CONTENTS

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]

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 common-sense 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]

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.

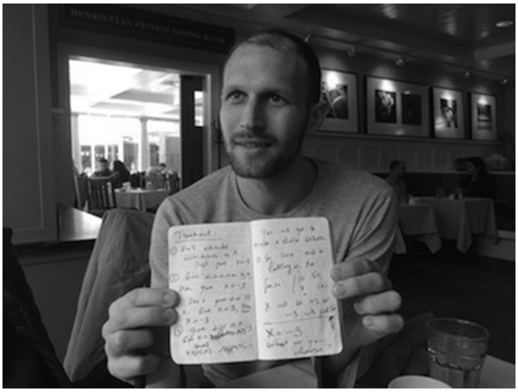

FIGURE 1.1. Image of the author holding a notebook in a restaurant. The image-captioning system believes the image shows "a person holding a cup."

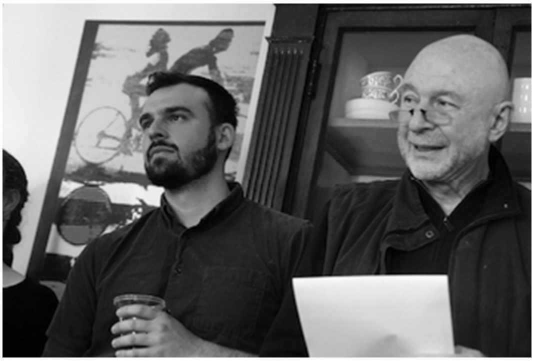

FIGURE 1.2. Image of the author's brother and father. The image-captioning system believes the image shows "a man holding a laptop."

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.

2

Rational illusions

There are strange flowers of reason to match each error of the senses.

—LOUIS ARAGON

Shortly after graduating from college, I went on a hiking trip with a friend in New Hampshire. After hiking since dawn, we stopped for a rest on a plateau and gazed at our goal: a mountain peak that loomed in the distance. As we were sitting there, two people came from that direction, and we asked them how long it would take to get to the peak. "About five minutes," they replied. Five minutes?! We stared at the peak in disbelief; it looked distant enough that it would take at least another half hour of hiking. And then, as we were staring, something monstrous appeared on the ascent to the peak: an enormous giant, towering over the trees, surely at least 30 feet tall. After the initial shock, I realized the trick that had been played on my visual system. Since we were above the alpine level, the trees were only about 3 feet tall. But because we were used to seeing much taller trees, our visual system inferred that the peak must be very far away. The giant, of course, was simply another hiker.

The illusion I experienced is illustrative of how size perception is influenced by contextual information. A well-studied example is the Ponzo illusion (Figure 2.1). The two converging lines resemble train tracks that converge into the distance, creating the impression of depth. Consequently, the lower line looks shorter than the upper line, as though it was placed closer to the observer. In fact, both lines are the same length.

While the Ponzo illusion is a contrived example, it relates to the real-world "moon illusion" that has been known since ancient times. At its zenith, the moon is perceived as smaller compared to when it sits at the horizon. The 2nd-century Roman astronomer Ptolemy argued that the moon illusion is caused by the greater apparent distance induced by terrain at the horizon, which seems to fill more space. In support of this argument, the moon illusion can be virtually eliminated by obscuring the terrain (e.g., viewing the moon in complete darkness or through an aperture). In fact, the moon illusion can be reversed by inverting the image so that the terrain appears closer to the zenith moon than the horizon moon.[1]

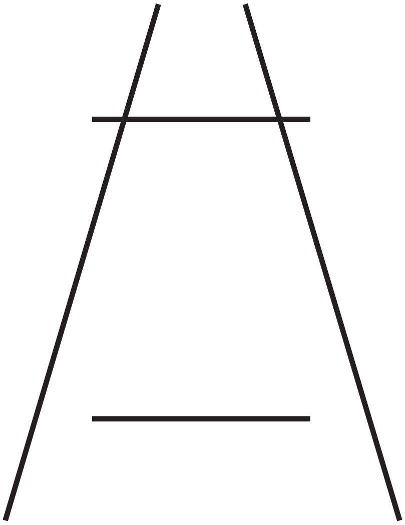

FIGURE 2.1. The Ponzo illusion.

Contextually induced illusions go far beyond size perception. They appear in the perception of color, location, brightness, speed, weight, and many other properties. The ubiquity of such illusions raises a general question: Why did we not evolve brains that perceive the world as it really is? One answer is that veridical perception is impossible given the limits of our sensory organs.[2] As a consequence, the sensory information that reaches the brain is often highly ambiguous. For example, the three-dimensional world is projected onto a two-dimensional retina. This means that the size and distance of an object are ambiguous: a retinal image could be produced by a small object up close or a large object far away (Figure 2.2). The visual system partially resolves this uncertainty by using contextual information (e.g., perspective cues) and background knowledge (e.g., the canonical sizes of objects).

