CONTENTS

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 INTERVIEW[1]

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