The Anatomy of Bias

How Neural Circuits Weigh the Options

Jan Lauwereyns

The MIT Press
Cambridge, Massachusetts
London, England

This book was set in Syntax and Times by SNP Best-set Typesetter Ltd., Hong Kong. Printed and bound in the United States of America.

Library of Congress Cataloging-in-Publication Data

Lauwereyns, Jan, 1969–
The anatomy of bias : how neural circuits weigh the options/Jan Lauwereyns.
p. cm.
Includes bibliographical references and index.
ISBN 978-0-262-12310-5 (hardcover : alk. paper) 1. Neural circuitry—Mathematical models. 2. Decision making—Physiological aspects. 3. Bayesian statistical decision theory. 4. Neural networks (Neurobiology)—Mathematical models. I. Title.
QP363.3.L38 2010
153.8'3–dc22

2009021061

10 9 8 7 6 5 4 3 2 1

1 Bayes and Bias

Eyes talked in-
to blindness.
Their—"a
riddle, what is pure-
ly arisen"—, their
memory of
floating Hölderlintowers, gull-
enswirled.

The words of the poet serve as the oracle, the omen or portent of things to come. The strange formulations create a peculiar anticipatory image that colors, biases, or otherwise influences the perception of objects and events. Here, the words are from the poem "Tübingen, January" by the great German-language poet Paul Celan (2001, p. 159), whose work was inextricably linked with the trauma of the Holocaust and achieved what Theodor Adorno had initially thought to be impossible—to make (some kind of) sense after Auschwitz. Of course, it is not easy to explain exactly how the notoriously difficult poetry of Paul Celan *makes sense*, but I think I will try to do just that, be it obliquely, and not right away.

What Good Is a Mystery?

Noisy or nebulous forms of communication will normally receive little sympathy from scientists and other lucid minds who wish to obtain clear information about the world around them. Yet, such clarity must itself be seen as a poetic construction, or an object of wishful thinking, given the level of complexity and the number of stochastic processes involved in our world. Sometimes the best way to approach things in words is simply to start speaking, mark the perimeter, and try to systematically move closer. Even Ludwig Wittgenstein, in his *Tractatus Logico-Philosophicus*, appeared to do

something of the sort, following his ominous preface with the stern dictum (Wittgenstein, 2001, p. 3) about when to speak (only when you can do so clearly) and when to be silent (in all other cases, i.e., forever?).

Wittgenstein happily ignored his own dictum with arguably the most mysterious, opaque, and poorly understood philosophical treatise of the twentieth century. He even admitted in the first sentence of the very same preface (p. 3) that "[p]erhaps this book will be understood only by someone who has himself already had the thoughts." The question remains whether anyone has, but that does not diminish the beauty and the attraction of the book, much like the incomprehensibility of Celan's poetry does not stand in the way of the pleasure of reading.

If it was Wittgenstein's intention to write the *Tractatus* in accordance with his dictum about speaking clearly, then his concept of clarity might have more to do with abundance of light and sharpness of contours than with the amount of activity in Wernicke's area, the brain's putative center for linguistic comprehension. This kind of speech, then, would truly represent objects or trains of thought, in a crystallized form, which we can marvel at, preserve, repeat, and return to, time and again. Is it only me, or does this sound like a definition of poetry?

The thing expressed in words would be given a special place, in the spotlight, clearly visible for anyone to see, like a strange prehistoric artifact in the British Museum. The visitor might have no clue what the artifact was used for or what it represented for whoever made it, but that bit of mystery is certainly compatible with a sense of enjoyment in contemplation. Indeed, associations with the abundance of light and the lack of short and easy answers bring us to some of the loveliest entries in the dictionary (e.g., "brilliant," "radiant," "amazing," and "wonderful"). With Gertrude Stein we can further offer the proposition that "if you enjoy it you understand it" (Stein, 2008, p. 10). Turning our attention to the scientific enterprise, we might point out that without an initial mystery there would be no hope for a happy end in the form of a correct theoretical model. Without the excitement of weird questions there would be no thrilling search for answers.

So I will charge ahead, and insert a few lines from some of my favorite poems in a scientific discourse, with the double purpose of prickling the senses and showing how the mere fact of having the senses prickled shapes the subsequent search. The borrowed elements of poetry might generate curiosity, arouse the mind, and heighten the acuity of cognitive processing. The occasional verse would serve as a target or template to guide the conceptual search, exerting top–down control over the way in which incoming signals (the actual scientific data) are selected, integrated, and abstracted. If my plan works, it

will serve its purpose as a rhetorical device, echoing the object of study. If it does not, the reader will still have read a few lines of poetry, and that will never bring sadness in my book.

But what, the verse-phobic scientist might wish to know, is the real purpose of your average oracle? In Greek mythology the hero or heroine who seeks advice, and instead receives only garbled instructions, invariably fails to benefit from the trip to Delphi. Worse, it often seems as though the very content of the oracle leads to the protagonist's doom, as when Oedipus, abandoned at birth, ends up killing a stranger and marrying the widow (his true father and mother, respectively). As a narrative technique, though, it works beautifully; it determines the architecture of the story, creating an expectation, opening up a mental slot that needs to be filled. The reader or listener knows what is coming and will not consider the story complete until Oedipus realizes the full horror of his fate.

Making a giant leap, a wild generalization, the complex structure with a prime (an oracle) and a response (the actual fate) looks like a prototypical example of the kind of thought process that some of this planet's most famous researchers (Fitch & Hauser, 2004; Hauser, Chomsky, & Fitch, 2002) have argued to be characteristically human, involving more than mere concatenation of elements or phrases, but hierarchical relations and long-distance dependency, as in an IF . . . —THEN . . . structure, in which you can keep embedding other conditions, in principle ad infinitum, but in practice only until your hard drive crashes or your brain forgets what you were talking about. (If, when reading the previous sentence, you get lost, then either I am a bad writer or you may not have what it takes to be human.)

In plain English: the oracle sets up an expectation that helps the reader or listener to interpret the story. Expectations that guide interpretation are an extremely powerful cognitive tool and may be the secret of our species' success. They imply the usage of some, however rudimentary or sophisticated, computational model that generates expectations on the basis of prior information. The process of interpretation can then provide feedback. With new information coming in, expectations are confirmed or disconfirmed, proving or disproving the validity of the computational model. Now, my language must start sounding familiar to readers versed in probabilistic approaches to the study of brain function (for excellent introductions to these, see Doya et al., 2007, and Rao, Olshausen, & Lewicki, 2002).

Expectations that guide interpretation are so basic as to pervade every domain of human thought. At the very heart of science there is the idea, perhaps most elaborately formulated by Popper (2002a), that there should be a kind of logic that underscores discovery, working from theory to hypothesis

to prediction, or from general to particular, until we arrive at a well-defined forecast, a concrete expectation, which can be tested empirically. If the prediction comes out, things remain as they are, and we may be happy, but we don't really learn anything; if the prediction proves to be false, we have to go back to the drawing board.

Though this logic for learning is the hallmark of science, similar cognitive sequences operate more implicitly in all our daily activities from choosing where to have lunch to deciding when would be the best time to wade through our e-mail. We constantly juggle beliefs and expectations—about which restaurant serves what quality of food and gets exactly how crowded or about when we can treat fifteen minutes as spare time because we can't really use it meaningfully in any other way. We may develop routines, have preferred restaurants, and tend to check e-mail when we arrive at the lab in the morning, suffering from a brief bout of cognitive inertia after an intense (mostly uphill) twenty-two-minute mountain bike ride. And sometimes we update our list of preferred restaurants, when, during the last visit to our former top choice, the waiter was rude, and the salad, for which we waited a full seventeen minutes, had a fruit fly in it. Or we stop checking e-mail in the morning, as it makes the cognitive inertia even worse, and instead we look briefly at the BBC News Web site, which bores us soon enough, so that we can move on to the real work in the lab with a properly blank and ready mind.

The principal advantage of using mental models that make predictions about the world must somehow resonate with Sir Francis Bacon's "*sciencia potentia est*"—knowledge equaling power—or, more fully, with aphorism III from the first book of *The New Organon*, first published in 1620 (Bacon, 2000, p. 33):

> Human knowledge and human power come to the same thing, because ignorance of cause frustrates effect. For nature is conquered only by obedience; and that which in thought is a cause, is like a rule in practice.

Bacon might have been borrowing from the bible's *Proverbs*, particularly 24 : 5. In the Authorized King James Version, from around the time Bacon was writing in Latin, it reads, "A wise man is strong; yea, a man of knowledge increaseth strength" (Anonymous, 2005). Biblical or not, there does seem to be a deep connection between knowing what will happen when and where, and the opportunity, if not the ability, to do something about it. Individuals, groups, and companies, and occasionally even the United Nations, benefit from devising strategies and aiming for fast and effective action that yields something good or prevents something bad.

It may be wise to be wise, even if this proposition looks suspiciously circular or in danger of infinite regress, with an endless series of "Why?" questions and a relentlessly growing sequence of embedded wisdoms ("Because it is wise to be wise to be wise to be…"). That it is better to know than not to know has generally been taken for granted ever since, or despite, the fall from Paradise (in the parlance of an Abrahamic religion). Clear though its benefits are, the process of knowing itself defies explanation, as if there rests a taboo on understanding our understanding. We may roughly say that it has something to do with getting access to fundamental laws of nature, but how we actually achieve this, and whether (or which of) those laws really exist, remains a matter of debate among the brightest minds of our species, and some of these are very skeptical about whether we will ever understand our understanding (one of the most forceful arguments being that by Penrose, 1989, with a wonderful twist on Gödel's theorem).

One aspect of knowing that has intrigued me personally is how personal the act of knowing often seems to be. Knowing feels more like a feeling than the rational processing of an undeniable truth. Along with understanding and remembering, knowing may have a component of belief in it or an acceptance that some mental images are good enough without questioning them any further, regardless of how they relate to any kind of actual scene, present or past. If I am right, we would tend to be pragmatic and rather minimalist when it comes to the amount of computational power we employ for any particular act of thinking, knowing, believing, and so forth. I see the human mind as a minimalist theorist, or a lazy thinker. By default, the mind would choose the theory of least resistance, or the cheapest concept (hence the early emergence of gods in every known human culture, easy theories as they are, perfectly to blame for everything). Rethinking and changes to the mental model of the world are inspired mainly by adverse runs when things do not go as expected. Simple models gradually get replaced by more complex models as a function of the amount of stress experienced. The level of accuracy required is dictated by the performance of the mental model; if all goes well with a simple but inaccurate model (e.g., Newtonian physics), we might as well keep relying on it.

The central point of this little excursion is that the use of knowledge is subjective, situational, dependent on the actual context in which an individual, a group, a company, or occasionally even the United Nations finds itself. "*Cogito ergo sum*," I would like to repeat after Descartes, but with emphasis on the subjectivity of being, pointing to the fact that both the phrase "I think" and the phrase "I am" have something in the subject slot, even if it is only

implied in Latin. With subjectivity comes perspective and limitation in time and in place. To deny the inherent subjectivity is to fall prey to the potentially damaging effects of distortion and bias. Only in the explicit acknowledgement of our own subjectivity, in the willingness to converse with and learn from people who have other viewpoints, can we hope to reach a more objective stance in which all the various idiosyncrasies in thought and feeling are given their due.

The counsel may sound obvious enough, or even slightly naive in its unchecked idealism. Yet how come the obviousness does not obviously translate to practical application? "Seeing is believing," the saying goes, as if the simple act of seeing were the best method of knowledge acquisition, in full denial of its limited validity as a truth procedure. When do we really perceive things, and when do we simply take to be true what fits our understanding of the way things are supposed to go? The process of apprehension with the mind is certainly one in which the quality of data analysis is too often overestimated. Here, I would like to turn to the opening words by Paul Celan. "Eyes" are "talked into blindness," he observed in his characteristically sparing use of words. It sounds disparaging, like a complaint addressed to those who see, but don't really *see*, believing the word (the prediction) rather than what is actually there to see.

How do we move on from the dilemma? If the acts of the mind are inherently subjective, in the sense that they are shaped by the semantic system of an individual, then how can we acknowledge this and work toward a truly objective view of things? Clearly, it will not be helpful to exaggerate the role of the subject. There is little be gained from a caricature view, be it postmodern, constructionist, or simply absurd, according to which there is no such thing as the truth or a knowable actual state of things out there. Somewhere in the middle between the denial of subjectivity and the refutation of reality, we can walk with Celan's poem, in which the "eyes talked into blindness" desire to resolve the riddle of "what is purely arisen." The eyes seem to be aware of their fallibility, and they wish to separate the wheat from the chaff. The peculiarities of our visions, "of floating Hölderlintowers, gull-enswirled," are too concrete and too arbitrary to be mere figments of the imagination. Somehow the mind must have bundled together bits and pieces of reality into a proposition about, or image of, some aspect of the world. The task is to understand how the subject's being in time and space restricts and sometimes skews the information available for processing.

For Paul Celan, this task was all the more urgent as he struggled with his chronic mental illness, a bipolar disorder of a rather malignant nature, with several violent outbursts over the years. Did he see things truly, or was he

crazy? The question would have haunted his mind. The end, unfortunately, was tragic; the poet jumped to his death in the river Seine in April 1970 (he had lived most of his adult life as an expat in Paris). Even in the early 1960s, when Celan wrote "Tübingen, January," he must have fully realized what was happening to him; his reference to Friedrich Hölderlin can count as exhibit A (Hölderlin is the archetypal "mad poet," who spent half of his sad life as a recluse in the attic of a mill, not too far from the center of the idyllic and well-preserved town in southern Germany).

The poem only gets starker in the remainder, not quoted here for fear that I would get stuck for another ten pages or so. I will simply note that Celan talks about drowning, the plunging of words, and how a visionary would only babble incomprehensibly "if he spoke of this time" (Celan, 2001, p. 159). It seems entirely possible to me that Celan, in a bout of obsessive–compulsive ideation, ended up believing in the necessary truth of the poem's prediction. If so, this would make it arguably one of the most miserable poems in recent history, as it hung for almost a decade over the patient's head like the sword of Damocles, until Celan finally gave in and committed suicide.