Illusions are a by-product of ambiguity resolution. The same strategy that aids perception can lead it astray in certain cases (in fact, as I discuss later, it is impossible to devise a strategy that will work well under all circumstances). According to this view, illusions are not bugs, but rather essential design features. If we were to design a robot to optimally perceive the world (within the limits of its sensory receptors), then we would expect it to experience illusions.[3] To unpack this argument, we need to dive deeper into what we mean by an optimally designed system. We can then ask to what extent such optimality principles provide a general explanation for perceptual illusions. Is there a logic of perception?[4]

Perception as inference

Putting our engineering hats on, let us consider how we would endow our robot with a perceptual system. Our robot's input is a retinal image, *I*, generated by some three-dimensional scene, *S*, which the robot can't directly

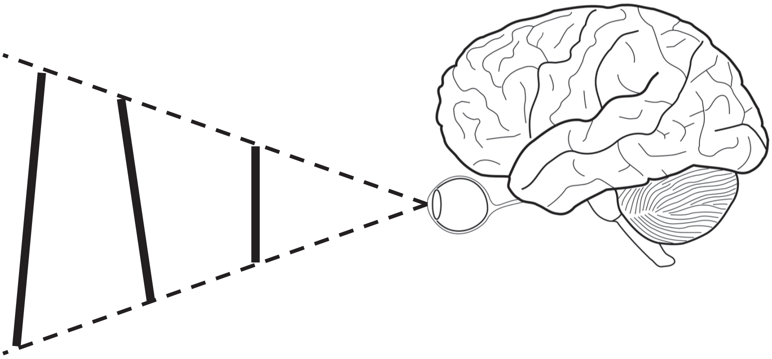

FIGURE 2.2. Sensory information is ambiguous. All of the vertical lines project to the same image on the retina.

observe. As an example, the scene structure could be (partially) specified by the size and distance of a particular object. As already mentioned, the image is typically consistent with many possible scenes. In the Ponzo illusion, for example, a line segment is consistent with an arbitrary size or distance, since we don't have an objective reference point. It's also possible for the image generation process to be "noisy" in the sense that it is influenced by effectively random processes, like whether photons striking the retina cause photoreceptive pigments to change shape, setting in motion the transduction of light into neuronal firing patterns. These different sources of uncertainty can be integrated into a single probability distribution, $P(I|S)$, which expresses the *likelihood* that image I was generated by scene S. Intuitively, the likelihood measures the "fit" between the image and the hypothetical scene.

Note that even if the world were completely deterministic, one can still use probability distributions to express uncertainty (what is sometimes referred to as *epistemic* or *subjective* uncertainty). In our usage, it is better to think of probabilities as degrees of belief rather than descriptions of randomness (the frequencies of repeating events).[5] In conventional usage, I might describe the probability that a coin lands heads by referring to the proportion of heads that I'd expect were I to repeatedly flip the coin. But it's also possible for a weather forecaster to tell you that tomorrow the probability of rain is 70%. Clearly the event "rain tomorrow" can only happen once, so it makes no sense to assign probabilities to a one-time event if we are restricting our usage to frequencies of repeating events. The forecaster is reporting a *belief* about whether or not it will rain tomorrow. This is why Pierre-Simon Laplace, a mathematician who contributed to the early development of probability theory, remarked that probability was "nothing but common sense reduced to calculus."

If our robot has to make a guess about the scene given the retinal image, then one reasonable solution is to report the scene with the highest likelihood (i.e., the scene that is most consistent with the retinal image). This is known as *maximum likelihood estimation*. There are, however, two problems with this solution. First, it neglects contextual information and background knowledge. If you know something about the sizes and distances of nearby objects, or if you know something about the canonical sizes of objects, then you should be able to utilize this information to improve your guess. This leads to the concept of "inductive bias," which I will elaborate later. The second problem is that maximum likelihood estimation neglects subjective uncertainty: if you only report a single guess (a *point estimate*, in statistical parlance), then there's no way to distinguish different levels of confidence in that guess. Suppose you had to bet on your guess; intuitively, you would be willing to bet more if your confidence was higher.

We can remedy the shortcomings of maximum likelihood estimation in two ways. First, we can integrate all of our robot's contextual information and background knowledge into a *prior* probability distribution, $P(S)$. Second, we can allow the robot to report subjective probabilities instead of point estimates. Specifically, it reports the *posterior* probability distribution, $P(S|I)$, the robot's degree of belief that scene S produced image I. One of the most powerful results in probability theory is Bayes' rule, which tells us that the posterior probability is simply the likelihood multiplied by the prior, and normalized so that the probabilities sum to 1:

$$P(S|I) = \frac{P(I|S)P(S)}{\sum_{S'} P(I|S')P(S')}. \tag{2.1}$$

This simple equation has had an enormous impact on theories of the brain and cognition (Box 2.1). We will witness some of its many applications across subsequent chapters in this book.

