

CONTENTS

<i>Acknowledgments</i>	ix
<i>Epigraphic Credits</i>	x
1	
Introduction: AI That Thinks	1
<i>The Two Most Frequently Asked Questions</i>	5
<i>The Dangers of AI</i>	6
<i>The Guided Tour</i>	6
<i>The Big Picture of Thinking AI</i>	8
2	
The Turing Test, Then and Now	9
<i>CAPTCHAs: Completely Automated Public Turing Tests</i>	10
<i>Thinking? Or Appearing to Think?</i>	10
<i>Turing's Two Imitation Games</i>	12
<i>Interpreting the Turing Test</i>	16
3	
Conversations with Computers: From ELIZA to Watson and Beyond	21
<i>The ELIZA Effect</i>	22
<i>Watson's Triumph on Jeopardy!</i>	25
<i>The Workings of Watson</i>	28
<i>The Jeopardy! Turing Test: Does Watson Play Like a Human?</i>	34
<i>Chat Systems Based on Deep Neural Networks: A Preview</i>	38

vi CONTENTS

4	Emergence: How AI Becomes Thinking AI	40
	<i>Emergent AI</i>	41
	<i>Reductionism Versus Emergence, or Reductionism and Emergence?</i>	43
5	The Low-Level Equivalence of Brains and Computers	44
	<i>Inputs, Outputs, and Digital Transformations</i>	45
	<i>Computer Programs Are Just Circuits</i>	49
	<i>Modeling the Brain: Neurons and Beyond</i>	52
	<i>Human Brains Can Simulate Computer Programs</i>	63
	<i>Computer Programs Can Simulate Human Brains</i>	70
	<i>Contrary Views on Brain-Computer Equivalence</i>	75
	<i>Low-Level Equivalence: The Good and the Bad</i>	75
6	Learning in the Brain	77
	<i>Hebbian Learning</i>	78
	<i>Higher-Level Learning</i>	82
	<i>Contrasting Hebbian Learning and Machine Learning</i>	82
7	Neural Networks: The Revolutionary and Evolutionary Tool	84
	<i>Artificial Neural Networks</i>	85
	<i>Training an Artificial Neural Network</i>	91
	<i>Artificial Neural Networks for Real Problems</i>	99
8	The Surprising Power of Deep Neural Networks	115
	<i>Features in Action: Recognizing a Handwritten Digit</i>	117
	<i>A Genuinely Deep Network</i>	130
	<i>The Human Brain and Deep Neural Networks</i>	146
	<i>The Big Problem of Bias</i>	150

CONTENTS vii

9	Reinforcement Learning and the Creativity of AlphaGo	152
	<i>Reinforcement Learning for the Game of Nim</i>	154
	<i>AlphaGo Zero</i>	166
	<i>Does the Brain Use Reinforcement Learning?</i>	180
10	Generative AI: Unprecedented Scale, Surprising Simplicity	182
	<i>The G: Generating the Next Word</i>	184
	<i>The P: Pre-Training ChatGPT</i>	189
	<i>The T: Transformer with Attention</i>	190
	<i>Chain of Thought: Reasoning Within LLMs?</i>	192
	<i>The Unprecedented Scale of ChatGPT</i>	193
11	Escape from the Chinese Room	196
	<i>Into the Chinese Room</i>	198
	<i>Escaping from the Chinese Room</i>	200
12	Flight and Life: Two Analogies for Thinking About Artificial Intelligence	210
	<i>Is It Possible for a Machine to Fly?</i>	211
	<i>Is It Possible for a Machine to Be Alive?</i>	213
	<i>Fusing the Analogies: Artificial Flight, Artificial Life, Artificial Intelligence</i>	220
13	Uniquely Human?	225
	<i>The Special-Human Questions</i>	226
	<i>Comparing Dualism, Biological Naturalism, and Computationalism</i>	229
	<i>Aliens and Orangutans: Other Minds with Reasoning and Consciousness</i>	232
	<i>Computationalism and Deep Neural Networks</i>	233
	<i>The Body-Minded Brain: Emotions and Sensations</i>	236
	<i>Computationalism and Conscious Awareness</i>	239

viii CONTENTS

	<i>Computationalism and Religion</i>	242
	<i>If AI Can Think, Are Human Lives Meaningless?</i>	243
14	Conclusion: AI Emergent, Humans Alive	246
	<i>Notes</i>	253
	<i>Bibliography</i>	263
	<i>Index</i>	273

1

Introduction: AI That Thinks

I've always promised that on 1 January 2000, I'm going to call everybody up and say, "See, you underestimated." Which will be true. I think we consistently continually underestimate what we can do with computers if we really try.

—GRACE HOPPER, IN A 1980 ORAL HISTORY INTERVIEW¹

I propose to consider the question, "Can machines think?" These are the bold and provocative opening words of a 1950 article in the philosophy journal *Mind*. But the author of this article was not a professional philosopher. The author was Alan Turing—British computer scientist, mathematician, code breaker, and war hero. Turing's article, "Computing Machinery and Intelligence," is one of the founding documents of artificial intelligence (or AI, as it has come to be known). The opening question of the article—"Can machines think?"—was published exactly halfway through the twentieth century, yet in the twenty-first century it remains unanswered and resonates with us more than ever. The question strikes at the heart of what it means to be human, challenging our instinctive notion that a lifeless lump of silicon cannot be equivalent to the living brain.

Today, it is obvious that machines—or "computers" as we now call them—can perform a vast range of challenging and useful tasks, far beyond what might have been envisaged by the readers of Turing's article in the 1950s. In particular, the first decades of the twenty-first century have seen an unprecedented revolution in artificial intelligence, propelled by the immense raw computing power of modern hardware combined with new algorithmic techniques. This AI revolution allows us to take a fresh look at the question of whether computers can think.

