1 Chapter 1: Probability is everywhere

Whenever humans perceive, make a prediction, or deliberate over a decision, we are reasoning with probabilities, even if we do not realize it. Specifically, we are using the information we have at hand to infer something else that interests us. The information we have is usually only partially informative, so our inference is not certain. For instance, the fact that a floor is shiny (the information we have) only *suggests* that it may be slippery (the focus of our interest). Using the available sensory information, the brain must determine the probability of each interpretation

(slippery or not). How can the brain make such judgments in the best possible manner? This book describes the optimal method for performing inference; this optimal method is a form of *Bayesian or probabilistic* inference.

Plan of the chapter: We outline the perceptual inference process, emphasizing the uncertainty that is inherent in perception. Using a simple example, we introduce the probabilities involved in perceptual inference – the likelihood, the prior, and the posterior – and how they are related through Bayes' rule. We then illustrate the ubiquity of perceptual inference in daily life with a series of examples involving visual and auditory perception. Our main goal is to provide an intuitive understanding of the perceptual inference process, which will serve as a foundation for the more rigorous mathematical treatments in the following chapters.

1.1 Perception as Probabilistic Inference

Humans and other animals are endowed with a collection of exquisite sensory organs through which they detect the environment. Organisms detect physical stimulus features as diverse as light wavelength (eyes), sound frequency (ears), temperature (skin), material texture (skin), chemical composition (nose, tongue), and body position (joint and muscle receptors). Our sensory organs form an integral part of ourselves, so much so that we usually take their presence for granted. To appreciate the role that our senses play, try to imagine life without vision, hearing, touch, smell, or taste.

Yet, the activation of sensory organs by physical stimuli is only the first step in the process of perception. We do not primarily care about the pattern of light wavelengths (colors) and intensities (brightness) entering our eyes, or about the pattern of acoustic energy, varying in amplitude and time, entering our ears. Rather, we care about the interpretation of those sensory inputs. In fact, our quality of life – and often our life itself – depends on our ability to come up with the correct interpretations. Does that pattern of light reflect the face of a friend? Is that acoustic waveform the sound of the wind, the howl of a dog, or the voice of our companion? In short, our interest lies not in sensory input per se, but in the information the input provides about the state of the world.

To make the interpretative transition from sensation (the activation of the sensory organs) to perception (a conclusion regarding the state of the world) is a sophisticated task. Broadly speaking, this book is about how the nervous system can optimally accomplish this task. We will examine this issue both at the level of behavior and at the level of neural activity. Our view, based on a large and rapidly growing body of experimental and theoretical work, is that perception is a process of probabilistic inference, in which the organism attempts to infer the most probable state of the world, using all relevant knowledge at its disposal.

The transition from sensation to perception requires *conditional probabilities*. A conditional probability is a probability of one event given another: for example, the probability that you are in a good mood given that it is raining outside. . We denote conditional probabilities as $p(B \mid A)$, read "the probability of B given A."

In perception, what is given (A) is the sensory input or sensory observations, for example the activation of the photoreceptors in our retina. Given the sensory observations, the observer is interested in a state of the world (B). For example, the observer might want to know how probable it is that a floor is slippery, given that the floor is shiny. Since the observer does not know the true state of the world, B is a hypothesis that the observer is entertaining, and we refer to B as the hypothesized world state. The conditional probability of interest to the observer is $p(B|A)$, the probability of a hypothesized world state given the sensory observations. Whether people are aware of it or not, we make conditional probability judgments very frequently in daily life (Fig. 1).

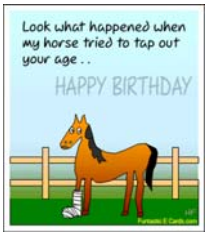
p(my teammate is open to receive my pass | peripheral visual information)

p(a predator is lurking | visual image)

p(that's the president's voice | acoustic information)

p(I will successfully jump this stream | its width, my ability)

p(I will get sick if I eat this apple | its look, smell)



p(my father will laugh when he reads this birthday card | his sense of humor)

p(this is the one | personality, behavior, appearance)

p(diagnosis | symptoms)

p(this book is worth reading | what I've read so far)

Figure 1. Probability judgments. The notation $p(B \mid A)$ is read "the probability of event B given event A."

Depending on the situation, the observer may be concerned with evaluating the conditional probabilities of just two hypothesized world states (B_1: the floor is slippery, and B_2: the floor is not slippery), multiple distinct world states, or even a continuum (infinite number) of world states. Ultimately, we would like to express the results of our inference by calculating the

probability of each world state, given the observation (Fig. 2). This would allow us to make an informed decision about the world.

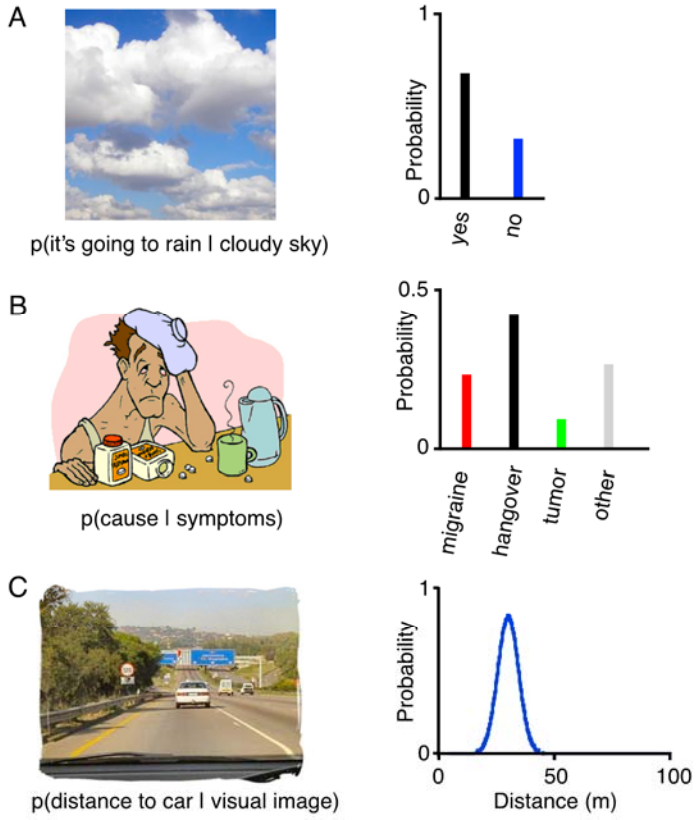


Figure 2. Probability distributions. **A.** Two-hypothesis reasoning. **B.** Multiple hypotheses. **C.** Continuous (infinitely many) hypotheses.

Conditional probabilities have a direction. It is important to understand that *p*(world state | observation), the probability the observer most wants to know, is not the same as *p*(observation | world state). Indeed, in general $p(A \mid B) \neq p(B \mid A)$. One way to envision probabilities is with areas of rectangles (Fig 3). The area of rectangle A is proportional to the probability of event *A*, denoted *p*(*A*) and the area of rectangle B is proportional to the probability of event *B*, denoted *p*(*B*). The area of overlap between two rectangles is the probability of both events occurring together, denoted *p*(*A*,*B*). Lastly, the ratio of overlap area to rectangle area is the conditional probability: $p(A|B) = p(A,B)/p(B)$ and $p(B|A) = p(A,B)/p(A)$.

For instance, suppose *B* = cloudy and *A* = raining. There are more cloudy days than rainy days, so the area of the blue rectangle is greater than the area of the green rectangle (Fig. 3, top panel). The overlap area is the same as the green rectangle area, showing that *p*(*B*|*A*) is 1: the

probability of clouds, given rain, is 100%. However, the overlap area is much less than the blue rectangle area, showing that $p(A|B) << 1$: the probability of rain, given clouds, is small. The difference between $p(A|B)$ and $p(B|A)$ is apparent in many real world examples (see Fig. 3). It is important not to confuse these two probabilities.

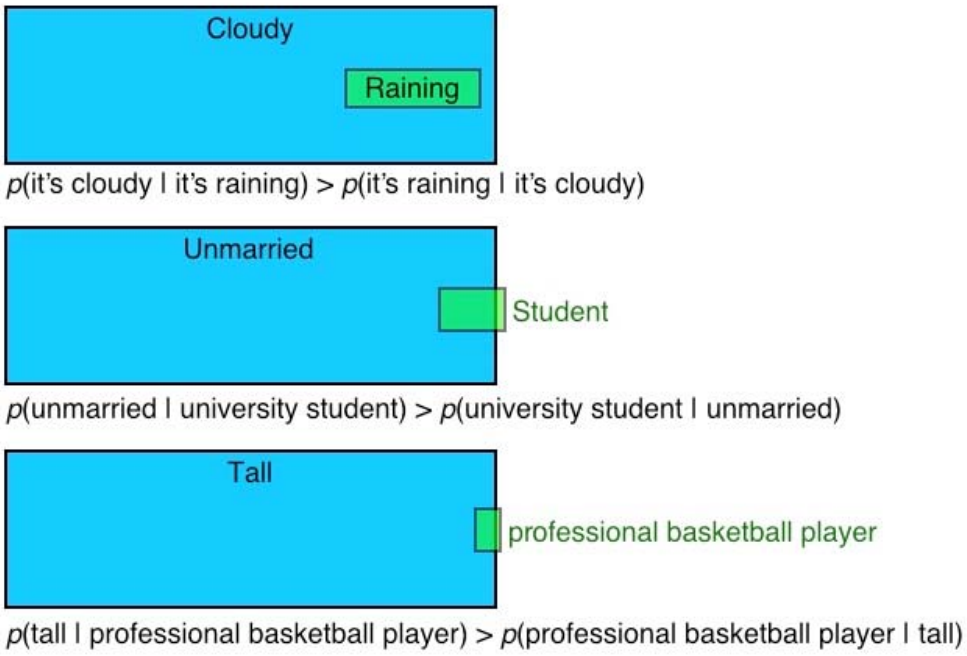


Figure 3. $p(A \mid B)$ does not in general equal $p(B \mid A)$. Rain is always accompanied by clouds [p(cloudy | raining) = 1], but clouds are not always accompanied by rain [p(raining | cloudy) << 1]. Almost all university students are unmarried, but most unmarried people are not university students. Nearly 100% of professional basketball players are tall, but the vast majority of tall people are not professional basketball players. The area of each rectangle represents the probability of the event (the areas are not drawn to scale).

The prosecutor's fallacy

The false belief that $p(A|B) = p(B|A)$ is called the prosecutor's fallacy or the conditional probability fallacy. In general, it is not true that $p(A|B) = p(B|A)$. For instance, most professional basketball players are tall, but most tall people are not professional basketball players. If A is "being a basketball pro" and B is "being tall", then this example illustrates that $p(B|A) > p(A|B)$. For each of the following examples, it should be clear that $p(A|B) \neq p(B|A)$. In some cases, it should also be clear which of the two probabilities is greater. Consider each example carefully:

- p(rain | clouds) $\neq p$ (clouds | rain)
- p(speaks French | born in Paris) $\neq p$(born in Paris | speaks French)

- p(patient has the condition | tests positive) ≠ p(tests positive | has condition)

The prosecutor's fallacy takes its name from the false argument, sometimes put forth in courts of law, that p(defendant is innocent | evidence) = p(evidence | defendant is innocent). For example, suppose that a partial, smudged fingerprint is found on a weapon left at a crime scene. A fingerprint database search reveals that a man who lives in the same city has a fingerprint that matches the one left on the weapon. A forensic expert testifies that only 1 in 1,000 randomly selected people would provide such a match. The prosecutor argues that, based on the forensic expert's testimony, the probability that the defendant is innocent is only 1 in 1,000. The prosecutor is confusing p(observation | innocent) – the testimony of the forensic expert – with p(innocent | observation). In fact, these conditional probabilities are rarely equal. Bayes' rule (section 1.1.4) permits the correct calculation of $p(A|B)$ from $p(B|A)$ and other relevant probabilities, and has been used for this purpose in some courts (Fenton, 2011).