Box 2.1. The Bayesian brain hypothesis

The idea that the brain represents and manipulates probability distributions might seem exotic at first glance. When I first started studying this question as an undergraduate and told a family friend (a computer science professor) about it, he pointed scornfully at his dog and said, "You think *he* is doing statistics?" But the idea becomes less exotic when we recognize that not only do we routinely report our uncertainty about things, but we can also act in accordance with this uncertainty (e.g., hedging our investments, buying insurance). Bayes' rule has attracted neuroscientists and psychologists because it offers a self-consistent framework for thinking about how uncertainty should be represented, updated, and acted upon.[6] Naturally this doesn't mean that it's correct as a hypothesis about the brain, but it has served as a useful starting point.

How does the brain represent probability distributions? One hypothesis is that populations of neurons implicitly encode distributions.[7] The basic intuition is that downstream neurons receiving signals (spikes) from this population have to reconstruct what information those signals encode, and to do this, they need to take into account the randomness of a neuron's signal-generating process. For example, imagine neurons that fire selectively when particular locations on the retina are stimulated with light. The task for downstream neurons is to reconstruct the stimulated location. Because the firing of neurons is noisy, the best that the downstream neurons can do is assign probabilities to each possible location; these probabilities will be higher to the extent that the noisy neurons selective for those locations are firing more strongly. There are a number of other schemes for representing probabilities with neurons, such as modeling neurons as generating random samples from a distribution[8] (see Chapter 12), or as signaling *prediction errors*[9] (the discrepancy between expected and observed input). These models have been successful at explaining why, for example, the randomness of neural firing seems to track uncertainty (in the case of the random sampling hypothesis), and why expectations about stimuli can sometimes suppress the activity of neurons selective for those stimuli (in the case of the prediction error hypothesis, also known as *predictive coding*).

Although Bayes' rule is conceptually simple, implementing it turns out to be tricky in situations where the denominator cannot be computed exactly (e.g., if there are a very large number of possible scenes). For example, inferring the size and position of even a single object could be computationally intractable. If we discretize the 2 dimensions of object size and 2 dimensions of spatial position into K bins, then to compute the denominator of Bayes' rule exactly would require summing over K^6 possible size-position configurations (a million with $K = 10$). In Chapter 12, we will see a surprisingly effective way to deal with this problem approximately—using random numbers!

Putting aside these computational issues for the time being, suppose now the robot had to act on its beliefs. It chooses an action A and gets rewarded according to $R(S, A)$, where S is the true state of the world (the scene that generated the observed image). The optimal decision rule is to choose the action that maximizes the *expected* reward under the posterior distribution, defined as:

$$\mathbb{E}[R(S,A)] = \sum_S P(S|I)R(S,A). \tag{2.2}$$

In other words, the robot should consider the posterior probability of each hypothetical scene, weigh it by the reward associated with acting on that hypothesis, and choose the action that leads to the highest weighted reward when summed across all hypotheses. Putting some technical subtleties aside, the Bayesian decision rule is optimal in the sense that no other decision rule will reliably lead to higher reward.[10]

As an example of how the Bayesian decision rule can be applied, consider the situation in which the robot's action is a guess (point estimate) about the scene, denoted by $\hat{S}$. I offer it $\$X$ if the estimate is correct, but if it is incorrect, then the robot has to pay me $\$X$. Should the robot take this bet? The expected reward in this case is $\$2P(\hat{S}|I) - 1$, which implies that the robot should only take the bet if the posterior probability of its guess is greater than 0.5 (otherwise the expected reward will be less than or equal to 0). This illustrates how the robot can use its uncertainty to calibrate its betting. This analysis also shows that the robot should (at least for this betting scenario) report the scene with highest probability, also known as the maximum *a posteriori* estimate, if forced to generate a point estimate.

There are several other theoretical arguments about why we would want our robot to be Bayesian. One is the so-called *Dutch book argument*: if the robot did not place its bets according to the Bayesian decision rule, one could create a bet (the Dutch book) that would guarantee the robot a net loss but that nonetheless the robot would accept.[11] Conversely, if the robot follows the Bayesian decision rule, it can guarantee that it won't lose money.[12]

Another theoretical argument is that expressing beliefs as probabilities, and updating them according to Bayes' rule, guarantees that our robot will satisfy an intuitive notion of rationality. Suppose the robot can assign a number, which we'll call a *plausibility*, to each possible hypothesis about the world. Intuitively, logically equivalent hypotheses should have the same plausibility; for example, if two different descriptions refer to the same object, then these two descriptions should be assigned the same plausibility. Small changes in hypotheses should yield small changes in plausibilities, and if a hypothesis is true, it should have a higher plausibility than if it is false. When appropriately formalized, these (and a few other) desiderata lead to the conclusion that plausibilities must be proportional to probabilities, and be updated according to Bayes' rule; any other choice of plausibilities will lead to violations of these criteria for rationality.[13]

Inductive bias

Central to the Bayesian framework is the notion of *inductive bias*: even before our robot has acquired sensory information, it has some prior beliefs about the world. Bayes' rule dictates that these prior beliefs should bias the posterior beliefs, discounting the sensory evidence. This means that a Bayesian robot will make systematic errors. But why would we design a robot that makes systematic errors? The answer is that making inferences about the world is impossible without an inductive bias. Errors are an inevitable consequence of a well-designed inferential robot.