The Role of the Prior

A sword above your head, dangling from a single hair, certainly should give you a vivid impression of imminent danger, and in the case of Damocles it effectively ruined his appetite, as if the sense of disaster in the making (even though the hair in question was good, strong horsehair) prevented his gustatory system from processing the riches that Dionysius II of Syracuse, a proper tyrant, had so generously offered him a taste of. Here, the anticipatory image did more than simply guide the interpretation; it fully dominated the experience and changed the course of action—as soon as Damocles noticed the lethal weapon on a virtual course to pierce his skull, he quickly asked his master if he could be excused from the table.

The legend can be read as a beautiful little parable of the interdependent dynamics of bias, sensitivity, and decision making, or how the fear for one thing dampens the perception of another and elicits an adaptive response, an instance of operant avoidance behavior. At some point in time, though, we would wish to move from parable to paradigm and develop ways to study these dynamics systematically in the formal language of science.

As it turns out, history has already shown that the influences of expectations and subjectivity in perception and cognition can indeed be studied to a surprisingly detailed degree thanks to computational tools based on probability theory. Statistics and the various techniques of likelihood estimation form

the perfect platform for the investigation of how we perceive things, think about them, and make decisions. Of course, statistics is crucial to most scientific endeavors, as it formalizes the processing of empirical data, but for the cognitive sciences, from psychology to computational neuroscience, and from neurophysiology to artificial intelligence, statistics takes an even more central position, being a method as well as a metaphor. The myriad acts of mind and events in the brain are all about processing data, and it may not be a crazy thought to think that what happens inside our skull is itself governed by a kind of applied statistics, with data archived in frequency distributions, and hypotheses accepted or rejected on the basis of the available evidence.

The point of departure in statistics and probability theory must be, for now and forever, the work of Thomas Bayes, and particularly his theorem—or rule, if you prefer—about how the likelihood of a particular something is weighted by its prior probability. The prior probability, or simply "the prior," is where subjectivity comes in, where knowledge, expectation, and beliefs can play their part. Bayes's theorem puts the role of the prior firmly in a formula, and its ramifications are the core focus of current psychophysical and neural models of decision making. However, before I turn to these, I would like to taste a sample of the original (Bayes, 1763, p. 4): "If a person has an expectation depending on the happening of an event, the probability of the event is to the probability of its failure as his loss if it fails to his gain if it happens." These are the posthumous words of an eighteenth-century Presbyterian priest, but they sound almost contemporary and are likely to make some kind of sense even to readers who are generally predisposed to get tired quickly from all things mathematical. Here, Bayes makes the straightforward proposal that the numerical data of microeconomics, in terms of likelihood of gain or loss, directly follow the statistics of the real world, in terms of likelihood of events happening or not. The statement may appear somewhat trivial, achieving nothing more than a mere duplication, creating a double, or a new representation of the original. On second thought, however, we may recognize the mechanism as one that enables the creation of a virtual model, a little "toy model of the universe," which would map the physical onto the mathematical, or the ways of the world onto circuits of the brain. Suddenly, this duplication project looks anything but trivial, rather impossibly difficult, a rational ideal. In between the lines we might read a task for the scientist, in comparing how an individual's virtual model deviates from the rational ideal. Any systematic deviation could reveal something peculiar about what it is like to be human, or what it is like to be a particular human at a particular set of coordinates in space-time. Invigorated by the undeniable ambition of this scientific project, we may wish to sample some more from the original Bayes, first in print in 1763 (p. 5):

> Suppose a person has an expectation of receiving N, depending on an event the probability of which is P/N. Then (by definition 5) the value of his expectation is P, and therefore if the event fail, he loses that which in value is P; and if it happens he receives N, but his expectation ceases. His gain is therefore N – P. Likewise since the probability of the event is P/N, that of its failure (by corollary prop. 1) is (N – P)/N. But (N – P)/N is to P/N as P is to N – P, i.e. the probability of the event is to the probability of it's [sic] failure, as his loss if it fails to his gain if it happens.

One thing I know for sure is that my own little virtual model cannot cope with this language, and so this is perhaps a good place to admit that I am one of those readers who are generally predisposed to get tired quickly from all things mathematical. My Great Step, or the Problematic Idea of This Book, of necessity will rely only on the most rudimentary type of equations and formulas—the type that are easy enough to capture in words or visual schematic representations. Where relevant, I will add references to the "real deal," papers and monographs with hard-core stuff aplenty for the reader with an insatiable appetite. In the meantime, I am afraid that difficult passages such as the one by Bayes above tend to literally drive me to distraction. I get derailed by surface features; here, I note the odd spelling, "its failure" and "it's failure," inconsistent, incorrect, and sloppy, and so why should I trust the incomprehensible argument?

Even the true masters of statistics, including such luminaries as Ronald Fisher and Karl Pearson, had trouble understanding the ramifications of the original proposal, suggested Stephen Stigler (1982), and the very same skeptic went on to dispute the conventional view that it was really Bayes's proposal in the first place. In a delightful little article for *The American Statistician*, Stigler (1983) upheld his Law of Eponymy, claiming that no discovery or invention is named after whoever really did the work; instead the first person who fails to give proper due would tend to scoop the honor. The piece reads like a true whodunit, with a plausible unsung hero emerging—Nicholas Saunderson, the famous and incredibly talented blind professor of mathematics at Cambridge. This also obliquely raises the question of who really discovered the Law of Eponymy (or who it was that Stigler failed to credit), and to avoid similar misconduct on my part, I will quickly admit having first read about Stigler's doubts in Gerd Gigerenzer's (2002) equally delightful *Reckoning with Risk*.

Back from the past, with our feet firmly on the ground, we will do wise to concentrate on the common understanding of Bayes's theorem in our time. Intuitively, the theorem simply says that likelihood is weighted by prior probability. The end result is a posterior probability, something like our best guess, given all the evidence. "The evidence," then, consists of an actual observation,

which constrains the likelihood of a particular hypothesis, in combination with a constant (or normalizing denominator) and the a priori likelihood of the hypothesis. Perhaps a concrete example is in order.

Consider your twelve-year-old daughter (or, if you know of no such instance, consider the twelve-year-old daughter of your neighbor, or of your neighbor's neighbor, or . . . ; the recursive process should be continued until you can think of a proper instance in your neighborhood). Have she and a boy in her class been kissing? Bayes's theorem can tell you how likely it is that this has occurred, given that she blushes (this is the posterior probability), on the basis of three sources of information: the generative model, the prior probability, and the marginal probability. The generative model says how likely an observation is (that she blushes), given that a particular hypothesis is true (that she and a boy in her class have been kissing). Note that this conditional probability, P(Blushing|Kissed), is the exact mirror image of the posterior probability, P(Kissed|Blushing). The prior probability gives the base rate of how likely it is that the (your) daughter and the boy in her class have been kissing, P(Kissed), whereas the marginal probability says generally how likely she is to blush, P(Blushing).

Obviously, if the twelve-year-old girl in question tends to blush quite often in general, but not necessarily when she has been kissing a boy in her class, then blushing tells you much less than if she rarely blushes, yet is sure to blush when she has been kissing a boy in her class. Bayes's theorem captures this mutual dependency of different types of probability, stating that P(Kissed|Blushing) equals the product of P(Blushing|Kissed) and P(Kissed) divided by P(Blushing). Let's say that that twelve-year-old daughter (of yours) generally tends to kiss a boy in her class with a likelihood of about one in twenty, [P(Kissed) = 0.05]. Now if she tends to blushes quite often, [P(Blushing) = 0.4], but not necessarily when she has been kissing a boy in her class, [P(Blushing|Kissed) = 0.7], the posterior probability equals 0.7 times 0.05 (i.e., 0.035), divided by 0.4, or precisely 0.0875. She happens to be blushing? This does not allow you to conclude that she has been kissing a boy in her class. The probability that she blushes, given a kiss of that kind, is less than one in ten. But if she rarely blushes, [P(Blushing) = 0.1], yet is very likely to blush when she has been kissing a boy in her class [P(Blushing|Kissed) = 0.9], the posterior probability equals 0.9 times 0.05 (i.e., 0.045), divided by 0.1, or precisely 0.45. She happens to be blushing? Chances are close to one in two that she has been kissing a boy in her class.

How can we be sure the numbers are correct? Do we simply take the theorem for granted, or can we work out some kind of proof in our own math-phobic way? I was one of the worst performers in math during high school,

but I am always willing to try something without too much effort or too many fancy tricks. First, looking at the theorem, we see just four terms, two of which are each other's counterparts. On the left of the equation we have the posterior probability, P(X|Y), and on the right we have its counterpart, P(Y|X), combined with two other terms, P(X) and P(Y). To phrase it exactly: P(X|Y) equals a fraction, which consists of a numerator determined by the product of P(Y|X) and P(X), and a denominator given by P(Y). Let us call this proposition 1, the theorem of Bayes. As a formula it should not look too ominous; all we need to do is multiply one thing by another and then divide the result by something else. That twelve-year-old daughter (of yours) can do it, so you can too.

Of course, the X and Y are just abstract placeholders, so we could easily rephrase proposition 1 by swapping them around, putting an X wherever we find a Y, and vice versa. Proposition 2 then reads as follows: P(Y|X) equals a fraction, which consists of a numerator determined by the product of P(X|Y) and P(Y) and a denominator given by P(X). Now, the funny thing is that we can plug proposition 2 into proposition 1, to remove one of the four terms and replace it with a new combination of the remaining three, each of which is now used twice in the formula left standing. Thus, we do a Bayes on Bayes, apply the rule in the rule, or rely a little on our beloved mechanism of recursion in the hope of proving the whole.

Having done the deed, we have this: P(X|Y) equals a fraction, which consists of a numerator determined by the product of *the right side of the equation in proposition 2* and P(X) and a denominator given by P(Y). This actually puts a fraction inside a fraction, making the formula look scarier if you use conventional notation than if you say it in words. Anyway, to simplify things, we can use the multiplication principle to carry P(X) and P(Y) to the left of the equation, leaving only *the right side of the equation in proposition 2* on the right. As for the left side, P(X) will end up being the denominator, whereas P(Y) will join up with P(X|Y) to form the product that defines the numerator. The multiplication principle and the rules about carrying terms across to the other side of the equation were properly etched in my memory. I will take the liberty of assuming that a similar kind of etching would have occurred for anyone picking up a book like the one before us.

Thus, we have a fraction on either side of the equation. On the left we have P(X|Y) times P(Y), to be divided by P(X). And on the right we have the right side of the equation in proposition 2, which claimed that P(X|Y) times P(Y) should be the numerator, and P(X) the denominator. The left says exactly the same as the right, quod erat demonstrandum!

Or not? Did I create nothing more than a tautology, showing that the rule is true if the rule is true? What would happen if you swap P(X) and P(Y) in

proposition 1 and then apply my recursive trick? Again the tautology works, saying that the new rule is true if the new rule is true, which it probably is not.

Well, at least we have acquired some practice in playing with the terms, and other than that, I would like to draw three conclusions from this little exercise: (1) It is fun to think for ourselves, we should do it more often; (2) recursion is not the answer to everything; and (3) we had better move on and stick with the original motto, saying this is not a book of mathematics. An unflappable skeptic might point out that conclusions 1 and 3 are somewhat contradictory, but I guess in this matter the limits of time and the reader's patience should prevail.

The proper proof of Bayes's theorem is actually quite straightforward, I must admit, though I did not manage to come up with it myself (for a more serious primer, see Doya & Ishii, 2007). All we need is a little detour via the definition of conditional probability, or the probability that one thing is true given that another is true, P(X|Y), that is, the type of term we are already familiar with from Bayes's theorem. The definition of conditional probability is based on another, in fact more basic, concept: joint probability, or the probability that two things are both true at once, P(X is true *and* Y is true), sometimes notated as P(X∩Y).

The definition of conditional probability, then, states that the probability of X, given Y, equals the joint probability of X and Y divided by the general probability of Y. That is, P(X|Y) = P(X∩Y)/P(Y). This naturally makes sense if you think about it. We might want to follow Gigerenzer's (2002) advice, and reason in natural frequencies for a minute. Take all the cases in which the twelve-year-old girl had been kissing a boy in her class; in how many of those cases did she also blush? This number, how many times blushing and kissing out of how many times kissing in general, basically gives you the desired conditional probability.

A good statistician should like to play around with any definition. So did Bayes, one of the very first of that species of human. If it is true that P(X|Y) = P(X∩Y)/P(Y), then we can also move P(Y) to the other side of the equation to give us P(X∩Y) = P(X|Y) P(Y). Defining the opposite conditional probability, P(Y|X), we get P(Y|X) = P(X∩Y)/P(X), which we can rewrite as P(X∩Y) = P(Y|X)P(X). So we are basically rewriting P(X∩Y) in two ways:

P(X|Y) P(Y) = P(X∩Y) = P(Y|X) P(X).

Now we can get rid of P(X∩Y), saying P(X|Y)P(Y) = P(Y|X)P(X). To get his theorem, Bayes then only had to move P(Y) to the right of the equation: P(X|Y) = P(Y|X)P(X)/P(Y). Quod erat demonstrandum! (For real, this time.)

By now I have spent much more time talking about, and circling around, Bayes's theorem than would have been needed to simply restate it in textbook format. Even if you were not schooled as a neuroscientist, you will know it by heart. Hopefully, you will also have developed some intuition about decision making as an inherently statistical enterprise, even if we shy away from numbers and formulas when weighing options and taking things to be true or not as we go about our business and do our mundane doings in daily life. However, Bayes's theorem says something more specific than that.