1.2 Conceptualizing inference through an example: the baggage claim

Many air travelers have waited expectantly in an airport baggage claim area, watching for their bags to drop down the chute into the circulating luggage carousel (Fig. 4A). Let us suppose that you are engaged in this ritual of modern-day air travel along with 99 other passengers from your flight, each of whom, like you, checked one item of luggage. A recording piped through the speakers reminds you that "Many bags look alike. Please check your bag carefully before exiting the terminal." Indeed, your bag is one of the most popular models on the market, a black rectangular case used by 5% of all travelers. Of course, if you look at your bag close-up, you will notice individual markings — a name tag, a piece of string you have attached to the handle, etc. — that allow you to unambiguously identify your bag. But at the distance you are standing from the luggage chute, you cannot tell your bag from the 5% of bags in general that have the same shape, size and color. Now let's suppose that the first bag from your flight to enter the luggage carousel indeed has the same shape, size, and color as your bag. Is it your bag?

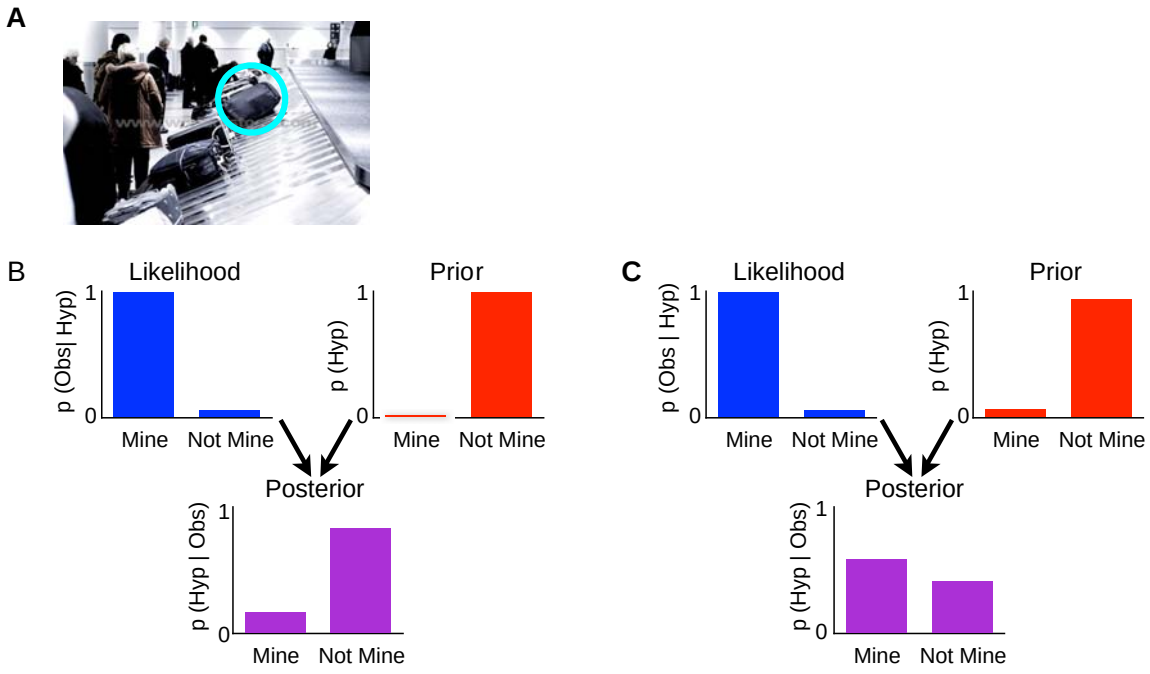


Figure 4. Expectation influences perception. **A.** The 1st bag and also the 86th bag match yours in shape, size, and color. **B.** Likelihood function, prior probability distribution, and posterior probability distribution upon viewing the 1st bag. Your posterior distribution indicates that the bag is probably not yours. **C.** Likelihood function, prior distribution, and posterior distribution upon viewing the 86th bag. The same likelihood as in (A) combined with a different prior expectation produces a posterior distribution that favors the hypothesis the bag is yours. In this and all subsequent figures in the book, likelihood functions are drawn in blue, prior distributions in red, and posterior distributions in purple.

This question cannot be answered with a definitive "yes" or "no." Rather, the question demands a probabilistic judgment. You may consider it more or less likely that the bag is yours, but cannot yet be sure. In lieu of certainty, perception is most often characterized by varying degrees of confidence, which can be expressed as probabilities ranging from impossible to certain, occupying some particular place along the stretch of numbers between 0 and 1 (0% to 100%). As you view the bag in the luggage carousel, you will have an intuitive sense of the probability that it is your bag, *p*(this bag is mine | shape, size, color). But how could you arrive at this probability estimate?

1.2.1 The likelihood function

At the root of perceptual uncertainty is the fact that different world states can generate the same sensory observation. Not only do "many bags look alike," but many objects, people, and events produce nearly identical observations of one kind or another (sights, sounds, etc.). Thus, the information provided by the senses is typically imprecise, open to multiple interpretations.

What information is contained in your observation? If the bag you are viewing is in fact your own, it will have the same shape, size, and color. Thus, *p*(observed shape, size, color | my

bag) = 1. But even if the bag you are viewing is not your own, it has some chance of matching the shape, size, and color of your bag. Since your bag is the model used by 5% of travelers, *p*(observed shape, size, color | not my bag) = 0.05. These two conditional probabilities are known as *likelihoods*. The likelihood of a hypothesis is the probability of the sensory observations if the hypothesis were true, or in other words, how expected the observations are if the hypothesis were true. A plot of the likelihood of every possible world state, known as the *likelihood function*, summarizes the degree to which the observation favors one world state interpretation over the other (Fig. 4B). The less informative the observation, the "broader" or "flatter" will be the likelihood function; the more informative the observation, the "narrower" or "sharper" will be the likelihood function. The observation will generally favor some interpretations more than others, but how much it does so will depend on the quality of the sensory information.

1.2.2 The prior probability distribution

Importantly, the likelihood function is not exactly what the observer wants to know. The likelihood function plots the probability of the observation given each hypothesized world state: *p*(observation | hypothesized world state). What the observer wants to know, however, is the probability of each possible world state, given the observation: *p*(world state | observation). To make this distinction clear, and to discover how to move from *p*(observation | world state) to *p*(world state | observation), let's consider how your perceptual inference will change over time as you wait at the baggage claim carousel.

When the very first bag from your flight enters the luggage carousel, and you notice the resemblance to your own bag, you will be hopeful but at the same time probably somewhat doubtful, that the bag in question is your own. Your skepticism is justified because not only do 5% of bags look like yours, but the probability that your bag would emerge as the first off the flight is just 1 in 100. After all, your flight carried 100 passengers, each of whom checked one bag. Now let's suppose that you have waited expectantly at the carousel, viewing each bag that emerges and checking more closely those that resembled your own, only to find yourself, 10 minutes and 85 bags later, still without having encountered your bag. At this point, let's suppose that the 86th bag emerges, and it again resembles your own. This time, you will be more confident than before that the bag is yours, despite the fact that the observation, and therefore the likelihood function, is identical for the first and the 86th bags. This illustrates that your perception, *p*(world state | observation) is not the same as the likelihood, *p*(observation | world state).

In short, perceptual inference is based not just on the observation (via the likelihood function), but also on expectation. We represent expectation by *prior probability*. The prior probability of a world state is based on all relevant information except the current observation. In the present example, your experience of waiting patiently as 85 bags emerged onto the carousel, together with your background knowledge that 100 bags were present on your flight, has informed you that the prior probability that your bag will emerge next is 1 in 15 (i.e., 6.7%), which is greater than the 1% that it was for the first bag. Although prior probabilities are

conditioned on experience and background knowledge, in the interest of brevity we usually omit the conditioning symbol (|) and write prior probabilities simply as *p(hypothesized world state)*, e.g., *p(the bag is mine)* and *p(the bag is not mine)*. We plot the prior probability of each hypothesized world state as a *prior probability distribution* (Fig. 4B,C).

1.2.3 The posterior probability distribution

The brain somehow has to combine likelihoods, *p*(observation | hypothesized world state), with prior probabilities, *p*(hypothesized world state), to generate the probabilities it most wants to know: *p*(hypothesized world state | observation). These latter probabilities are called *posterior probabilities* to indicate that, unlike prior probabilities, they are formed *after* the observation. The *posterior probability distribution* (Fig. 4B,C) represents the brain's belief in each possible world state, based on all relevant information (i.e., observation and expectation).

How should the brain combine likelihoods with priors to generate posteriors? Remarkably, it turns out that the optimal method is based simply on the multiplication of the likelihood and the prior:

$$p(\text{hypothesized world state}_i \,|\, \text{observation}) = \frac{p(\text{observation}\,|\,\text{hypothesized world state}_i)\,p(\text{hypothesized world state}_i)}{\sum_{k=1}^{N} p(\text{observation}\,|\,\text{hypothesized world state}_k)\,p(\text{hypothesized world state}_k)}$$

where the sum in the denominator is over all possible world states, $k = 1, 2, \ldots N$ (in the luggage example, $N = 2$). This simple but powerful relationship is known as Bayes' rule, and provides the basis for all topics in this book. The numerator shows that the posterior probability of a given world state is simply proportional to the product of its prior probability and its likelihood. The constant term in the denominator ensures that the posterior probabilities of the different world states sum to one (indicating that exactly one of the states is the correct one).

Continuing with our perceptual example, let us evaluate the posterior probability that the first bag that you see emerge onto the luggage chute is your own. We first enumerate the possible world states, which in this case we will call hypothesis H_1 (the bag is mine), and H_2 (the bag is not mine). Next, we write down the prior probabilities of each hypothesis, given our knowledge that this is the first bag to appear. We then write the likelihoods that express the probability of the sensory observation, (shape, size, and color of the luggage seen) given each hypothesis

$$p(H_1) = 0.01 \quad p(\text{observation}\,|\,H_1) = 1$$
$$p(H_2) = 0.99 \quad p(\text{observation}\,|\,H_2) = 0.05$$

Since the prior probability is 1% that the first bag is yours, it is 99% that the first bag is not yours. Note that, since the visual image shows a bag that matched yours in shape, size, and color, we set the likelihood to 1 for H_1. This is logical, since if it were your bag, the visual image will surely match the shape, size, and color of your bag. Finally, we enter the prior probabilities and likelihoods into Bayes' rule, to calculate the posterior probabilities of the hypotheses:

$$p(H_1 \mid \text{observations}) = \frac{1 \cdot 0.01}{1 \cdot 0.01 + 0.05 \cdot 0.99} = 0.168$$

$$p(H_2 \mid \text{observations}) = \frac{0.05 \cdot 0.99}{1 \cdot 0.01 + 0.05 \cdot 0.99} = 0.832$$

There are several important considerations to appreciate at this point:

1) First and foremost, it is important to realize that we have learned from the observation, updating our prior probability for H_1 (0.01) to a posterior probability that is much greater (0.168). Our posterior probably for H_1 has increased because the observation was more consistent with H_1 than with H_2. In general, the more strongly the observation favors one hypothesis over the other, the more we will learn.

2) Nevertheless, we are still more confident that the bag is not ours (83.2%) than that it is ours (16.8%). Despite the favorable observation, we believe that the bag is most probably not ours, because we started with such a low prior probability for H_1. In essence, the observation of a bag that looks like yours does not sufficiently favor H_1 to overcome our well-justified prior bias against H_1.

