
How to Make Cognitive Illusions Disappear

Social psychology was transformed by the “cognitive revolution.” Cognitive imperialism has been both praised and lamented. But a second revolution has transformed most of the sciences so fundamentally that it is now hard to see that it could have been different before. It has made concepts such as probability, chance, and uncertainty indispensable for understanding nature, society, and the mind. This sweeping conceptual change has been called the “probabilistic revolution” (Krüger, Daston, & Heidelberger, 1987; Krüger, Gigerenzer, & Morgan, 1987). The probabilistic revolution differs from the cognitive revolution in its genuine novelty and its interdisciplinary scope. Statistical mechanics, Mendelian genetics, Brownian motion, radioactive decay, random drift, randomized experimental design, statistical inference—these are some of the fruits of that transformation. Social psychology was no exception. It currently bears the marks of both the cognitive revolution and the probabilistic revolution.

Probabilistic and statistical concepts were piggybacked onto cognitive concepts. Some of the most popular theories and research programs owed their genesis to an analogy between social cognition and “intuitive statistics.” In 1967, for instance, Harold Kelley proposed that the layperson attributes a cause to an effect in the same way as a statistician of the Fisherian school would, by (unconsciously) calculating an analysis of variance (ANOVA). Research on the ANOVA mind soon became mainstream social psychology (Kelley & Michaela, 1980). It is well documented in the history of science that statistics *transformed* almost everything it touched. So has causal attribution (Chapter 1). Just as statistical calculations are those of an *individual* statistician, attribution and social cognition were investigated as the calculations of individual minds, confirming the individualism in social psychology (Newcombe & Rutter, 1982).

More recently, Bayesian statistics, rather than Fisherian statistics, has been used as a yardstick to evaluate social cognition, and as measured by this new yardstick, many people’s judgments seemed to be flawed by fallacies and errors in statistical reasoning. “Hot” motivational terms were replaced by the “cold”

cognitive language of intuitive statistics. Self-serving perceptions and attributions, ethnocentric beliefs, and many types of human conflict were analyzed as passionless information-processing errors, due to basic shortcomings in intuitive statistical reasoning (e.g., Borgida & Brekke, 1981; Nisbett & Ross, 1980; Sherman, Judd, & Park, 1989). Social cognitive psychologists started to study (what they believed to be) errors in probabilistic reasoning, such as the base-rate fallacy, the conjunction fallacy, and overconfidence bias, and adopted the explanatory language of Kahneman and Tversky's "heuristics," such as representativeness and availability. Some, such as Strack (1988), even pointed to Kahneman and Tversky's heuristics as primary evidence of the end of the "crisis" of social psychology and of new, rising confidence and decisive progress in the field.

Heuristics and Biases

The "heuristics and biases" program of Kahneman, Tversky, and others has generated two main results concerning judgment under uncertainty: (1) a list of so-called biases, fallacies, or errors in probabilistic reasoning, such as the base-rate fallacy and the conjunction fallacy, and (2) explanations of these biases in terms of cognitive heuristics such as representativeness. Table 12.1 gives a taste of the conclusions drawn from this program.

Kahneman and Tversky (1982) see the study of systematic errors in probabilistic reasoning, also called "cognitive illusions," as similar to that of visual illusions. "The presence of an error of judgment is demonstrated by comparing people's responses either with an established fact (e.g., that the two lines are equal in length) or with an accepted rule of arithmetic, logic, or statistics" (p. 493). Their distinction between "correct" and "erroneous" judgments under uncertainty has been echoed by many social psychologists: "We follow conventional practice by using the term 'normative' to describe the use of a rule when there is a consensus among formal scientists that the rule is appropriate for the particular problem" (Nisbett & Ross, 1980, p. 13).