The main point to take away from the Bayesian way of looking at the world is this: Our beliefs about the world should be updated by combining new evidence with what we believed before, "the prior." The role of the prior is to color, or to help interpret, new information. Does blushing mean that the twelve-year-old girl has been kissing a boy in her class? Our prior beliefs about her kissing behavior and blushing tendencies will help us draw better conclusions than we would reach if we were to consider only the current evidence. Decision making stands to benefit from keeping track of how often things happen in the real world. Even the most rudimentary records and nonparametric statistics (relying only on rank ordering as in "This happens more often than that") are likely to improve perception, categorization, and all the more complex forms of cognitive processing to define things as they are. The study of decision making translates into the problem of uncovering exactly how an individual, a group, a company, or occasionally even the United Nations makes use of different kinds of information about the likelihood of things and events. The role of the prior, here, determines the potentially idiosyncratic characteristics of how decisions might tend to go one way rather than another. The prior determines bias.

I have dropped the Heavy Word. Bias, what does it mean exactly? According to *A Dictionary of Statistical Terms* by Kendall and Buckland (1957, p. 26), it is as follows:

> Generally, an effect which deprives a statistical result of representativeness by systematically distorting it, as distinct from a random error which may distort on any one occasion but balances out on the average.

The statement is a bit terse, as we might expect from an old-school statistical dictionary (picked up practically for free at a sale of unwanted books at the City Library of Wellington), but the gist is clear. Bias implies something systematic, nonrandom, which pulls the statistical result, or decision, in a particular direction, away from the neutral. The loss of neutrality, or the introduction of subjectivity, carries a load of negative connotations in daily life as reflected in the lemma for bias in *The Pocket Oxford Dictionary of Current English* (Thompson, 1996, p. 75):

> n. 1 (often foll. by towards, against) predisposition or prejudice. 2 Statistics distortion of a statistical result due to a neglected factor. 3 edge cut obliquely across the weave of a fabric. 4 Sport *a* irregular shape given to a bowl. *b* oblique course this causes it to run.—v. (-s- or -ss-) 1 (esp. as biased adj.) influence (usu. unfairly); prejudice. 2 give a bias to.

Except in weaving or bowling, bias is associated with a number of known villains: prejudice, distortion, and the adverb for lack of fairness. It is only one step shy of racism, sexism, nepotism—a host of—isms that good citizens, if not good politicians, would like to stay away from. Whether I am a good citizen I will leave for others to judge, though on the topic of *The Pocket Oxford Dictionary of Current English* I should note that the copy inexplicably found its way to my personal library (it had belonged to the first author of Shimo & Hikosaka, 2001, a rogue reference offered in compensation for the excessively long duration of borrowing the pocket dictionary, which, by the way, even if I pocketed it, fits in no pocket of mine). Good citizen or not, I think the word "bias" did not get a fair shake. Obviously, the unfair treatment of others represents an ugly disease in human society, one that we should prevent and remediate in any way possible, but to simply equate bias with something bad may be throwing the baby out with the bathwater.

Bias is part and parcel of Bayesian reasoning, emphasizing the crucial role of the prior in the assessment of probabilities, representing beliefs about how one thing might be more likely than another. The prior is exactly the term that modelers of Bayesian inference in perception employ to characterize observer biases (e.g., Mamassian, 2006; Mamassian & Landy, 1998). Such biases are perfectly rational if they correspond with the statistical regularities of the environment. Perhaps we should dust our vocabulary and heed once more the words of our favorite fourteenth-century Buddhist priest from Kyoto (Kenkō, 2001, p. 13):

> The same words and subjects that might still be employed today meant something quite different when employed by the poets of ancient times. Their poems are simple and unaffected, and the lovely purity of the form creates a powerful impression.

The word "bias" deserves to be exonerated, polished, and used properly. Instead of dismissing bias as a form of evil, I propose we should acknowledge bias as a fundamental property of human thinking, perceiving, and decision making. If we are to eradicate the social crime of prejudice, we should establish whether and how prejudice derives from bias and whether and how bias derives from statistical regularities in the world. The derivation of prejudice would represent the real evil, the one we would wish to redress. However, perhaps the best way to do so is not by denial but by explicit formulation of

bias and the extent to which it is rational. Put differently, we need to learn to distinguish "good bias" from "bad bias." In the meantime, the best way to start the enterprise is by considering the basic role of bias in decision making. For this we need to wrestle with probability distributions.

Before doing so, I would like to pay one final tribute to the lovely and pure, if slightly incomprehensible, ancient words of Thomas Bayes by reciting my all-time favorite title for a scientific monograph (one that Bayes published anonymously in 1736): *An Introduction to the Doctrine of Fluxions, and a Defence of the Mathematicians Against the Objections of the Author of The Analyst.* Bayes's defense was for Newton against Berkeley (the author of *The Analyst*), and though I searched hard for it on the Internet, I found no free online version. I can only imagine what Bayes's argument sounds like, but the word "fluxions" appeals to me, and the defense of mathematics against analysts somehow rings a bell for me, with contemporary neuroscience, and its computational approaches, on the defense against present-day psychoanalysis, represented by post-Lacanian thinkers such as Slavoj Žižek—a battle that may be outside the field of vision for many neuroscientists but is nowhere near dying down in cultural studies, including literary theory, a field that interests me for idiosyncratic reasons. So I, too, wish to introduce fluxions, but then fluxions in the form of movement between computation and metaphorical intuition, not to have one pushing the other out of the ring but to get the best of both worlds if that is at all possible—hence the poetry as well as the wrestling with distributions.

Wrestling with Distributions

One of the first things I learned, to my dismay, when I started collecting real, heavy-duty, hard-core scientific data was how massively variable they were. In my maiden project, a partly tongue-in-cheek exploration of the eye movements of a poetry critic, I was not particularly worried about that, thinking it was a crazy project anyway (Lauwereyns & d'Ydewalle, 1996). But when I then started collecting data for my PhD thesis in a very classic visual search paradigm, I was positively baffled by the fact that so simple a task as pressing a button when you find a letter Q among distractor letters O on the computer screen could lead to such vastly different response times, with some of my victims (first-year students in psychology) taking forever to find the target (more than a second, occasionally even two) and others finding it right away (in less than half a second). Even for the same participant, the data often looked very messy, with response times all over the place, sometimes 300 milliseconds, sometimes 800.

Apparently, people were unable to exactly replicate what they did, though I asked them to do the same thing for hundreds of trials in a single session of less than an hour, keeping all factors constant as best I could: the same participant, the same apparatus, the same task, the same events, the same time of day . . . It dawned on me, slowly, that concepts such as the variability and the standard deviation were crucial to the science I found myself in. My data looked messy, but no messier than those from other laboratories—I was still able to draw publishable conclusions (Lauwereyns & d'Ydewalle, 1997). Investigating the mechanisms of visual perception and decision making involved wrestling with distributions; there was no escaping it. I remembered a remark made in one of my undergraduate classes about how someone, Francis Galton probably, had once said that variability was a blessing in statistics, more important even than the mean of a distribution.

Rather than remaining petrified in the face of variability, I had to record it, chart it, and make it visible in numbers and graphs. How often does a particular event happen? How frequent is it? How frequent is it relative to other events? Moving from observation to abstraction, I was working with probability distributions before I knew it. Measuring the response times of my participants, I would mindlessly apply the descriptive and inferential statistics that are standard procedure in the research field, computing the means and the standard deviations and performing fancy analysis of variance—a few clicks and button presses and out came a set of results that psychologists of a previous generation would have labored on for hours, days even. I was able to make perfectly sanctioned statements about factors that did or did not have a statistically significant effect on response time, without really understanding how I could say what I was saying. Things changed only when I had to teach the materials to others; I took a good look at the textbooks, practiced a great deal in the labs that I was volunteered to be in charge of, and finally started seeing some light at the end of a dark statistical tunnel.

After a while, you can even develop some kind of aesthetic appreciation for the beauty of distributions. Figure 1.1 is, hopefully, a case in point, showing two sets of three distributions—continuous probability distributions, to be precise, which depict the likelihood of all possible outcomes for a given measure, say, response time in my visual search task or weekly ticket sales for movies. The horizontal dimension gives the possible outcomes—ticket sales from zero to a hundred million dollars. The vertical dimension provides the actual probability associated with each outcome, which must be very low indeed for a hundred million dollars, just once in a blue moon, that is, the opening week of a Batman movie, thriving on the ghostly appearance of an actor who had died of an apparent overdose a half year before the movie's release.

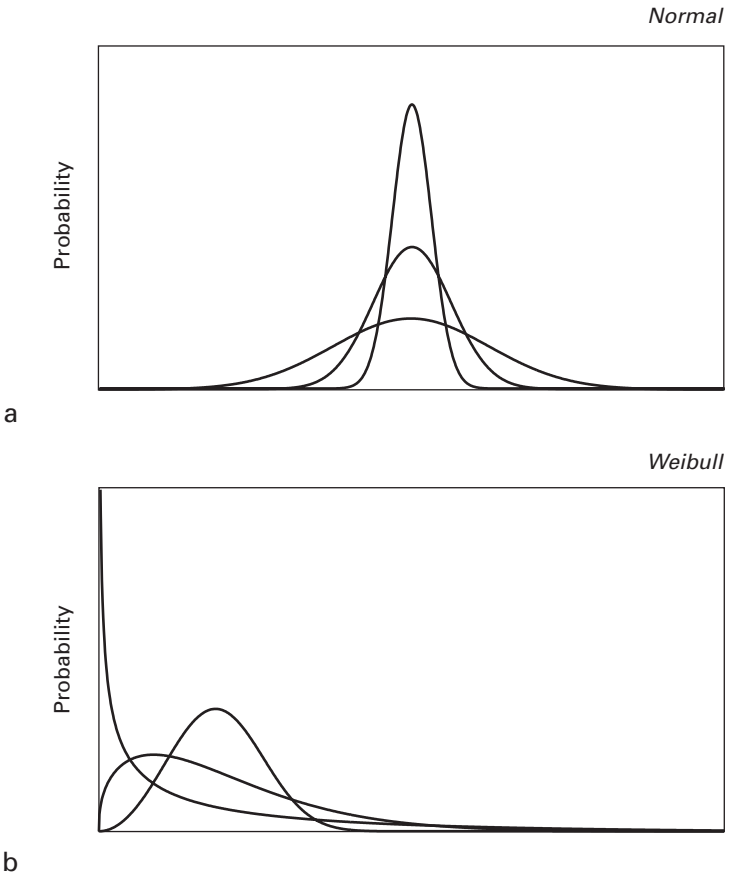


Figure 1.1
Two different types of probability distribution. (a) Three examples of the normal distribution, with the same mean but a different variance. (b) Three examples of the Weibull distribution, with different shapes, morphing from exponential to normal.

The horizontal dimension should contain all possible outcomes, whereas the sum of all probabilities should be one, if we talk in proportions, or a hundred, if we deal in percentages. Let us take a closer look at a few distributions. The three examples in the top panel a of figure 1.1 are instances of the normal distribution, sometimes termed Gaussian in honor of the German mathematician Carl Friedrich Gauss but perhaps more commonly known as "the bell curve." The normal distribution is determined by a location parameter (the mean) and a scale parameter (the variance). The examples in figure 1.1a have the same mean but a different variance—the higher the variance, the flatter the appearance. Note how all three distributions look symmetrical. Perhaps

they look too neatly symmetrical? We should be able to think of other shapes for distributions.

It is in fact possible to quantify shape, as shown with the three examples of the Weibull distribution in figure 1.1b. This distribution was named after the Swedish engineer Wallodi Weibull, who specialized in the study of the strength and durability of materials, often facing data that simply could not be fitted into a normal distribution. The Weibull distribution, like the normal, has just two parameters: Again, one is for scale, but the second determines shape instead of location. This makes the Weibull distribution particularly flexible. Depending on how you set the shape parameter, it morphs into an exponential distribution (like the example in figure 1.1b that swings down from top left) or a normal distribution (like the symmetrical example with the rightmost peak), or anything in between (like the asymmetrical example that sits in the middle).

Wrestling with distributions, then, comes down to categorizing events and deciding whether a particular observation belongs to this or that distribution. We can work in two directions. In most cases we will start from a particular situation, or experimental manipulation, and then collect data to compare the distributions in one case against the other. This is the default approach for a scientific experiment, comparing an experimental condition against a neutral or control condition, which ideally is as similar as possible to the experimental condition except with respect to one factor or dimension—the factor under investigation. For instance, we might be interested in the effects of caffeine on visual search performance. We could perform the experiment in several ways, working with the same or different participants in the two conditions, working with different types of visual search task, caffeine solution, dosage, and so on, and we would have to think hard about how to ensure that a host of other things (e.g., placebo effects, practice with the task, fatigue, boredom) do not contaminate our data, but the bottom line is that we would create an experimental condition with caffeine and a control condition without caffeine. Then all we have to do is write down our observations of response times in two distributions and measure to what extent these overlap. If the overlap is complete, we can safely conclude that the caffeine had no effect. If the two distributions show some degree of separation, we can start thinking that the caffeine managed to do something after all, like increasing the speed of visual search performance. Statistics will give us numbers to support decisions about when the data from two conditions show a “significant” difference. The entire rationale for drawing these conclusions is quite complex when spelled out, but the basic idea is simple: We want to avoid mistakes, so

we use estimates of how likely we are to make a mistake if we claim that our observed distributions prove the two conditions to produce different results. (Usually statistical software does this for us in the form of *p* values.) If the likelihood of error is only one in twenty, or even less, common practice says we can go ahead and make claims about there being a difference.

Perhaps a slightly counterintuitive approach is to work in the opposite direction, from observations back to guesses about which distribution they belong to. Yet this is probably a good characterization of what must happen in the brain when decisions are computed on the basis of the available evidence in terms of activity levels of neurons that represent different alternatives. To see how this works, let us consider a neurophysiological experiment in which we record the electrical impulses of, say, a neuron in secondary (or higher order) visual cortex while the subject is presented with visual stimuli. The subject will usually be a cat or a monkey, but occasionally a human, undergoing neurosurgery (e.g., Quiroga et al., 2005), and the visual stimuli could be anything from former presidents of the United States to random groups of dots moving this way or that.

