. 2016. [↑](#footnote-ref-441)
369. Muthukrishna et al. 2014. [↑](#footnote-ref-442)
370. Minsky 1988; Hawkins 2021; M. Bennett 2023. [↑](#footnote-ref-444)
371. Mountcastle 1957; Buxhoeveden and Casanova 2002; Cadwell et al. 2019. [↑](#footnote-ref-445)
372. Sharma, Angelucci, and Sur 2000. [↑](#footnote-ref-446)
373. Kolarik et al. 2014. [↑](#footnote-ref-447)
374. Chebat, Schneider, and Ptito 2020. [↑](#footnote-ref-448)
375. Godfrey-Smith 2016. [↑](#footnote-ref-449)
376. Exceptional sites have been found where groups of octopuses live together; Scheel et al. 2017. [↑](#footnote-ref-450)
377. Since its enormous size makes it so easy to work with in the lab, the squid giant axon played a starring role in the midcentury experiments that established the biophysics of the action potential. [↑](#footnote-ref-451)
378. Sivitilli and Gire 2019. [↑](#footnote-ref-452)
379. Crook and Walters 2014. [↑](#footnote-ref-453)
380. Godfrey-Smith 2016. [↑](#footnote-ref-454)
381. Kuuspalu, Cody, and Hale 2022. [↑](#footnote-ref-455)
382. Budelmann 1998. [↑](#footnote-ref-456)
383. S. L. Bush 2012. [↑](#footnote-ref-457)
384. D. J. Brown 2013. [↑](#footnote-ref-458)
385. Galen 1968. [↑](#footnote-ref-460)
386. Descartes 1662. [↑](#footnote-ref-461)
387. Dennett 2017a; Schneider 2021; D. J. Chalmers 2022 [↑](#footnote-ref-462)
388. Bailey 2003; Dennett 2009 [↑](#footnote-ref-464)
389. (Harris 2012; Sapolsky 2023). [↑](#footnote-ref-465)
390. Bostrom 2003; Feldman 2019. [↑](#footnote-ref-466)
391. Tables and chairs seem to be invoked by philosophers more often than anything else in an appeal to “reality,” presumably because they are omnipresent when such debates are taking place. [↑](#footnote-ref-467)
392. Mogavero and Laskar 2022. [↑](#footnote-ref-470)
393. Lorenz 1963. [↑](#footnote-ref-471)
394. Eguíluz et al. 2000. [↑](#footnote-ref-472)
395. Bell 1964. [↑](#footnote-ref-473)
396. This is not in any way an appeal to “quantum consciousness.” The idea that quantum effects are needed to account for neural computation is a fringe position with no mainstream neuroscientific support. At body temperature, thermal noise is so many orders of magnitude greater than quantum uncertainty that a direct role for quantum phenomena in neural processing is doubtful. Nonetheless, quantum mechanics establishes important “ground truth” in any argument about determinism, the limits of physical prediction, and the nature of “ground truth” itself, as will soon be discussed. [↑](#footnote-ref-474)
397. Magnasco 2022. [↑](#footnote-ref-475)
398. Though something like this has actually been tried, with some success; Spiegel and Rodríguez 2017. [↑](#footnote-ref-477)
399. You may notice that this looks a lot like evolution taking place in imaginary worlds. [↑](#footnote-ref-478)
400. Sapolsky 2023. [↑](#footnote-ref-479)
401. Let’s assume he didn’t mean to gender it. Schopenhauer 2011. [↑](#footnote-ref-480)
402. Free will in the sense described here is broadly consistent with the views of German neuroscientist Björn Brembs, who points out how, curiously, studies of insect cognition have been the most illuminating; Brembs 2011. [↑](#footnote-ref-481)
403. Yong 2022. [↑](#footnote-ref-482)
404. Classic studies explore the self-recognition “mirror test” in chimpanzees, human infants, and other animals; Gallop 1970; Amsterdam 1972. [↑](#footnote-ref-483)
405. Hamilton, Hofstadter, and Dennett 1984. [↑](#footnote-ref-484)
406. D. Chalmers 1995. [↑](#footnote-ref-486)
407. Trnka and Lorencova 2022. [↑](#footnote-ref-487)
408. Henrich, Heine, and Norenzayan 2010. [↑](#footnote-ref-488)
409. Descartes’s answer, the pineal gland, takes the prize for originality. [↑](#footnote-ref-489)
410. Kimmerer 2013, 2017. [↑](#footnote-ref-490)
411. Varro et al. 1934; J. P. Lewis 2013. [↑](#footnote-ref-491)
412. Tracy Gleason, *Murray: The Stuffed Bunny*, in Turkle 2011. [↑](#footnote-ref-492)
413. An exception is neuroscientist Michael Graziano, whose “Attention Schema Theory” and social perspective on consciousness largely agrees with mine. See Graziano 2013. [↑](#footnote-ref-493)
414. A. Harris 2019. [↑](#footnote-ref-494)
415. Freedom Baird, then a grad student at MIT, proposed holding Furby upside-down as a kind of “emotional Turing Test” in 1999; the experiment was actually conducted with 7-year-olds, albeit informally, on the Radiolab podcast in 2011. See Forbes 2021. [↑](#footnote-ref-496)
416. “Declaration of Independence: A Transcription” n.d.. [↑](#footnote-ref-497)
417. There is an element of social performativity in these judgments, also; it’s ironic, for instance, that in the treatment of rats, we hold labs up to a far higher standard today than exterminators. Perhaps that can be understood relationally too: when rats are antagonists, we have less empathy for them than when they are, in effect, collaborators. [↑](#footnote-ref-499)
418. Fein et al. 2019. [↑](#footnote-ref-500)
419. Rovelli 2021. [↑](#footnote-ref-501)
420. Ornes 2019. [↑](#footnote-ref-502)
421. P. J. Lewis 2021; Gao 2018; Okon and Sebastián 2020; Yu and Nikolić 2011; D. J. Chalmers and McQueen 2021. [↑](#footnote-ref-503)
422. Wheeler 1978. [↑](#footnote-ref-504)
423. Jacques et al. 2007. [↑](#footnote-ref-505)
424. Rovelli 2021. [↑](#footnote-ref-506)
425. de La Mettrie 1748b. [↑](#footnote-ref-508)
426. Isaacs 1987; van der Hart, Lierens, and Goodwin 1996; Pederzoli, Tressoldi, and Wahbeh 2022; Perrotta 2019; Dein and Illaiee 2013. [↑](#footnote-ref-510)
427. 100 billion neurons recorded at one bit per millisecond. [↑](#footnote-ref-511)