Social psychology is not the only area in which the "heuristics and biases" program has made strong inroads. Experimental demonstrations of "fallacious" judgments have entered law (e.g., Saks & Kidd, 1980), economics (e.g., Frey, 1990), management science (e.g., Bazerman, 1990), medical diagnosis (e.g., Casscells, Schoenberger, & Grayboys, 1978), behavioral auditing (see Shanteau, 1989), philosophy (e.g., Stich, 1990), and many other fields. There is no doubt that understanding judgment under uncertainty is essential in all these fields. It is the achievement of the "heuristics and biases" program to have finally established this insight as a central topic of psychology. Earlier pioneers who studied intuitive statistics (Hofstätter, 1939; Peirce & Jastrow, 1884; Wendt, 1966) had little impact. Even Ward Edwards and his colleagues (e.g., Edwards, 1968), who started the research from which Kahneman and Tversky's "heuris-

Table 12.1 A sample of conclusions from the *heuristics-and-biases* program

In making predictions and judgments under uncertainty, people do not appear to follow the calculus of chance or the statistical theory of prediction. Instead, they rely on a limited number of heuristics which sometimes yield reasonable judgments and sometimes lead to severe and systematic errors.

Daniel Kahneman & Amos Tversky, 1973, p. 237

It appears that people lack the correct programs for many important judgmental tasks. . . . We have not had the opportunity to evolve an intellect capable of dealing conceptually with uncertainty.

Paul Slovic, Baruch Fischhoff, & Sarah Lichtenstein, 1976, p. 174

The genuineness, the robustness, and the generality of the base-rate fallacy are matters of established fact.

Maya Bar-Hillel, 1980, p. 215

The biases of framing and overconfidence just presented suggest that individuals are generally affected by systematic deviations from rationality.

Max Bazerman & M. A. Neale, 1986, p. 317

[Overconfidence bias] has proved so robust that it is hard to acquire much insight into the psychological processes underlying it.

Baruch Fischhoff, 1988, p. 172

. . . mental illusions should be considered the rule rather than the exception.

Richard H. Thaler, 1991, p. 4

[We are] a species that is uniformly probability-blind, from the humble janitor to the Surgeon General. . . . We should not wait until A. Tversky and D. Kahneman receive a Nobel prize for economics. Our self-liberation from cognitive illusions ought to start even sooner.

Massimo Piattelli-Palmarini, 1991, p. 35

tics and biases" program emerged, had no comparable influence on cognitive and social psychology.

Despite its influence, I will argue that the "heuristics and biases" program is merely an important transitional stage, which must be transformed if long-term progress is to be made. I will review some serious shortcomings of that research program and show how they can be overcome.

In this chapter I do three things. First, I discuss the validity of the normative yardstick that is used to define people's judgments as systematic errors in probabilistic reasoning. I will argue that most so-called errors or cognitive illusions are, contrary to the assertions in the literature, in fact *not* violations of probability theory. In their normative claims, Tversky and Kahneman, and social psychologists following in their footsteps, have neglected conceptual distinctions that are fundamental to probability and statistics. Second, I show that if we pay attention to these conceptual distinctions, we can make apparently stable "cognitive illusions" disappear, reappear, or even invert. Third, the interesting fact that intuitive reasoning is highly sensitive to conceptual distinctions made by statisticians (but ignored by many psychologists) leads to a revised understanding of judgment under uncertainty.

Why Biases Are Not Biases

In the “heuristics and biases” program, a bias or error in probabilistic reasoning is defined as a systematic discrepancy between a person’s judgment and a norm. What is that norm? It is often referred to as “the normative theory of prediction” (Kahneman & Tversky, 1973, p. 243), as the “normative principles of statistical prediction” (Ajzen, 1977, p. 304), or simply as an “accepted rule” of statistics. Many have understood this rhetoric to imply that there exists precisely one “correct” answer to the cab problem, engineer–lawyer problem, Linda problem, and other problems posed to participants—an answer sanctioned by the authority of the eminent mathematicians, probabilists, and statisticians of this century. The claim that all these problems have *one* correct answer is crucial. If they did *not* have one and only one answer, it would make little sense first to identify “errors” and “cognitive illusions” and then to use these cognitive illusions to understand the principles of inductive reasoning, in the way that visual illusions are used to understand the principles of normal perception. This two-step program, identifying errors and explaining them, in analogy to perceptual research, is the basic idea behind the heuristics-and-biases program (Kahneman & Tversky, 1982, p. 493).