Thus, for example, we pull up Jimmy Carter on the screen and check what the neuron does in response. We might notice that the neuron becomes particularly active, or tends to fire many spikes, whenever it is Jimmy Carter, but not George H. W. Bush. Can we work the other way around?

We could continue running the experiment but now avoid looking at the screen. Some stimuli are being presented, but we have no clue who or in what order. If we listen only to how often the neuron spikes, can we deduce which former president must have been presented on the screen? How many spikes must the neuron fire for us to conclude that it was Jimmy Carter? Metaphorically speaking, these are exactly the types of questions that other neurons in the brain would be faced with when weighing the input they get from neurons in secondary visual cortex.

The logic unfolded is probably the most powerful approach in contemporary neuroscience when one is trying to model the mechanisms and algorithms seen in neural circuits for decision making (see Gold & Shadlen, 2001, for a bright introduction). The approach was pioneered in the 1960s and 1970s by David Green, R. Duncan Luce, and John Swets (e.g., Green, 1964; Luce, 1963; Luce & Green, 1972; Swets, 1961) and found its definitive formulation in *Signal Detection Theory and Psychophysics*, a book published by Green and Swets in 1966, one of the very few unmistakable classics in this area (the reprint in my collection dates from 1988). The original concern seemed to be all about the purity of signal processing:

> The approach discussed here clearly isolates the inherent detectability of the signal from certain attitudinal or motivational variables that influence the observer's criteria for judgment. . . . To the stimulus-oriented psychophysicist, this analysis is a methodological study, but one that is clearly pertinent since it claims to provide an unbiased estimate of what, for the stimulus-oriented psychophysicist, is the major dependent variable. (Green & Swets, 1988, p. 31)

Green and Swets were clearly rooting for the ideal observer, though they quickly realized that their "Theory of Ideal Observers" (chapter 6) provided an excellent opportunity for "Comparison of Ideal and Human Observers" (chapter 7). In later work, it seemed that John Swets in particular became more and more interested in the broad merits of understanding bias rather than developing a bias of his own against the topic of bias (Swets, 1973, 1992).

Let us borrow the concepts of signal detection theory for a visual schematic representation in figures 1.2, 1.3, and 1.4. The best place to start is by considering the simplest possible decision-making task. Going back to our example with the neuron in secondary visual cortex, we can envisage a forced-choice situation in which the owner of the neuron is simply asked to indicate, in a number of trials, whether a target or "signal" is present, "yes" or "no," just two alternatives. We might get our observer to press a button whenever he or she sees Jimmy Carter. If there is no target, the observer should refrain from pressing the button.

Now we can start recording spikes and button presses and try to relate the former to the latter in our search for a neural correlate of perceptual decision making. Signal detection theory is an invaluable tool in this enterprise, as it allows us to distinguish between two basic ways in which decision making can be influenced. Without diving too deep into the algorithmic depths of the theory, I promise we will be able, a few pages from now, to marvel at its principal strength in teasing apart mechanisms of bias (see figure 1.3) and of sensitivity (see figure 1.4). To fully appreciate how these two ways are fundamentally different, but not mutually exclusive, we first need to come to terms with the basic framework.

As the observer (the owner of the neuron under investigation) makes a decision about the presence or absence of Jimmy Carter, there are logically four possible outcomes: (1) a correct rejection, which occurs when the observer, presented with George H. W. Bush, reports there is no signal; (2) a hit, which occurs when the observer correctly reports the presence of a signal; (3) a miss, which occurs when the observer fails to detect Jimmy Carter actually present among the noise; and (4) a false alarm, which occurs when the observer erroneously reports the presence of a signal.

Green and Swets suggested that these four outcomes could be accounted for with a model that incorporates a "noise distribution," a "signal distribution," and a "criterion" (see figure 1.2, with indications of the four possible outcomes). The two distributions can be thought of as one probability distribution broken down in two "subdistributions," one indicating the likelihood of observing a particular number of spikes given the presence of a signal (i.e., signal distribution) and its complement indicating the likelihood of observing a particular number of spikes given the presence of only noise (i.e., noise distribution). How does our observer decide whether seven spikes should be taken as evidence of Jimmy Carter?

The terminology brings Bayes's theorem back to mind, and indeed, working our way inside figure 1.2, we can recognize the different components of the theorem at play. So let us say that we get a reading of seven spikes during a particular trial in our experiment with the observer looking for Jimmy Carter. What is the likelihood that the stimulus is indeed Jimmy Carter, given a reading of seven spikes? To compute P(Carter|Seven), as we have duly learned by heart, we should work out P(Seven|Carter) times P(Carter), divided by P(Seven).

P(Carter) refers to the entire signal distribution and its relation to all possible cases. In our two-choice task, there are only two possibilities: Carter (signal)

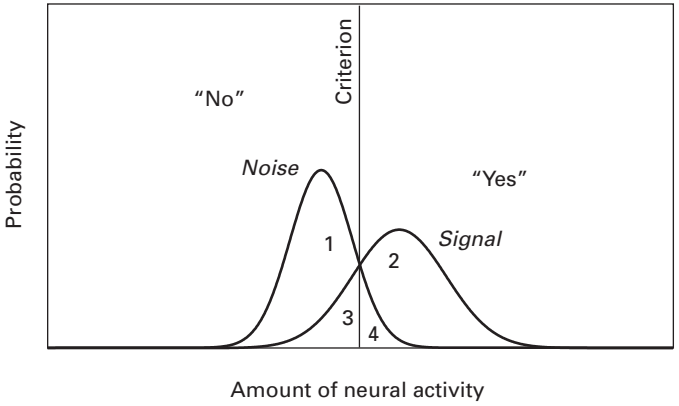


Figure 1.2
An application of signal detection theory. The horizontal dimension represents the number of spikes fired by a neuron. Shown are two hypothetical normal probability distributions, for the neural responses in the case of "noise" versus "signal." The vertical line cutting through the two distributions represents a criterion for signal detection, saying "yes" for spike values above criterion and "no" for those below criterion. There are four possible outcomes: *1* represents the "correct rejections" (saying "no" when there was in fact no signal), *2* shows the "hits" (saying "yes" when there was indeed a signal), *3* points to the area of "misses" (saying "no" although there was actually a signal), and *4* indicates the area of "false alarms" (saying "yes" even if there was really nothing but noise).

or not-Carter (noise). This implies that P(Carter) equals 1 – P(Noise). Thus, we can compare the size of the entire signal distribution to that of the noise distribution—in figure 1.2 they are the same size, that is, we have just as many signal as noise trials in our experiment, or P(Carter) = 0.5. This component represents the prior, and it is easy to see that the basic likelihood of the signal, or our belief of how likely it is, has a large impact on all computations to follow. The numbers would certainly be very different if the experiment had only one Jimmy Carter appearing in every hundred trials.

We can find P(Seven|Carter) by considering the signal distribution to be a complete probability distribution on its own—or multiplying the signal distribution with a factor that brings the total sum of all its cases to one. If P(Carter) = 0.5, we simply need to multiply by two. Now we can trace the curve of the signal distribution until we reach the value of seven on the horizontal axis. The associated value on the vertical axis, multiplied by the appropriate factor, gives us P(Seven|Carter).

To read the general probability of seven, P(Seven), we should not multiply the signal or noise distributions but instead simply trace each of the two distributions until we reach the value of seven on the horizontal axis, read the associated probabilities, and compute the sum of both values.

Now we already have all the components that we need for solving the equation, but in figure 1.2 we might as well look up P(Carter|Seven) more directly, by locating the value of seven on the horizontal axis, then moving up vertically until we hit the curve of the signal distribution, and reading the associated probability. Next we do the same for the noise distribution. Effectively, we find the same two probabilities that we made use of to compute P(Seven). This time we can consider these to make up a total of one, that is, we need to multiply P(Seven) by a factor that brings it to one. Now we can multiply the individual probabilities by the same factor to give us the sought-after number: P(Carter|Seven).

One way or another, the visual scheme presented in figure 1.2 does incorporate the truisms of Bayes's theorem. However, rather than performing these somewhat tedious computations, there is nothing to stop you or me from working more intuitively with the logic and leaving the numerical applications, proofs, and annotations for another day in another life. One thing glaringly absent in Bayes's theorem is an instruction on how to interpret whatever probability we do compute. The theorem might help us wrestle with distributions, but it does not specify what we are to do with the posterior probability once we have computed it. Somehow, we should try to link the posterior probability to a decision or an action. We need a decision rule.

Actually, finding a good decision rule should not be too difficult. A simple adagio would be to try to maximize gain and make sure that our decisions on

the presence or the absence of a signal, with the four possible outcomes of hit, miss, correct rejection, and false alarm, combine to our profit. Here, signal detection theory provides us with its most ingenious addition to Bayesian thinking—the concept of a criterion or threshold. In one sense, this is nothing new, merely a formal application of a common practice in statistics, where categorical decisions are imposed on continuous distributions in the form of conventional criteria for "statistical significance." However, in signal detection theory, the criterion is introduced as a borderline that cuts across the signal and noise distributions, enabling one to clearly visualize how the positioning of this borderline determines the likelihood of each of the four possible outcomes.

In figure 1.2 we see that the criterion is taken as the borderline between "yes" and "no" responses. For spike counts higher than criterion, to the right of the borderline, our observer would conclude that Jimmy Carter was shown on the screen. For spike counts below criterion, the answer would be "no." Any case belonging to the portion of the signal distribution to the right of the criterion would then produce a hit (area 2 in the figure), but if the observation of a spike count above criterion actually belonged to the noise distribution, our observer would make a false alarm (area 4 in the figure). Conversely, we can see how this scheme relates misses (area 3) and correct rejections (1) to the positioning of the criterion.

In figure 1.2 the criterion is placed right at the crossroads between the two distributions, at the point where the spike count is equally likely to reflect a signal or noise. To the left of the criterion, the noise distribution dominates, with spike counts that more likely reflect noise than a signal, whereas the signal distribution rules to the right of the criterion. In fact, the criterion is quite strategically (rationally!) placed to minimize the likelihood of an erroneous decision, be it a miss or a false alarm.

Research on eye movement control in macaque monkeys suggests that this kind of categorical threshold idea makes for a plausible neurophysiological mechanism. Doug Hanes and Jeff Schall (1996) showed that the activity of neurons in the frontal eye field (the prefrontal cortical structure for voluntary control of eye movement) consistently peaked at around a hundred spikes per second right before the initiation of an eye movement, regardless of how long it took for the neural firing rate to grow to that peak, and regardless of how long it took for the monkey to initiate the eye movement. When the spike rate was at 100 spikes per second, the eye movement took off. Data from similar experimental paradigms, recorded from neurons in superior colliculus (the major subcortical station that drives eye movement initiation), provided additional support for the existence of an absolute threshold (Krauzlis & Dill, 2002; Paré & Hanes, 2003).

Perhaps the threshold idea is not a crazy one. Applying a rule like that is certainly not difficult and really involves no thinking. All that is needed is enough heat from electrical impulses to wake up the next layer of neurons. It is at about the right level of simplicity to be useful in mapping decision-making properties onto neural circuits. But returning to a safer level of abstraction for the time being, we can explore the effects of positioning the criterion somewhere other than strategically in the middle between noise and signal. Figure 1.3 reproduces the neutral case of figure 1.2 and brings up two other cases for comparison: one in which the criterion is shifted to the right, and another case with a criterion shift in the opposite direction. It is easy to appreciate that the position of the criterion determines the likelihood of different types of errors. With a rightward criterion shift, as in panel b, we avoid false alarms but are much more likely to miss actual signals. In contrast, we reduce the misses at the expense of false alarms if we shift the criterion to the left as in panel c.

When misses and false alarms are equally costly in economical or evaluative terms, the most rational strategy will be to place the decision criterion so that both types of error are minimized as much as possible. In other situations, it may be important to avoid misses—like when we interpret the data from a diagnostic test for pancreatic cancer—whereas false alarms carry less weight. The decision criterion would then better be shifted to the left, minimizing the area of the signal distribution that falls on the wrong side of the criterion, at the expense of an increased number of false alarms. In yet other situations—when we point a rifle at a cloud of dust, kicked up by, potentially, an armed insurgent—it would be crucial to avoid making a false alarm and shooting an innocent victim. Here, the preferred option should be to shift the criterion to the right.

The shifts of criterion install observer biases, leading to different actions or decisions even if the underlying signal and noise distributions retain the same outlook. With rightward shifts, our observer applies a conservative criterion, requiring more evidence than in the neutral case before agreeing that Jimmy Carter is present. On the other hand, more liberally minded observers might shift the criterion to the left and be happy to decide, on the basis of relatively few spikes, that the signal is there all right. The criterion sets the amount of evidence or information required for a decision, and any decision that is predisposed in favor of, or against, accepting a signal can rightfully be called "biased," and yet might still be appropriate, reasonable, or even rational.