Experts—including philosophers, cognitive scientists, and computer scientists—still disagree on the answer to this question. And even in areas where experts agree, media hype can make it difficult for nonexperts to form their own opinions. The difficulties are compounded by our natural instinct to regard creativity, intuition, consciousness, and other aspects of thinking as uniquely human. An understanding of modern AI systems can demystify these ideas, hopefully allowing nonexperts to draw their own fact-based conclusions on the extent to which machines can think. *Thinking AI* does not provide a binary yes/no response to the question of whether computers can think, because a binary answer is not appropriate for nuanced philosophical questions. The reader will instead find a careful analysis of how computer programs can think like humans, and under what circumstances. Ultimately, the book provides evidence that computer programs can exhibit creativity, intuition, reasoning, and understanding—and they can do so using methods that resemble human brain processes.

The book's argument has six elements:

1. Computer programs can, in principle, produce outputs that perfectly emulate phenomena such as creativity, intuition, consciousness, and thinking.
2. Modern AI systems do not emulate these phenomena perfectly, but their emulations are of very high quality in some cases.
3. Sometimes, modern AI systems emulate thinking and related phenomena via methods that have clear parallels with human brain processes.
4. The parallels between computer programs and brain processes often rely on *emergence*: the capability of a system to exhibit sophisticated properties based only on the interaction of a large number of simple components. The human mind can be viewed as arising from the interaction of neurons via emergence, and we'll encounter examples demonstrating that certain types of intelligence can arise from computer programs via emergence also. We call this phenomenon *emergent AI*.
5. By understanding the technical details of how modern AI systems employ emergence, nonexperts are better positioned to address one of the most important and challenging questions facing our society today: In what ways can computer programs appear to think like humans?

6. After combining both technical and philosophical evidence, it is reasonable to conclude that the emergent humanlike thinking of AI is plausible, but that human society will continue to have meaning and importance.

The notion of emergence (item 4 above) is a unifying theme of the book. A major goal of *Thinking AI* is to demonstrate that humanlike mental processes can emerge from interactions of data manipulated by a computer program. This claim may seem preposterous at first, but we will establish the plausibility of emergent AI step by step, as the fundamental tools of the AI revolution are revealed.

To present the argument outlined above, *Thinking AI* thus aims to examine the can-machines-think question from an *informed* perspective: informed by detailed understanding of how some modern AI systems work. By understanding the theoretical algorithms and practical engineering that produce modern artificial intelligence, we can make informed judgments about how the “thinking” performed by a computer program compares to the thinking done by a biological brain. Consider some specific examples in which modern AI systems might be regarded as thinking:

- In the early 2010s, the computer system Watson defeated human champions on the game show *Jeopardy!*. This task involves understanding spoken English questions in real time, then responding with sensible, correct answers on topics that encompass the full range of human knowledge. What kind of “understanding” is really involved here? How does Watson “think” about the question before formulating a response?
- Computers have been able to beat the best human players at chess and checkers since the 1990s. In the late 2010s, computers defeated the top humans at the even more complex game of Go, considered by many to be the ultimate board game challenge for computer programs. Even in games such as Texas hold ’em poker—which combine skill, chance, and psychology—computers are able to achieve victories over human champions. The algorithms employed by these programs are subtle and illuminating, a far cry from the brute-force exploration of every possible move that might be expected. State-of-the-art systems achieve most of their skill by practicing against themselves, and thus learning how to improve. So what kinds of “learning” and “thinking” are employed by computers that defeat humans at chess, Go, and poker?

- Before the 2010s, computer performance on the task of *object recognition*—identifying one or more objects in a photograph—was dismal. From 2012 onward, a new variant of an old idea began to transform the landscape of AI: deep neural networks. Object recognition was one of the problems conquered by deep neural nets. Modern systems can recognize thousands of objects with surprisingly specific descriptions such as “Brittany spaniel” and “steel arch bridge.” So, what does a neural network “know” about objects in the world, and how can it “recognize” them?
- The 2020s saw the arrival of a novel architecture built on top of neural nets: the *transformer*. Transformers enabled chatbots, such as OpenAI’s ChatGPT, to respond to prompts with prose of unprecedented quality and relevance. ChatGPT can, for example, explain the themes in a literary work and answer PhD-level chemistry exam questions. So, how does a chatbot “think” about its prompt and “create” a response?

These four examples of AI systems—answering *Jeopardy!*-style questions, playing games of skill, recognizing objects in images, and responding helpfully to text prompts—will be our reference points for understanding the modern revolution in artificial intelligence. To appreciate those systems, we will take a tour through some of the algorithmic advances in twenty-first-century AI, focusing especially on artificial neural networks (ANNs), which have become the dominant approach in the AI community. In particular, multiple chapters are devoted to the two pillars supporting the AI revolution of the 2010s: deep neural networks and reinforcement learning. These two concepts provide the clearest demonstrations of one of the book’s main theses—that a technical understanding of modern AI systems provides some clear parallels between computer programs and human thought processes. We will also examine some ideas from neuroscience and philosophy, to further understand the potential connections between the way computer programs think and the way humans think.

In explaining AI algorithms and engineering, no background knowledge will be assumed. Readers will learn *how* modern AI systems achieve their results, without the need for mathematical equations or computer programming.

The Two Most Frequently Asked Questions

As already mentioned, this book will not provide a simple binary response to questions about whether computer programs can replicate all mental aspects of humanity. There are too many nuances and subtleties in such questions. Nevertheless, it will be helpful to give direct and high-level answers summarizing our approach to the two biggest questions: Can computer programs appear to think like humans, and can they be sentient or conscious?

Can a Computer Program Appear to Think Like a Human?