428. Even though we can’t get anywhere close to recording from every neuron in the human brain, this may be a surprisingly achievable goal, given how correlated brain activity is. Neuroscientists can already approximately reconstruct what you’re seeing based on functional MRI recording; Kay et al. 2008; Naselaris et al. 2009; Nishimoto et al. 2011. [↑](#footnote-ref-512)
429. L. Saxe, Dougherty, and Cross 1985; Honts, Raskin, and Kircher 1994; Iacono 2008. [↑](#footnote-ref-513)
430. D. J. Brown 2013 [↑](#footnote-ref-515)
431. Fumagalli and Priori 2012; Dax and Freudenberg 1948; Banay and Davidoff 1942; Braslow 1999. [↑](#footnote-ref-516)
432. Panahi et al. 2017. [↑](#footnote-ref-517)
433. Gazzaniga 1967; Sperry 1968; Gazzaniga 2005. [↑](#footnote-ref-518)
434. de Haan et al. 2020. [↑](#footnote-ref-519)
435. Preilowski 1975; Sperry 1965. [↑](#footnote-ref-520)
436. Obvious costs of a big brain are: high energy consumption (the human brain accounts for about 20% of metabolism despite only weighing about 2% of body mass), and high-risk childbirth due to the difficulty of fitting the infant’s skull through the mother’s pelvis. Additionally, our big brains are delicate and accident-prone, require a very long childhood to learn adult skills, and require a great many genes to build and maintain. [↑](#footnote-ref-521)
437. Stroop 1992. [↑](#footnote-ref-522)
438. Volz and Gazzaniga 2017. [↑](#footnote-ref-524)
439. A. Harris 2019. [↑](#footnote-ref-525)
440. Gazzaniga 1998. [↑](#footnote-ref-526)
441. Volz and Gazzaniga 2017. [↑](#footnote-ref-527)
442. Due to state regulations they each needed to be licensed separately. Coordination between strangers to perform collective motor tasks has also been studied in the lab; Schmid and Braun 2020. [↑](#footnote-ref-528)
443. Van de Crommert HW, Mulder, and Duysens 1998. [↑](#footnote-ref-529)
444. Johansson et al. 2005. [↑](#footnote-ref-530)
445. Chater 2018. [↑](#footnote-ref-531)
446. Hall et al. 2013. [↑](#footnote-ref-532)
447. Chater 2018. [↑](#footnote-ref-533)
448. Graziano 2018. [↑](#footnote-ref-536)
449. Botvinick and Cohen 1998. [↑](#footnote-ref-537)
450. One set of such findings not discussed here is the controversial work of Benjamin Libet and colleagues, who were able to detect brain signals predicting voluntary action significantly before subjects reported that they had made the decision to act. While there are many methodological challenges with this work and a number of very different interpretations of the results, the absence of a homunculus implies that *no* specific part of the brain can accurately exactly say when a decision is made, especially if that decision originates elsewhere. Libet et al. 1993. [↑](#footnote-ref-538)
451. Understanding the precise nature of these necessary and sufficient conditions is an ongoing area of research. [↑](#footnote-ref-539)
452. Leamer 1996; Goodwin 1991. [↑](#footnote-ref-541)
453. Rowland 2005. [↑](#footnote-ref-542)
454. Walter Freeman and Watts 1950. [↑](#footnote-ref-543)
455. Yamins et al. 2014. [↑](#footnote-ref-544)
456. Miłkowski 2018; Bruder 2017; Laird, Lebiere, and Rosenbloom 2017; A. Saxe, Nelli, and Summerfield 2020. [↑](#footnote-ref-545)
457. Whittington and Bogacz 2019. [↑](#footnote-ref-546)
458. Athalye et al. 10–15 Jul 2018. [↑](#footnote-ref-547)
459. Waldrop 2019; Heaven 2019. [↑](#footnote-ref-548)
460. Guo et al. 17–23 Jul 2022. [↑](#footnote-ref-549)
461. Williams and Zipser 1990. [↑](#footnote-ref-551)
462. In this discussion, I’ll pretend that time occurs in discrete steps, as it in fact mostly does in the present-day world of digital computing and artificial neural nets. Real brains and neurons don’t work in discrete time steps, but the sequential nature of information processing as a signal passes from neuron to neuron still applies. A more faithful model of “neural time” would be based only on local interactions, and would thus only impose local time ordering, as with functional programming. [↑](#footnote-ref-552)
463. Huang, Liu, and Weinberger 2016. [↑](#footnote-ref-553)
464. DeepMind’s PixelRNN architecture, published in 2016, offered a minimal proof of concept, modeling (hence predicting) the contents of a static image a single pixel at a time. Pixels were scanned in reading order, from top left to bottom right. Of course, spiders (and people) decide on their own scanning pattern; van den Oord, Kalchbrenner, and Kavukcuoglu 20–22 Jun 2016. [↑](#footnote-ref-554)
465. Harland and Jackson 2000. [↑](#footnote-ref-555)
466. Montague and Sejnowski 1994; Saponati and Vinck 2023. [↑](#footnote-ref-557)
467. Keller and Mrsic-Flogel 2018. [↑](#footnote-ref-558)
468. This can be proven using straightforward extensions of the universal function approximation theorems introduced in Part V. Hornik, Stinchcombe, and White 1989; Zakwan, d’Angelo, and Ferrari-Trecate 2023. [↑](#footnote-ref-559)
469. Pouget 2014. [↑](#footnote-ref-560)
470. Matthews 1982. [↑](#footnote-ref-561)
471. Merker 2007. [↑](#footnote-ref-563)
472. Cassell 1998. [↑](#footnote-ref-564)
473. Grandin 1980, 1993. [↑](#footnote-ref-565)
474. Humphrey and Weiskrantz 1967. [↑](#footnote-ref-567)
475. Humphrey 2023, p. 40. [↑](#footnote-ref-568)
476. Humphrey 1972. [↑](#footnote-ref-569)
477. Hartmann et al. 1991. [↑](#footnote-ref-570)
478. chewbster 2008. [↑](#footnote-ref-571)
479. M. Bennett 2023. [↑](#footnote-ref-573)
480. Yong 2022. [↑](#footnote-ref-574)
481. There were some minor exceptions; for instance, through frequent reinforcement, it was shown to be possible for an existing memory to be modified to include new information. [↑](#footnote-ref-575)