3) Another important point is that the posterior probability, $p(H_1 \mid \text{observation}) = 16.8\%$, does not equal the likelihood, $p(\text{observation} \mid H_1) = 100\%$. As explained above, in general $p(A \mid B) \neq p(B \mid A)$.

4) Finally, note that in this example the hypothesis with the maximum likelihood (known as the maximum likelihood estimate, or MLE) – H_1 - is not the hypothesis with the maximum posterior probability (the maximum a posteriori estimate, MAP) – H_2. This situation is not uncommon in perceptual inference. Sometimes the MLE and the MAP are the same, but often they are not.

Now suppose that we continue to wait for our bag to appear, failing to see it among the first 85 bags to enter the carousel. To calculate the posterior probability that the 86th bag, which also matches ours in shape, size, and color, is our own, we follow the same procedure, but with new prior probabilities of 1/15 for H_1 and 14/15 for H_2 (Fig. 4C).

Exercise: Verify that the posterior probabilities when evaluating the 86th bag will be $p(H_1 \mid \text{observation}) = 0.588$, and $p(H_2 \mid \text{observation}) = 0.412$.

Thus, our confidence that the bag we are viewing is our own has now increased dramatically, from 16.8% (first bag seen) to 58.8% (86th bag seen), despite the fact that in the two cases the observation, and therefore the likelihood functions, are the same (Fig. 4). The posterior distribution depends not only on the sensory data but also on the prior distribution.

Revisiting The prosecutor's fallacy

Using Bayes' rule, we can now provide a fully worked example to illustrate the prosecutor's fallacy (Box in Section 1.1). Recall that a partial, smudged fingerprint is found on a weapon left at a crime scene. Some of the people who live in the city happen to have their fingerprints on file, and a fingerprint database search reveals that a man who lives in the same city has a fingerprint that matches the one left on the weapon. A forensic expert testifies that only 1 in 1,000 people would provide such a match, and on this basis the prosecutor argues that the defendant's probability of being innocent is only 1 in 1,000.

Let's suppose that the city has 1,000,001 adult inhabitants. Given only that the defendant lives in the city, his prior probabilities of being innocent (H_1) or guilty (H_2) are therefore:

$p(H_1) = 1{,}000{,}000 / 1{,}000{,}001$
$p(H_2) = 1/1{,}000{,}001$

The observation that the defendant's fingerprint matches that at the crime scene results in the likelihoods:

$p(\text{observation} \mid H_1) = 1/1{,}000$
$p(\text{observation} \mid H_2) = 1$

Using Bayes' theorem, we find that:

$$p(H_2 \mid \text{observation}) = \frac{1 \cdot \frac{1}{1{,}000{,}001}}{\frac{1}{1{,}000} \cdot \frac{1{,}000{,}000}{1{,}000{,}001} + 1 \cdot \frac{1}{1{,}000{,}001}} = \frac{1}{1{,}001}$$

The defendant is almost surely innocent, despite the prosecutor's argument! Note that our inference can be verified as follows: The city contains 1,000,001 people, 1,000,000 who are innocent and 1 who is guilty. Consequently, if we had the fingerprints of everyone in the city, we'd expect 1,001 matches, only 1 of which is from the guilty citizen. The probability of guilt, given a fingerprint match, is therefore 1/1,001.

1.3 Bayesian inference: a closer look

Having introduced the elements of Bayesian inference, we now explore more deeply. We first derive Bayes' rule. Next, we discuss factors that influence the likelihood function and prior distribution. Lastly, we provide a brief historic perspective on the origins of inference.

1.3.1 Derivation of Bayes' rule

Where does Bayes' rule come from? There are several derivations that all lead to the same rule. Here we derive Bayes' rule using two basic rules of probability: the product and sum rules. We use the baggage claim scenario in our derivation, but the derivation can easily be extended to scenarios with any number of world states.

Suppose you observe a bag that looks like yours. It is possible that the bag is yours and it looks like yours, or that the bag is not yours but it looks like yours anyway. These two situations can be thought of as pairs of events (world state, observation) that might have occurred:

1. (the bag is mine, it looks like mine) = (H_1, observation)
2. (the bag is not mine, it looks like mine) = (H_2, observation)

We can express the probability of each event pair by applying the *product rule*, which states simply that $p(A, B) = p(B|A)p(A)$. Thus,

1. $p(H_1, \text{observation}) = p(\text{observation}|H_1)p(H_1)$
2. $p(H_2, \text{observation}) = p(\text{observation}|H_2)p(H_2)$

Because $p(A, B) = p(B, A) = p(A|B)p(B)$, we can also write:

1. $p(H_1, \text{observation}) = p(H_1|\text{observation})p(\text{observation})$
2. $p(H_2, \text{observation}) = p(H_2|\text{observation})p(\text{observation})$

Dividing expression 1 by expression 2 in each case, we obtain:

$$\frac{p(H_1|\text{observation})}{p(H_2|\text{observation})} = \frac{p(\text{observation}|H_1)p(H_1)}{p(\text{observation}|H_2)p(H_2)}$$

Since the bag must either be yours or not, these two mutually exclusive probabilities must sum to 1. This is a result of the *sum rule* for probabilities. Thus:

$$p(H_1|\text{observation}) + p(H_2|\text{observation}) = 1$$

The solution to these two equations is Bayes' rule:

$$p(H_1|\text{observation}) = \frac{p(\text{observation}|H_1)p(H_1)}{p(\text{observation}|H_1)p(H_1) + p(\text{observation}|H_2)p(H_2)}$$

$$p(H_2|\text{observation}) = \frac{p(\text{observation}|H_2)p(H_2)}{p(\text{observation}|H_1)p(H_1) + p(\text{observation}|H_2)p(H_2)}$$

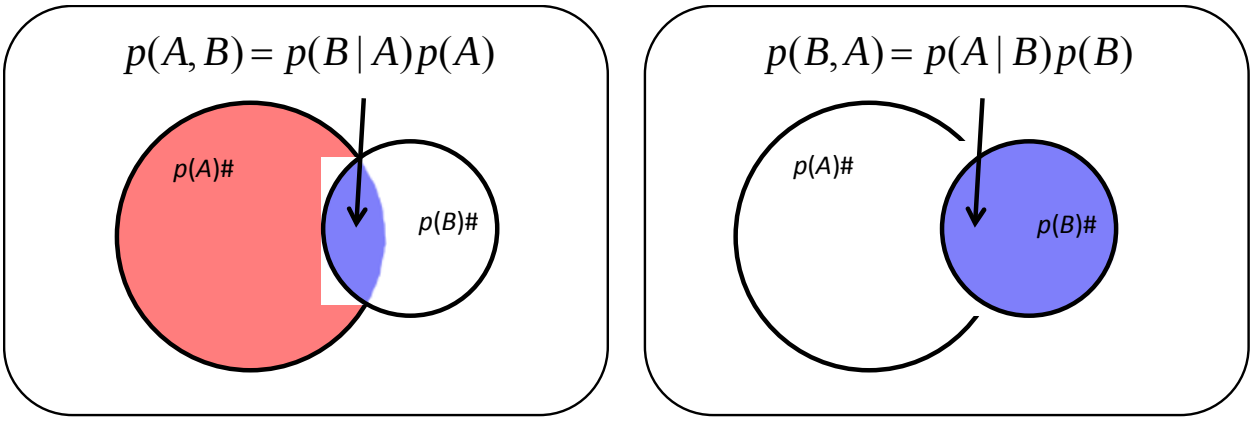


Figure 5. Bayes' rule can be derived by expressing the overlap area (purple) in two equivalent ways.

An Alternate Form of Bayes' Rule

Bayes' rule can be written in several forms, which are all mathematically equivalent. One common form results from the two expressions for $p(H_1, \text{observation})$ shown above:
$p(H_1, \text{observation}) = p(\text{observation}|H_1)p(H_1)$, and
$p(H_1, \text{observation}) = p(H_1|\text{observation})p(\text{observation})$.
It follows that: $p(\text{observation}|H_1)p(H_1) = p(H_1|\text{observation})p(\text{observation})$
Dividing by $p(\text{observation})$ yields a compact form of Bayes' rule:

$$p(H_1|\text{observation}) = \frac{p(\text{observation}|H_1)p(H_1)}{p(\text{observation})}$$

By comparing this form of Bayes' rule to the one we derived earlier, we see that:

$$p(\text{observation}) = p(\text{observation}|H_1)p(H_1) + p(\text{observation}|H_2)p(H_2)$$

In words, the probability of the observation is the probability that H_1 is true AND that the observation would occur if H_1 were true, OR that H_2 is true AND that the observation would occur if H_2 were true (following the product and sum rules of probability, each "AND" is a multiplication, and the "OR" is an addition). Thus, the probability of the data is a weighted sum of the likelihoods of the hypotheses, where the weights are the prior probabilities of the hypotheses. This type of summation is an example of a procedure called marginalization, which we consider in more detail later in the book (e.g., Ch. 6).

1.3.2 Factors affecting the likelihood function

In real life, many factors influence the likelihood function. When you first glimpsed the 86th bag, your view of it may have been partially blocked by other baggage circulating on the

carousel, or by people standing in front of you. Your partial view may have revealed only that the bag was black, but nothing about the bag's shape or size. Based on this initial scant visual information, your likelihood function would have been broader, reflecting the fact that, for example, 40% of travel bags are black (Fig. 6A). When the bag later came into clear view, your likelihood function sharpened as you gained access to the bag's shape and size, in addition to its color (Fig. 6B; this is the likelihood function that we used above). As the bag comes closer, you can distinguish textural details such as wrinkles and scratches. This will result in further sharpening of the likelihood function (Fig. 6C). Many environmental conditions, including distance, obstructions, and others can reduce the quality of sensory data and thereby broaden the likelihood function (Fig. 7A).

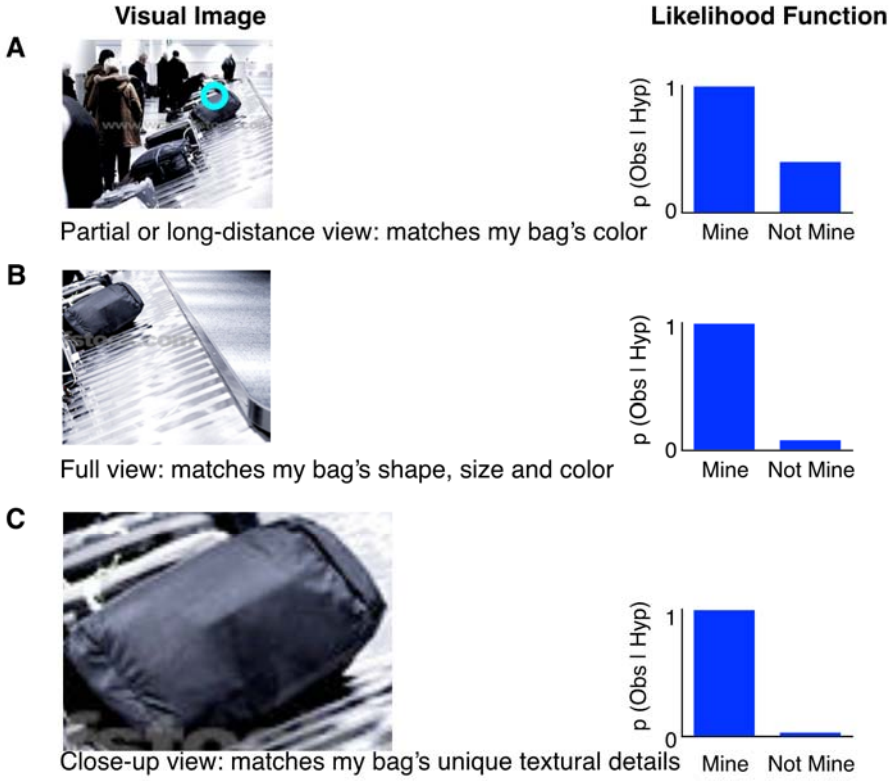


Figure 6. A closer look at the likelihood function. A closer view of the bag reveals features that were not apparent from a distance. This improvement in the quality of the observartion can dramatically sharpen the likelihood function. **A.** 40% of bags match yours in color. **B.** 5% of bags match yours in shape, size, and color. **C.** Fewer than 1% of bags have the same textural details (wrinkles, bulges, scratches) as your bag.