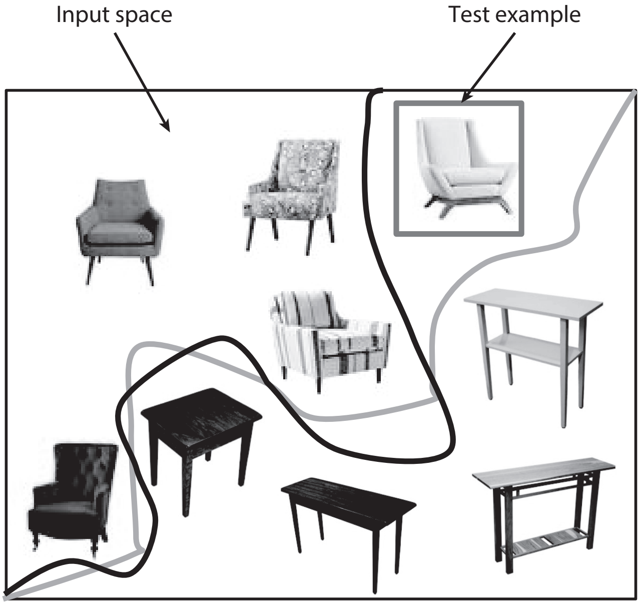

FIGURE 2.3. A chair classifier draws a boundary (possibly complex and non-linear) between chair and not-chair examples in some sensory input space (represented here in 2D). A finite number of examples can be separated by an infinite number of boundaries, assuming the input space is continuous. If none of these boundaries are preferred over others, then the classifier will be unable to classify a test example.

To illustrate this point, I'll borrow an example from the computer scientist Eric Baum.[14] Suppose I build a chair classifier into my robot's visual system. It takes sensory input (e.g., images) and outputs a binary judgment (chair vs. not-chair). Let's suppose I assemble an arbitrarily large (but finite) set of training examples. The classifier is powerful enough to infer a boundary in the input space that separates all the chair examples from all the not-chair examples (Figure 2.3). This means that it can achieve perfect performance on its training set. If the classifier has no inductive bias and the input space is continuous, there exist an infinite number of boundaries that are equally good, in the sense that they all perfectly separate the training examples and thus achieve perfect performance. The implication of this fact is startling: if I give the classifier a new example, it will be unable to determine whether it is a chair or not a chair, no matter how accurate it was on the training set, and no matter how many examples it has collected (short of infinity). There are an infinite number of boundaries that achieve perfect performance and classify the new example as a chair; but there are also an infinite number of boundaries that achieve perfect performance and classify the new example as not-chair. Not having an inductive bias means that the robot has no reason to prefer one of these boundaries over any of the others. Generalization is impossible without an inductive bias!

So inductive bias is necessary, but which inductive bias should we have? It's possible that we could choose a bad inductive bias which would cause us to generalize very poorly, perhaps even worse than random guessing. One way to finesse this problem is to *learn* the inductive bias through repeated experiences with related problems. I will discuss this idea further in the next chapter when we come to the topic of hierarchical learning.

Understanding perceptual illusions

Inductive bias is a key concept for understanding how we perceive the world. Consider, for example, the Kanizsa triangle in Figure 2.4A.[15] One interpretation of this image is that three "Pac-men" are positioned so that their missing sectors form the apices of an equilateral triangle, while three V-junctions are positioned so that their endpoints also form the apices of an equilateral triangle, such that the endpoints intersect the imaginary edges of the triangle formed by the Pac-men. Note that this interpretation does not require us to posit any triangles at all; we simply use triangles to succinctly describe the arrangement of the objects in the image. And yet, we vividly perceive an "illusory surface" formed by a triangle that seems to be implied by the arrangement of the other objects, consistent with an alternative interpretation in which a "camouflaged" equilateral triangle is occluding another triangle flanked by three black circles.

Why do our brains prefer the occlusion interpretation? Arguably because it would require a highly improbable coincidence to arrange the Pac-men and V-junctions in just the right way, whereas the occlusion of one surface by another is quite common. The occlusion interpretation should be more generalizable across different viewing conditions and slight perturbations of the scene. For example, rotating the occluding surface should leave the effect intact (Figure 2.4B). Even after rotating the bottom two Pac-men so that their "mouths" point slightly upwards, one continues to see an illusory surface, as though the triangle was folded at the corners (Figure 2.4C).[16] Slightly rotating the Pac-men and V-junctions in random directions, as though one bumped a table overlaid with a fragile arrangement of cutouts, diminishes the effect (Figure 2.4D), but even in this case one can discern a surface that has been folded and partially cut.