482. Nádasdy et al. 1999; A. K. Lee and Wilson 2002. [↑](#footnote-ref-576)
483. Hartley 1834; Killgore 2010; Rasch and Born 2013. [↑](#footnote-ref-577)
484. No, your cat doesn’t have a barrel cortex. Rats make serious use of their whiskers, actively whisking with them and building up a detailed mental representation of the somatosensory environment, while animals like cats seem to have them mainly to gauge whether they will fit through an opening… or maybe just for looks. [↑](#footnote-ref-578)
485. Friston 2005; Kilner, Friston, and Frith 2007. [↑](#footnote-ref-579)
486. Antonio R. Damasio 1999; Eslinger and Damasio 1985; A. R. Damasio, Damasio, and Chui 1980. [↑](#footnote-ref-580)
487. Völlm et al. 2006; Wurm, von Cramon, and Schubotz 2011; van Veluw and Chance 2014. [↑](#footnote-ref-581)
488. W. Freeman 1971; Snow 1949. [↑](#footnote-ref-582)
489. For a general overview of this “paradoxical” functional improvement following brain lesion, see (Kapur 1996). [↑](#footnote-ref-583)
490. Gunz et al. 2010; Pearce, Stringer, and Dunbar 2013. [↑](#footnote-ref-584)
491. Henneberg 1988; Henneberg and Steyn 1993. [↑](#footnote-ref-585)
492. Ghirlanda, Lind, and Enquist 2017. [↑](#footnote-ref-586)
493. Clark 2023. [↑](#footnote-ref-588)
494. Huttenlocher 1979. [↑](#footnote-ref-589)
495. Steklis 2012; Owren, Amoss, and Rendall 2011; Belin 2006. [↑](#footnote-ref-590)
496. Duchenne 1876; Ekman, Davidson, and Friesen 1990. [↑](#footnote-ref-591)
497. Kobayashi and Kohshima 1997. [↑](#footnote-ref-592)
498. It’s becoming increasingly clear that many historical debates pitting the individual selection against group-level selection schools of Darwinism against each other are misguided. Both occur at once. Wilson 1997; O’Gorman, Sheldon, and Wilson 2008. [↑](#footnote-ref-593)
499. Churchland 2019. [↑](#footnote-ref-594)
500. Hrdy 2009. [↑](#footnote-ref-595)
501. *How Smart Are Dolphins?* 2011. [↑](#footnote-ref-598)
502. Tweti 2008; Pepperberg and Pepperberg 2009; Craige 2010; Colbert-White 2013; also *Cosmo the Funny Parrot* n.d. [↑](#footnote-ref-599)
503. Chiang and Allora & Calzadilla 2015. [↑](#footnote-ref-600)
504. Everett 2009. [↑](#footnote-ref-601)
505. Rutz et al. 2023. [↑](#footnote-ref-602)
506. Perruchet and Rey 2005. [↑](#footnote-ref-603)
507. Mercier and Sperber 2011, 2018. [↑](#footnote-ref-604)
508. Abramson et al. 2018; Filatova et al. 2011; Hynninen, n.d. [↑](#footnote-ref-605)
509. Gibson 1993; Case 1985; Wilden 1972; Nguyen, Sagot, and Dupoux 2022. [↑](#footnote-ref-606)
510. Slobodchikoff, Paseka, and Verdolin 2009. [↑](#footnote-ref-607)
511. A caveat: it is certainly possible for actions whose consequences can’t be perceived by an individual to nonetheless be learned on an evolutionary timescale as instincts, if they confer a survival advantage. [↑](#footnote-ref-608)
512. Gibson Gieseking et al. 2014; Greeno 1994. [↑](#footnote-ref-609)
513. Kudo and Matsumoto 2001; Webster and Kit 1992. [↑](#footnote-ref-611)
514. Levesque, Davis, and Morgenstern 2011. [↑](#footnote-ref-613)
515. Kocijan et al. 2022. [↑](#footnote-ref-614)
516. Translate uses a much smaller model than general-purpose chatbots. With a large model, it would likely perform Winograd Schema tasks with human-like accuracy. [↑](#footnote-ref-615)
517. Adiwardana et al. 2020. [↑](#footnote-ref-616)
518. Thoppilan et al. 2022. [↑](#footnote-ref-617)
519. Lenat, Prakash, and Shepherd 1985; McCarthy et al. 2002. [↑](#footnote-ref-619)
520. Jones et al. 2022. [↑](#footnote-ref-620)
521. More technically, the prediction is given by the softmax of the dot product of every word in the vocabulary with the vector sum of the surrounding words. [↑](#footnote-ref-621)
522. Mikolov et al. 2013. [↑](#footnote-ref-622)
523. B. Russell 1937. [↑](#footnote-ref-623)
524. Most of taste is really smell. [↑](#footnote-ref-624)
525. This notation, familiar to all of us who took the dreaded SAT in the 1990s, reads “king is to queen as man is to woman.” [↑](#footnote-ref-625)
526. For a more detailed treatment of how unsupervised multilingual learning can be harnessed to perform translation, see Artetxe et al. 2017. [↑](#footnote-ref-626)
527. Coincidentally, in the language of vectors, the word “translation” literally means just such a parallel shift. [↑](#footnote-ref-627)
528. Rovelli 2023. [↑](#footnote-ref-628)
529. Caliskan, Bryson, and Narayanan 2017. [↑](#footnote-ref-630)
530. Boniol et al. 2019. [↑](#footnote-ref-631)
531. Leaf 1946. [↑](#footnote-ref-632)
532. As in: (1) Thou shalt not make unto thee any graven image, (2) Thou shalt not take the name of the Lord thy God in vain, (5) Thou shalt not murder, (6) Thou shalt not commit adultery, (7) Thou shalt not steal, (8) Thou shalt not bear false witness against thy neighbor, (9) Thou shalt not covet thy neighbor’s house, (10) Thou shalt not covet thy neighbor’s wife or his slaves, or his animals, or anything of thy neighbor. [↑](#footnote-ref-633)
533. Kuczmarski 2018. [↑](#footnote-ref-634)
534. See also p. 114 of Rae et al. 2021. [↑](#footnote-ref-635)
535. Vaswani et al. 2017. [↑](#footnote-ref-637)
536. This description is a bit simplified, as the goal here is to understand the model conceptually; normalization and multi-head attention aren’t described. For readers familiar with linear algebra, the original paper offers a clear and complete description. [↑](#footnote-ref-638)
537. Hochreiter and Schmidhuber 1997. [↑](#footnote-ref-639)
538. Simplified alternatives to the LSTM were also introduced in the 2010s, most notably the Gated Recurrent Unit (GRU); Chung et al. 2014. Broadly speaking, they share both the advantages and drawbacks of LSTMs. [↑](#footnote-ref-640)