In addition to environmental conditions, the sensory capabilities of the observer play an important role in shaping the likelihood function. A different viewer, with worse eyesight than yours, would experience more sensory uncertainty. In short, the "eyes of the beholder" affect the quality of sensory inputs and therefore the shape of the likelihood function (Fig. 7B).

More subtly, the observer's background knowledge also plays an important role in

shaping the likelihood function. To an observer who remembered only the shape and size of his bag, but not its color, the same visual scene would be less informative (perhaps 5% of bags match yours in shape, size, and color, but 12% match it in shape and size); to an observer who had never before noticed the scratch on his bag, the sight of a scratch might have misinformed the likelihood calculation; an observer with little travel experience might have employed in the construction of the likelihood function an inaccurate estimate of the proportion of bags that look like his own; and so on. Thus, the shape of the likelihood function, like the shape of the prior distribution, depends on the background knowledge of the observer.

Figure 7. Sources of sensory degradation. These factors reduce the quality of visual inputs, causing likelihood functions to broaden. **A.** Physical features of the environment. **B.** The observer's sensory acuity. **C.** The observer's central nervous system. Most of the factors shown in A and B have analogs in the other senses. For example, in the case of audition, distance, soft speech, ambient noise, and ageing ears all result in low-quality inputs. In every sensory system, neural limitations such as faulty background knowledge and neural noise (see Chapters 11 and 12) also pose a challenge to perception.

1.3.3 Factors affecting the prior

Like likelihood functions, prior distributions are based on background knowledge. Prior probabilities therefore evolve over time as the observer acquires new knowledge, and because of this priors also tend to differ from one observer to another. To take an obvious example, an adult will in general have more knowledge about the dangers of crossing a busy street, or being bitten by a snake or a dog, than will a young child, because the adult has more experience with the world.

Priors can change on multiple timescales. They can evolve gradually as the observer gains experience with the world, or much more rapidly during the course of an evolving situation. In the luggage example, the prior probability that the next bag would be yours changed after every observation of a bag entering the carousel. In general, our priors update as we observe a changing situation.

Differences in prior probabilities can arise between one person and another due to differences in a lifetime of experience, or they can arise in a momentary, situation-dependent fashion. Consider a person who arrives late at the baggage claim, after the bags from her flight have already begun to circulate. This person can look at the bags on the carousel, but will not be able to know how many bags have already been retrieved and taken away by other passengers who have left the area. Consequently, she will not have as accurate a prior distribution as will a person who arrived early enough to see each bag enter the carousel. In general, those with greater relevant knowledge have more realistic priors, facilitating accurate perception.

1.3.4 Historic background: perception as unconscious inference

Bayes' rule is so-named after the English mathematician Thomas Bayes (1702-1761), who was interested in problems of inverse probability, essentially how to calculate $p(B|A)$ when we know $p(A)$ and $p(A|B)$. Bayes' *An Essay Towards Solving a Problem in the Doctrine of Chances*, published posthumously in 1763, introduced the foundation for the conditional probability calculus, a field of statistical reasoning now called Bayesian inference.

Bayes' rule was later derived independently by the French mathematician and physicist, Pierre Simon Marquis de Laplace (1749-1827). Laplace applied the formula with great effect to problems in a wide range of disciplines. Importantly, Laplace also recognized the pervasiveness of probability, stating that "the most important questions of life ... are indeed, for the most part, only problems in probability. One may even say, strictly speaking, that almost all our knowledge is only probable" (Laplace, 1995). Indeed, today Bayesian statistical inference is playing a rapidly growing role in an extraordinarily diverse set of disciplines covering nearly all fields of science and engineering: neuroscience, psychology, evolutionary and molecular biology, geology, astronomy, statistical data analysis, economics, robotics, and computer science, to name but a few.

The idea that perception is a form of unconscious inference, however, arose independently of Bayes and Laplace. Several scientists contributed to this notion. The early Arab physicist and polymath, Ibn Alhacen (965-c.1040 CE), recognized presciently that "...not everything that is perceived by sight is perceived through brute sensation; instead, many visible characteristics will be perceived through judgment…in conjuction with the sensation of the form that is seen." Thus, "...familiar visible objects are perceived by sight through defining features and through previous knowledge..." (Alhacen, De aspectibus, Book 2, translated by Smith, 2001).

Much later, the German physician and physicist Hermann von Helmholtz (1821-1894) again expressed the idea that perception is a form of unconscious inference, stating eloquently

that "Previous experiences act in conjunction with present sensations to produce a perceptual image'' (Physiological Optics, 1867). The ideas of Alhacen and Helmholz fit beautifully with the view that perception is a form of Bayesian inference (Fig. 8).

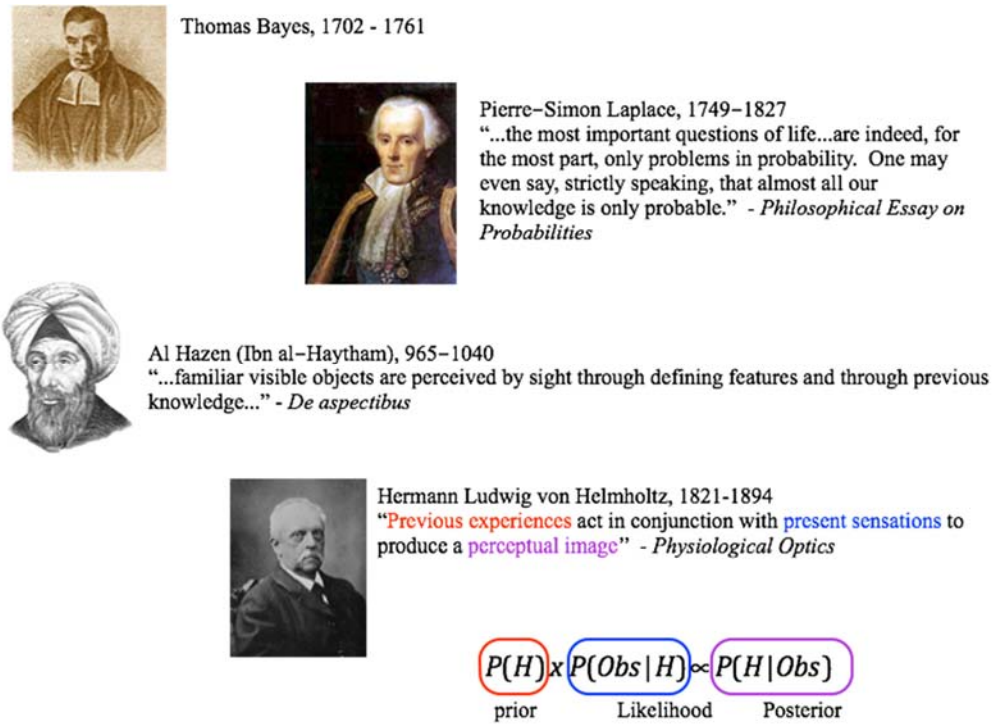


Figure 8. Luminaries in the development of Bayesian inference and the view that perception is unconscious inference.

1.4 Bayesian inference in visual perception

We will now further illustrate perceptual inference with a variety of examples, drawn from everyday life. Our goal is for the reader to develop an intuitive understanding of likelihoods, priors, and posteriors, and an appreciation for the remarkable explanatory power of Bayesian inference as a model of perception. We will not use mathematics in these examples, but will instead explore each example qualitatively and graphically. We will see that each has unique features, yet each is based upon the joining of likelihood function and prior through Bayes' rule to generate a posterior perceptual inference. We hope that these examples begin to reveal both the richness of perceptual inference, and the wide applicability of the Bayesian perceptual framework.

1.4.1 Recognizing a friend

Suppose you see a person walking in the distance, and wonder whether he is your friend (Fig. 9A). Whatever conclusion you reach, you will have some degree of confidence, and your degree of confidence may change over time as you continue to view the scene. We perceive visual scenes with little conscious effort, but the processing the brain engages in is sophisticated. Even the best computer vision systems fail to match the accuracy of human visual perception.

Why is scene recognition so challenging? The sensory input captured by the nervous

system (the visual image in this case) is compatible with multiple interpretations. The visual image could be that of your friend or of another person. The image is not entirely uninformative - it provides sufficient information to recognize that the object in question is in fact a person, and it provides information regarding the approximate shape (height, girth, etc.) of the person. Over time, the moving image also provides information about the person's gait. Nevertheless, the person is far away and your view of him is partially blocked by other people. Thus, recognition is difficult in part because the sensory information is limited, and also in part because many pieces of distributed information must be combined into a joint estimate.

The likelihood function summarizes how well a visual image allows you to distinguish one possible world state from another (friend or not, Fig 9). Many factors affect this likelihood function, including height (your friend is tall) hair color (brown), way of holding the head (tilted). We will leave aside at present how we might arrive at the exact form of the likelihood function. For now, it is sufficient to understand that the likelihood function represents the full information content of the image relevant to the question at hand (is that my friend?). Specifically, it represents the probability that your friend would give rise to the visual image you currently sense, compared to the probability that another person would give rise to the same visual image. The width of the resulting likelihood function defines the difficulty of a perceptual problem.

Importantly, as we have already seen, the likelihood function is not sufficient to solve the problems we want to solve. The likelihood function plots the probability of the observation given each hypothesized world state: *p*(observation| world state). What we want to know is the posterior distribution: the probability of each possible world state, given the observation: *p*(world state | observation). To determine the posterior probability, we combine the likelihood function with a prior distribution. Let's consider two different scenarios, each one associated with the same visual image:

- Scenario 1: You had arranged to meet your friend on the street shown, and at the time shown, when you see the person walking towards you who looks like your friend.
- Scenario 2: When you see the person walking towards you who looks like your friend, you are surprised, because you thought your friend was still away on vacation and not planning to return to town until the following week.

The sensory input is identical (Fig. 9A), but your perceptual inference would differ dramatically under these two scenarios. Under Scenario 1, you would probably conclude that the person walking towards you is your friend; under Scenario 2, you would probably conclude that he is not. Clearly, as we have already seen, expectation plays a crucial role in the perceptual inference process.

Bayes' rule shows how to optimally combine expectation, represented by the prior distribution, with the observation, represented by the likelihood function, to calculate a posterior distribution (Fig. 9). The posterior distribution represents our perceptual inference, and is based on all knowledge we have (present observation and previous experience).