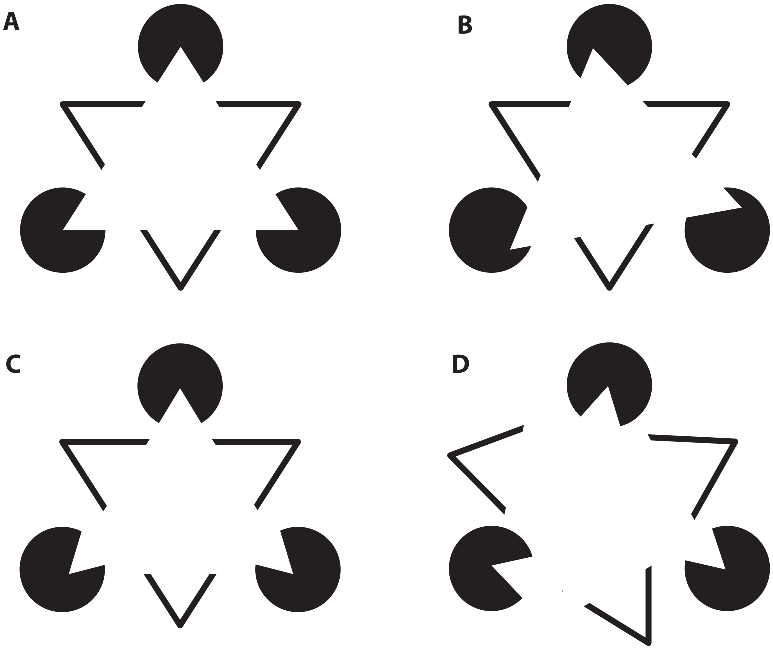

FIGURE 2.4. (A) The Kanizsa triangle. (B) Rotating the invisible occluding surface leaves the illusion intact. (C) If the bottom two Pac-men are rotated slightly, one sees a bent occluding surface. (D) If the image is a suspicious coincidence of Pac-men and V-junctions, then jittering the arrangement largely destroys the illusory surface.

These examples illustrate how our visual system has strong inductive biases about the structure of the world. Though many different scene interpretations are consistent with our sensory inputs, we strongly prefer certain interpretations over others.

How far can we go with this framework? As we will see later, not all errors are naturally derived from inductive biases. Before we get to that point, it will be instructive to go through a few more examples in greater detail. These examples were chosen to illustrate two fundamental principles that arise naturally from the rules of probability:

1. *Explaining away*: When several hypotheses can potentially explain the same observations, additional evidence for one of the hypotheses reduces belief in the other hypotheses.[17]
2. *Integration*: When several sources of information about a hypothesis are available, each source influences beliefs about the hypothesis in

proportion to the source's precision (the accuracy of the information it provides).

Explaining away and lightness illusions

When a surface is illuminated by a light source, a proportion of it (the *reflectance*) bounces off the surface. Some of this light (the *luminance*) reaches our retina, activating photoreceptive neurons. One of the problems facing the visual system is to reconstruct the reflectance from retinal measurements of luminance. *Lightness* is the subjective estimate of reflectance. Whereas reflectance and luminance are physically measurable quantities, lightness is a perceptual quantity that can only be measured by self-report.

Reflectance is ambiguous, because a particular luminance could be produced by a highly reflective (shiny) surface under low levels of illumination (dim light), or by a less reflective surface under high levels of illumination (note the analogy to the size-distance ambiguity discussed earlier in this chapter). The brain uses a number of visual cues to resolve this ambiguity. A classic example of this ambiguity resolution is the Craik-O'Brien-Cornsweet illusion (Figure 2.5).[18] Two adjacent squares have identical surfaces, whose reflectance (and hence luminance) decreases gradually from left to right. Despite the fact that they are identical, the left square is perceived as slightly darker than the right square.

One explanation for this error is that the brain has a strong inductive bias to assume that a light source will not uniformly illuminate a surface unless it is placed directly above the object (an improbable coincidence). More typically, objects and light sources are not aligned with one another, and therefore the light hitting the surface of an object will manifest as a gradient of illumination. This gradient "explains away" the reflectance gradient; the visual system can explain the luminance gradient in terms of differences in illumination, and since it would be improbable for there to be both reflectance and illumination gradients, the brain prefers only one of these explanations.