I should point out that there are other ways to conceptualize shifts of criterion. My favorite, and a neurophysiologically plausible way, is to shift both distributions, while keeping the threshold in place (Lauwereyns et al., 2002a,

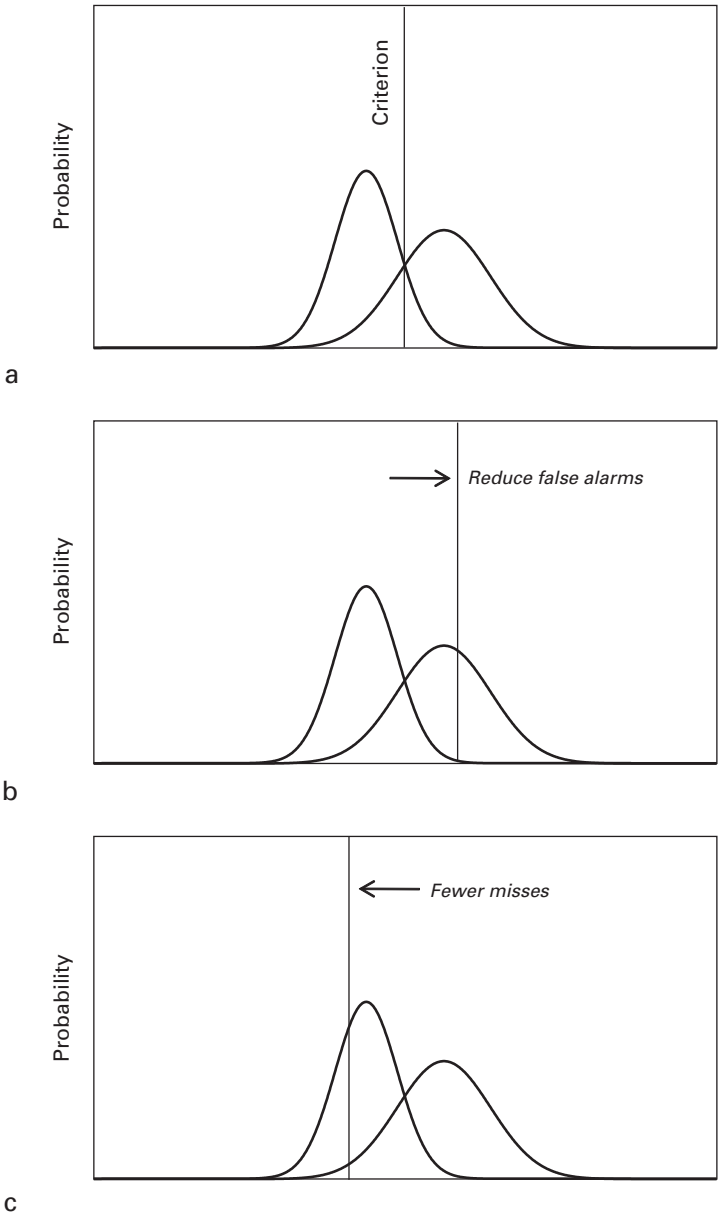


Figure 1.3
Variations to the tune of bias. (a) Take this to be the neutral case, the same as in figure 1.2. (b) The criterion is shifted to the right, more "conservative" than the neutral case, requiring a higher neural firing rate before accepting that there is a signal, and so reducing the likelihood of making a false alarm. (c) The criterion is shifted to the left, more "liberal" than the neutral case, already happy with a lower neural firing rate to say "yes," which brings down the likelihood of missing a signal.

2002b). In practice, this explanation is perfectly interchangeable with the variations to the tune of bias as shown in figure 1.3, and I consider the two proposals to be equivalent. I will introduce the neurophysiological data in detail in chapter 2.

Yet another way to implement bias, definitely a theoretical possibility but as yet not seen in neurons, is to move the criterion to the left or right so that it stays perfectly in the middle between signal and noise. This would be done by more literally applying the role of the prior, that is, by enlarging or reducing the signal distribution relative to the noise distribution (without changing the shapes or the means of either distribution). With an enlarged signal distribution, for instance, the midway crossover point between the two distributions would shift to the left. The enlarged signal distribution could reflect an actual increase in the likelihood of a signal (as when we now present Jimmy Carter on two thirds of the trials in the experiment), or it might reflect an observer's overestimation of the true likelihood—a distorted image of reality. But the conjecture that the shapes and means of the distributions remain the same makes it hard to translate this possibility into a neurophysiologically plausible model of bias in decision making. The real decision making would have to be done outside of the model, with different weights of the signal distribution in the workings of some mysterious Master of Shadows, a decision maker hidden from view. When it comes down to neurons, I would like to see them actually do something if they are to contribute to decision making.

Figure 1.4 shows an entirely different mechanism influencing decision making. Here, the movements and variations occur to the tune of sensitivity. Again we start from the neutral case, the one introduced in figure 1.2. The task of decision making is particularly challenged by the overlap between signal and noise distributions. The overlap implies uncertainty and increases the likelihood of error. Arguably the ideal way to improve decision making, then, would be to try to reduce the overlap, or improve the signal-to-noise ratio so that the two distributions are more clearly distinguished. Assuming that each of the two distributions has a normal shape, we could heighten the sensitivity for a signal by fine-tuning so that both distributions have a crisper appearance with smaller standard deviations, as shown in figure 1.4, panel b. Alternatively, the signal-to-noise ratio can be improved by moving the two distributions further apart, changing the means but not the standard deviations, as shown in panel c. In both cases, we can easily place the criterion at an optimal spike level that succeeds nicely in segregating signals from noise.

To effectively improve the Jimmy Carter-to-noise ratio, we might finally allow our observer to put his or her glasses on. Or we could dim the lights in the room so that the screen stands out. Real-life examples of improved

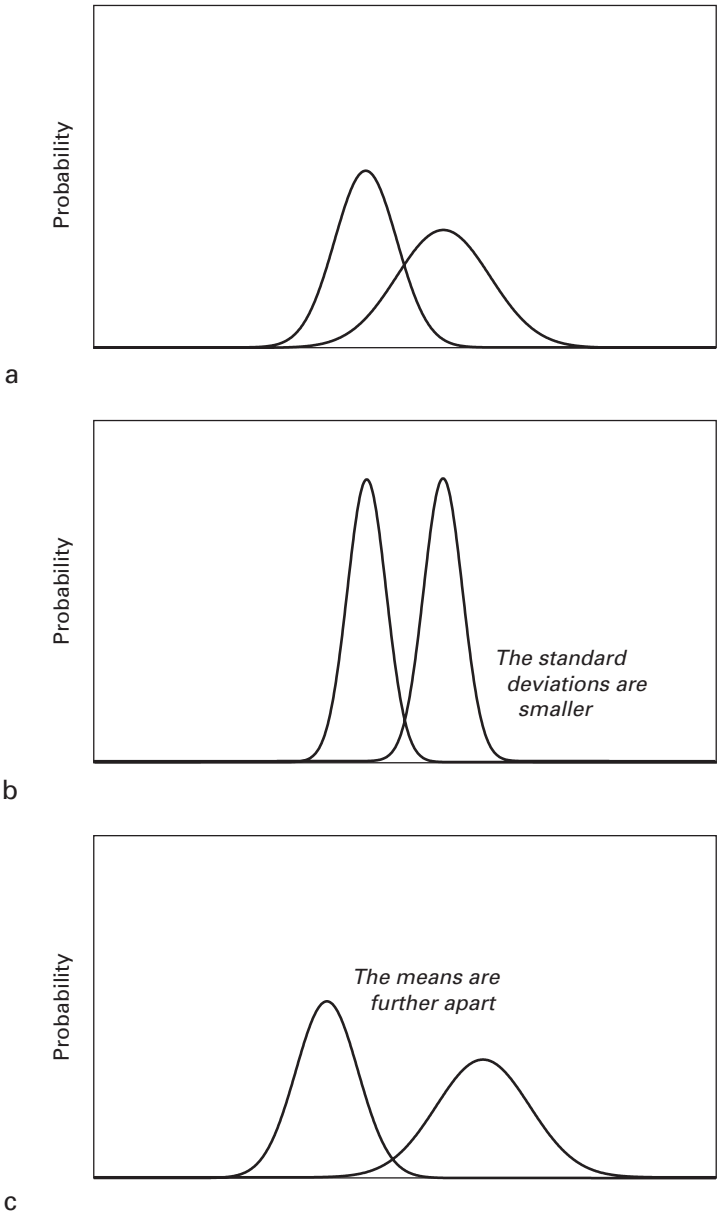


Figure 1.4
Variations to the tune of sensitivity. (a) We start from the same position as in figure 1.2 and 1.3a. (b) Here, the overlap between the two distributions is reduced, facilitating the extraction of signals from noise by decreasing the variance (giving smaller standard deviations). (c) In this case the overlap is reduced by changing the location parameter, leaving the scale parameter untouched. That is, now the means of the two distributions are further apart, whereas the standard deviations remain as they were.

sensitivity often reflect some kind of physical change, creating better conditions for signal reception, but the enhanced decision making can sometimes also be achieved by internal, cognitive operations like paying attention, thinking twice about the data, or double-checking the numbers.

In any case, with whatever degree of bias or sensitivity, the conceptualization of decision making with signal detection theory quickly raises many questions, about the shapes of the distributions, the nature of the decision rules, and so forth, but the framework has the considerable merits of simplicity, specificity, and testability. We will be able to derive predictions from it in terms of neural signatures. One limitation of the signal detection theory, however, is that the logic with categorical decisions is difficult to convert into predictions about response time during decision making. Yet, response times are just possibly the most powerful behavioral measure of what is going on in the brain. They might tell us more than only the accuracy of "yes" or "no," or the trade-off between response speed and decision quality. Time could well provide a quantitative measure of the decision process, of how hard it was, how much thinking it took, and how many neurons had to have their say. Can we work toward some kind of integration of signal detection theory and response time measurement?

Time and the Measurement of Mind

The desire for a quantitative approach to the study of brain and behavior has surfaced only recently, but perhaps it is fair to say that the wishful thinking had been there for centuries, at least since the amazingly modern proposals of Doctor Mirabilis, Roger Bacon—possibly, and horribly, better known as the thirteenth-century model for Sean Connery in the film version of *The Name of the Rose* (see Cregg, 2003, for a preferable biography). To really make progress with numbers, though, and to stimulate the desire further, humans first needed to invent reliable clocks that could tick away the seconds, and then, toying in the lab, stumble on the brilliant idea that these clocks could be useful for the as yet unnamed venture of neuroscience. As it happened, the original proposal emerged in my native language, Dutch, exactly a hundred years before I was born—good enough reason, I would like to think, for a fetishist attachment. Here it is, radically unfiltered and incomprehensible to most (Donders, 1869, p. 119):

> Maar is dan ten opzichte der psychische processen iedere quantitatieve behandeling uitgesloten? Geenszins! Een gewichtige factor scheen voor meting vatbaar: ik bedoel den tijd, die tot eenvoudige psychische processen wordt gevorderd.

The idea that we could measure the mind in seconds is certainly an outlandish one, not necessarily understood any better if it is formulated in English. W. G. Koster made a complete translation for the second volume of *Attention and Performance* that he edited in the year that I, of course, do not associate with any landing on the moon, it being a hundred years after a hundred years before I was born (Donders, 1969). Reasoning with time is not easy, but Franciscus Cornelis Donders suggested we should try. Here is my own translation of the excerpt:

> But is then every quantitative approach impossible with respect to mental processes? Not at all! An important factor seemed amenable to measurement: I mean the time taken up for simple mental processes.

In French, we would have spent thousands of inebriating pages *In Search of Lost Time*, but this is the best of Dutch, fully exhibiting its pragmatic quality. Even if the theory was wanting, and the rationale idiosyncratic or simply absent, the intuition that it might be useful was all the incentive required to commence with experiments. Thus, Donders went on to develop his infamous subtraction method, still a standard tool today, comparing the response times of his subjects as they performed different tasks, with systematic variation of the level of complexity for stimulus processing and response preparation. The more complex the mental process, the more time it took to give a correct response.

Somehow the differences in response times did, in fact, correspond with the complexity of cognitive operations, and one way or another, the fact that thinking took time had to be an important observation. For one thing, it suggested that the mechanisms of thought left a material trace, one that was not easily reconciled with Cartesian dualism and its profound divide between the immaterial world of the mind and the physical reality of the body. However, even if the hard-core dualism was easily rejected in principle, most researchers remained vulnerable to its lure in more implicit ways, in assumptions of what the brain did and where the cognitive operations took place (Bennett & Hacker, 2003). It might be true that thinking took time and left material traces, but this was a long way from explaining exactly what kind of cognitive operations took how much time and why.

Around the hundredth birthday of Donders's famous article, some researchers started getting more serious about deriving knowledge of the cognitive architecture from distributions of response times. Sternberg (1969a, 1969b) explored the use of search times in memory tasks to tease apart parallel versus serial processing. He asked his subjects to search their memory for items from a set they had learned by heart (or were trying to keep online in their head).

Some types of memory search gave flat response time curves, independent of the number of items in memory, whereas other types of search produced a steep increase in response time depending on the set size—the more items, the slower the response. The flat slopes suggested parallel processing, Sternberg concluded, and steep slopes indicated that the subject had to work one by one, serially considering each item in memory.

Treisman and colleagues (Treisman & Gelade, 1980; Treisman & Sato, 1990; Treisman & Souther, 1985) applied a similar logic to analyze response times in visual search tasks, and went one step further in the interpretation of the underlying cognitive operations with the feature-integration theory. Parallel search, seen in flat slopes, would occur for "singleton" targets, which differed from the distractors in only one visual dimension—like when you look for a red item among a set of green distractors. Serial search would be required whenever the target was defined on the basis of a combination of features—like when you look for a red square among blue squares and red circles. To combine visual features, you would need something called "attention" to glue the features together at one location at a time. Searching for a combination of features, then, meant that you would have to allocate attention to one location, let attention do its gluing there, decide whether the element at that location matches the target, and move on to the next location if it does not. The search time would literally depend on the number of times you have to shift attention to a new location.

The theory suffered badly from a load of incompatible data in dozens of new studies from other labs (see Wolfe, 2001, for a succinct review), but as a first shot it was not bad at all, and I have always admired its wonderful precision in translating response times to a fairly precise drawing of the underlying cognitive scheme. My own little experiment on visual search, in which I learned to wrestle with distributions, would have had no meaning if there were no feature-integration theory to shoot down.

Arguably the most forceful plea for response time analysis was put forward by R. Duncan Luce (1986), who had helped establish the threshold concept and was very familiar with the tenets of signal detection theory. He explained, for all who could follow, that it was feasible to exploit the shapes of response time distributions in an effort to deduce the covert operation of separable parameters relating to actions of the mind. For a long time, I thought this was the most esoteric of all things in psychology and statistics, something I would like to be able to understand if only I had the brain power for it. But then I encountered R. H. S. Carpenter's LATER model (Carpenter, 1981, 1999, 2004; Carpenter & Williams, 1995; Reddi & Carpenter, 2000), and I found myself actually coming to grips with it, or even liking it to the point that

I started working with it myself (Lauwereyns & Wisnewski, 2006). Here was a model that looked innocent enough, at least if you consider the schematic drawings, and it managed to work with just a handful of parameters, one which looked suspiciously like what I thought of as bias and another that just had to be sensitivity.