Yes: Scientists understand the functioning of biological neurons in the human brain sufficiently well that they can be modeled by a computer. Current models are accurate enough to mimic the inputs and outputs of simple neurons with reasonable fidelity, reproducing the shape and frequency of electronic signals known as spike trains. It is reasonable to assume, at least in principle, that computational neuroscientists can continue to increase the resolution of these models, eventually reaching much higher levels of detail and accuracy. Therefore, in principle, it would be possible write a computer program that simulates every neuron in the human brain and mimics its outputs with sufficient accuracy to reproduce the brain's outputs from any given inputs. We call this the *low-level equivalence* between brains and computers; chapter 5 explains it in detail.

But: This argument via low-level equivalence is unsatisfying for at least two reasons. First, it depends on a chain of statements that are true in principle but may be impossible in practice. For various practical reasons, we may never be able to write a computer program that can simulate the brain's outputs. Second, an exact copy of a brain, whether it is made from biological tissue or a software simulation, will contain a vast quantity of detail that may be incomprehensible to researchers seeking to understand brain processes. Low-level equivalence implies brain simulations can produce outputs that appear to be insightful, creative, and human. But the overwhelming, inscrutable details of the simulation may not reveal *how* such outputs are produced. Therefore, it is probably more interesting and useful to investigate how AI systems can perform certain higher-level brain functions, without resorting to low-level simulation. *Thinking AI* explains how modern AI systems achieve this, using the powerful tools of deep neural networks and reinforcement learning.

And: Chapter 2 explains why questions such as “Can a computer program think?” and even “Can a computer program appear to think?” are too simplistic. Part of the problem is that no one has succeeded in satisfactorily defining *think*. But it’s also problematic to expect a binary yes/no answer to these questions. In chapter 2, we instead settle on the following question as the central one for this book: *In what ways can computer programs appear to think like humans?*

Can a Computer Program Be Conscious or Sentient?

The short answer is yes: the low-level equivalence between brain processes and computer programs implies that computer programs can emulate conscious awareness, by replicating the brain processes giving rise to consciousness in humans. But for the reasons explained in chapter 13, this type of conscious awareness from an AI is relatively uninteresting. It’s more fruitful to ask whether AI systems can exhibit understanding or creativity in human-like ways—and in chapters 7–10 we will see concrete examples in which the answer to this question is yes.

The Dangers of AI

There are many excellent books devoted to the interactions of AI with society and to the potential threats it presents.² So, the dangers of AI won’t be a focus of this book. However, I do believe that any discussion of AI should acknowledge its risks, and I’ll briefly offer my own opinion on two key points here. First, is there an existential threat from AI? AI will result in profound changes to our society, and we should certainly be vigilant about possible dangers. But, at the time of writing, there are no direct threats to human survival, and I’m optimistic we can keep it that way.

Second, will AI systems that “think” exacerbate the problems of bias and unfairness in our society? My answer here is definitely yes, unless we continuously work to mitigate the bias and unfairness inherent in AI. The benefits of AI appear to significantly outweigh its risks. But the AI community must strive for the benefit of all and the detriment of none.

The Guided Tour

A guided tour of the book may be useful at this point. After the overview provided by the current chapter, we move on to a subfield of AI that has

always been closely associated with the question of whether machines can think: computer programs that attempt to converse with humans. The next two chapters explore these ideas. Chapter 2 examines the *Turing test*—the classical method for testing whether a machine can think, via conversation with a human. Although widely recognized as inadequate today, the Turing test is still relevant and is a valuable starting point for our analysis. Chapter 3 explores some early AI programs that converse with humans, throwing the Turing test into relief. It then turns to Watson, one of the first AI breakthroughs in the twenty-first century. With its carefully engineered combination of GOFAI (good old-fashioned AI) and machine learning, Watson became a *Jeopardy!* champion. For us, it provides insight into both similarities and differences in the way that computer programs “think.”

Chapter 4 expands on the book’s unifying concept: the notion of *emergence*. As mentioned above, emergence is the capability for novel phenomena to arise from the interactions of many small components. The key examples for us are the human mind (produced by the interaction of billions of neurons) and certain types of humanlike AI (produced, for example, by the interactions of artificial neurons in a neural network).

We then move on to two chapters discussing the biology of the brain. Chapter 5 covers some basics in neuroscience and computer architecture. By connecting these concepts we establish that, at the level of their most basic components (neurons and logic gates respectively), brains and computers are equivalent. This is the concept of *low-level equivalence* mentioned above. Chapter 6 gives a brief glimpse into how biological brains learn, preparing us to understand the similarities and differences between biological and machine learning described in later chapters.

The next four chapters focus on the two pillars of the AI revolution mentioned earlier: deep neural networks (chapters 7–8) and reinforcement learning (chapter 9). In each case, we examine AI systems that have parallels to the biological neural networks in our own brains. Chapter 10 follows up with an explanation of the game-changing chatbots that arrived in the 2020s, again noting some explicit connections to humanlike thought.

The book then switches gear, moving into three chapters that consider philosophy of AI in the light of the systems described earlier. Chapters 11 and 12 consider classic thought experiments and analogies that deepen our understanding of what it means for a computer program to think. This leads us to examine, in chapter 13, the status of human minds compared to those of nonhuman animals and computer programs. Ultimately, we adopt *computationalism*—the equivalence of brain processes and computer programs—as

our preferred philosophy of cognition. The connections between computationalism, consciousness, and the value of being human are also addressed.

Finally, chapter 14 assembles from throughout the book examples of programs that appear to think or understand like living brains, assessing the weight of the evidence for “thinking AI.” We conclude that the emergent thinking of AI is plausible, but that human society will remain meaningful: Humans will continue to value interactions with other humans, enhanced by interactions with artificial intelligence.