If other cues are available, the brain may resolve the ambiguity differently. Figure 2.6 shows a three-dimensional version of the Craik-O'Brien-Cornsweet illusion using bricks that appear to be painted different shades of gray. In fact, both bricks have identical reflectance gradients, as in the two-dimensional version of the illusion. When we apply the same reflectance gradients to two cylinders, the illusion is attenuated, because now object shape (surface curvature) offers another plausible explanation of the luminance gradient.

The same line of reasoning can explain why a surface appears darker when placed on a light background: the inferred surface reflectance is lower if the

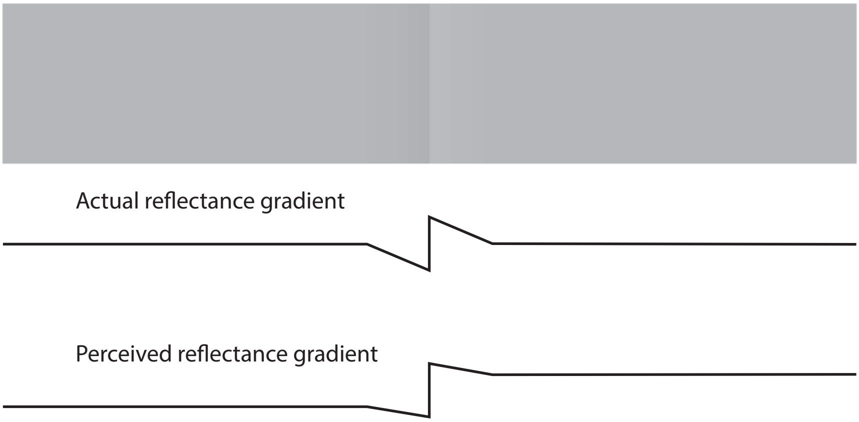

FIGURE 2.5. The Craik-O'Brien-Cornsweet illusion. *Source:* Wikipedia.

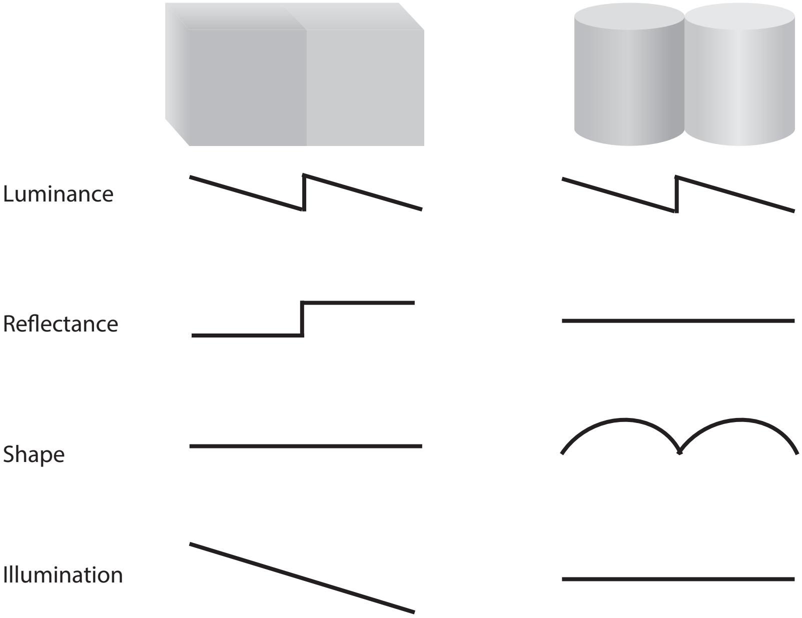

FIGURE 2.6. The brick on the left appears darker than the brick on the right, as though they were painted with different shades of gray, but in fact both bricks have the same reflectance. This illusion is attenuated for the cylinders, because the shape (surface curvature) "explains away" the luminance gradient. Below each stimulus is the scene interpretation favored by the visual system. Adapted from Adelson (2000) and Knill and Kersten (1991). Courtesy of Springer Nature.

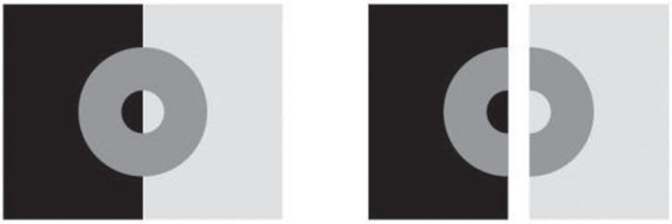

FIGURE 2.7. Koffka rings.

inferred scene illumination is higher. Intuitively, shining a light on an object will not make the object appear to have a lighter shade, even though the luminance reaching your retina is greater. Illumination explains away reflectance. Nonetheless, this explaining away can be overridden by other cues. For example, the Koffka rings in Figure 2.7 illustrate how surface cues can produce the appearance of uniform reflectance. When two half-rings are connected to form a ring, the contrast illusion disappears, because the scene is more plausibly explained by a single surface occluding two backgrounds with different reflectances, rather than two surfaces with different reflectances occluding a single background with uniform reflectance.