539. Baddeley and Graham 1974; Moscovitch and Winocur 2002; Cowan 2001; Logie 2011. [↑](#footnote-ref-642)
540. Foer 2005. [↑](#footnote-ref-643)
541. Thalmann, Souza, and Oberauer 2019; Ellis 2001; Norris and Kalm 2021. [↑](#footnote-ref-644)
542. Cherry 1953. [↑](#footnote-ref-645)
543. Carlile and Corkhill 2015. [↑](#footnote-ref-646)
544. Rahimi, Afouras, and Zisserman 2022. [↑](#footnote-ref-647)
545. Patten 1990. [↑](#footnote-ref-648)
546. Whittington, Warren, and Behrens 2021. [↑](#footnote-ref-649)
547. This is analogous to the way “simple cells” and “complex cells” in the visual cortex have receptive fields that look like oriented ripples. Such ripples are the most natural building blocks of images (in mathematical language, eigenfunctions). [↑](#footnote-ref-650)
548. Kozachkov, Kastanenka, and Krotov 2023. [↑](#footnote-ref-651)
549. Fields et al. 2014. [↑](#footnote-ref-652)
550. This means that the size of a Transformer goes up like the square of the size of the context window; clearly such scaling runs into hard limits before long. [↑](#footnote-ref-653)
551. Bhojanapalli et al. 2021. [↑](#footnote-ref-654)
552. Zhongli Li et al. 2021; Patel, Bhattamishra, and Goyal 2021; Pardos and Bhandari 2023. [↑](#footnote-ref-656)
553. Hao 2020; Nelson and The Conversation US n.d.. [↑](#footnote-ref-657)
554. Farnsworth and Beecham 1999; Viswanathan et al. 2011; Bartumeus et al. 2014. [↑](#footnote-ref-658)
555. Giannou et al. 23–29 Jul 2023; Merrill and Sabharwal 2023. [↑](#footnote-ref-659)
556. B. Edwards 2022. [↑](#footnote-ref-660)
557. Geneva and Zabaras 2022. [↑](#footnote-ref-661)
558. Jason Wei et al. 2022. [↑](#footnote-ref-663)
559. Kojima et al. 2022 [↑](#footnote-ref-664)
560. J. Kaplan et al. 2020; Bubeck and Sellke 2022. [↑](#footnote-ref-665)
561. This cliff, though, has no top. In the words of Vannevar Bush, it offers an “endless frontier”; Park 2019. [↑](#footnote-ref-666)
562. The Pirahã people, for example, lack the cultural concept of numbers, and can’t do problems like this; Frank et al. 2008; P. Gordon 2004. [↑](#footnote-ref-667)
563. Daniel Kahneman, winner of the Nobel Prize in Economics in 2002 for work in this vein, popularized this idea in his 2011 book, *Thinking, Fast and Slow*; Kahneman 2011. [↑](#footnote-ref-670)
564. Hagendorff, Fabi, and Kosinski 2022. [↑](#footnote-ref-671)
565. Mercier and Sperber 2018. [↑](#footnote-ref-672)
566. Along similar anthropocentric lines, Aristotle thought of humans as the unique possessors of a “rational soul,” layered atop the merely “vegetative soul” also possessed by plants, and the “sensitive soul” (a.k.a. System 1) also possessed by other animals. It was a fairly obvious move to meld this schema with religious faith by identifying the rational soul with the (also uniquely human) Christian soul, as Descartes did. To many modern philosophers and cognitive scientists, it still seems counterintuitive to imagine that the machinery of the rational soul could be the same as that of the “irrational” sensitive soul. This is another manifestation of the GOFAI fallacy that rational behavior and formal logic are one and the same. [↑](#footnote-ref-673)
567. The CPU of my first computer, the Texas Instruments TI-99/4A, contained only about 8,000 transistors. An Apple M2 chip has more than *16 million times* as many. These enormous transistor counts are for parallel processing, and nowadays are increasingly used to run large neural nets. [↑](#footnote-ref-674)
568. Chittka 2022. [↑](#footnote-ref-676)
569. Dennett 1984. [↑](#footnote-ref-677)
570. Fabre 1921. [↑](#footnote-ref-678)
571. Keijzer 2013. [↑](#footnote-ref-679)
572. Chittka and Spaethe 2007; MaBouDi et al. 2020; Nityananda and Chittka 2015; Dukas 2009; Chittka and Thomson 1997. [↑](#footnote-ref-680)
573. Koch 2008. [↑](#footnote-ref-681)
574. Mazokhin-Porshnyakov and Kartsev 2000. [↑](#footnote-ref-682)
575. Eldan and Li 2023. [↑](#footnote-ref-683)
576. Bees have about one million neurons, and given the very complex structures of the Kenyon cells in the mushroom body, it seems safe to assume at least an order of magnitude more parameters. It is also known that both the relative and absolute size of a bee’s brain correlates with its ability to learn Collado et al. 2021; Lanuza et al. 2023. To be clear, though, it’s highly unlikely that bees could learn to understand short stories, even if their brains were theoretically complex enough to do so. Their *Umwelt* is radically different from ours, and while they are good learners, worker bees only live for a few weeks. [↑](#footnote-ref-684)
577. Cruse and Hübner 2008; Spaethe, Tautz, and Chittka 2006. [↑](#footnote-ref-685)
578. Simons and Chabris 1999. [↑](#footnote-ref-686)
579. Garcia et al. 2017; Paulk et al. 2009; Honkanen et al. 2023. [↑](#footnote-ref-687)
580. Bernadou, Kramer, and Korb 2021; Holldobler and Wilson 2009; Page 2012. [↑](#footnote-ref-688)
581. K. Frisch 1955; Karl von Frisch and Seeley 2013. [↑](#footnote-ref-689)
582. After reading the “eight-legged cats” article about *Portia* Harland and Jackson 2000, my sister Clea has taken to calling her cat, Guar, a “four-legged spider.” Guar is an enthusiastic stalker of prey, both real and pretend, though not necessarily the brightest spark. [↑](#footnote-ref-690)
583. The French term, coined by encyclopedist Denis Diderot, is *l’esprit d’escalier*—literally, “staircase wit.” [↑](#footnote-ref-691)
584. Weir 2012. [↑](#footnote-ref-692)
