1

Introduction

ARE WE SMART?

Against stupidity the gods themselves contend in vain.

—FRIEDRICH SCHILLER

This book is motivated by a fundamental puzzle about human cognition: How can we apparently be so stupid and so smart at the same time? On the one hand, the catalog of human error is vast: we perceive things that aren't there and fail to perceive things right in front of us, we forget things that happened and remember things that didn't, we say things we don't mean and mean things we don't say, we're inconsistent, biased, myopic, overly optimistic, and—despite this litany of imperfections—overconfident. In short, we appear to be as far as one can imagine from an ideal of rationality.[1]

On the other hand, there is an equally vast catalog of findings in support of human rationality: we come close to optimal performance in domains ranging from motor control and sensory perception to prediction, communication, decision making, and logical reasoning.[2] Even more puzzlingly, sometimes the very same phenomena appear to provide evidence both for and against rationality, depending on the theoretical lens through which the phenomena are studied.

This puzzle has been around for as long as people have contemplated the nature of human intelligence. It was aptly summarized by Richard Nisbett and Lee Ross in the opening passage of their classic book on social psychology:

> One of philosophy's oldest paradoxes is the apparent contradiction between the great triumphs and the dramatic failures of the human mind. The same organism that routinely solves inferential problems too subtle and complex for the mightiest computers often makes errors in the

> simplest of judgments about everyday events. The errors, moreover, often seem traceable to violations of the same inferential rules that underlie people's most impressive successes.[3]

As indicated by Nisbett and Ross, the puzzle of human intelligence is reflected in our conflicted relationship with computers. On the one hand, it has long been advocated that error-prone human judgment be replaced by statistical algorithms. In 1954, Paul Meehl published a bombshell book entitled *Clinical Versus Statistical Prediction*, in which he argued (to the disbelief of his clinical colleagues) that the intuitive judgments of clinical psychologists were typically less accurate than the outputs of statistical algorithms. This conclusion was reinforced by subsequent studies and expanded to other domains.[4] For example, in his 2003 book *Moneyball*, Michael Lewis popularized the story of the baseball manager Billy Beane, who showed (to the disbelief of his managerial colleagues) that statistical analysis could be used to predict player performance better than the subjective judgments of managers.[5] Today, the idea that computers can outperform humans, even on tasks previously thought to require human expertise, has become mundane, with stunning victories in Go, poker, chess, and Jeopardy.[6]

More info
= worse

And yet, despite these successes, computers still struggle to emulate the scope and flexibility of human cognition.[7] After the Go master Lee Sedol was defeated by the AlphaGo computer program, he could get up, talk to reporters, go home, read a book, make dinner, and carry out the countless other daily activities that we do not even register as intelligence. AlphaGo, on the other hand, simply turned off, its job complete. Even in the domains for which machine learning algorithms have been specifically optimized, trivial variations in appearance (e.g., altering the colors and shapes of objects) or slight modifications in the rules will have catastrophic effects on performance. What seems to be missing is some form of "common sense"—the set of background beliefs and inferential abilities that allow humans to adapt, almost effortlessly, to an endless variety of problems.

The lack of common sense in modern artificial intelligence (AI) systems is vivid in the domain of natural language processing. Consider the sentence "I saw the Grand Canyon flying to New York."[8] When asked to translate into German, Google Translate returns "Ich sah den Grand Canyon nach New York fliegen," which implies that it is the Grand Canyon that is doing the flying, in defiance of common sense. In fact, the problem of commonsense knowledge was raised at the dawn of machine translation by the linguist Yehoshua Bar-Hillel, who contrasted "The pen is in the box" with "The box is in the pen."[9] Google Translate returns *Stift* (the writing implement) for both instances of "pen," despite its obvious incorrectness in the latter instance.

These errors reflect the fact that modern machine translation systems like Google Translate are based almost entirely on statistical regularities extracted from parallel text corpora (i.e., texts that have already been translated into multiple languages). Because the writing implement usage of "pen" is vastly more common than the container usage, these systems will fail to appreciate subtle contextual differences that are transparent to humans.

Similar issues arise when computers are asked to answer questions based on natural language. The computer scientist Terry Winograd presented the following two sentences that differ by a single word:[10]

1. The city councilmen refused the demonstrators a permit because they feared violence.
2. The city councilmen refused the demonstrators a permit because they advocated violence.

Who does "they" refer to? Humans intuitively understand that "they" refers to the councilmen in sentence 1 and the demonstrators in sentence 2. Clearly we are using background knowledge about councilmen, demonstrators, permits, and violence to answer this simple question. But building AI systems that can flexibly represent and use such knowledge has proven to be extremely challenging.⑪ 7

As a final example, consider the abilities of a modern image-captioning system.[12] When given the image in Figure 1.1, it returns the caption, "I think it's a person holding a cup." Apparently, the system has implicitly used a heuristic that if it sees a cup and a person in the image, then the image probably shows a person holding a cup. But now consider the image in Figure 1.2, which the same system identifies as "a man holding a laptop." Although the cup is heavily occluded, humans have no trouble recognizing that the person on the left is holding one. And of course the "laptop" is a piece of paper![13]

The lesson from this cursory examination of AI systems is that it is much easier to engineer systems that achieve superhuman performance on specific tasks like Go than it is to engineer systems with human-like common sense. This tells us something very important about the nature of human intelligence: our brains are evolved for "breadth" rather than "depth." We excel at flexibly solving many different problems approximately rather than solving a small number of specific problems precisely. Common sense enables us to make sophisticated inferences on the basis of the most meager data—single sentences or images. And the fact that this ability appears to us so effortless—the very fact that common sense is "common" to the point of being almost invisible—suggests that our brains are optimized for fast, subconscious inference and decision making.