The idea that the brain seeks "explanations" of its visual inputs, and evaluates the quality of explanations using Bayes' rule, can be contrasted with "low-level" accounts of illusions, which attribute them to various kinds of image-filtering operations implemented by the brain. For example, some lightness illusions have been attributed to neurons that are excited in response to luminance in a particular region of visual space and are suppressed in response to luminance in neighboring regions. These "center-surround" neurons collectively have the effect of enhancing edges in the neural representation of an image, which in some cases produces illusory lightness. In principle, all illusions could be accounted for by positing appropriate neural filtering operations. The basic problem facing such accounts is that they cannot do justice to the bewildering flexibility of the visual system—it would require a baroque collection of filtering operations akin to Ptolemaic epicycles. How, for example, would these operations "know" about the three-dimensional shape of objects and adjust their filter parameters in just the right way?

The Bayesian framework lifts this problem to a higher level of abstraction by placing the explanatory burden on the internal models assumed by the brain. As long as these internal models are sufficiently flexible (e.g., different luminance patterns can be explained by different combinations of illumination, shape, and reflectance), then Bayes' rule (and, more generally, the probability calculus) offers a coherent mechanism for reasoning about them. This does not deny the existence of neural filtering operations, but rather constrains what kinds of operations the brain would need to implement.

Integration and ventriloquism

Nearly everybody is familiar with the ventriloquist act: a puppeteer coordinates her voice and the puppet's movement in such a way that the puppet appears to be talking. We really feel as though the voice is emanating from the puppet. How is this feat accomplished?

From the observer's perspective, we can formalize this scenario as a problem of multi-sensory cue combination. The observer needs to integrate auditory and visual cues about the spatial location of the speaker. In a laboratory setting, this can be studied by presenting human subjects with auditory and visual cues simultaneously, and then asking them to indicate the location of the audiovisual source on a one-dimensional axis (e.g., horizontal positions on a computer screen) or to judge whether the source location was to the left or right of some preceding reference event. The standard finding is that vision "captures" sound: localization of the source is systematically biased towards the visual location.[19]

A simple mathematical model can explain this phenomenon (Figure 2.8). Assume for simplicity that the prior distribution over location is uniform, so that no location is preferred a priori. Then Bayes' rule says that the posterior over location S given sensory information I is proportional to the likelihood:

$$P(S|I) \propto P(I|S) = P(I_A|S)P(I_V|S). \tag{2.3}$$

The sensory information consists of two parts—auditory (I_A) and visual (I_V)—each corresponding to a location sampled from distributions centered on the true (but unknown) location S. The widths of these distributions depend on the precision of each sensory modality, with broader distributions for less precise modalities. At a physical level, precision derives from many different sources (e.g., the structure of the retina and cochlea), but we can operationally define precision as the average accuracy of spatial localization when a cue from a single modality is presented. Humans are less accurate at localizing auditory cues compared to visual cues. Consequently, the posterior distribution is biased towards the mean of the visual location distribution—a simple laboratory analog of the ventriloquist illusion.

The Bayesian model makes another striking prediction: the ventriloquist illusion can be reversed! Specifically, if the precision of the location information provided by the visual cue is sufficiently degraded (e.g., by blurring or enlargement), then the auditory cue will have higher relative precision, causing sound to capture vision.[20]

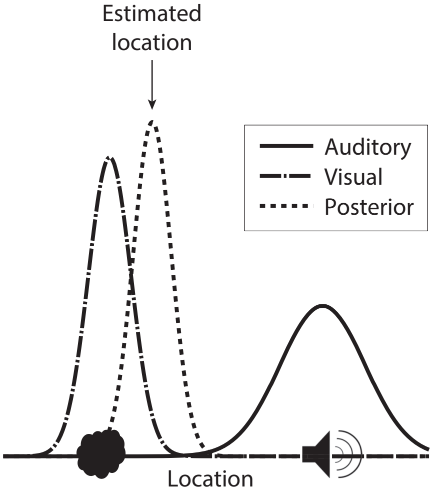

FIGURE 2.8. Ventriloquism as optimal cue combination. Auditory and visual signals are sampled from observation distributions and then combined via Bayes' rule to produce a posterior distribution. The location with the highest posterior probability is usually taken as the subjective estimate of the source object's location. Because visual precision is higher than auditory precision, the subjective estimate is biased towards the mean of the visual distribution.