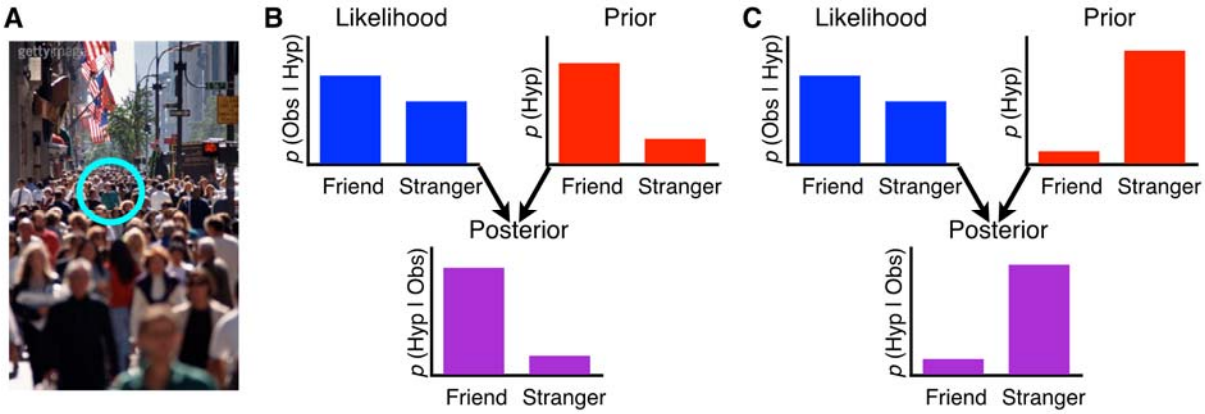


Figure 9. Recognizing a friend. **A.** A crowded visual scene offers a low-resolution view of a person who resembles your friend. **B.** You consider the probability that the visual image would result from your friend to be greater than the probability that it would result from a stranger (likelihood function). You expected to meet your friend (prior distribution). Therefore, you believe the person in question is probably your friend (posterior distribution). **C.** In this alternate scenario, you thought your friend was out of town, so your prior distribution sharply favors the *stranger* hypothesis. Given the same observation (likelihood function), you conclude that the person in question is probably not your friend.

1.4.2 Slippery when wet

As humans move through the world, we rely on our senses, particularly vision, to avoid hazards. In the modern world, hazards come in many forms, for instance an object in our path, a rapidly approaching car, or a downward step such as a curb. Another hazard of modern life is the slippery floor. Is the floor slippery (Fig. 10)? If it is – or might be – caution is warranted: small and slow steps. If it is not, we can safely proceed with long, purposeful strides. An important question is how perception can distinguish the two cases.

Many floors are both shiny and slippery, whereas many other floors are neither. Observing a shiny surface thus results in a relatively sharp likelihood function. The construction of the likelihood function requires background knowledge - in this case, our knowledge of the way various floors tend to reflect light. In general, to determine likelihoods, the observer needs to have an (implicit) understanding of the process by which different world states generate sensory data. In this case, the observer needs an intuitive understanding of optics: that a slippery floor tends to reflect light to a greater degree. To recognize the dependence of the likelihood on the observer's background knowledge, we sometimes write the likelihood as p(observation| world state, B), where B signifies information obtained through previous experience. This makes explicit that likelihood functions depend on background knowledge.

As in the problems discussed above, we need to calculate the posterior probability of each hypothesized world state, p(world state | observation). This calculation involves multiplying priors and likelihoods. Recall that the prior probability, p(slippery), reflects the observer's expectation regarding the slipperiness of the floor, independently of the visual observation.

Before even entering a room, what probability would the observer assign to the hypothesis that the floor will be slippery? How does the observer acquire such priors?

The background knowledge that informs priors may have been acquired over a lifetime of previous experience, or very recently. If the observer has entered the same room in the recent past, and it has not been slippery, the observer's prior distribution will sharply favor the not-slippery hypothesis; if the observer has no previous experience with the room, but on the way to it passed through other rooms in the same building, and these were not slippery, the prior will again sharply favor the not-slippery hypothesis. Even if the observer is entering a building for the first time, the prior will favor (but not as sharply) the not-slippery hypothesis, provided that the majority of floors in the observer's experience are not slippery. By contrast, if the observer sees a newly posted *slippery when wet* sign, the prior will favor the opposite hypothesis. To recognize the dependence of the prior on the observer's background knowledge, we sometimes write the prior as p(world state | *B*), where *B* again signifies information obtained through previous experience.

Since both prior and likelihood depend upon background knowledge, the posterior, too, depends on background knowledge. To recognize this dependency, we sometimes write the posterior probability of each world state as p(world state | observation, *B*).

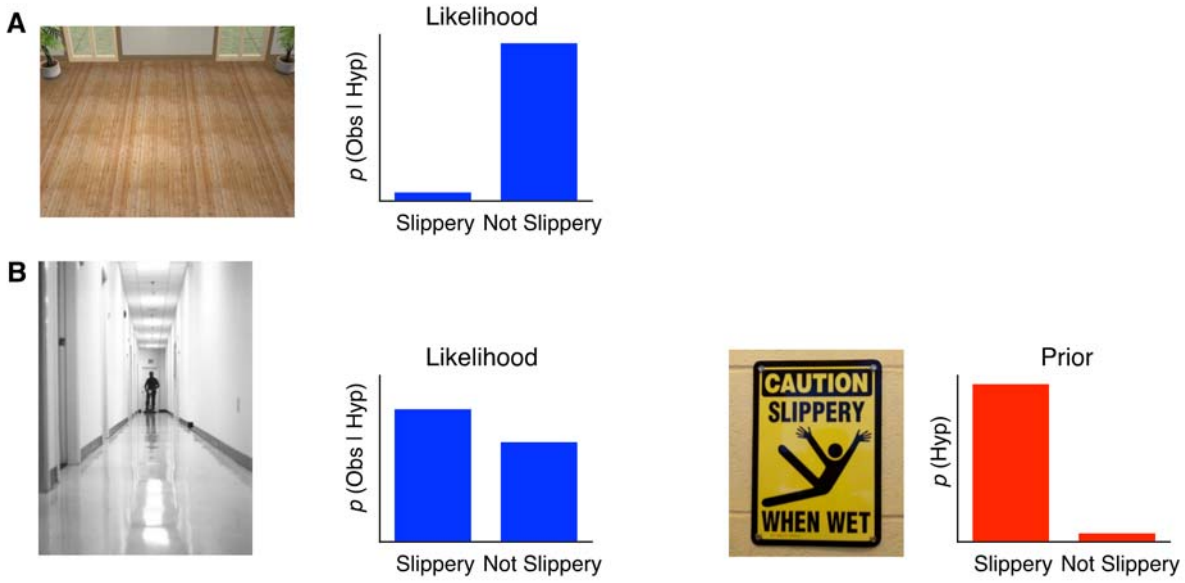


Figure 10. Perception of slipperiness. **A**. The likelihood function resulting from the visual image of this wood floor favors the "not slippery" world state: *p*(observation| not slippery) >> *p*(observation| slippery). **B**. The shiny floor results in a likelihood function that favors the "slippery" world state: *p*(observation| slippery) > *p*(observation| not slippery). A "slippery when wet" sign would result in a sharp prior in favor of the "slippery" world state.

1.4.3 Camouflage

Flat likelihood functions present challenges to perception, and camouflage and mimicry can be

seen as strategies aimed at flattening likelihood functions. Many species have evolved traits and behaviors that serve to disguise their presence or their identity. Consider, for instance, the peppered moth caterpillar. Remarkably, individuals of this species assume the color of the tree bark on which they live (Fig. 11A). By blending in with the background, these caterpillars protect themselves from predatory birds. The visual image provides little indication of the caterpillar's presence. Camouflage is not exclusive to prey; predators, too, benefit from it. Consider the image of a lioness as she lies in waiting for her prey. Crouching low in the high golden grass, whose color closely resembles her own, she is nearly invisible until the moment she strikes. Although they can run fast, lions and other large cats lack stamina for long chases. Their success in hunting therefore depends on their ability to approach prey unnoticed. Examples of camouflaged predators and prey abound in the animal kingdom (Fig 11).

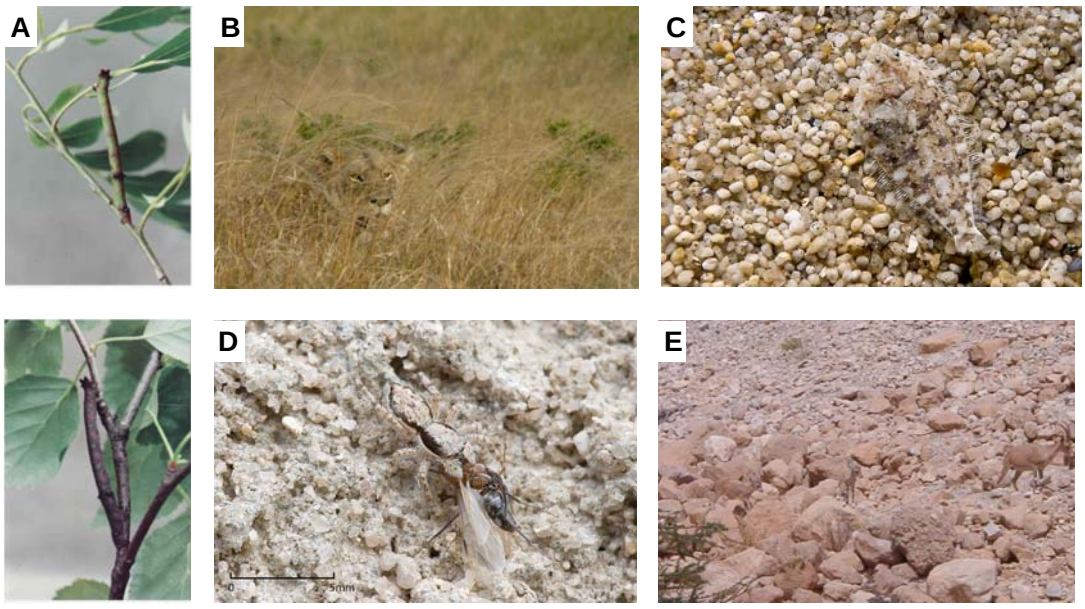


Figure 11. Likelihood function-flattening features in the animal kingdom. **A.** The peppered moth (Biston betularia) caterpillar changes its color to blend in with the background (above: willow; below: birch); from Noor et al., 2008. **B.** In tall golden grass of Kenya's Masai Mara National Reserve, a well-camouflaged lioness lies in waiting for wildebeest prey. **C.** A flounder against the sea floor. **D.** A well-camouflaged jumping spider with its captured ant prey (Dar es Salaam, Tanzania). **E.** Ibex in the Israeli desert.

When a well-camouflaged animal is viewed, (Fig. 12A, the observer's likelihood function does not clearly favor the animal's presence. Of course, as is always the case, the shape of the likelihood function results, not exclusively from the visual image, but also from the sensory abilities and acumen of the observer. A lion that to one observer is nearly perfectly camouflaged may be visible to another observer who has better visual acuity (Fig. 12). To an observer who knows from experience that peppered moth caterpillars tend to be slightly wider than the twigs of the tree they inhabit, the same visual scene (Fig 11A) will result in a sharper likelihood function

than it does for an observer who does not have this background knowledge. As long as the observer's likelihood function is not perfectly flat, the observer will learn something from the sensory input.