Carpenter wrote the most enjoyable introduction to the LATER model in his article for the *Journal of Consciousness Studies* (1999). The model starts from the observation that the eye movements of human observers are curiously slow if we consider the underlying anatomy for visual processing and eye movement control. Between the retina (when a stimulus excites the photoreceptor cells in the back of the eye) and the repositioning of the eyeball (when we move our eyes to bring the stimulus in central vision and examine it more closely), there should in principle be only a few synaptic steps involved, or a sequence of maybe five or six neuron-to-neuron transmissions. This should take up a few tens of milliseconds at the most. Instead, the response times with eye movements normally clock in at about two hundred milliseconds, and often even more. To explain this procrastination, Carpenter suggested, there must be a central decision-making mechanism that converts the available sensory evidence into an eye movement via a decision process with random variability. He even offered a philosophical perspective on the biological advantages of this random behavior, from escaping boredom and promoting creativity to outwitting our opponents and really willing freely.

Given the central role of random procrastination, the model is aptly named LATER. The acronym, however, stands for linear approach to threshold with ergodic rate, a rather ominous whole, in which I suspect the E was forced a bit for poetic reasons. Nobody really knows what "ergodic" means or whether it stems from "a monode with given energy" or "a unique path on the surface of constant energy" (Gallavotti, 1995). In practice, "ergodic" must be borrowed from the ergodic hypothesis in thermodynamics, which, brutally simplified, claims that, if you just measure long enough, you will find that a particle spends an equal amount of time in all possible states. In statistics, the ergodic hypothesis is taken to imply that sampling from one process over a very long period of time is equivalent to sampling from many instances of the same process at the same time. The process should be stable, no decay, no learning. With this caveat, then, the LATER model addresses decision making in a static context when the observer performs at full capacity.

The LATER model conceives of decision making as a process represented by a continuous, straight line, the "decision line," that rises to a threshold, or cutoff level—when the decision line crosses this threshold, the decision becomes effective, the motor execution is initiated, and the response time can

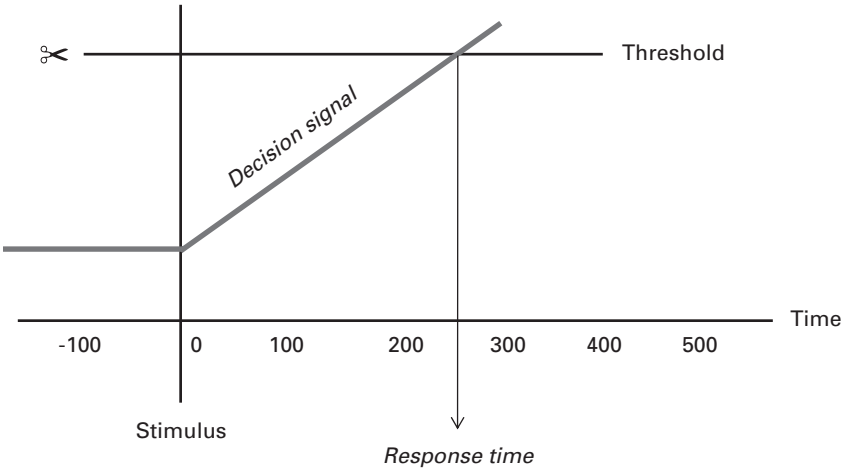
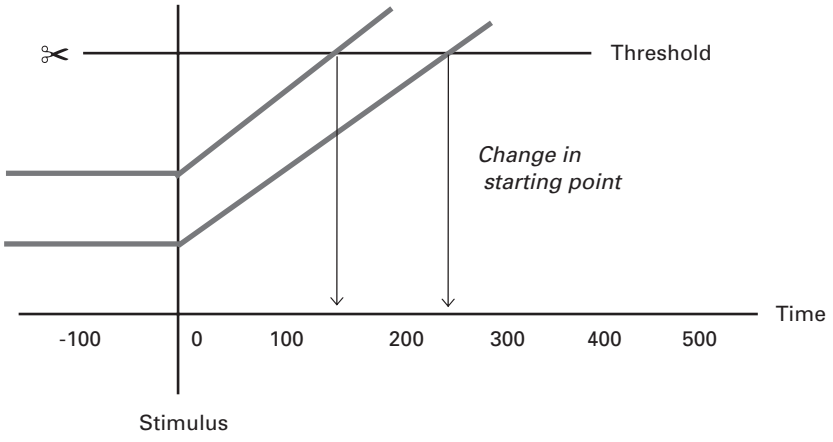


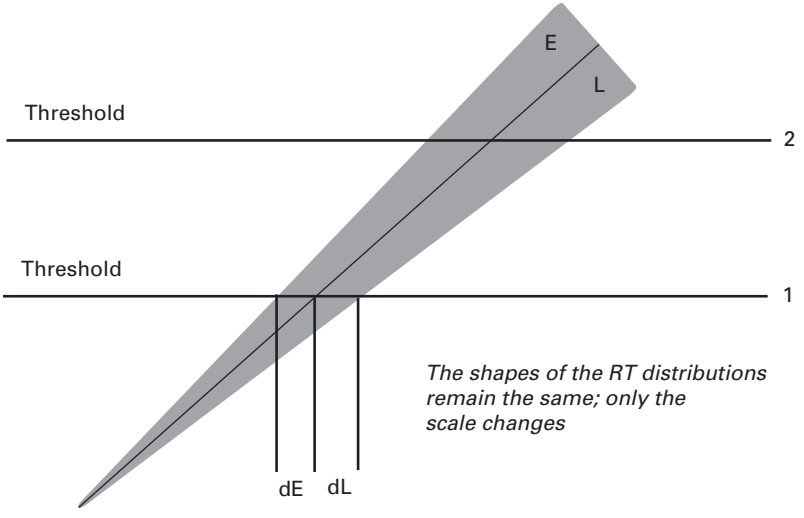
Figure 1.5
A linear model of response time in decision making. The horizontal axis represents time; the thick gray line gives the "decision signal" (a putative neural correlate of decision making). A decision is reached when the decision signal crosses a fixed threshold. In this example, the decision signal grows linearly from the time of stimulus onset and crosses the threshold in about 250 milliseconds.

be recorded (see figure 1.5). The model is based on just a few parameters: the starting point of the decision process (i.e., the distance between the intercept of the decision line and the threshold, assuming that the threshold is fixed); the average steepness, growth rate, or gradient of the decision process (i.e., the slope of the decision line—there is no shortage of synonyms); and the variance of the gradient.

The primary attraction of the LATER model is that it makes specific predictions about how changes to the parameters affect the shapes of response time distributions. With figures 1.6 and 1.7, I provide an unorthodox explanation that deviates substantially from the actual way in which the LATER model checks for changes to response time distributions. The true LATER model employs the reciprocal (or inverse) of response time—a little trick aimed at morphing the typically skewed response time distribution into a nicely symmetrical and normal one—and then draws the transformed distribution using a so-called "reciprobit plot," which pictures normal distributions as a straight line. The lines then swivel or move in parallel, depending on which parameter is changed. It works elegantly and makes for a straightforward statistical analysis, but to the untrained eye it can seem a bit confusing because a parallel change in the visual scheme of the model (moving the starting point up or down) translates into swiveling in the reciprobit plot, and vice versa. The

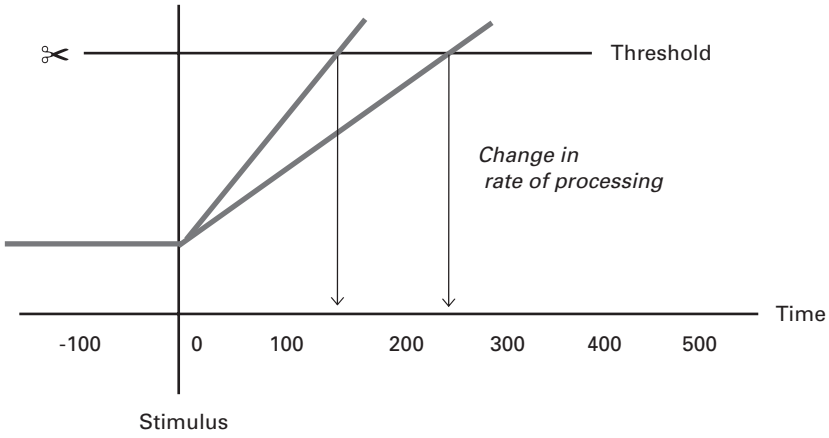


a

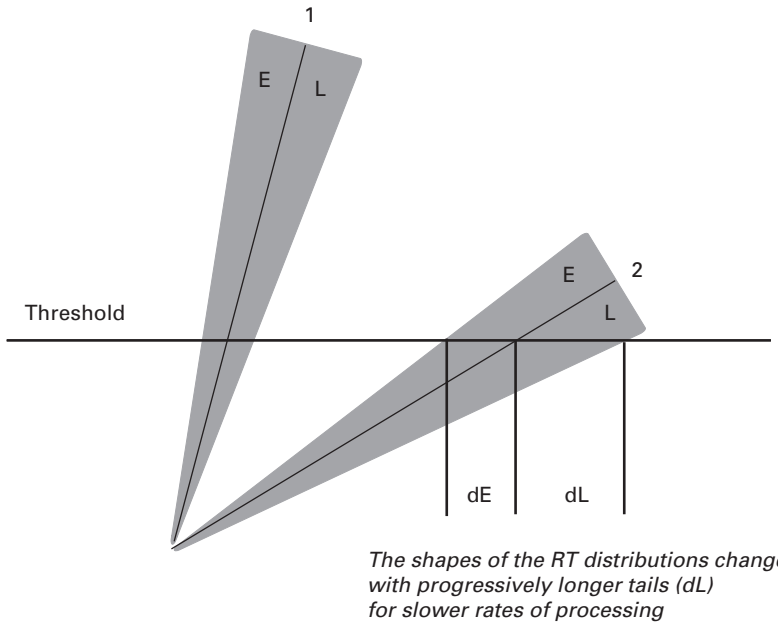


b

Figure 1.6
Effects of bias on response time. (a) The scheme is the same as in figure 1.5, including the decision line for the neutral case. Compared to the neutral case, the new decision line has a starting point closer to the threshold, allowing it to reach the criterion faster. Time is given in milliseconds. (b) Changing the starting point of the decision signal does not affect the ratio of the early (E) versus late (L) part of the response time (RT) distribution.



a



b

Figure 1.7
Effects of sensitivity on response time. (a) The scheme is the same as in figures 1.5 and 1.6a, including the decision line for the neutral case. Compared to the neutral case, the new decision line shows a steeper slope, allowing it to reach the threshold faster. Time is given in milliseconds. (b) Variation in the gradient (or steepness) of the decision signal influences the ratio of the early (E) versus late (L) part of the response time (RT) distribution: The tail of the distribution gets stretched out with shallow gradients.

problem is that the double transformation, using the inverse of response time and plotting in a funny way, warps the mind beyond our gut instincts (or implicit associations).

To satisfy my own intuitive inclinations, I tried to work out a rationale without the double transformation. Figure 1.6a shows what happens if we move the starting point closer to the threshold—a sure form of bias, or predisposition to reach a particular conclusion. It is easy to see that with the shorter distance to cover, response times will decrease considerably. Here, we go from 250 milliseconds down to 150. To get an idea of what this implies for the shapes of the response time distributions, figure 1.6b shows the decision line with its variance (the gray region; the thin line in the middle represents the mean). As we change only the distance between the starting point and the threshold, we might as well picture the situation with one distribution traveling up to threshold 1 or 2. Now we can consider how the early half of the response time distribution ("dE") relates to the late half ("dL"). The ratio of the two does not change. That is, the shape of the distribution stays the same; it only gets magnified if the decision line has to travel a greater distance.

Figure 1.7 applies the same logic for changes to the gradient of the decision line. Panel a shows what happens to the decision process if we have a steeper growth rate to the threshold, presumably due to more efficient information processing, or a clearer signal reception, that is, heightened sensitivity. With respect to the distance between the starting point and the threshold nothing has changed, but again we see a clear improvement in response time, from 250 milliseconds down to 150. Panel b works out the ramifications for the shapes of the response time distributions. Case 1, with a very steep growth rate, shows that the distribution approaches a symmetrical shape, with a ratio between the early half ("dE") and the late half ("dL") of not much less than one. Venturing into the absurd, we can even imagine a straight vertical decision line with a mean response time of zero, which would have a ratio between dE and dL of exactly one, or even more absurd, a line tilting to the left, with negative response times that imply a ratio of higher than one. Of course, in reality we can only tilt to the right, but the point should be clear: The ratio between dE and dL changes with the slope of the decision line. If we look at case 2, with a much shallower slope, we see that dL increases relative to dE, that is, the ratio of dE/dL dives well below one and the tail of the response time distribution gets stretched out.

However strange it sounds, or downright mystical, there must be some truth to Luce's (1986) dictum that we can read cognitive architecture out of response time distributions. With the LATER model, I found it was easy to relate mechanisms of bias and sensitivity to specific parameters that influence the shapes of

response time distributions. This is not to say that everything is perfect with the LATER model. It cannot account for errors in decision making, and the idea that the decision process grows linearly must surely be a particularly vulnerable abstraction (see Smith & Ratcliff, 2004, and Bogacz et al., 2006, for comparisons of the strengths and weaknesses of different models). I would like to think of the LATER model as the simplest of all, and therefore the best place to start, even if it means tweaking the experimental paradigm so that errors are logically impossible (Lauwereyns & Wisnewski, 2006). But sooner or later, it may be necessary to extend the LATER model by adding parameters (e.g., Nakahara, Nakamura, & Hikosaka, 2006) or to develop nonlinear models that can account for error and exhibit a more neurophysiologically plausible growth rate—sometimes also called "drift rate" to emphasize that the rate does not necessarily grow (Ratcliff, Van Zandt, & McKoon, 1999).