But what does the “heuristics and biases” investigation of judgment under uncertainty have to do with probability and statistics? The short answer to this question is: all too little. The probabilistic rules against which cognitive and social psychologists have measured the proficiency of their participants are in fact a highly (and, I shall argue, often misleadingly) selected sample of those routinely used, consulted, and discussed by working probabilists and statisticians. When claiming “errors” and “fallacies,” cognitive and social psychologists have largely ignored conceptual and technical distinctions *fundamental* to probability and statistics.

What in the heuristics-and-biases literature is called the “normative theory of probability” or the like is in fact a very narrow kind of neo-Bayesian view that is shared by some theoretical economists and cognitive psychologists and to a lesser degree by practitioners in business, law, and artificial intelligence. It is *not* shared by proponents of the frequentist view of probability that dominates today’s statistics departments, nor by proponents of many other views; it is not even shared by all Bayesians, as I shall show shortly. By this narrow standard of “correct” probabilistic reasoning, the most distinguished probabilists and statisticians of our century—figures of the stature of Richard von Mises and Jerzy Neyman—would be guilty of “biases” in probabilistic reasoning.¹ Let me illustrate this point with some of the best-known demonstrations of “fallacies.”

1. Despite the widespread rhetoric of a single “normative theory of prediction,” it should be kept in mind that the problem of inductive reasoning still has no universal solution (the “scandal of philosophy”) but many competing ones. The controversies

Overconfidence Bias

Confidence in general knowledge is typically studied with questions of the following kind:

Which city has more inhabitants?
(a) Hyderabad, (b) Islamabad
How confident are you that your answer is correct?
50% / 60% / 70% / 80% / 90% / 100%

The participant chooses what he or she believes is the correct answer and then rates his or her confidence that the answer is correct. After many participants answer many questions, the experimenter counts how many answers in each of the confidence categories were actually correct. The typical finding is that in all the cases in which participants said, “I am 100% confident that my answer is correct,” the relative frequency of correct answers was only about 80%; in all the cases in which they said, “I am 90% confident that my answer is correct,” the relative frequency of correct answers was only about 75%, and so on (for an overview, see Lichtenstein, Fischhoff, & Phillips, 1982). This systematic discrepancy between confidence and relative frequency is termed “overconfidence.”

Little has been achieved in explaining this “bias.” A common proposal is to explain “biases” by other, deeper mental flaws. For instance, Koriat, Lichtenstein, and Fischhoff (1980) proposed that the overconfidence bias is caused by a “confirmation bias.” Their explanation was this: After one alternative is chosen, the mind searches for further information that *confirms* the answer

between the Fisherians, the Neyman-Pearsonians, and the Bayesians are evidence of this unresolved rivalry. For the reader who is not familiar with the fundamental issues, two basic themes may help introduce the debate (for more, see Hacking, 1965). The first issue relevant for our topic is whether probability is *additive* (that is, satisfies the Kolmogorov axioms, e.g., that the probabilities of all possible events sum up to 1) or not. The above-mentioned points of view (including that of the heuristics-and-biases program) subscribe to additivity, whereas L. J. Cohen’s (e.g., 1982) Baconian probabilities are nonadditive (for more on nonadditive theories, see Shafer, 1976). In my opinion, Cohen correctly criticizes the normative claims in the heuristics-and-biases program insofar as not all uses of “probability” that refer to single events must be additive—but this does not imply that Baconian probability is the only alternative, nor that one should assume, as Cohen did, that all minds reason rationally (or at least are competent to do so) in all situations. I do not deal with this issue in this chapter (but see Gigerenzer, 1991d). The second fundamental issue is whether probability theory is about *relative frequencies in the long run* or (also) about *single events*. For instance, the question “What is the relative frequency of women over 60 who have breast cancer?” refers to frequencies, whereas “What is the probability that Ms. Young has breast cancer?” refers to a single event. Bayesians usually assume that (additive) probability theory is about single events, whereas frequentists hold that statements about single cases have nothing to do with probability theory (they may be dealt with by cognitive psychology, but not by probability theory).

given, but not for information that could falsify it. This selective information search artificially increases confidence. The key idea in this explanation is that the mind is not a Popperian. Despite the popularity of the confirmation bias explanation in social psychology, there is little or no support for this hypothesis in the case of confidence judgments (see Chapter 7).