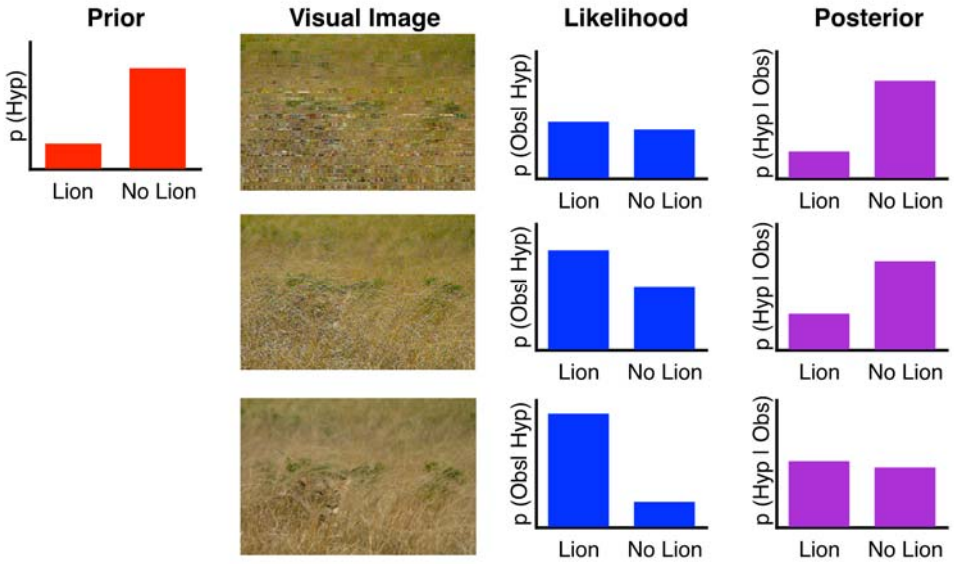


Figure 12. Effect of visual acuity on the posterior distribution. Three prey who hold identical (20%) prior expectation for the presence of a lion (upper left) differ in visual acuity, and therefore experience different likelihood functions when confronted with the same visual scene. The flatter the likelihood function, the more the posterior distribution resembles the prior distribution. **Top**: To this animal with poor visual acuity, the visual scene evokes a nearly flat likelihood function. The animal's posterior distribution is therefore similar to its prior distribution; it has learned little from the visual observation. **Middle:** An animal with intermediate visual acuity has a likelihood function that is not flat. This animal's posterior distribution differs slightly from its prior distribution. **Bottom:** For this animal with excellent visual acuity, the scene results in a sharp likelihood function in favor of the Lion's presence. The animal's posterior distribution indicates slightly greater than 50% probability that a lion is present..

Indeed, along with camouflage, evolution has given rise to sophisticated sensory systems – and cognitive abilities - that function to reduce uncertainty about the locations of other animals. In an arms race of sorts, animals have evolved progressively more sophisticated sensory systems to detect their (often progressively better-hidden) opponents. The evolution of mammalian visual, auditory and olfactory systems are cases in point, as is the evolution of highly specialized detection systems such as the ultrasonic echolocation used by insect-eating bat species. It can be argued that animals' perceptual systems have evolved to produce sharper likelihood functions.

1.5 Bayesian inference in auditory perception

So far, we have considered visual examples. However, nothing about inference is specific to vision. In this section and the next, we consider audition. Humans live in an acoustically rich

environment: birds chirp, the wind howls, dogs bark, car horns blare, music plays, and, perhaps most importantly, we talk to one another. Whether we are identifying the source of a sound (is that a dog barking?), perceiving its location (where is that barking dog?), or interpreting its meaning (what was that word you just said?), we use perceptual inference, combining likelihoods and priors to generate posterior probabilities.

1.5.1 Birds on a wire

Humans often rely at least in part on our sense of hearing to locate objects. We and other mammals localize sounds sources by using sophisticated yet unconscious calculations, including comparing the intensity and time of arrival of sounds at the two ears. Nevertheless, our ability to localize sounds is not perfect, and therefore, as with all perception, we combine prior probabilities with our acoustic likelihoods to reach the most precise perceptual inference we can.

Suppose that you are walking outside on a beautiful sunny morning, when you notice the silhouettes of 5 birds perched on a wire (Fig. 13A). Suddenly, one of the birds (you cannot see which) bursts into melodious song. Which bird sang? Your auditory system rapidly processes the acoustic observation into a broad likelihood function. This likelihood function is a continuous function over location; that is, the sound you heard is compatible with a source at many possible (indeed, an infinite number of) locations. Nevertheless, certain locations are associated with higher likelihoods than others. Interestingly, the location of highest likelihood may well not coincide with the exact location of any bird. This situation is common in acoustic perception, and can be due to many factors. For instance, the bird that chirped was probably not facing you directly; sound can deflect off nearby objects before it reaches your ears; and noise in your own nervous system can cause the likelihood function to shift slightly in location from trial to trial even when the identical sound is repeated.

Unlike the likelihood function, the visual information is not continuous, but rather discrete. You see five individual birds. Thus, the prior distribution, based on your visual observation, is nonzero at just five discrete locations (here we are assuming that your visual perception is highly accurate for this high-contrast scene). Note that the prior probabilities are taken to be equal across the five birds, and the prior probability that the sound source would occupy an empty location on the wire is zero. This simply means that, prior to hearing the song, we consider it equally probable that any one of the birds will sing. Using Bayes' rule, we can now easily calculate the posterior distribution for the location of the sound source. For each of the hypothesized sound source locations, we multiply the likelihood by the prior. The resulting posterior distribution indicates that the singing bird was probably the second one from the right.

Intuition tells us that if the birds had been closer together on the wire, our inference would be less certain. This result indeed emerges from our Bayesian procedure, as shown in Fig. 13B. Here we show the singing bird at the same location, but with three of the other birds closer to it than they were before. Our prior distribution reflects the new positions of the birds, but the acoustic observation, and therefore the likelihood function, are the same as before. The posterior

distribution in this case is broader and lower than before, indicating that, although the same bird is still the most probable singer, our uncertainty has grown.

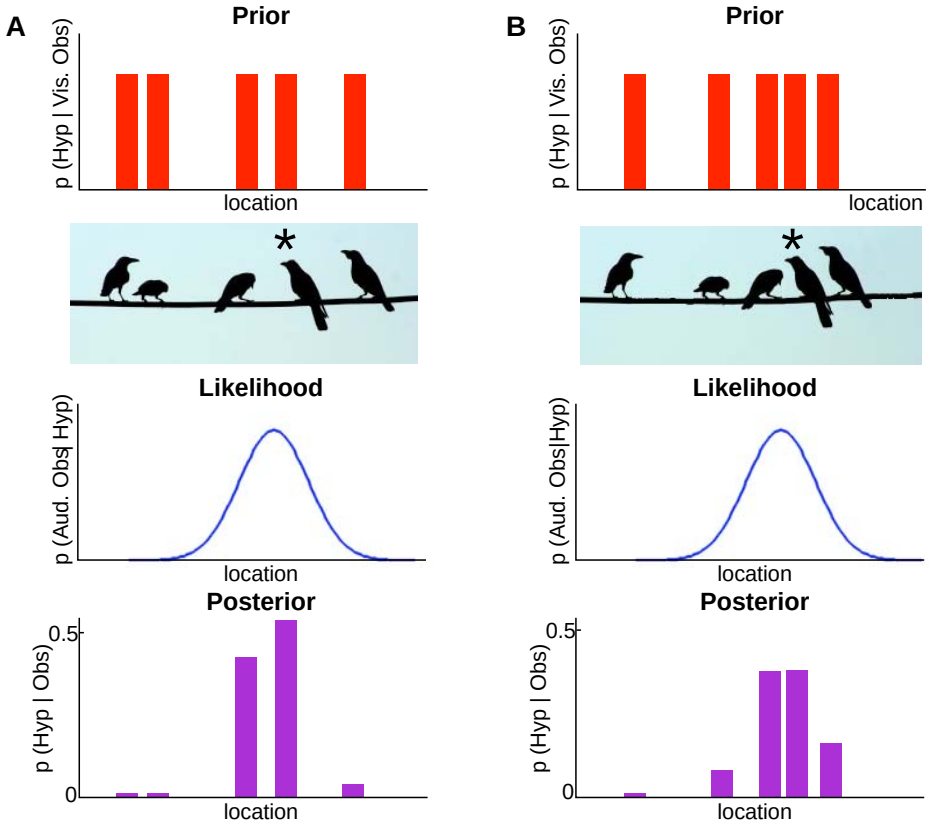


Figure 13. Sound source localization. **A.** The visual image of the birds provides the basis for a prior distribution over sound source location. The broad likelihood function reflects the imprecision of the acoustic observation. The posterior distribution favors the hypothesis that the second bird from right sang (*). **B.** Perceptual uncertainty increases if the birds crowd closer together.

Before leaving this example, we would like to draw the reader's attention to two alternative approaches to solving the problem that would have lead to the same answer. In one alternative approach, we could have started, before looking at the wire, with a flat prior over hypothesized bird locations, reflecting the fact that, before looking, we had no idea as to where any birds would be perched. We could then have incorporated the subsequent visual observation into a likelihood function, and combined this with our flat prior distribution to produce a posterior distribution over the birds' locations. Indeed, it was this original *posterior* distribution from the visual input that we used here as a *prior* distribution for our analysis of the auditory observation. This illustrates a general important feature of Bayesian inference: it can be done iteratively, the posterior distribution from one inference being used as the prior distribution for the next.

We will learn about a second alternative approach to this problem in Chapter 4, which again would reach the same answer: Starting with a flat prior over position, we could incorporate simultaneously both the visual and the acoustic observations as likelihood functions, in a procedure known as *cue combination*. In this approach, we would not use the visual information to generate a prior distribution for the subsequent auditory observation, but would instead combine a visual likelihood function with five discrete peaks with a continuous acoustic likelihood function. In essence, when we have two or more independent sources of information we can choose whether to incorporate the different sources sequentially, with the posterior from each observation being used as the prior for the next, or all at once, with all the observations entering through a likelihood function. Thus, there is often a blurring of boundaries between likelihood functions and priors, with the choice of how to incorporate the information left up to the Bayesian modeler. This flexibility is not a problem but a benefit of Bayesian inference. The internal consistently of the rules of Bayesian inference ensures that, as long as all the information is incorporated, the resulting posterior distribution will be the same regardless of the route taken. Within the Bayesian framework, there are often multiple ways of arriving at the same, useful solution.

1.5.2 Mondegreens

Although our brains do it automatically and apparently effortlessly, speech perception requires sophisticated inference on a variety of levels. Most obviously, we must correctly perceive the spoken word. It is easy to misinterpret even a single word spoken in isolation, particularly in the presence of ambient noise (the drone of a car engine, street sounds, chatter from other nearby speakers, and so on) and/or when the speaker is soft-spoken. Under such conditions, akin to low-contrast vision, likelihood functions are broad. It is instructive to keep a list of such occurrences. In recent conversations with others, we have misheard *Mongolia* as *magnolia*, *fumaroles* as *funerals*, *hogs* as *hawks*, *census* as *senses*, *a moth* as *I'm off*, *maple leaf* as *make believe*, *peaches and strawberries too* as *peaches and strawberries stew*.

As the last three of these examples illustrate, an additional challenge to speech perception arises because the pauses between spoken words are often no longer than the pauses between syllables within a single word. Thus, it is by no means a trivial task to infer where one word ends and the next begins. This difficulty leads to errors in which syllables from different words combine improperly in our perception. As a child, the author Sylvia Wright enjoyed listening to the popular 17th-century Scottish ballad, The Bonny Earl o'Moray, read to her frequently by her mother. She was particularly fond of the sad but beautiful lines describing the murder of the Earl and of his love, the Lady Mondegreen:

> Ye Highlands and ye Lowlands,
> Oh, where hae ye been?
> They hae slain the Earl o'Moray,
> And Lady Mondegreen.

In fact, the words heard by the young Sylvia Wright were not those that her mother spoke. The last line of the ballad actually reads: "And laid him on the green." The unfortunate dead Earl was placed on the grass, alone – Lady Mondegreen existed only in Sylvia Wright's mind! Sylvia Wright's creative but mistaken interpretation of the spoken ballad reflects a parsing error. She heard the sounds "laid hi-" as "lady," and "-m on the green" as "Mondegreen."