585. As one would expect given that its text-specific function is a recent cultural development, it has other functions too, and is tellingly implicated in attention generally, per Chen et al. 2019. [↑](#footnote-ref-694)
586. Pinker 2010. [↑](#footnote-ref-695)
587. Christiansen and Chater 2022. [↑](#footnote-ref-696)
588. Petkov and Ten Cate 2020; Christiansen et al. 2002; Conway and Pisoni 2008. [↑](#footnote-ref-697)
589. G. Kaplan 2023; E. A. Smith 2010. [↑](#footnote-ref-698)
590. Fine-grained control of the vocal tract has also been posited as a language pre-adaptation, and it may indeed be. Parrots and cetaceans, both of which are well adapted to language like us, are also gifted at sound synthesis. (In fact some brainy bird species are able to make a wider array of vocalizations than we are, producing uncanny imitations of ringtones, vehicles, and power tools, as well as blood-curdling “syrinx shrieks” loud enough to serve as acoustic weapons.) However, nonhuman primates are highly dextrous, and can learn the elements of sign language. Deaf communities have no problem communicating this way. The limits to language acquisition among nonhuman primates thus appear to arise from more fundamental limits to sequence learning and/or sociality. [↑](#footnote-ref-699)
591. The caveat is that these modalities have to communicate, directly or indirectly, with the parts of our cortex that process language. So, for instance, we have neurons that sense the state of every part of our gut, squeeze the food through, and perform emergency evacuations when needed, but we don’t have non-technical language to describe these experiences or the sense that we voluntarily control them, because those circuits aren’t wired up to the interpreter in any way that allows us to talk about what they do beyond vague generalities like “feeling full” or “having stomach cramps.” [↑](#footnote-ref-700)
592. Marjieh et al. 2023. [↑](#footnote-ref-701)
593. Zeghidour et al. 2022. This description simplifies away some details. For instance, SoundStream actually creates *two* token streams: high-frequency “acoustic tokens” and low-frequency “semantic tokens.” [↑](#footnote-ref-703)
594. Borsos, Marinier, et al. 2023. [↑](#footnote-ref-704)
595. Borsos, Sharifi, et al. 2023. [↑](#footnote-ref-705)
596. Chomsky 1959, 1980. [↑](#footnote-ref-706)
597. Ibbotson and Tomasello 2016. [↑](#footnote-ref-707)
598. Ghani et al. 2023; Agostinelli et al. 2023. [↑](#footnote-ref-708)
599. Hopper 1996. [↑](#footnote-ref-709)
600. Pagel et al. 2013; Heggarty et al. 2023; Ruhlen 1994. [↑](#footnote-ref-711)
601. Ćwiek et al. 2022. [↑](#footnote-ref-712)
602. Alper and Averbuch-Elor 2023. [↑](#footnote-ref-713)
603. Adams 1979. [↑](#footnote-ref-714)
604. Anderson 2004; “The World” n.d.. [↑](#footnote-ref-715)
605. Somé et al. 2013; Nag 2017; Dombrowsky-Hahn and Slezak 2004. [↑](#footnote-ref-716)
606. Agüera y Arcas 2023. [↑](#footnote-ref-717)
607. Frankopan 2016. [↑](#footnote-ref-718)
608. J. A. Evans 2010; Fortunato et al. 2018. [↑](#footnote-ref-719)
609. Campbell et al. 2013. [↑](#footnote-ref-720)
610. Drude, Group, and Others 2003. [↑](#footnote-ref-721)
611. Ebrahimi and Kann 2021. [↑](#footnote-ref-723)
612. Daniel Everett, a brilliant linguist whose years-long work among the Pirahã in the Brazilian rainforest did much to dismantle Chomsky’s armchair theories about universal grammar, started off as a Christian missionary in precisely this vein. In Everett’s wonderful book, *Don’t Sleep, There Are Snakes* (Everett 2009) he describes not only the unusual features of the Pirahã language, but the way life among them, far from producing native converts, ultimately caused him to give up his own faith. [↑](#footnote-ref-724)
613. Hoffmann et al. 2022. [↑](#footnote-ref-725)
614. Revell 2017. [↑](#footnote-ref-726)
615. Villalobos et al. 2022. [↑](#footnote-ref-728)
616. Lebovitz, Levina, and Lifshitz-Assaf 2021; Omrani et al. 2022; Bender et al. 2021; Chiang 2023. [↑](#footnote-ref-729)
617. Hayes 2023. [↑](#footnote-ref-730)
618. Ray Perrault and Jack Clark 2024. [↑](#footnote-ref-731)
619. A related but more optimistic parallel would be the space race, also pitting the US against the USSR, and motivated by the same desire for military and ideological world domination. While the weapons side of the space race fully overlapped with the nuclear arms race, nuclear arms have no upside, while space exploration has (in my opinion) great upsides, including its role in the genesis of the modern environmental movement. It’s unfortunate that the end of the Cold War also brought space exploration to a virtual standstill for decades. [↑](#footnote-ref-732)
620. See Part IX. [↑](#footnote-ref-733)
621. Aware I’m dating myself here. [↑](#footnote-ref-734)
622. Self 1993. [↑](#footnote-ref-735)
623. T. Brown et al. 2020. [↑](#footnote-ref-736)
624. Agüera y Arcas and Norvig 2023. [↑](#footnote-ref-737)
625. Von Oswald et al. 23–29 Jul 2023. [↑](#footnote-ref-738)
626. Giannou et al. 23–29 Jul 2023. [↑](#footnote-ref-739)
627. Wark, Lundstrom, and Fairhall 2007; Newell et al. 2009. [↑](#footnote-ref-740)
628. Pearl and Mackenzie 2018. [↑](#footnote-ref-741)
629. Merrill et al. 2024. [↑](#footnote-ref-742)
630. See Part III. [↑](#footnote-ref-743)
631. AIs performing experiments to learn could be extremely helpful in, for example, high-throughput cellular biology research; but of course, we’d want to be exceptionally careful about robots pushing vases off shelves, or worse, out of idle curiosity. [↑](#footnote-ref-744)
632. Eldan and Li 2023. [↑](#footnote-ref-745)