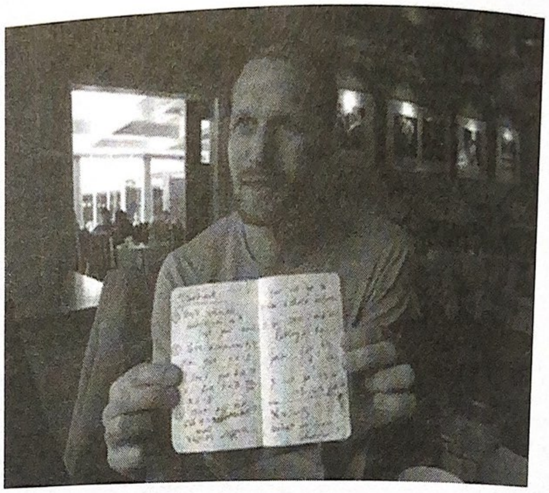

FIGURE 1.1. Image of the author holding a notebook in a restaurant. The image-captioning system believes the image shows "a person holding a cup."

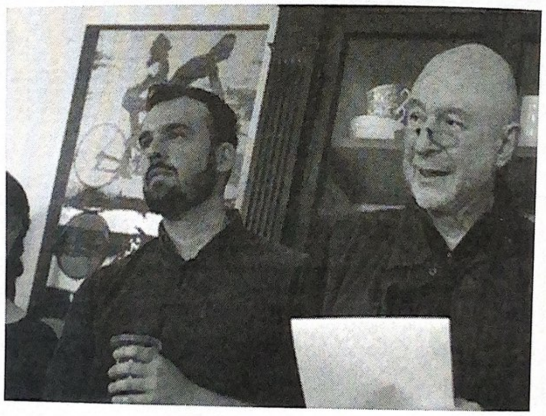

FIGURE 1.2. Image of the author's brother and father. The image-captioning system believes the image shows "a man holding a laptop."

These features of human cognition are shaped by the constraints of the environment in which we live and the biological constraints imposed on our brains. The complexity of our society and technology places a premium on flexibility and scope. We constantly meet new people, visit new places, encounter new objects, and hear new sentences. We are able to generalize broadly from a limited set of experiences with these entities. We have to do all of this with extremely limited energy and memory resources (compared to conventional computers), and under extreme time constraints. To negotiate these demands, our brains make trade-offs and take aggressive shortcuts. This gives rise to errors, but these errors are not haphazard "hacks" or "kluges," as some have argued.[14] They are inevitable consequences of a brain optimized to operate under natural information-processing constraints. The central goal of this book is to develop this argument and show how it reveals the deeper computational logic underlying a range of errors in human cognition.

One might rightfully be concerned that the outcome of this endeavor will be a collection of "just-so" stories—ad hoc justifications of various cognitive oddities.[15] Like Dr. Pangloss in Voltaire's satirical novella *Candide*, we could start from the assumption that "this is the best of all possible worlds" and, given enough explanatory flexibility, explain why all these oddities spring from "the best of all possible minds." However, the goal of this book is not to argue for optimality per se, but rather to show how thinking about optimality can guide us towards a small set of unifying principles for understanding both the successes and failures of cognition. Unlike just-so stories, we will not have bespoke explanations for individual phenomena; the project will be judged successful if the *same* principles can be invoked to explain diverse and superficially distinct phenomena.

I will argue that there are two fundamental principles governing the organization of human intelligence. The first is *inductive bias*: any system (natural or artificial) that makes inferences on the basis of limited data must constrain its hypotheses in some way *before* observing data. For those of you encountering this idea for the first time, it may seem highly unintuitive. Why would we want to constrain our hypotheses before observing data? If the data don't conform to these constraints, won't we be shooting ourselves in the foot? The answer, as I elaborate in the next chapter, is that if all hypotheses are allowable, a huge (possibly infinite) number of hypotheses will be consistent with any given pattern of data. The more agnostic an inferential system is (i.e., the weaker its inductive biases), the more uncertain it will be about the correct hypothesis. Naturally, this gives rise to errors when the inductive biases are wrong. Chapters 2 through 9 are devoted to exploring the implications of this fact, showing the ways in which many different errors that people make are consistent with particular inductive biases. Critically, these are only errors with respect to an *objective* description of reality, to which people do not have direct access.[16] From the *subjective* perspective of an inferential system, the use of inductive biases is not an error at all—it is an indispensable property of a rationally designed inferential system.

The second principle is *approximation bias*: any system (natural or artificial) that makes inferences and decisions with limited resources (time, memory, energy) must make approximations. In particular, optimal inductive inference and planning are intractable for most resource-bounded systems: executing the computations needed to obtain the correct answer requires more time, memory, and energy than is available to these systems. Thus, approximate algorithms are necessary which attain efficiency at the cost of precision. These approximate algorithms give rise to different forms of error, which I explore in Chapters 10 through 12. For example, I show how the need to represent information efficiently leads to distortions in perception, and how the need to calculate probabilities efficiently leads to algorithms that exploit randomness. Again, these are errors with respect to an objective description of reality, whereas they may be optimal from the subjective perspective of the computational system.