As with many "cognitive illusions," overconfidence bias seems to be a robust fact waiting for a theory. This "fact" was quickly generalized to account for human disasters of many kinds, such as deadly accidents in industry (Spettell & Liebert, 1986), confidence in clinical diagnosis (Arkes, 1981), and shortcomings in management and negotiation (Bazerman, 1990) and in the legal process (Saks & Kidd, 1980), among others.

The Normative Issue Is overconfidence bias really a "bias" in the sense of a violation of probability theory? Let me rephrase the question: Has probability theory been violated if one's *degree of belief (confidence) in a single event* (i.e., that a particular answer is correct) is different from the *relative frequency* of correct answers one generates in the long run? The answer is "no." It is in fact not a violation according to several interpretations of probability.

Let us look first at the now dominant school of probability: the frequentists (the frequentist interpretation of probability has been dominant since about 1840; see Daston, 1988; Porter, 1986). Most readers of this chapter will have been trained in the frequentist tradition and, for instance, will have been taught that the probabilities of Type I and Type II errors are long-run frequencies of errors in repeated experiments, not probabilities of single outcomes or hypotheses. For a frequentist like the mathematician Richard von Mises, the term "probability," when it refers to a *single event*, "has no meaning at all for us" (1928/1957, p. 11). For predictions of single events, as studied in present-day overconfidence research, he put the issue in crystal-clear terms: "Our probability theory has nothing to do with questions such as: 'Is there a probability of Germany being at some time in the future involved in a war with Liberia?'" (p. 9). In this view, probability theory is about frequencies, not about single events. To compare the two means comparing apples with oranges.

Even the major opponents of the frequentists—subjectivists such as Bruno de Finetti—would not generally think of a discrepancy between confidence and relative frequency as a "bias," albeit for different reasons. For a subjectivist, probability is about single events, but rationality is identified with the internal consistency of subjective probabilities. As de Finetti emphasized, "however an individual evaluates the probability of a particular event, no experience can prove him right, or wrong; nor, in general, could any conceivable criterion give any objective sense to the distinction one would like to draw, here, between right and wrong" (1931/1989, p. 174).

Other theories and interpretations of probability are also at odds with the claim that overconfidence is a bias, that is, a violation of probability theory. But I will stop here and summarize the normative issue. A discrepancy between confidence in single events and relative frequencies in the long run is not an error or a violation of probability theory from many experts' points of

view. It only looks like it from a narrow interpretation of probability that blurs the distinction between single events and frequencies fundamental to probability theory. (The choice of the word "overconfidence" for the discrepancy put the "fallacy" message into the term itself.)

How to Make the Cognitive Illusion Disappear If there are any robust cognitive biases at all, overconfidence in one's knowledge would seem to be a good candidate. "Overconfidence is a reliable, reproducible finding" (von Winterfeldt & Edwards, 1986, p. 539). "Can anything be done? Not much" (Edwards & von Winterfeldt, 1986, p. 656). "Debiasing" methods, such as warning the participants of the overconfidence phenomenon before the experiment and offering them money to avoid it, have had little or no effect (Fischhoff, 1982).

Setting the normative issue straight has important consequences for understanding confidence judgments. Let us go back to the metaphor of the mind as an intuitive statistician. I now take the term "statistician" to refer to a statistician of the dominant school in this (and in the last) century, not one adopting the narrow perspective some psychologists and economists have suggested. Assume that the mind is a frequentist. Like a frequentist, the mind should be able to *distinguish* between single-event confidences and frequencies in the long run.