633. F. Jackson 1982. [↑](#footnote-ref-747)
634. The phrase “not even wrong” (attributed to physicist Wolfgang Pauli) is sometimes used to refer to such arguments. [↑](#footnote-ref-748)
635. Henschen 1893; Milner 1995; Silvanto 2014; Butter et al. 1997. [↑](#footnote-ref-749)
636. Philippe 1901; Bonnet 1769. [↑](#footnote-ref-750)
637. Hassabis et al. 2007. [↑](#footnote-ref-751)
638. Players alternately put down stones that each have a white and black side, with their own color facing up, so a move is notated with a letter and a number giving the square’s coordinates, from A1 to H8. When you place a stone, any of your opponent’s stones in a straight line from one of yours is flipped over to your color, and you win if by the time the board is full there are more stones with your color facing up. [↑](#footnote-ref-752)
639. K. Li et al. 2022. [↑](#footnote-ref-753)
640. M. Mitchell 2023. [↑](#footnote-ref-754)
641. Koh, Fried, and Salakhutdinov 2023; Miyazawa, Kyuragi, and Nagai 2022; Suzuki and Matsuo 2022. [↑](#footnote-ref-755)
642. Brewer, Cook, and Bird 2016. [↑](#footnote-ref-756)
643. Thaler, Arnott, and Goodale 2011. [↑](#footnote-ref-757)
644. Garner and Keller 2022. [↑](#footnote-ref-758)
645. Layfield et al. 2021; Pisoni et al. 2008; Berrettini et al. 2008; Almomani et al. 2021; Nagels et al. 2023. [↑](#footnote-ref-759)
646. Merrill and Sabharwal 2023. [↑](#footnote-ref-761)
647. Bubeck et al. 2023. [↑](#footnote-ref-762)
648. Merken 2023. [↑](#footnote-ref-763)
649. Pauling 1968. [↑](#footnote-ref-764)
650. Margulis 2004; Margulis and Sagan 2007. [↑](#footnote-ref-765)
651. Kadavath et al. 2022; Yin et al. 2023. [↑](#footnote-ref-766)
652. Jerry Wei et al. 2024. [↑](#footnote-ref-767)
653. J. Li, Yu, and Ettinger 2022; Merrill et al. 2024. [↑](#footnote-ref-768)
654. A caveat, though: there is a nascent alternative take on fundamental physics, “Constructor Theory,” formulated *in terms of* counterfactuals; Marletto 2022. While the project of deriving all of physics from it remains incomplete, it has been used to derive key results in thermodynamics (Marletto 2016), information theory (Deutsch and Marletto 2015), and biology (Marletto 2015). [↑](#footnote-ref-769)
655. See Part VI. [↑](#footnote-ref-770)
656. Mercier and Sperber 2018. [↑](#footnote-ref-771)
657. F. Shi et al. 2019. [↑](#footnote-ref-772)
658. Y. Shi et al. 2019. [↑](#footnote-ref-773)
659. AudioLM has a fixed sampling interval, like the frames per second in a movie, but also like a movie, above some sampling rate the experience (or interaction) is for all practical purposes continuous. [↑](#footnote-ref-774)
660. Bhirangi et al. 2024; Ghadirzadeh et al. 2021. [↑](#footnote-ref-775)
661. Heinrich et al. 2020; Yoon et al. 2019. [↑](#footnote-ref-776)
662. Goldin 2013; Gerla et al. 2014. [↑](#footnote-ref-777)
663. Hu et al. 2021. [↑](#footnote-ref-779)
664. Peddinti, Povey, and Khudanpur 2015; Munkhdalai, Faruqui, and Gopal 2024; Ma et al. 2024. [↑](#footnote-ref-780)
665. Zhan 1964. [↑](#footnote-ref-781)
666. Mercier and Sperber 2018. [↑](#footnote-ref-782)
667. Goering et al. 2021. [↑](#footnote-ref-783)
668. Thoppilan et al. 2022. [↑](#footnote-ref-784)
669. Cundy and Ermon 2023. [↑](#footnote-ref-785)
670. Zelikman et al. 2024. [↑](#footnote-ref-786)
671. Gemini, in 2024, adopted the “multiple drafts” approach. [↑](#footnote-ref-787)
672. Long 2023; Yao et al. 2024. [↑](#footnote-ref-788)
673. Mollick 2024. [↑](#footnote-ref-789)
674. Mink 2021. [↑](#footnote-ref-790)
675. Heffernan 2004. [↑](#footnote-ref-791)
676. Roose 2023b. [↑](#footnote-ref-792)
677. Roose 2023a. [↑](#footnote-ref-793)
678. Kleinman 2024. [↑](#footnote-ref-794)
679. “My column about the experience was probably the most consequential thing I’ll ever write—both in terms of the attention it got (wall-to-wall news coverage, mentions in congressional hearings, even a craft beer named Sydney Loves Kevin) and how the trajectory of A.I. development changed.” Roose 2024. [↑](#footnote-ref-795)
680. Personal communication. [↑](#footnote-ref-796)
681. Urban 2015. [↑](#footnote-ref-798)
682. Adams 1980. [↑](#footnote-ref-799)
683. To anyone else over 40 still with me here: “jailbreaking,” originally meaning the removal of carrier restrictions on smartphones, is now a term of art for getting large language models to do things the AI companies didn’t intend; and the Raspberry Pi is a small computer on a circuit board popular for general prototyping, especially involving the “Internet of Things.” [↑](#footnote-ref-800)
684. Agüera y Arcas and Norvig 2023. [↑](#footnote-ref-801)
685. Churchland 2016. [↑](#footnote-ref-802)
686. See Parts IV and VIII; the latter hinted at this, in describing how a sequence model trained to do translation is also a general language model. The technical findings are in Guth and Ménard 2024. [↑](#footnote-ref-803)
687. Szathmáry and Smith 1995. [↑](#footnote-ref-806)
688. Valentin Fedorovich Turchin 1977; V. F. Turchin 1995. [↑](#footnote-ref-807)
689. M. Bennett 2023. [↑](#footnote-ref-808)
690. Colapinto 2007. [↑](#footnote-ref-809)
691. Jacobsen 2024. [↑](#footnote-ref-810)
692. This account focuses on prediction rather than energy, though prediction, computation, and energy are all related at a deep level. We have far to go in making AI models more energy-efficient, but it’s notable that even today, AI models may be more energy-efficient than brains simply due to the speed with which they operate; Tomlinson et al. 2024. [↑](#footnote-ref-811)
693. Walker 2023. [↑](#footnote-ref-812)
694. Sloman and Fernbach 2018. [↑](#footnote-ref-813)