Despite its limitations, however, the LATER model manages to provide an astonishing fit to response time distributions in some tightly controlled situations. In these cases, we can hope to apply the most powerful triangulation, measuring behavioral responses concurrently with neural activity on a trial-by-trial basis in one and the same experimental paradigm. Historically speaking, triangulation might have evolved as a method to measure the distance between shore and ship. Taking readings at two different angles on the shore, we should be able to work out where the lines will meet the ship in the distance. Applied to neuroscience, we can think of the experimental paradigm as the shore, and the behavioral and neural readings as our two angles that seek to meet the mind in the distance. No doubt most scientists will agree that this is the obvious best way to proceed. It is disappointing, however, how rarely it is applied in practice. All too often one of the two readings, usually the behavioral, is sketchier than it might have been.

The default approach seems to be to roughly compare one condition with another in terms of behavior as well as neural activity. Say we compare the ability of our observer to detect Jimmy Carter with glasses on versus off, and we measure the activity of a neuron in the observer's medial temporal lobe in the same two conditions. If we do what most researchers do, we will compute only four data points: the percentage of correct responses with glasses on versus off and the neuron's average Carter-to-noise ratio with glasses on versus off. Our observer makes fewer errors with glasses on and—lo and behold—the neuron fires more for Jimmy Carter than for anyone else… *We have a neural correlate of perception!* Or, the neuron fires less for Jimmy Carter than for anyone else… *We have a neural correlate of perception!* Whatever the neuron does, we will publish a paper in a very nice journal, but did we really compute a correlation?

We should be able to do better than that. We can record our observer's response times and use the trial-by-trial variability to check whether the gradient of his or her decision line is steeper with glasses on than with glasses off, as we might expect under conditions of heightened sensitivity. At the same time, we can check whether trials with shorter response times correlate with higher (or lower) firing rates for that wonderful neuron in medial temporal cortex.

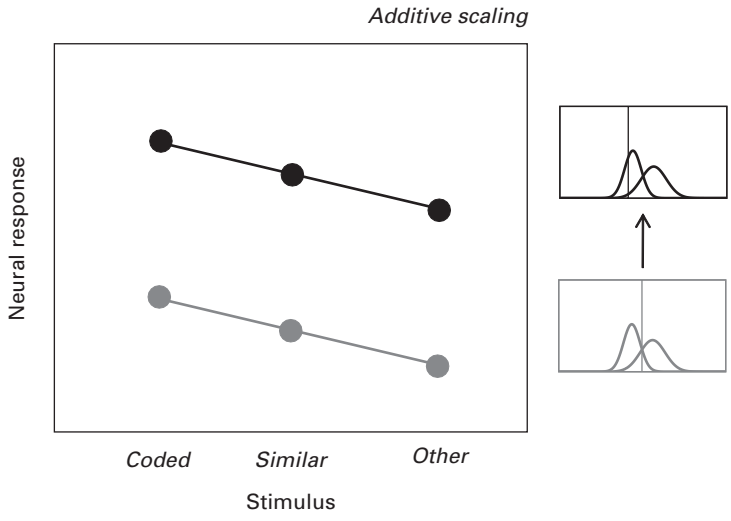
The goal must be to establish the most detailed correlation of neural activity and behavioral responses, preferably in conditions that allow us to compute separable parameters in the responses. Then, and only then, can we hope to present a complete account of how neural circuits weigh the options. The best strategy will be to incorporate the LATER model or other tools of behavioral analysis in the design of experimental paradigms. As these tools become more sophisticated, researchers will be better equipped to examine, among other things, the neural signatures of bias and sensitivity in decision making.

Neural Signatures of Bias and Sensitivity

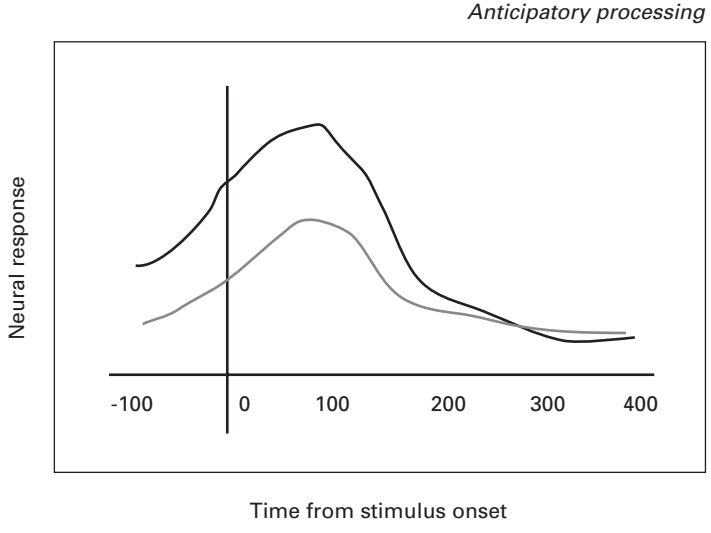
Combining the concepts of signal detection theory with the LATER model, there emerge a few solid markers that we can apply in the search for neural mechanisms of bias and sensitivity. At the moment, these markers are merely hypothetical, speculative, and conjectural, or subject to some of the most dreaded adjectives in science—they are the offspring of two very different ways of thinking about decision making, and so they may wither with the demise of either theoretical parent. Nevertheless, the markers are the proper kind of instrument for our search as they are wonderfully precise about what we should see in neural activity under the regime of bias versus sensitivity.

Figure 1.8 does the deductive work for the case of bias. In panel a, the data from a hypothetical neuron (or neural population) are drawn from a factorial design with 2×3 conditions: There are three possible types of stimulus (coded, similar, and other) and two possible treatments (biased or neutral). The different stimuli are needed to get an idea of what the neuron basically responds to—what kind of information does it normally "encode"?

The most thorough way to characterize a neuron's response properties would be by drawing a tuning curve in the way Vernon B. Mountcastle and colleagues originally conceived it (LaMotte & Mountcastle, 1975; Mountcastle, LaMotte, & Carli, 1972; Talbot et al., 1968), by systematically charting the changes in neural responses as a function of changes to a stimulus parameter. The stimulus that elicits the strongest level of neural activity, or the apex of the tuning curve, must be the prototypical stimulus, the one



a



b

Figure 1.8
The neural signature of bias. (a) The horizontal axis of the main panel marks three different types of stimulus as a function of a neuron's basic tuning (or strength of response, listed from high to low): coded, similar, and other. The gray data represent a neutral case; the black data are driven by bias and show an additive increase as compared to the neutral case. The two inset figures to the right are borrowed from figure 1.3. (b) Average neural activity levels are shown over time (horizontal axis) relative to the onset of the stimulus (vertical line at time zero). The difference between the data driven by bias (black line) versus those from the neutral case (gray line) is already apparent before stimulus onset, reflecting anticipatory processing.

encoded, represented, or otherwise conveyed to the rest of the brain by the neuron in question—I call it the "coded" stimulus. (I prefer to avoid the term "preferred stimulus" because preferences sound too much like a matter of choice and the application of wishes, and that is not only too anthropomorphic for a neuron but also confusing with the real wishes in chapter 2.)

From early efforts in drawing tuning curves, it was immediately clear that, from the neuron's perspective, not all noise is equal. Tuning curves tend to show a peak that does not suddenly emerge from the flat but is supported by noticeable slopes, often in the shape of a bell. Panel a in figure 1.8 takes a shortcut, sampling just three positions on the tuning curve: the apex ("coded"), somewhere in the middle ("similar"), and the bottom ("other"), where noise is really just noise or maybe the perfect "antipreferred" stimulus, the exact opposite of what would get a neuron's juices to flow.

As an aside, we should note that, of course, the idea of a tuning curve should not be restricted to sensory stimulus coding. In fact, I prefer to think of tuning curves as part of the same family as receptive fields, mnemonic fields, and movement vectors, all the different charts and plots that characterize a neuron's firing rate with respect to any physical parameter in the experimental paradigm, be it spatial or nonspatial, present or past.

The gray data represent the neutral case. A shift of criterion, we noted, would be equivalent with a parallel movement for both the signal and the noise distribution in the framework of signal detection theory. Translated to the three positions on the tuning curve, this means we should see an additive increase when the observer is biased in favor of the "coded" stimulus that takes the apex of the neuron's tuning curve. The black data, driven by bias, seem to have undergone a parallel (linear) movement upwards from the neutral case, regardless of the actual stimulus, as if the tuning curve simply rides on top of an elevated baseline. This is, of course, also compatible with the LATER model, which further specified that the change in baseline, or starting point for the decision line, would be fully in place at the moment the first sensory evidence of the stimulus arrives. From this, we can distill an important second marker with respect to the temporal dynamics of bias effects. We should expect to find evidence of anticipatory processing, or a way in which the neuron effectively manages to change its "starting point" before the stimulus is presented. Figure 1.8b depicts a very visible way of elevating the baseline, with neural activity ramping up toward the expected arrival of a stimulus, more so when biased than when neutral.

Perhaps the most canonical way in which we can open the door for bias to influence decision making in a given experimental paradigm is to play with the probability of events. In doing so, we manipulate the prior probability, to

use the terminology of Bayes's theorem. Carpenter and Williams (1995) had human observers make eye movements to peripheral visual stimuli during several blocks of trials. In each block, the likelihood that a stimulus would appear in any one trial was kept constant, from very likely (95%) to very unlikely (5%). The response time data matched nicely with the predictions of the LATER model, suggesting that with very likely stimuli, the decision line's journey toward the threshold was much shorter, and so the response much faster, than with very unlikely stimuli.

In the late 1990s researchers in several laboratories performed essentially the same experiment with monkeys while recording the activity levels of single neurons (Basso & Wurtz, 1997; Dorris & Munoz, 1998; Platt & Glimcher, 1999). In each case, the activity level of neurons was enhanced for stimuli or saccades whose prior probability was higher than that of other stimuli or saccades, even before the visual stimulus was presented (Dorris & Munoz, 1998) or before the monkey received an instruction about which of two possible stimuli was the actual target (Platt & Glimcher, 1999). The study by Platt and Glimcher deserves special mention as it showed data from three different experiments—not just the probability manipulation—and couched the entire data set in a then-unheard-of language, applying concepts from economics to the analysis of neural activity (read Glimcher, 2003, for the full introduction to the science of "neuroeconomics").

We will certainly have to return to the paper by Platt and Glimcher, but in the meantime, seeing as the Law of Eponymy is out of the window, we might as well highlight the massive contribution by Robert H. Wurtz, one of the authors of the cited 1997 paper on probability. In addition to mentoring a host of important researchers (including Michael E. Goldberg, Okihide Hikosaka, Douglas P. Munoz, William T. Newsome, Barry J. Richmond, and Marc A. Sommer, to name a nonrandom few), Wurtz compiled an impressive set of studies, showing time and again that neurophysiology—more specifically, the extracellular recording of action potentials from single neurons in awake and task-performing animals—can be applied with great effect to the study of elusive mechanisms operating somewhere in the big divide between sensory processing and motor control. The discoveries included "attention" (Goldberg & Wurtz, 1972), "memory" (Hikosaka & Wurtz, 1983), and "internal monitoring of movements" (Sommer & Wurtz, 2002).

For anyone who has witnessed or conducted this type of experiment, it is hard not to be amazed by the immediacy and precision with which it provides a window to the mental events that take place in the infamous black box, or the dark recesses beneath the skull. The first time I saw it, I was profoundly disoriented, unable to imagine how anyone could begin to invent a paradigm

like that. In a brief and very readable story, specked with Nobel prizes, Charles Gross (1998) traced the historical origins of the technique back to Adolf Beck at the University of Kracow in the 1880s, who worked with rabbits and dogs, mapping visually evoked responses with field potentials in occipital cortex. It took another few geniuses, including E. D. Adrian and Stephen W. Kuffler, to move from field potentials to measuring things like "the receptive fields of single neurons in the cat's striate cortex" (Hubel & Wiesel, 1959; the paper is easily and freely accessed in digital form, courtesy of the *Journal of Physiology*). Gross's (1998) historical account ends at this point, but the science went on evolving.

By the 1960s, Edward V. Evarts (1966, 1968) was able to record from awake and task-performing monkeys. In September 1969, Wurtz published his first article in *Journal of Neurophysiology*, on the "visual receptive fields of striate cortex neurons in awake monkeys." It was the first of 66 in the same journal ("*The Journal of Wurtz*"), spanning four decades of total focus on the neurophysiological underpinnings of visual processing and eye movements in monkeys. Since the original paper in that special year of 1969 (the paper was published after, but submitted before, I was born), there registered no essential changes to the experimental paradigm: The monkey sat in front of a screen, looked for dots, and made eye movements, while Wurtz and his collaborators recorded the activity of single neurons.

Today, the technique remains arguably the most powerful method to study information processing in the brain, providing a temporal and spatial resolution far beyond what can be reached with other methods while subjects are making decisions. In principle, the technique allows researchers to compute trial-by-trial correlations between behavioral response times and neural activity, measured on a continuous time scale (down to milliseconds, enough to pick up each individual action potential) and at the level of single neurons (down to micrometers). Much of what I have learned about neural mechanisms of decision making is based on the firing rates of individual neurons, and so Wurtz-like papers (more commonly called "single-unit studies") will feature heavily among my references. This is not to say all is well with the technique.