Sylvia Wright later coined the term "mondegreen" to refer to a misheard word or phrase. Given the inherent phonetic ambiguity of spoken language, examples of mondegreens abound. When Queensland, Australia was inundated by tropical cyclone Tasha, the Morning Bulletin of Rockhampton (Jan 6, 2011) reported that, as a result of the flooding, "More than 30,000 pigs have been floating down the Dawson River since last weekend." This information, based on an interview between the reporter and the owner of a local piggery, was staggeringly incorrect. The owner had spoken, not of "30,000 pigs," but of "30 sows and pigs" floating downstream! The Morning Bulletin published a correction the next day.

Books and many websites are devoted to listing peoples' favorite mondegreens, particularly those resulting from misheard song lyrics, which we can all enjoy because we share access to the songs. It is instructive to visit websites on which listeners post their particular misheard versions of the same songs. The many different misheard versions of a line such as "Lucy in the sky with diamonds" presumably reflect both the phonetic ambiguity (broad likelihood function) and improbable content (low prior probability) of those lyrics. With respect to prior probabilities, "There's a bathroom on the right" is surely a more common sentence than "There's a bad moon on the rise", and "submarine" is arguably more plausible than "summer breeze" as a mode of transport (Fig. 14).

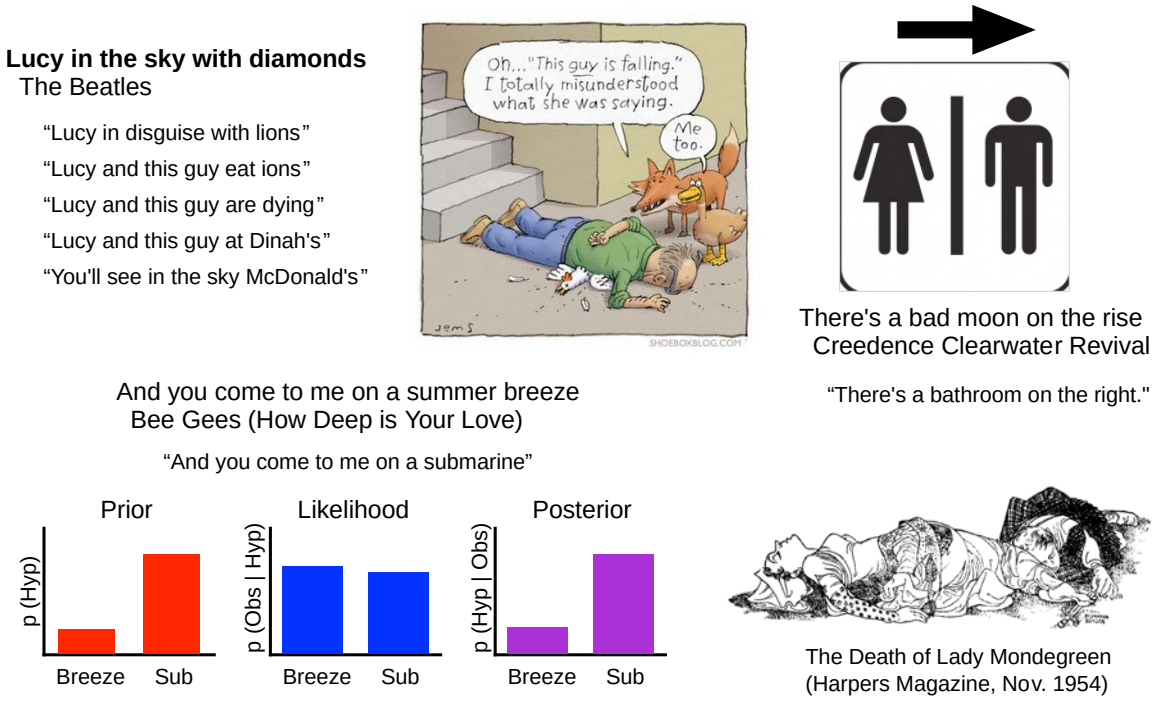


Figure 14. Mondegreens result from phonetic ambiguity (broad likelihood functions) coupled with low expectation for the actual phrase that was sung or spoken (prior distribution in favor of the "wrong" hypothesis).

An important lesson to take away from mondegreens is that speech perception is based, as is any perceptual inference, in the combination of likelihood functions and prior distributions. We generally perform this task very well, but of course occasional mistakes are inevitable. Indeed, "The more unintelligible the original lyrics, the more likely it is that listeners will hear what they want to hear" (O'Connell, 1998). In the terms of Bayesian inference, the flatter the likelihood function, the greater will be the influence of the prior distribution on the resulting posterior distribution (just as in Fig. 12). Thus, the same feature of perception that usually serves us so well – our incorporation into perceptual inference of our prior expectation – backfires to create mondegreens when we are faced with an unexpected word (low prior) that sounds like (broad likelihood) another, more expected (high prior) word.

Keeping this in mind, it is easy to evoke mondegreens in others: simply select two different words or phrases that sound alike, ensure that your listener has a prior distribution in favor of one of the words or phrases, then speak the other. You may wish to try the following demonstration with a friend. Tell the friend that "You know, humans are very good at speech recognition; in fact, we can understand speech much better than even the best computer programs can. We really know how to wreck a nice beach. Now, what did I just say? We really know how to….?" If you spoke the words "wreck a nice beach" naturally, at your typical speed, and in a typical, not extremely clearly enunciated fashion, your friend will probably have perceived "recognize speech," rather than the words you actually spoke. In Bayesian terms, the

broad likelihood function experienced by your friend will combine with a sharp prior distribution (given the previous content of your discourse) to favor the "recognize speech" hypothesis.

Even when the listener perceives every word correctly, she faces a final crucial challenge: to identify the intended meaning of the string of words. Once again, this often requires evaluating multiple hypotheses. Consider the sentence, "The bridge is being held up by red tape." This sentence, even when perfectly heard, is nevertheless consistent with two interpretations; that is, it evokes a broad likelihood function, due not to phonetic ambiguity but rather to syntactic ambiguity. In fact, sentences such as this occur quite commonly in English (Fig. 15). When we hear such a sentence, or read it, we naturally combine the likelihood functions with a prior distribution, and usually reach the correct perception. We are sometimes bemused – and amused - momentarily, however, as both interpretations cross our minds. This occurred recently to one of the authors when a friend told him of a wilderness trip he took with his parents. "There was wildlife everywhere," he exclaimed, "In fact, I saw a bear walking with my mother" (Fig. 16). Although this book will not focus on this type of semantic inference, we point it out to illustrate that uncertainty and inference play a role at all levels of brain function.

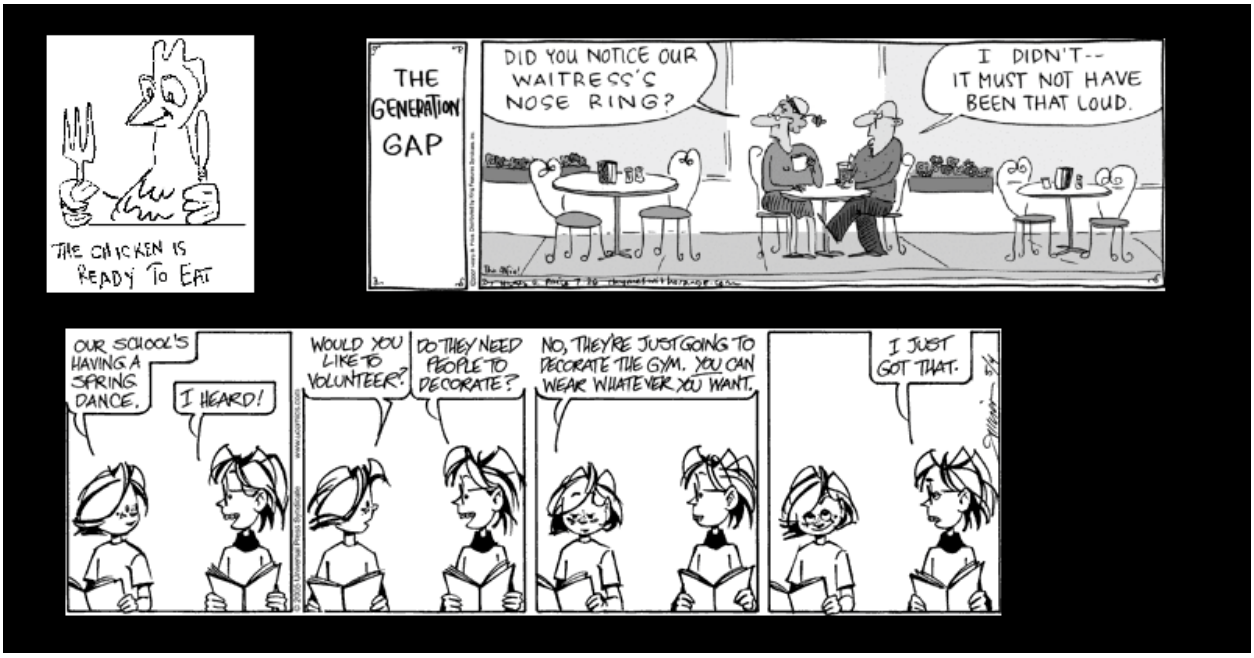


Figure 15. Syntactic ambiguity in language.

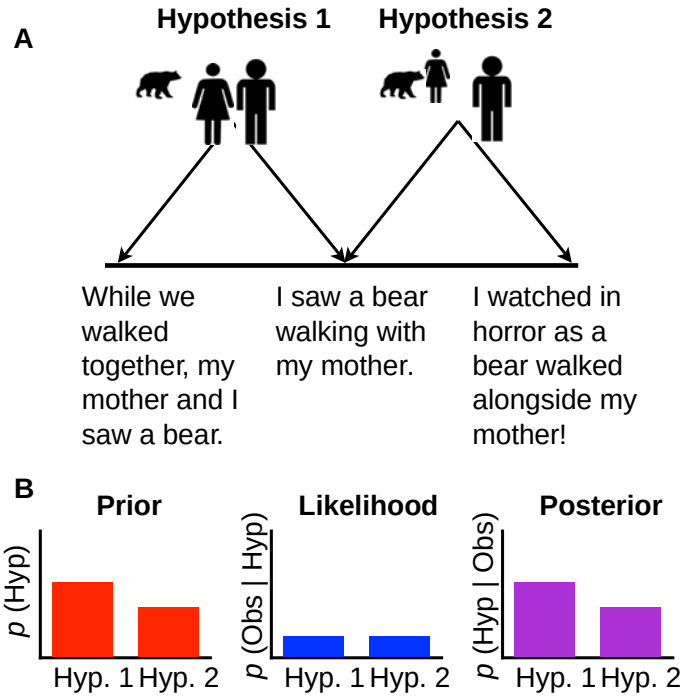


Figure 16. Perceptual inference under syntactic ambiguity. **A.** Each world state could be described in many different ways, a few of which are shown. The speaker happened to choose an expression that could describes both world states: "I saw a bear walking with my mother." **B.** Bayesian perceptual inference. Background knowledge suggests that bears are less likely to walk alongside people than to be seen by them at a distance, so the prior distribution favors Hypothesis 1. The likelihood function shows that the spoken sentence has about equal probability under the two hypotheses. The posterior distribution therefore favors Hypothesis 1.

1.6 Concluding remarks

In this chapter, we have introduced the concept that perception is inherently probabilistic, and as such it is optimally characterized as a process of Bayesian inference. Regarding Bayesian inference, we have learned the following:

- The likelihood function summarizes the information content of the sensory observation, relevant to distinguishing one world state from another.

- Perception is not based entirely on sensory observation, but also on expectation grounded in previous experience. We express expectation as a prior distribution over world states.

- Bayes' rule calculates the posterior probability of each possible world state from the likelihoods and prior probabilities of the world states.