The Big Picture of Thinking AI

What will a reader gain from reading this book? Here are the four capabilities I hope every reader will acquire:

- **Understanding the arc of the AI revolution in the first quarter of the twenty-first century.** This traces the evolution from programs that failed the Turing test and performed miserably at object recognition, to systems that converse fluently and appear to understand images.
- **Understanding how modern AI systems work, without requiring any technical background.** This includes an understanding of deep neural networks, reinforcement learning, and the transformer neural network architecture.
- **Absorbing the key ideas in philosophy of AI—in the context of real AI systems.** This includes classic debates about the Chinese room, the definition of life, the importance of consciousness, the relevance of religion, and the uniqueness of humanity. And of course, the big one: In what ways can computer programs appear to think like humans?
- **Deciding for yourself, based on evidence and understanding:** Do existing AI systems reveal glimpses of the living brain? Could future AI programs think like a human, without directly simulating the human brain? Is “thinking AI” a reality?

That is the voyage I hope we will navigate together.

INDEX

- Aaronson, Scott, 14, 15
Abelson, Robert, 197, 199, 258
action potential, 55
activation function, 89, 133
adversarial example, 235–236, 256
African American, 150, 151
agent (in Minsky mind model), 42
AI, *see* artificial intelligence
AI winter, 86, 255
AlexNet, *see* neural network, AlexNet
algorithm, 4, 28, 31, 33, 35, 175, 183, *see also* value-learning algorithm; DQN algorithm; stochastic gradient descent; parsing
 biased, 151
 deep learning, 115, 136, 139
 machine learning, 1, 86, 232
 object recognition, 131
aliens, 232
all-or-nothing (neuron spike), 56, 59, 64, 78
AlphaFold, 179–180
AlphaGo, 170–175, 183, 202, 209, 234, 235, 249
AlphaGo Zero, 175–179, 233, 235, 249
amino acid, 179
AND, *see* logic gate, AND
Aniston, Jennifer, 248, *see* neuron, Jennifer Aniston
ANN, *see* neural network, artificial
aperiodic molecule, 215
apology, 18
artificial flight, 211–212, 220, 251
artificial intelligence
 bias in, 6, 115, 150–151, 251, 256
 dangers of, 6, 150, 251, 253
 emergent, *see* emergent AI
 nonemergent, *see* nonemergent AI
 strong, *see* strong AI
 weak, *see* weak AI
artificial life, 217–220, 251
artificial neural network, *see* neural network
associative memory, 43
asynchronous circuit, 53, 68
Atari, 167–172, 249
attention, *see* self-attention
attention head, 258
axon, 54–67, 77, 78
axon hillock, 54–59

backpropagation, 83, 94, 96–97, 109, 130, 136, 190, 255
bamboo-copter, 211
bar feature, 121–130, 135–142, 146–147, 248
Barto, Andrew, 166
basal cognition, 259
bathtub analogy, 57–61, 63, 71
Bayne, Tim, 253
BBS, *see Behavioral and Brain Sciences Behavioral and Brain Sciences*, 201, 204
behaviorist, 19
Bengio, Yoshua, 86
Berkeley, University of California at, 196, 201, 204
betting (in *Jeopardy!*), 28

- bias, *see* neural network; artificial intelligence
- binary, 49
- biological naturalism, 227, 230–231, 243, 251
- bipolar cell, 112
- bit, 49
- Black, *see* African American
- board game, 42, 244, *see also* chess; Go
- Boden, Margaret, 196, 201
- body-minded brain, 236–239
- boojum reflector, 218, 219
- Boole, George, 51
- Boolean
- algebra, 51, 57, 64
 - formula, 51
 - function, 69
 - variable, 50, 89, 91, 92, 99
- BPE, *see* byte-pair encoding
- brain
- analog nature of, 56–57
 - convolutional nature of, 111–113
 - emergence in, 41, 43
 - human, 1, 35, 38, 146–150, 167, 189, 192, 194, 225, 227, 248, *see also* low-level equivalence
 - learning in, 77–83
 - male and female, 18
 - modeling, 52–61, *see also* connectome
 - randomness in, 48
 - reinforcement learning in, 180–181, 250
 - simulating computer programs, 63–70
 - simulation, 5, 42, 70–75, 202, 204–207, 217, 223, 232, 247, *see also* functional simulation
- brain in a vat, 238, 260
- Breakout* (video game), 168, 169, 235, 249
- Breuer, Josef, 82
- brittleness (of AI algorithm), 169, 174, 235–236, 257
- Broussard, Meredith, 253
- browning, elizabeth barrett, 84
- brute-force analysis, 3, 171, 222
- Buolamwini, Joy, 115, 253
- byte-pair encoding, 257
- C. elegans*, 72–74
- Caltech, 148, 149
- Cambridge University, 73, 167
- cap feature, 121–130, 135
- capitalist, *see* Marxist-capitalist
- CAPTCHA, 10
- Carnegie Mellon University, 10, 196
- cat, 146, 147, 229
- catamaran, 131, 132, 143, 146, 150, 248
- Cavic, Milorad, 26, 32
- cell
- biological, 70, 112, 147, 213, 215, 223, 229, 250, *see also* bipolar; cone; grandmother; granule; horizontal; neuron; off-center; on-center; photoreceptor; pyramidal; rod; spiny
 - in Game of Life, 218–220
 - membrane, 60
 - off-center, 147
 - on-center, 147
 - W3, 256
- chain of thought, 192–193, 240–241, 250
- Chalmers, David, 255
- Chan, Bert, 219
- chaotic behavior, 255
- chatbot, 4, 13, 15, 17, 31, 48, 182–195, 197, 207, 249, *see also* ELIZA; ChatGPT
- ChatGPT, 4, 15, 21, 38, 39, 182–195, 202, 209, 234, 235, 246, 249
- chess, 3, 25, 34, 166, 169, 171, 210, 220
- childlike strategy, *see* Nim
- China, 211
- Chinese
- character, 199–206
 - language, 198–209
- Chinese computer room scenario, 206–208