A cautious ethical note must be attached. The invasive nature of the technique drives researchers to work with animals other than humans—a move that is not appreciated by everyone in the same way. The present monograph is hardly the place to elaborate on the issue, but the minimal stance, implied also in the U.S. Animal Welfare Act, should be to look for alternatives wherever possible. Whether other techniques are viable replacements depends on the topic under investigation and on the level of precision required. In some cases, we can take brain scans to measure the cerebral blood flow in humans

as they perform tasks, via functional magnetic resonance imaging (fMRI) or positron emission tomography (PET). Especially, fMRI has come to the fore quite vigorously in the past ten years or so. There will be a good portion of fMRI studies among my references as well. With fMRI, we trade temporal resolution (in seconds) as well spatial resolution (in millimeters) for a major improvement in external validity—working with the right species, drawing no blood, and applying decision-making tasks that go from anything a fruit fly can do to things that only the smartest of us can do. Though the relation between cerebral blood flow and neural activity is yet to be determined precisely, there can be no doubt that the so-called blood-oxygen-level-dependent (BOLD) signal in fMRI does in fact provide a reliable parametric estimate of the extent of neural processing in a given brain structure (see Logothetis, 2008, for the state of the art).

In other situations, we might wish to measure "brain waves," or global electrical activity, from the scalp, via electroencephalography or its newer cousin, magnetoencephalography, with good temporal resolution (down to milliseconds) but poor spatial resolution (at the level of entire lobes at best). We can also learn a great deal from how the brain responds to drugs or more damaging assaults, either induced experimentally in an animal or occurring naturally as when one of us suffers a stroke or gets injured in an accident. Some kind of convergent approach seems the obvious best solution, working from multiple angles and with different methods simultaneously. Neuroscience certainly benefits from its wide variety of research tools and paradigms, and I will draw on any of them in my attempt to provide a coherent account of how neural circuits underscore decision making.

Coming back to the Wurtz-like studies, one recent development that certainly has my sympathy is a gradual shift toward a different species—monkeys still dominate the decision-making scene, but rats are gaining fast (e.g., Houweling & Brecht, 2008; Kepecs et al., 2008; Pan et al., 2005; Roesch, Calu, & Schoenbaum, 2007). Switching to rats creates a magnificent opportunity for a more integrated systems-neuroscience approach, including pharmacological, anatomical, genetic, and intracellular electrophysiological techniques that are too costly with monkeys, in both ethical and financial terms. Relying on single-unit studies with rats versus fMRI studies with humans, it should be possible to significantly reduce the future need for monkey research. At present, however, we should fully acknowledge the crucial role of monkey single-unit studies in the accumulation of our database on the neural mechanisms of decision making.

Thus reinvigorated, we pick up the paper by Basso and Wurtz (1997) again, and appreciate its demonstration of how "the role of the prior" is translated

into systematic variation of anticipatory processing in individual neurons. The monkey was required to make an eye movement to a visual target that could appear at one out of a predetermined set of possible locations, indicated by placeholders. The set size changed on a trial-by-trial basis, from one to eight possible locations. Basso and Wurtz recorded from buildup neurons in the superior colliculus (a type of neuron first identified by Munoz & Wurtz, 1995). Was it a strategic choice to record from buildup neurons? In hindsight, it definitely seemed the perfect pick. If we would like to find evidence for bias in anticipatory processing of a form like that presented in figure 1.8, then it makes total sense to focus on neurons that naturally tend to ramp up their activity levels in preparation for task events. And, sure enough, the baseline activity of a typical buildup neuron, in response to a placeholder in its receptive field, increased as target uncertainty decreased, well before the actual target appearance.

More recently, similar manipulations of target uncertainty have yielded the predicted type of differences in the baseline activity of other neural structures not specifically associated with buildup activity. A single-unit study in monkeys showed the effect in lateral intraparietal cortex during a motion-discrimination task (Churchland, Kiani, & Shadlen, 2008), whereas an fMRI study showed increased activity in extrastriate and anterior temporal lobe regions when human observers needed to compare the orientation of Gabor patches against one alternative rather than two (Summerfield & Koechlin, 2008). The effects of probability, then, do seem to accord with the neural signature of bias as pictured in figure 1.8. The search is on for other determinants of bias.

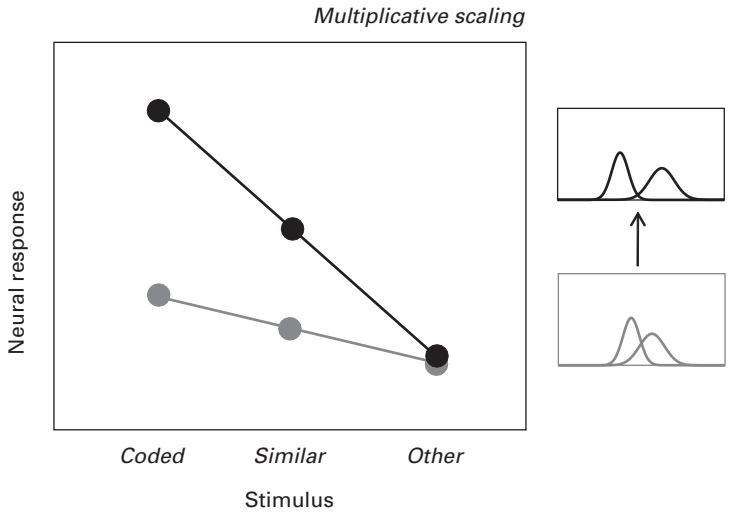
As a corollary of the studies on probability, there appears to be a conspicuously systematic, inverse relation between the number of alternatives and the ease of decision making. In signal detection theory and the LATER model, however, the actual number of alternatives seems to be abstracted away, as decision making is translated into a contest between signal and noise—to be or not to be, in the parlance of the Prince of Denmark. In defense of the reductive attitude, we could point to the etymology of the very word "decision," from the Latin *decidere*, or *de* + *caedere*, "to cut off." It may not be entirely clear what the ancient Romans were in the business of cutting off, but I would prefer to take the least bloody interpretation, as in putting an end to nothing more material than a thinking process, or the internal agonizing over different options. Perhaps the threshold theory really is thousands of years old. In any case, the general implication seems to be that decision making is all about reaching something final or definitive, a conclusion, a solution, an outcome, a statement, a proposition, a value on the color map, a number from one to a hundred, a judgment of character, a sentence with three subordinate clauses,

some *thing*, of which there usually is only one, even if it is a highly convoluted one, or something tricky, operating on a metalevel, like the decision that there will be no decision, as when a legal court rules that it has no jurisdiction over the matter at hand.

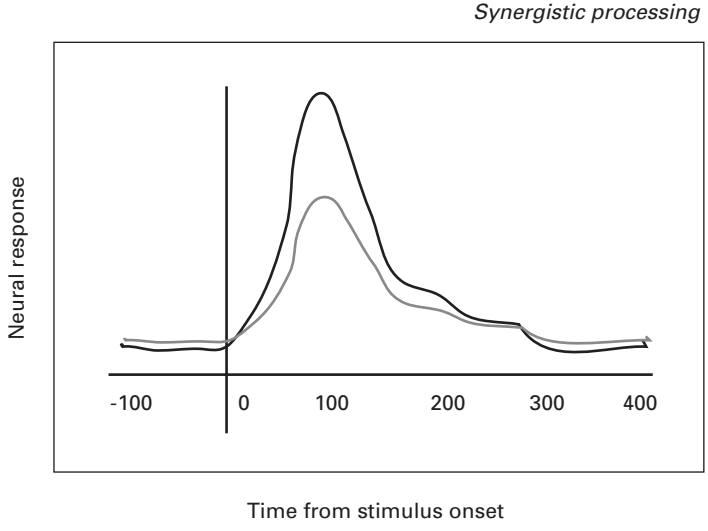
The rigorous one-track mind, holding onto the necessary singular of *the* decision, can capture the weighing of multiple alternatives as a parallel competition among different "decision units"—as if we have not one but several LATER functions running at the same time, one for each alternative, such that what constitutes a "signal" for unit A is actually part of the "noise" for unit B. As soon as one of the LATER functions reaches its threshold, it would take home the big prize, and be proclaimed the winner—the decision outcome. Or maybe we employ another algorithm to identify the solution. Or maybe the different decision units interact or influence each other. In any case, the notion of different decision units, working in parallel or interactively, helps us bend the two-choice logic of signal detection theory and the LATER model to suit the needs of any decision-making situation. In chapter 6, "Less Is More," I will return to the multiplicity of decision processes at any one point in time, no matter how dormant or awake they are.

However, now it is about time we take a look at figure 1.9 and learn to recognize the neural signature of sensitivity. The presentation format is the same as that of figure 1.8, with a hypothetical neuron in a factorial design. In panel a, we have the three types of stimulus (coded, similar, and other), and two treatments—this time, increased sensitivity (black data) versus neutral (gray data). Increased sensitivity is achieved by reducing the overlap, or widening the distance, between the signal and noise distributions according to the proposals of signal detection theory. Applied to the tuning curve, this produces an enlarged ratio of the response to the "coded" stimulus relative to any "other" stimulus. Put differently, in absolute terms the effect of sensitivity on neural firing should be larger for the "coded" stimulus than for any "other" stimulus. This corresponds to a multiplicative scaling effect, as if the tuning curve is multiplied by a constant sensitivity factor greater than one—effects of this kind are sometimes tagged as "gain changes," although the term "gain" remains ambiguous with respect to the additive versus multiplicative nature of the effect, blurring the difference between bias and sensitivity.

Distinguishing between these two mechanisms, as do the LATER model and signal detection theory, is a sine qua non if we wish to unravel how the brain provides us with the computational power to make decisions. Distinguishing between additive and multiplicative effects on tuning curves should be very useful indeed and might surely be practiced more often. Of course, the additions and multiplications will rarely work out perfectly in the quirky



a



b

Figure 1.9
The neural signature of sensitivity. (a) Here, the black data, driven by increased sensitivity, show a multiplicative effect as compared to the neutral case. The two inset figures to the right are borrowed from figure 1.4. (b) The temporal dynamics of the neural response to a "coded" stimulus. The difference between the data driven by increased sensitivity (black line) versus those from the neutral case (gray line) emerges only after stimulus onset, reflecting synergistic processing.

reality of empirical data. For instance, bias might add a smaller absolute amount for the "coded" stimulus than for "other" stimuli when the neural firing rate hits its maximum capacity. The addition is then obscured by a "ceiling effect," leading to a signal-to-noise ratio that would even deteriorate under the bias regime as compared to the neutral case. Conversely, sensitivity might outdo multiplication and show exponential growth. Nevertheless, such deviations and complications can be formulated precisely in computational terms. The main point is that the logic of bias versus sensitivity does translate into differential movements on tuning curves. The peculiar fluxions are there for us to check up on in the data and to use strategically as markers and diagnostics for the involvement of this or that underlying neural mechanism.

In figure 1.9a, then, we note that the black data, driven by increased sensitivity, swivel upwards from the neutral case, with a degree of change that depends on the actual stimulus being presented. Here, there must be some kind of ad hoc interaction between the signal processing and the mechanism of sensitivity. This makes perfect sense, of course—the changed signal-to-noise ratio can only become visible when there is, in fact, a signal to be processed in the first place. In the LATER model, we can easily appreciate what the interaction does for the temporal dynamics of sensitivity effects. The steeper gradient can only take effect once there is some sensory information to work with, that is, from the moment of stimulus onset. The incoming sensory information and the increased sensitivity work together, simultaneously, interactively, synergistically—taking "synergy" in its early sense, derived from the Greek *sunergos*, "working together," instead of the "rather blowsy word" it is "these days, with its implications of corporate merger for profit-enhancing capacity" (dixit the word-cleaning poet, Michael Palmer, 2008, p. 28). Figure 1.9b shows how the neural response to a "coded" stimulus reaches a much higher amplitude with increased sensitivity (black line) as compared to the neutral case (gray line). Yet, the difference in the neural response emerges only after stimulus onset.

In practice, we can easily examine the neural correlates of increased sensitivity by physically modifying the degree of similarity between signal and noise. The most thorough investigation of this kind was, and is, being conducted using a perceptual discrimination task with different levels of motion coherence, in which the subject has to report the dominant direction among a set of moving dots. The task is easy enough when all dots move in the same direction (100% coherence) and obviously is impossible when the dots move in random directions (0% coherence). Between these two poles, performance improves steadily with higher coherence levels. In terms of response times, the improvement should be attributed to changes to the gradient of the decision

line in the LATER model, as confirmed in a study by Reddi, Assress, and Carpenter (2003). However, the coherence levels also affect neural activity in exactly the way we would expect, with neurons reaching higher firing rates for easily discriminated stimuli that match the "coded" direction. (In addition to the already cited single-unit work by Churchland, Kiani, & Shadlen, 2008, in which probability and motion coherence played in concert, the landmark studies employing motion coherence were performed by Britten et al., 1992; Newsome, Britten, & Movshon, 1989; Roitman & Shadlen, 2002; and Shadlen et al., 1996; for an fMRI version, see Heekeren et al., 2006.) The neural signature of increased sensitivity bears out perfectly in these data, both the multiplicative scaling and the synergistic processing.

Arguably the most thorough analysis of increased sensitivity in neural firing rates as well as response times was performed by Ratcliff and colleagues (2007) on the basis of data from superior colliculus neurons while the monkey performed a brightness-discrimination task, in which some levels of brightness were easy to discriminate (98% white pixels, very "bright," or 2% white pixels, very "dark"), others hard (45% or 55% white pixels). Response times were fast and neural firing rates high for easy "coded" stimuli, and again the effects in neural processing emerged only after stimulus onset. But more than this, the study reached an unprecedented level of detail in modeling the trial-by-trial variation of both response times and neural activity—a great achievement, exactly the type of triangulation that forms my ideal of neuroscience.

Armed with the analytic tools to distinguish bias versus sensitivity, familiar with the Bayesian way of thinking about decision making, no longer afraid of signal detection theory and the LATER model, always on the lookout for anticipatory versus synergistic processing, and eager to compare additive versus multiplicative scaling, we are now ready to investigate how neural circuits really weigh the options, in what kind of conditions, under what sort of circumstances, and to what degree of inevitability. Having duly sniffed at the formulas, we can finally take a look at how they function.