This view has testable consequences. Ask people for their estimated frequencies of correct answers and compare them with true frequencies of correct answers, instead of comparing the latter frequencies with confidences. We are now comparing apples with apples. Ulrich Hoffrage, Heinz Kleinbölting, and I carried out such experiments. Participants answered several hundred questions of the Islamabad-Hyderabad type (see above), and, in addition, estimated their frequencies of correct answers.

Table 12.2 (top row) shows the usual "overconfidence bias" when single-event confidences are compared with actual relative frequencies of correct answers. In both experiments, the difference was around 13 to 15 percentage points, which is a large discrepancy. After each set of 50 general knowledge

Table 12.2 How to make the *overconfidence bias* disappear

Difference between	Experiment 1 (n = 80)	Experiment 2 (n = 97)
Mean confidence and relative frequency of correct answers ("overconfidence bias")	+13.8	+15.4
Estimated frequency and frequency of correct answers	-2.4	-4.2

Note: To make values for frequency and confidence judgments comparable, all frequencies were transformed to relative frequencies. Values shown are differences multiplied by a factor of 100. Positive values denote "overconfidence" (Gigerenzer, Hoffrage, & Kleinbölting, 1991).

questions, we asked the same participants, "How many of these 50 questions do you think you got right?" Comparing their estimated frequencies with actual frequencies of correct answers made "overconfidence" disappear. Table 12.2 (second row) shows that estimated frequencies were practically identical with actual frequencies, with even a small tendency toward underestimation. The "cognitive illusion" was gone. Similar results were obtained when participants estimated the relative frequencies of correct answers in each confidence category. In all cases in which participants said they were "100% (90%, 80%, . . .) confident," they estimated that, in the long run, they had a lower percentage of answers correct, and their estimates were close to the true relative frequencies of correct answers (May, 1987, reported similar results). Eliminating the experimenter's normative confusion between single events and frequencies made the participants' "overconfidence bias" disappear.

The general point is (i) a discrepancy between probabilities of single events (confidences) and long-run frequencies need not be framed as an "error" and called "overconfidence bias," and (ii) judgments need not be "explained" by a flawed mental program at a deeper level, such as "confirmation bias." Rather, people seem to be able intuitively to make conceptual distinctions similar to those that professional statisticians make. How they do it can be accounted for by the theory of "probabilistic mental models" (PMM), which explains both confidence and frequency judgments in terms of frequentist probability cues (Chapter 7). PMM theory is a frequentist theory of judgment and uncertainty; it can predict overconfidence, good calibration, and underestimation within the same participant.

Conjunction Fallacy

The original demonstration of the "conjunction fallacy" was with problems of the following kind (Tversky & Kahneman, 1983, p. 299):

Linda is 31 years old, single, outspoken and very bright. She majored in philosophy. As a student, she was deeply concerned with issues of discrimination and social justice, and also participated in antinuclear demonstrations.

Participants were asked which of two alternatives was more probable:

- Linda is a bank teller (T)
- Linda is a bank teller and is active in the feminist movement (T&F)

Eighty-five percent of the participants chose T&F in the Linda problem (see Table 12.3). Tversky and Kahneman, however, argued that the "correct" answer is T, because the probability of a conjunction of two events, such as T&F, can never be greater than that of one of its constituents. They explained this "fallacy" as induced by the representativeness heuristic. They assumed that judgments were based on the match (similarity, representativeness) between the description of Linda and the two alternatives T and T&F. That is, since

Table 12.3 Linda problem: How to make the conjunction fallacy disappear

Linda problem	Single-event probability	Frequency
Tversky & Kahneman (1983)		
Which is more probable?	85	—
Probability ratings	82	—
Probability ratings T*	57	—
Betting	56	—
Fiedler (1988)		
Exp. 1	91	22
Exp. 2	83	17
Hertwig & Gigerenzer (1999)		
Studies 1 and 3	83	0
Studies 2 and 4	88	13