695. Scott 2017; Agüera y Arcas 2023. [↑](#footnote-ref-815)
696. Čapek 1920. [↑](#footnote-ref-816)
697. Bryson 2010. [↑](#footnote-ref-817)
698. Piketty 2017. [↑](#footnote-ref-819)
699. Merchant 2023. [↑](#footnote-ref-820)
700. Per William Blake’s poem, *Jerusalem*: “And did the Countenance Divine, / Shine forth upon our clouded hills? / And was Jerusalem builded here, / Among these dark Satanic Mills?”; Blake 1810. [↑](#footnote-ref-821)
701. Dattani et al. 2023. [↑](#footnote-ref-822)
702. Agrawal, Gans, and Goldfarb 2023; Ben-Ishai et al. 2024. [↑](#footnote-ref-823)
703. Hoffman 2023. [↑](#footnote-ref-824)
704. Ford 2015. [↑](#footnote-ref-825)
705. King 1967. [↑](#footnote-ref-826)
706. Bastani 2019. [↑](#footnote-ref-827)
707. Although we have no shortage of needs and wants, much the same would hold for computationally simulated human minds, as will soon be discussed. [↑](#footnote-ref-828)
708. Carlsmith 2022; Yudkowsky 2023b. [↑](#footnote-ref-830)
709. Jacobsen 2024. [↑](#footnote-ref-831)
710. Lovelock and Margulis 1974. [↑](#footnote-ref-832)
711. A. J. Watson and Lovelock 1983; Lenton and Lovelock 2001. [↑](#footnote-ref-833)
712. Pörtner and Belling 2022. [↑](#footnote-ref-834)
713. Richards et al. 2023. [↑](#footnote-ref-835)
714. Yudkowsky 2023a. [↑](#footnote-ref-836)
715. Bostrom 2014. [↑](#footnote-ref-838)
716. I think I’m in earnest here. British comedians like Douglas Adams (author of *The Hitchhiker’s Guide to the Galaxy*) and Charlie Brooker (creator of *Black Mirror*) seem to be our era’s most prescient futurists. Jules Verne and H.G. Wells started off as humorists too. Perhaps the future is just funny. [↑](#footnote-ref-839)
717. In 2023, Bostrom apologized for some of his posts to the Extropian listserv, such as: “Blacks are more stupid than whites. I like that sentence because I think it is true. But recently I have begun to believe that I won’t have much success with most people if I speak like that. […] [W]hile speaking with the provocativeness of unabashed objectivity would be appreciated by me and many other persons on this list, it may be a less effective strategy in communicating with some of the people ‘out there.’” In the post, Bostrom goes on to assert that he is no racist, as he holds no personal animus against Black people. Following an investigation, Oxford agreed, “we do not consider you to be a racist or that you hold racist views, and we consider that the apology you posted in January 2023 was sincere.” Not everyone agrees. Bostrom 2023. [↑](#footnote-ref-840)
718. Empirically, this relationship doesn’t appear to hold; being smart will not make you rich; Pluchino, Biondo, and Rapisarda 2018. [↑](#footnote-ref-841)
719. Heinlein 1966. [↑](#footnote-ref-842)
720. Friedman 1975; Boaz 2023. [↑](#footnote-ref-843)
721. Original last name: O’Connor, legally changed, probably without irony, in 1990. [↑](#footnote-ref-844)
722. More 2003. [↑](#footnote-ref-845)
723. Goertzel 2000. [↑](#footnote-ref-846)
724. Gebru and Torres 2024. [↑](#footnote-ref-847)
725. “May not have been best micromachinist in Luna and certainly wasn’t cybernetics psychologist. But I knew more about all these than a specialist knows—I’m general specialist. Could relieve a cook and keep orders coming or field-repair your suit and get you back to airlock still breathing. Machines like me and I have something specialists don’t have: my left arm. You see, from elbow down I don’t have one. So I have a dozen left arms, each specialized […].” Does a multifunction bionic arm count as cheating? [↑](#footnote-ref-848)
726. Of course picking fruit, hunting game, and weaving baskets takes work, and involves risk. However, the amount of work needed for subsistence varies dramatically by ecology. There is no law of nature specifying a fixed relationship between the effort involved and its “payoff,” especially since both are subjective. [↑](#footnote-ref-849)
727. Espejo 2014. [↑](#footnote-ref-850)
728. AI doomer Eliezer Yudkowsky, who founded LessWrong in 2009, was “one of the more interesting young Extropians” Goertzel 2000. [↑](#footnote-ref-851)
729. Author of *L’Homme machine* (de La Mettrie 1748a); see Part VI. [↑](#footnote-ref-853)
730. Lind 1776. [↑](#footnote-ref-854)
731. Schofield, Pease-Watkin, and Quinn 2014. [↑](#footnote-ref-855)
732. OK, here’s the rest of it: “[…] and the several Sets of Appellatives, Neutral, Eulogistic and Dyslogistic, by which each Species of MOTIVE is wont to be designated : to which are added EXPLANATORY NOTES and OBSERVATIONS, indicative of the Applications of which the Matter of this Table is susceptible, in the Character of a Basis or Foundation, of and for the Art and Science of Morals, otherwise termed Ethics, —whether Private, or Public alias Politics —(including Legislation) —Theoretical, or Practical alias Deontology —Exegetical alias Expository, (which coincides mostly with Theoretical), or Censorial, which coincides mostly with Deontology : also of and for Psychology, in so far as concerns Ethics, and History (including Biography) in so far as considered in an Ethical Point of View.” Bentham 1817. [↑](#footnote-ref-856)
733. “Bentham’s Bulldog,” for instance, is a popular Substack writer in the sizable intersection among all of these communities Bulldog n.d.. [↑](#footnote-ref-857)
734. Tversky 1969. [↑](#footnote-ref-858)
735. It would, in fact, be contradictory for Utilitarianism to be both moral and mechanical, because morality makes little sense without counterfactuals. If everything happens as it’s bound to happen, then it’s equally pointless to debate how either people or governments composed of people “should” behave; Sapolsky 2023. [↑](#footnote-ref-859)