- The posterior probability of a world state is not the same as the likelihood of the world state. In general, $p(A \mid B) \neq p(B \mid A)$.

- Flat likelihood functions pose a challenge to perception. The flatter the likelihood function, the more the posterior distribution resembles the prior distribution. If the likelihood function is perfectly flat, then the posterior distribution is identical to the prior distribution, and the observer has learned nothing from the observation.

- The procedures of Bayesian inference apply equally to situations in which the hypothesized world states are discrete or in which they are continuous.

- Like visual perception, auditory perception is well described as a process of unconscious Bayesian inference.

- Bayesian inference can be done iteratively, a process in which the posterior distribution from one inference is used as the prior distribution for the next. For example, a posterior distribution based on a previous observation in one modality (e.g., vision) can be used as a prior distribution for a subsequent inference based on a later observation in another modality (audition).

- Speech is fraught with phonetic and syntactic ambiguity, giving rise to flat likelihood functions in many instances. As with other perceptual inference, posterior distribution s in speech perception reflect the influence of likelihoods and priors. When likelihoods are nearly flat, priors exert a greater influence on the posterior distribution. This can cause misinterpretations such as mondegreens.

1.7 References

Bayes, T. (1763). An Essay Towards Solving a Problem in the Doctrine of Chances. Philosophical Transactions of the Royal Society of London, 53: 370–418.

Brugger P, Brugger S (1993) The Easter bunny in October: Is it disguised as a duck? Perceptual and Motor Skills 76: 577-578.

Burdon D (Jan. 6, 2011) Pigs float down the Dawson. The Morning Bulletin.

Fenton N (Nov. 3, 2011) Improve statistics in court. Nature 479:36-37

Helmholtz, Hermann Ludwig von (1925) Treatise on Physiological Optics, III: The Perceptions of Vision (1910); Southall JPC, ed. RochesterN.Y.: Optical Society of America.

Laplace, Pierre Simon, Philosophical Essay on Probabilities, translated from the fifth French edition of 1825 by Andrew I. Dale. Springer-Verlag: New York (1995).

Noor MA, Parnell RS, Grant BS (2008) A reversible color polyphenism in American peppered moth (Biston betularia cognataria) caterpillars. PLoS One 3:e3142.

O'Connell, PL. Sweet Slips Of the Ear: Mondegreens. New York Times (Aug. 9, 1998).

Rosenhouse, Jason, The Monty Hall Problem: The remarkable story of math's most contentious brain teaser. Oxford University Press: New York (2009).

Smith, A. Mark, Ed. (2001) Alhacen's Theory of Visual Perception: A Critical Edition, with English Translation and Commentary, of the First Three Books of Alhacen's *De aspectibus*, the Medieval Latin Version of Ibn al-Haytham's *Kitab al-Manazir*. Transactions of the American Philosophical Society, volume 91, parts 4-5.

Wright, S. (Nov, 1954) The Death of Lady Mondegreen. Harper's Magazine, 209: 48-51.

1.8 Further reading

History and applications of Bayesian Inference

McGrayne, Sharon Bertsch, The theory that would not die: how Bayes' rule cracked the enigma code, hunted down Russian submarines & emerged triumphant from two centuries of controversy. Yale University Press: New Haven (2011).

Evolution of perceptual systems and camouflage: a Bayesian perspective

Geisler, W. S. and Diehl, R.L. (2003) A Bayesian approach to the evolution of perceptual and cognitive systems. Cognitive Science 27: 379-402.

Perception as Unconscious Inference

Hatfield, G. (2002). *Perception as Unconscious Inference.* In Perception and the Physical World: Psychological and Philosophical Issues in Perception, ed. by Dieter Heyer and Rainer Mausfeld (New York: Wiley), 115–143.

Phonetic and syntactic ambiguity

Smith, R. *Milk drinkers turn to powder and other pun-ishing headlines.* Globe and Mail (Sept. 24, 2009)

Red tape holds up new bridge, and more flubs from the nation's press, Gloria Cooper, ed.; collected by the Columbia journalism review (1987)

1.9 Problems

1. If A is the event "a person is old," and B is the event "a person suffers from Alzheimer's disease," is $p(A|B)$ less than, equal to, or greater than $p(B|A)$? Why?

2. Generate three comparisons of your own, of the type $p(A|B) \neq p(B|A)$. In each case, state which probability is greater, and explain why.

3. At a particular university, 15% of all students are in humanities, 58% of all students are undergrads, and 19% of undergrads are in humanities. What is the probability that a random humanities student is an undergrad?

4. 1% of the population suffers from disease D. A diagnostic test for D is being piloted. The probability that someone without D tests positive (false-alarm rate) is 2%. The probability that someone with D tests negative (miss rate) is 3%.
   a) Make a quick guess of the probability that someone who tests positive actually has D.
   b) Calculate this probability. If it is very different from your answer to a), what went wrong in your intuition?
   c) (*) (Due to Huihui Zhang, Beijing University) Suppose now that there is an extra variable we have ignored, namely whether someone goes to the doctor to have a diagnostic test done. This probability is higher if someone has the disease (because there will likely be symptoms) than if someone does not have the disease. Assume a 5-to-1 probability ratio for this. Now recalculate the probability that someone who tests positive actually has D. Is it closer to your original intuition?

5. If you look carefully at the probability graphs in the figures of this chapter, you will notice that the prior probabilities of the different hypotheses sum to one, as do the posterior probabilities of the different hypotheses. The likelihoods, however, do not generally sum to one. Why do likelihoods not have to sum to 1, whereas priors and posteriors do? Hint: think carefully about the definitions of likelihood, prior and posterior.

6. Prove using Bayes' rule that when the likelihood function is perfectly flat, the posterior distribution is identical to the prior distribution.

7. (See Section 1.2.) Prove using Bayes' rule that if you see a bag on the luggage carousel that does not match yours (for instance, a small red bag, when yours is large and black), the posterior probability that it is yours is zero.

8. * (See Section 1.2.) You are one of 100 passengers waiting for your bag at an airport luggage carousel. Your bag looks the same as 5% of all bags. Derive a general expression for the probability that the bag you are viewing (which matches your bag visually) is your own, as a function of the number of bags you have viewed so far. How many bags must you view (without finding your own) before the posterior probability that the bag you are viewing (which matches your own visually) is greater than 70%?

9. (See Section 1.2.) In the luggage carousel example, we defined the visual observation as the shape, size, and color of the bag seen, and we therefore took p(observation | H_1) to equal 1 when the observation matched the shape, size, and color of your bag. But the exact "look" of the bag on the luggage carousel involves more than just its shape, size, and color. For instance, as the bag enters the carousel, it may come to rest at any one of many different orientations. Now suppose we were to redefine the "observation" to be shape, size, color and *orientation* of the bag seen. To keep things simple, let's assume that there are 360 possible angles (one for each degree around the circle) and two possible sides (right-side up or upside down), for a total of 720 possible orientations with which a bag may come to rest on the carousel. If we further assume that each orientation is equally likely, then the probability of the observation given hypothesis 1 is no longer one, but rather 1/720. Similarly, the probably of the observation given hypothesis 2 would no longer be 0.05, but rather 0.05/720. Since the likelihoods have changed, must not the posterior distribution change as well? Explain.

10. Suppose you are waiting to catch a particular bus in a city that has just 10 bus routes; the route followed by each bus is indicated by an integer in the corner of its front display. You see the bus below from a distance, and naturally wonder whether this is the bus you are waiting for. A) Based on the visual image of the difficult-to-discern bus route number (see arrow), and your intuitive understanding of how different numbers might appear, construct a likelihood function that plots p(visual observation | hypothesized bus route), for all numbers from 1 to 10. B) As it turns out, you happen to know that only buses 3, 4, 5, and 6 travel down the street you are on. Furthermore, you know that buses 3 and 4 come twice as frequently as buses 5 and 6. Based on this background knowledge, construct your prior distribution for the bus number. C) Use Bayes' rule to calculate your posterior distribution for the number of the bus.

11. Rephrase in terms of Bayesian perceptual inference the following statement written by Ibn Alhacen approximately 1,000 years ago: "...when sight perceives a rose-red color among the flowers in some garden, it will immediately perceive that the things in which that color inheres are roses because that color is specific to roses...But this does not happen when sight perceives a myrtle-green color in the garden. For when sight perceives only the myrtle-green in the garden, it will not perceive the myrtle-green to be myrtle simply from the perception of the green, because several plants are green, and, in addition, several plants resemble myrtle in greenness and shape." (*De aspectibus, book 2*, as translated by Smith, 2001).

12. Why is it that we identify ourselves at the very beginning of a phone conversation, even to people we already know, but we do not do this when we meet in person? Express your answer within the framework of Bayesian perceptual inference.

13. When a conversation companion speaks softly, or when a conversation occurs in the presence of significant ambient noise, we sometimes cup our ears and/or look carefully at the speaker's lips. Why, in Bayesian perceptual terms, do we do this?

14. To explore how a noisy environment engenders uncertainty, consider the word "lunch." Suppose that you see this word written (or hear it spoken), with the letter "l" blocked out:_unch (e.g., by ambient auditory noise). List all source words that are compatible with what you see. Now repeat, but with the letter "n" blocked: lu_ch. Finally, consider the case in which both the l and the n are blocked: _u_ch. In terms of conditional probabilities relevant to perception, what is the effect of blocking out the l, n, and both?

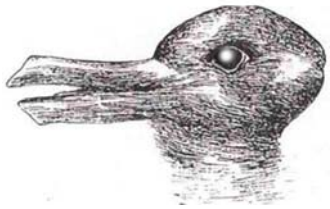
15. The NATO phonetic alphabet, used by many military, maritime, and other organizations during radio communications, represents each letter with a word: A (Alpha), B (Bravo), C (Charlie), D (Delta), E (Echo), F (Foxtrot), and so on. What purpose does this serve in radio communications? Explain with respect to conditional probabilities. In particular,

consider a radio communication under conditions of considerable background noise, in which the sender wishes to spell the word "FACE." Compare *p*(auditory signal heard by the receiver | FACE spelled by the sender) vs. *p*(auditory signal heard by the receiver | another word, such as FADE, spelled by the sender), when the sender uses the regular alphabet, and again when the sender uses the NATO phonetic alphabet.

16. English speakers sometimes incorrectly perceive English words when they listen to songs sung in a foreign language with which they are unfamiliar, and listeners also mistakenly perceive words in music that is played backwards. Provide a Bayesian explanation for these phenomena.

17. Suppose you see someone you do not know, getting only a brief look at him from a distance of about 10 meters. If your interest is in estimating this person's age, how would you proceed? What factors, including and in addition to the person's appearance, would affect your estimation? Provide a Bayesian description of your reasoning. As part of your answer, draw examples of your likelihood function, prior distribution, and resulting posterior distribution.

18. A research article entitled "The Easter bunny in October: Is it disguised as a duck?" explained that "Very little is known about the looks of the Easter bunny on his non-working days." To investigate, the authors showed an "ambiguous drawing of a duck/rabbit…to…265 subjects on Easter Sunday and to 276 different subjects on a Sunday in October of the same year." The authors report that "Whereas on Easter the drawing was significantly more often recognized as a bunny, in October it was considered a bird by most subjects." The drawing shown by the authors in their study was similar to the following:



Provide a Bayesian perceptual explanation for the authors' results.

19. The images below show a Charlie Chaplin face mask. The left image is a side viewing revealing that the mask is hollow. The middle image is a front view. The right image is a back view of the hollow side of the mask:

Provide a Bayesian explanation for why the right image looks like a normal, convex face, when in reality it is the hollow (concave) side of the mask (images from www.richardgregory.org).