- Chinese room argument, 198–209, 231, 251
 and brain simulation, 204–207
 and deep neural networks, 207–208
 and Turing test, 202
 and Watson, 202–204
- choreography, 243
- circuit
 electronic, 49–52, 57, 63, 64, 67, 89, 183, 247
 neural, 37, 68, 184, 190, 191
- Clinton, Bill, 148, 150
- cognitive science, 24, 70, 227, 240
- color image, 104
- computational problem, 45
- computationalism, 7, 225–227, 231–245, 251
- computer program
 and artificial life, 217
 and emergent learning, 81
 and human cognition, 198
 and intentionality, 231
 as circuit, 49
 conversing with humans, 21
 equivalence to brain processes, 44
 learning, 84
 simulated by neurons, 67
 simulating brains, 70
 surpassing human intelligence, 225
 thinking like a human, 3, 5, 9, 12, 41, 43, 246
- conditioned reflex, 153
- conductance, 77
- conductivity, 41
- cone cell, 112
- connectionism, 86
- connectome, 72–74
- consciousness, 2, 6, 11, 37, 75, 76, 82, 226, 233, 247
 and computationalism, 239–242
 definition of, 240
- conversation, 7, 10, 12–39, 183–186, 244
- convnet, 102, *see also* neural network, convolutional
- convolutional layer, *see* layer
- convolutional neural network, *see* neural network, convolutional
- Conway, John, 218
- copper wire, 40
- CRA, *see* Chinese room argument
- creativity, 2, 17, 169, 176, 183
 emergent, 249
- Cullingford, Richard, 258
- current-best move, 163–164
- Damasio, Antonio, 236, 237, 241
- dance, 243
- Darwin, Charles, 214, 246
- Darwinism, 225
- dbPedia, 30–32
- decision tree, 50, 89, 90
- Deep Blue, 25, 34, 170, 210
- deep learning, 115, 124, 130, 133, 145–146,
 see also neural network, deep
- deep neural network, *see* neural network, deep
- DeepMind, 166–180, 233, 249
- DeepQA, 34
- Dehaene, Stanislas, 241
- dendrite, 54–67, 77, 78
- Deng, Jia, 84
- Dennett, Daniel, 206, 237, 260
- Descartes, René, 214, 226, 260
- Dickinson College, 137, 138
- digital transformation, 45–48, 53, 54, 63, 67, 68, 74–75
- DNA, 112, 113, 215–217
- dog, 131, 143, 227, 229, *see also* Pavlov's dog
- dopamine, 180–181, 250
- downstream
 dendrite, 77, 78
 neuron, 54, 55, 57, 133
 signal, 78
- DQN algorithm, 167–172, 235, 249
- Drage, Eleanor, 253
- Drosophila*, 73
- dualism, 214, 225–230, 243, 251
 substance, 229

- ear
 - feature, *see* feature, ear
 - neuron, *see* neuron, ear
- edge feature, 138, 147, 248, 256
- Eduardo (in imitation game), 17
- EduardoBot, 19
- elevated state, 244–245
- Eliot, Thomas Stearns, 182, 189
- ELIZA
 - chatbot, 21–24, 48, 183, 197, 201
 - effect, 24
- embodied cognition, 236
- emergence, 2, 7, 40–43, 71, 76, 81, 176, 188, 207, 217, 223, 229, 230, 244, 248,
see also emergent AI
 - of intentionality, 231
- emergent AI, 8, 41–43, 81, 139, 146, 147, 169, 174, 176, 177, 189, 191, 204, 207–209, 225, 231, 247–251
- emotions, 11, 228, 237–239
- empathy, 17–19
- empiricism, 233–234
- <EndPrompt>, 184, 185, 188, 190, 191
- <EndReply>, 184–186, 188, 190
- English language, 3, 11, 15, 29, 47, 184, 197–205
- epilepsy, 149
- epiphenomenon, 241
- ethics, 149, 222, 253
- Eugene Goostman (chatbot), 15
- evil demon, 260
- evolution, 113, 139, 146, 147, 214, 223, 227, 229, 230, 244, 246
- excitatory synapse, 60, 69
- exploration probability, 160, 164
- exploratory move, 164–165

- face
 - feature, *see* feature, face
 - neuron, *see* neuron, face
- face detection system, 151
- face recognition system, 24
- feature (in a neural network), 110–111, 117–130, 135–145, 168, 172, *see also* cap; slash; bar; edge
 - ear, 143, 145, 146, 249
 - face, 143, 145–148, 249
 - sound, 105, 106
 - visual, 42, 138, 139, 146, 147
 - visualization of, 137
- feedback, 53, 68, 232, 255
- feminism, 253
- filter, 110
- firing
 - frequency, 61, 63, 71
 - of neurons, 61
 - threshold, 56–69, 79, 81, 87–89, 120, 248
 - firing frequency, 64
- fixed-wing aircraft, 212, 220–222
- flight, *see* artificial flight
- floppy ear neuron, *see* neuron
- forklift, 132, 143, 150
- formal symbol, 45, 176, 196, 200–209, 231, 251
- Frankenstein, 210, 259
- frequency, *see* firing frequency
- Freud, Sigmund, 82
- fully connected layer, *see* layer
- functional simulation, 42, 70–75, 247, *see also* brain simulation
- functionalism, 259