736. Redelmeier, Katz, and Kahneman 2003. I hear colonoscopies have gotten less unpleasant in the decades since this experiment was done. [↑](#footnote-ref-860)
737. I’m using positive signs here to match the pain dial. As Utilitarians, we’d more properly think of pain as *negative* utility, and say that if X and Y are both negative numbers, X+Y must be *less* than X. [↑](#footnote-ref-861)
738. Even here, there are plenty of exceptions. Numerous faith traditions, from Buddhism to Christianity to Islam, involve giving away one’s possessions, embracing poverty, renouncing one’s freedom, and willingly enduring pain or even death. [↑](#footnote-ref-862)
739. Easterlin 1974; Mentzakis and Moro 2009; Easterlin 2001. [↑](#footnote-ref-863)
740. Jingna and Zhang 2024. [↑](#footnote-ref-864)
741. Singer 1973. [↑](#footnote-ref-866)
742. Poet-musician Gil Scott-Heron had sung mordantly in 1970, “A rat done bit my sister Nell (with Whitey on the moon) / Her face and arms began to swell (and Whitey’s on the moon) / […] Was all that money I made las’ year (for Whitey on the moon?) / How come there ain’t no money here? (Hm! Whitey’s on the moon)”; Scott-Heron 1970; Cherkaev 2021. [↑](#footnote-ref-867)
743. If only they were still alive, Jorge Luis Borges and Douglas Adams could collaborate to write this short story. [↑](#footnote-ref-868)
744. Calculus alert for the mathematically inclined. Mathematically, a stronger statement can be made: that the vector field of paths taken will have zero vorticity. This follows from the fact that this vector field is the gradient of the value function. Needless to say, no real human, plant, animal, or corporation’s behavior is vorticity-free. [↑](#footnote-ref-869)
745. Ng and Russell 2000. [↑](#footnote-ref-870)
746. Doctorow 2024. [↑](#footnote-ref-871)
747. Kurzweil 2005. [↑](#footnote-ref-873)
748. Agüera y Arcas 2023. [↑](#footnote-ref-874)
749. See Parts V and VI. [↑](#footnote-ref-875)
750. D. J. Chalmers 2022. [↑](#footnote-ref-876)
751. Copernican Turns (Kant 1991) have occurred repeatedly in astronomy, as when we realized that other planets have moons too, that Sol is just another star, and that the Milky Way is just another galaxy. [↑](#footnote-ref-877)
752. There are many technical arguments regarding the computational complexity of a simulated world compared to the world simulating it, and even whether it would be possible for worlds to simulate *each other*; Wolpert 2024. [↑](#footnote-ref-878)
753. Our bodies and brains have evolved for a finite lifespan, so learning and memory might pose profound challenges for a brain simulation living a subjective lifetime well in excess of a century. [↑](#footnote-ref-879)
754. Piccinini 2021. [↑](#footnote-ref-880)
755. Moravec 1988. [↑](#footnote-ref-881)
756. Hsiao 2012; “Hawaii HB693” n.d.. [↑](#footnote-ref-882)
757. Ehrlich 1968; Worstall 2014. [↑](#footnote-ref-883)
758. Agüera y Arcas 2023. [↑](#footnote-ref-885)
759. Foot 1967; J. J. Thomson 1976. [↑](#footnote-ref-886)
760. Singer 2005. [↑](#footnote-ref-887)
761. MacAskill 2022. [↑](#footnote-ref-888)
762. In fairness, Bostrom writes in his more recent book, *Deep Utopia* Bostrom 2024, “for the record, that I’m not a total utilitarian, or indeed any kind of utilitarian, although I’m often mistaken for one, perhaps because some of my work has analyzed the implications of such aggregative consequentialist assumptions. (My actual views are complicated and uncertain and pluralistic-leaning, and not yet properly developed.)” This may—or may not—reflect a change of heart. In *Superintelligence*, he certainly gives a lot of weight to “aggregative consequentialist assumptions.” [↑](#footnote-ref-889)
763. The idea is to create a “Dyson sphere,” a thin shell of computing fabric fully enclosing the sun, thus capable of converting its entire energy output into computation. [↑](#footnote-ref-890)
764. To be sure, this idea has far older roots. In 1603, at the dawn of the Scientific Revolution, Francis Bacon wrote an essay whose title can be translated as *The Masculine Birth of Time, Or the Great Instauration of the Dominion of Man over the Universe*; Bacon 1603. The essay expressed Bacon’s fervent wish for science and technology “to stretch the deplorably narrow limits of man’s dominion over the universe to their promised bounds.” [↑](#footnote-ref-891)
765. This service is offered by the Alcor Life Extension Foundation. Extropian-in-chief Max More was Alcor’s CEO between 2011 and 2020. [↑](#footnote-ref-892)
766. In fact a standard Silicon Valley trope involves describing an idea that worked perfectly well, but only for a limited number of “users”; invariably, the raconteur then says they moved on to a “more scalable” idea to achieve “greater impact.” [↑](#footnote-ref-893)
767. Peter Singer, Nick Bostrom, and Sam Bankman-Fried have all been prominent Effective Altruists. [↑](#footnote-ref-894)
768. Faux 2023. [↑](#footnote-ref-895)
769. Doctorow and Stross 2012; Paura 2019. [↑](#footnote-ref-896)
770. Christian 2020. [↑](#footnote-ref-898)
771. Asimov 1950. [↑](#footnote-ref-899)
772. Co-granddaddy Arthur C. Clarke based the homicidal behavior and ultimate nervous breakdown of HAL 9000 on a similar premise; Clarke 1968. [↑](#footnote-ref-900)
773. LessWrong n.d.. [↑](#footnote-ref-901)
774. Bostrom 2020. [↑](#footnote-ref-902)
775. I’m not convinced that most Utilitarians actually *would*. After all, they are just as human as the rest of us. [↑](#footnote-ref-903)