The story does not end there, however. It turns out that multisensory integration (and hence the ventriloquist illusion) breaks down when the discrepancy between auditory and visual information is large.[21] Thus, the simple model in which auditory and visual cues are obligatorily integrated seems to fail under these conditions. One explanation is that people are making inferences about the causal structure of the cues.[22] Integration occurs when people infer that a single object (hidden cause) produces both auditory and visual cues. This single-cause hypothesis becomes increasingly improbable when the location information provided by the two cues is highly discrepant. Instead, a multiple-cause hypothesis becomes more probable, according to which the two cues are generated by different objects. Indeed, when asked directly, people are more likely to report that a single object generated both cues under low-discrepancy conditions.[23]

Cognitive illusions

A key idea of this chapter is that many perceptual illusions are fundamentally cognitive, in the sense that they draw upon high-level knowledge. At the same time, high-level cognition is itself susceptible to many illusions. One view of cognitive illusions attributes them to heuristics that, while typically useful, generate systematic errors.[24] These heuristics are analogous to the image-filtering operations discussed above in the context of lightness illusions, and they run into some of the same explanatory problems. They can account for specific illusions, but they don't offer a coherent account of cognitive flexibility: What allows us to adapt to the wide range of circumstances that we regularly face? Just as it did for perceptual illusions, the Bayesian framework places the explanatory burden on internal models combined with probabilistic reasoning. Cognitive flexibility derives from the richness of the mind's internal models and the versatility of Bayes' rule. To illustrate this point, I will consider a few examples here that parallel the perceptual examples in the previous section. We will encounter many other examples in subsequent chapters.

Explaining away and the fundamental attribution error

Most of human behavior is a function of personal disposition (e.g., how nice a person is) and situational factors (e.g., cultural norms). In 2004, journalists uncovered evidence of prisoner abuse at Abu Ghraib prison in Iraq, leading to the dishonorable discharge of several soldiers. When interviewed on CBS, Brigadier General Mark Kimmitt made the argument that this was the action of a few bad apples:

> So what would I tell the people of Iraq? This is wrong. This is reprehensible. But this is not representative of the 150,000 soldiers that are over here . . . I'd say the same thing to the American people . . . Don't judge your army based on the actions of a few.[25]

In other words, Kimmitt is making a dispositional inference about the soldiers, discounting the situational factors that may have influenced their behavior. The focus on the responsibility of individual actors, rather than situational factors, is characteristic of human social judgment—so characteristic that this tendency has been designated the *fundamental attribution error*.[26] The reason it is called an error is because people seem to inadequately take into account the power of situational factors, even when they are highly relevant. Inspections of Abu Ghraib prison by the International Committee of the Red Cross led to the conclusion that the abuses were not isolated acts, but rather part of a "pattern and broad system."[27]

This conclusion echoes Hannah Arendt's famous "banality of evil" argument that Adolf Eichmann's crimes during the Holocaust were not the idiosyncratic acts of an unusually evil individual.[28] The Nazis had created a legal system that justified and normalized acts considered crimes by people outside the Nazi system. Eichmann was just doing his job. Arendt's argument sparked a huge amount of controversy, in part because it seems to contradict our strong inclination to assign responsibility to individuals rather than to situations.[29]

Stanley Milgram's studies of obedience to authority reinforce this observation. Milgram took ostensibly normal people off the street of New Haven and asked them to act as a "teacher," delivering electric shocks to another individual (the "learner") whenever the learner gave an incorrect response to a question.[30] The shocks increased in voltage for each incorrect answer. In reality, there were no shocks, and the learner was a confederate pretending to be shocked, but from the teacher's perspective everything was quite real. Despite their visible discomfort, most teachers, when commanded by the experimenter, continued delivering shocks with higher and higher voltages, as the learner's expressions of pain and protest gave way to screams and ultimately complete silence. I recall watching videos of these experiments as a high school student and being astonished that "normal" people could, under sufficient pressure, commit what looked like murder. My astonishment derived from the fundamental attribution error: my mind resisted the inevitable conclusion that these people were not dispositionally "bad." Like Eichmann, they were just doing their job.[31]

Psychologists have investigated the fundamental attribution error more directly by asking people to make dispositional inferences about actors after receiving various kinds of situational information. In one classic study, university students read an essay (supposedly written by a classmate), which gave a favorable or unfavorable opinion about Fidel Castro.[32] In the free-choice condition, the students were told that their classmate was free to write a pro or con essay, whereas in the forced-choice condition, the students were told that their classmate was instructed to write a pro or con essay. Not surprisingly, students in the free-choice condition judged the attitude of their classmate towards Castro to be consistent with the opinion expressed in the essay. Critically, this was still true (albeit more weakly) for the students in the forced-choice condition. In other words, they failed to completely explain away the situational constraints in forming a dispositional inference.

The normative question here is whether people discount enough relative to a rational standard of inference. If one assumes that instructions deterministically control behavior, then the answer is no: discounting should be complete, but empirically it is not. We know, however, that instructions are rarely so potent. When asked about the probability that a classmate with an anti-Castro attitude would write a pro-Castro essay in the forced-choice condition, students judged this probability to be 0.85. Thus, instructions are not considered sufficient to produce behavior.[33] Indeed, 35% of people in the

INDEX