- Gage, Philip, 257
- Game of Life, *see* Life, Game of
- gate, *see* logic gate
- Gemini, 15, 38, 39
- gender, 13
 - bias, 115, 151
 - game, 12, 17, 19
- gene, 215, 223
- generative AI, 21, 182–195, 246
- Gibson, William, 238
- glia, *see* glial cells
- glial cells, 71, 255
- Glover, Denis, ix
- Go, 3, 169–171, 182, 183, 209, 233, 235, 249
- God, 214, 225

- GOFAL, 7, 24, 33–34, 38
Google, 10, 15, 38, 73, 194
GPT, 184, 190
GPT1, GPT2, GPT3, GPT3.5, 186, 190,
194, 235, 258
GPU, 136
gradient descent, 93, *see also* stochastic
gradient descent
grandmother neuron, *see* neuron,
grandmother
granule neuron, 55

hallucination (by an LLM), 192
handwritten digit, 110
recognizing, 117–130
Harvard University, 73, 74
Hassabis, Demis, 166, 180
Hebb, Donald, 78
Hebbian learning, *see* learning
Helmstaedter, Moritz, 73
hidden human, 14
hidden layer, *see* layer
Hinton, Geoffrey, 86, 130–131
Hopfield, John, 43
Hopfield network, 43
Hopper, Grace, 1
horizontal cell, 113
Howard Hughes Medical Institute,
73
Hubel, David, 147
Hui, Fan, 170, 173
human face neuron, *see* neuron, face
Hume, David, 77, 82, 233, 234
humor, 244, 247
hurricane, 41, 72
Huxley, T. H., 225

IBM, 25–36, 170
ImageNet, 130, 131
imitation game, 12–20, *see also* Turing
test
inflation, 41
inhibitory synapse, 60, 66, 69
innatism, 233
insight, 5, 115, 145–146, 209

intentionality, 196, 230–231
intuition, 2
intuition pump, 206
ion channel, 60, 71, 77

James, William, 82
Java (programming language), 205–206
Jennings, Ken, 25, 27
Jeopardy!, 3, 7, 21, 25–28, 42, 46, 48, 74,
202, 209, 233, 250
Turing test, 34–38
jōseki, 176–177, 249
judge (in imitation game), 12
non-expert, 16
Jumper, John, 180

Kahneman, Daniel, 37
Kasparov, Garry, 25
kernel, 110
Kim, Boo Sung, 256
knee pad, 132
Krizhevsky, Alex, 130

labeled, 92, 146, *see also* training data
Laboratory of Molecular Biology, 72,
216
Labrador, 131, 146
ladder sequences, 178
language, *see also* large language model;
English; Chinese; Spanish
natural, 29, 182, 184, 198, 200, 206,
233, 244, 257
programming, 30, 205
large language model, 14, 186–195, 250,
253
LAT, *see* lexical answer type
Law of Effect, 153
layer, 99
convolutional, 102, 107–113, 130,
133, 136–140, 142, 143, 147,
168, 175
fully connected, 100, 102–104, 133,
134, 136, 168, 172
hidden, 100–108, 112, 117–129,
133, 136, 147

- layer (*continued*)
 - max pool, 133, 134, 141
 - meaning in deep network, 133
 - shape of, *see* shape
 - softmax, 134
 - weight, 133
- leakage channel, 60
- leaky
 - bathtub model, *see* bathtub analogy
 - integrate-and-fire, 60
- learning, 3, 91, 248, *see also* brain
 - algorithm, *see* algorithm; stochastic gradient descent
 - deep, *see* deep learning
 - emergent, 248
 - Hebbian, 42, 78–83, 247
 - machine, *see* machine learning
 - rate, 97
 - reinforcement, *see* reinforcement learning
 - supervised, *see* supervised learning
 - unsupervised, *see* unsupervised learning
- LeCun, Yann, 86
- Lee, Sedol, 173, 174, 177, 249
- Legg, Shane, 167
- Lenia, 219–220, 223
- lexical answer type, 35–36
- Li, Fei-Fei, 84, 131, 253
- Library of Congress, 29
- life
 - artificial, *see* artificial life
 - biological, 213–217, 223, 225
 - definition of, 213–217
- Life, Game of, 218–220, 223
- lionfish, 131, 132, 146, 248
- LLM, *see* large language model
- locality, 103–107, 112, 113, 250
- Locke, John, 233, 234
- Loebner Prize, 16
- logic gate, 7, 49, 50, 63, 67, 68, 70, 232
 - AND, 49, 52, 63–65, 89
 - NOT, 49, 66–67, 89
 - OR, 49, 65–66, 81, 89
- logistic activation function, 89
- long-term potentiation, 255
- Lotus Temple, 148, 149
- love, 18, 244
- low-level equivalence, 5, 7, 44–76, 236, 238, 240, 243, 247
- LTP, *see* long-term potentiation
- machine
 - alive, 213–220
 - animal as, 226, 227
 - as computer, 1, 11, 13, 154, 196
 - flying, 211–212
- machine learning, 7, 24, 33, 82–83, 94, 166, 231–232
- macromolecule, 215, 223
- many experts model, 31
- Marxist-capitalist imitation game, 18–20, 202
- masked self-attention, *see* self-attention
- Matrix, The* (movie), 238
- Max Planck Institutes, 73
- max pooling layer, *see* layer
- McCarthy, John, 196, 198
- McCorduck, Pamela, 9, 17
- McCulloch, Warren, 44
- McInerney, Kerry, 253
- mechanistic (view of mind or body), 214–215, 226, 227, 243
- memory
 - computer, 29, 49, 53, 68, 223, 231
 - human, 71, 82, 167
- mind
 - emergence of, 2, 41, 43, 72, 76, 229
 - human, 37, 226, 227, 231, 247
 - in nonhuman animals, 227
 - of a weather system, 41, 71
- Mind* journal, 1, 11
- mind-body duality, 214
- minibatch, 94–98
- Minsky, Marvin, 42, 86, 196, 198, 204
- MIT, 73, 86, 148, 151, 196
- Mitchell, Melanie, 21, 261

- Mollick, Ethan, 253
monkey, 148, 180–181, 256
motion
 detection, 73, 147
 in organisms, 220, 222
mouse retina, 73–74, 256
mRNA, 216
multilayer network, *see* neural network
multiplication, 45–46, 220
Murdoch, Iris, 246, 251
music, 243

National Physical Laboratory, 154
nativism, 233–234
natural language, *see* language, natural
Nature (journal), 148, 167, 169, 171, 173, 175, 176
neocortex, 61
nerve, 236
 optic, *see* optic nerve
nerve bundle, 54, 236
nervous system, 54, 229, 236
neural network
 bias parameter, 89
neural network, 67, 68, 83–114, 232, 248, 250
 AlexNet, 130–131, 136, 183
 bias parameter, 87, 88, 91, 94, 96, 97, 100, 124
 biological, 85
 convolutional, 86, 101–113, 117, 123, 195, 250
 deep, 4, 38, 115–152, 167–179, 182, 187, 189, 194, 202, 207–209, 233–236, 247, 248
 multilayer, 99–101
 parameter count, 194
 RD2, 117–130
 residual, 175
 shallow, 86, 130
 size, 194
 VGG16, 116, 130, 133–146, 234, 235
 visualization of, 136–142
 weights in, *see* weight
 Y8, 131, 136–146, 248
neuron, 41, 229, 244
 artificial, 87–89, 117–146, 186–195
 visualization of, 136–142
 biological, 2, 5, 7, 37, 44, 52–72, 77–81, 146–150, 180–181, 194, 243, 247
 ear, 146
 face, 143, 144, 146, 147, 248, 256
 floppy ear, 144, 145, 248
 grandmother, 148–150
 Jennifer Aniston, 148–150, 248
 pointy ear, 144, 145, 248
neuroscience, 4, 7, 18, 38, 48, 74, 85–89, 148, 150, 167, 180, 181, 247
 computational, 56, 60, 61, 85
neurotransmitter, 78, 180, 250
New Delhi, 148
Newell, Alan, 196, 198
next word trick, 184
NEXTTOKEN, 184–191
Nim, 154–167, 172, 176, 180
 childlike strategy, 159
 final state, 155
 initial state, 155
Nobel Prize, 37, 43, 86, 130, 147, 180, 215, 217
node (as artificial neuron), 87
non-emergent AI, 250–251
nondeterminism
 in digital transformation, 48, 205
NOT, *see* logic gate, NOT

O’Neil, Cathy, 253
object recognition, 4, 85, 130, 132, 246
off-center cell, 113
olfactory system, 64, 79–81
on-center cell, 113
open peer commentary, 201
OpenAI, 4, 15, 38, 193–194
optic nerve, 147
OR, *see* logic gate, OR
orangutan, 233, 244

- ornithopter, 212, 222
Owen, Richard, 225
Oxford University, 130
- pain, 237–239
Papert, Seymour, 86
parameter, 33, *see also* neural network,
 parameter count; neural network, bias;
 learning rate; weight; tied weight; step
 size; exploration probability
parameter sharing, 108, *see also* tied weights
parsing algorithm, 32
patch (of an image), 107–111, 122, 126,
 133–143
Pavlov, Ivan, 153
Pavlov's dog, 153
Penrose, Sir Roger, 75, 255
perceptron, 86
Phelps, Michael, 26, 32
philosophy, 1, 3, 4, 9, 11, 19, 75, 76, 82, 175,
 196, 206, 240, 251, 255
 ancient Greek, 225
 of AI, 7, 198, 217
 of cognition, 226–245
 of mind and body, 213–214
photoreceptor cell, 112
physics, 49, 213, 222, 227
Pitt, Brad, 148, 150
Pitts, Walter, 44
plasticity, 77, 248, 255
Plato, 225, 251
Platonic forms, 225
pluralism, 238–239, 243
pointy ear neuron, *see* neuron
poker, 3
potassium, 59, 71
pre-training, 184, 189–190, 193
Princeton University, 29, 73, 131
program, *see* computer program
programming language, *see* language,
 programming
protein, 60, 179, 215–216, 223
protein folding, 179
psychiatry, 137
psychology, 3, 37, 180, 250
 animal, 153
 comparative, 229
Putnam, Hilary, 260
pyramidal cell, 55, 61
Q-function, 168
Ramachandran, Vilayanur S., 225
randomness, 45, 93
 in digital transformation, 48, 205
rationalism, 233
RD2, *see* neural network, RD2
reasoning, 38, 76, 77, 139, 149, 189,
 192–193, 204, 209, 214, 228, 229, 232,
 233, 240, 250
receptive field, 106, 107, 113, 134, 140
recognition, *see also* face; object; speech
 emergent, 248–249
rectified linear unit, 89, 133
reductionism, 42–43, 72, 81, 176, 188, 191,
 207
refractory period, 61
regularization, 97
reinforcement learning, 4, 38, 152–181,
 190, 193, 207, 232, 249, 250
religion, 222, 225, 227, 242–244, 246
reproduction, 213, 218–220
residual network, *see* neural network
response (of a neuron), 119, 126, 137, 139
resting potential, 60
retina, 112, 113, 147, *see also* mouse retina
ribosome, 215–217
RL, *see* reinforcement learning
RNA, 215
robot, 11, 153
 “I'm not a robot”, 10, 13
rod cell, 112
rose, 64–66, 79–81
Rush, Benjamin, 137–145
Russell, Stuart, 261
Rutter, Brad, 25
SAM, *see* Script Applier Mechanism
Schank, Roger, 197, 199, 258

- schema networks, 257
- Schrödinger Erwin, 214–215
- script (for ELIZA), 23
- Script Applier Mechanism, 197–200
- Searle, John, 196, 198–200, 208, 227
- self-attention, 184, 190–192, 194, 207, 249
- Sennrich, Rico, 257
- sentience, 6
- Service, Robert, 35–38
- shape (of convolutional layer), 108, 110
- Shelley, Mary, 210, 259
- sigmoid activation function, 89
- signal processing, 103, 243
- silicon, 1, 49, 196, 232
- Silver, David, 173
- Simon, Herbert, 196, 198
- Simonyan, Karén, 130
- simulated data, 93
- simulation, *see* brain simulation; functional simulation
- single-cell organism, 229
- six-stone Nim, *see* Nim
- slash feature, 121–130, 135
- smell, *see* olfactory system
- society of mind, 42
- society, interactions with AI, 3, 6, 8, 150, 151, 222, 243, 244, 251, 253
- sodium, 59, 71, 196
- softmax function, 134, *see also* layer, softmax
- soma, 54, 78, 255
- soul, ix, 226
- Space Invaders, 167, 168
- Spanish language, 47
- SPARQL, 30
- sparse (neural network architecture), 102–105, 107
- sparsity, 103–104, 109, 112, 113, 250, *see also* sparse
- special-human questions, 226, 228
- speech recognition, 85, 103–110, 130
- spike, neural, 55–71, 78
 - frequency, *see* firing frequency
 - train, 61, 62
- spiny neuron, 55
- step size (for temporal difference update), 161, 162
- stochastic gradient descent, 93–110, 130, 146
- stride, 141, 256
- strong AI, 198, 201, 203
- structured content, 29
- subconscious, 209, 241
- substance dualism, *see* dualism
- Suleyman, Mustafa, 167, 253
- supervised learning, 93, 166, 172, 177, 179, 232
- Sutskever, Ilya, 130
- Sutton, Richard, 152, 166, 256
- Sweeney, Latanya, 115, 150, 151
- Sydney Opera House, 148–150
- symbol, *see* formal symbol
- synapse, 54–73, 77–82, 180, 250, *see also*
 - inhibitory; excitatory
- synaptic cleft, 54, 56
- synaptic plasticity, *see* plasticity
- synchronization, 53, 68
- System 1 and System 2, 37–38, 250
- tabula rasa, 233
- tabular solution, 160
- Tegmark, Max, 255
- Tel-Aviv University, 148
- temporal difference update, 160, 162, 165, 180, 181, 250
- test image, 132, 137, 138, 141, 143–145
- test set, 94–98
- thinking AI, 8, 38, 40, 193, 235, 243, 244, 251
- thinking, emergent, 247–248
- Thorndike, Edward, 153
- threshold, *see* firing threshold
- tied set, *see* tied weight
- tied weight, 103, 107–113, 117, 123, 126, 128, 135, 147, 195, 250
- token (in LLM), 184–191, 249
- Top 5 (scoring system), 132, 133, 135
- training, 33, 91, 109, 168, 173, 176, 179
 - an artificial neural network, 91–99

- training data, 84, 92, 103, 117, 124, 135, 136, 139, 172, 183, 184, 189
 - excessive, 234–235
 - for an LLM, 190
 - labeled, 93
- training set, *see* training data
- transformer, 4, 184, 190–192, 194
- Turing, Alan, 1, 9, 12, 15, 154, 231
- Turing Award, 86, 130, 166, 196, 256
- Turing test, 7, 9–20, 22, 34, 47, 74, 154, 200, 202–203, 206, 208, 231
- twin earth thought experiment, 260

- UCLA, 148
- unconscious, 37
- understanding, emergent, 249–250
- unit (as artificial neuron), 87
- University College London, 167
- unstructured content, 29
- unsupervised learning, 42, 146, 248
- upstream
 - axon, 77, 78
 - neuron, 54
- urea, 213

- value function, 158, 159, 163, 165, 171, 172, 176, 180, 181
- value-learning algorithm, 158–166, 180
- vat, *see* brain in a vat
- VGG16, *see* neural network, VGG16
- visual cortex, 147
- vital force, 213–214, 222
- vitalism, 213–215
- vocabulary (in LLM), 184, 186
- voltage, 49, 59, 60, 63, 64, 71, 78, 79

- Wall Street Journal, 202
- Watson, 3, 7, 21, 25–38, 131, 202–204, 209, 250
- Watson, Thomas J., 25
- weak AI, 198
- weather forecasting, 41, 71
- weight (in neural network), 87, 117, 172, 175, 190, 194, 250, *see also* tied weight
 - initialized randomly, 94, 95, 139, 145, 176
 - set manually, 124
- weight layer, *see* layer, weight
- weight sharing, 107, *see also* tied weights
- weighted sum, 87
- Weizenbaum, Joseph, 22–24
- wheelbarrow, 132, 143, 146, 150, 248
- White, John Graham, 72
- width (of convolution), 106
- Wiesel, Torsten, 147
- Wikipedia, 30, 190
- Wilberforce, Bishop Samuel, 225
- Williams, Iwan, 253
- window size, 256
- wing, 212, 220–222
- Wired magazine, 173
- Wöhler, Friedrich, 213
- WordNet, 29–32, 131

- Y8, *see* neural network, Y8
- Yahoo, 10
- Yale AI Project, 258
- Yale University, 197, 201
- Yosinski, Jason, 131

- zero-padding, 122
- Zisserman, Andrew, 130