Note: Numbers are violations (in %) of the conjunction rule. The various versions of the Linda problem are (i) which is more probable (see text), (ii) probability ratings on a 9-point scale, (iii) probability ratings using the alternative "Linda is a bank teller whether or not she is active in the feminist movement" (T*) instead of "Linda is a bank teller" (T), (iv) hypothetical betting, that is, participants were asked "If you could win \$10 by betting on an event, which of the following would you choose to bet on?" Fiedler asked participants to rank order T, T&F, and other alternatives with respect to their probability. In his first frequency version the population size was always 100, in the second it varied. Hertwig and Gigerenzer asked participants to rank order T, T&F, and F with respect to their probability, or estimate their frequency. Tversky and Kahneman (1983, p. 309) had reported a facilitating effect of frequency judgments for a different problem.

Linda was described as if she were a feminist and T&F contains the term "feminist," people believe that T&F is more probable.

This alleged demonstration of human irrationality in the Linda problem has been widely publicized in psychology, philosophy, economics, and beyond. Stephen J. Gould (1992, p. 469) put the message clearly:

I am particularly fond of [the Linda] example, because I know that the [conjunction] is least probable, yet a little homunculus in my head continues to jump up and down, shouting at me, "but she can't just be a bank teller; read the description." . . . Why do we consistently make this simple logical error? Tversky and Kahneman argue, correctly I think, that our minds are not built (for whatever reason) to work by the rules of probability.

I suggest that Gould should have had more trust in the rationality of his homunculus.

The Normative Issue Is the "conjunction fallacy" a violation of probability theory, as has been claimed in the literature? Has a person who chooses T&F as the more probable alternative violated probability theory? Again, the answer

is "no." Choosing T&F is *not* a violation of probability theory, and for the same reason given previously. For a frequentist, this problem has nothing to do with probability theory. Participants were asked for the probability of a *single event* (that Linda is a bank teller), not for frequencies. For instance, the statistician Barnard (1979) commented thus on subjective probabilities for single events: "If we accept it as important that a person's subjective probability assessments should be made coherent, our reading should concentrate on the works of Freud and perhaps Jung rather than Fisher and Neyman" (p. 171).

Note that problems that are claimed to demonstrate the "conjunction fallacy" are structurally different from "confidence" problems. In the former, subjective probabilities (that Linda is a bank teller or a bank teller and a feminist) are compared with one another; in the latter, they are compared with frequencies.

To summarize the normative issue, what is called the "conjunction fallacy" looks like a violation of *some* subjective theories of probability, including Bayesian theory. It is not, however, a violation of a major view of probability, the frequentist conception.

How to Make the Cognitive Illusion Disappear What if the mind were a frequentist? If the untutored mind is as sensitive to the distinction between single cases and frequencies as a statistician of the frequentist school is, then we should expect dramatically different judgments if we pose the above problem in a frequentist mode, such as the following:

- There are 100 persons who fit the description above (i.e., Linda's).
How many of them are:
- (a) bank tellers
 - (b) bank tellers and active in the feminist movement.

Participants are now asked for frequency judgments rather than for single-event probabilities. If the mind solves the Linda problem by using a representativeness heuristic, changes in information representation should not matter because they do not change the degree of similarity. The description of Linda is still more representative of (or similar to) the conjunction "teller and feminist" than of "teller." Participants therefore should still exhibit the conjunction fallacy. Table 12.3, however, shows that with frequency judgments, the "conjunction fallacy" largely disappears. The effect is dramatic, from some 80% to 90% conjunction violations in probability judgments to 10% to 20% in frequency judgments, with one study even reporting 0%.

What accounts for this striking effect of frequency judgments? Hertwig and Gigerenzer (1999) analyzed how participants understood the phrase "which is more probable?", for instance, by asking them to paraphrase the problem to another person who is not a native speaker of the language in which the problem was presented. The results indicate that most participants did not understand "probability" in the sense of mathematical probability but as one of the many other legitimate meanings that are listed in, for example, the Oxford English Dictionary (e.g., meaning credibility, typicality, or that there is evi-

dence). The term frequency, unlike probability, narrows down the spectrum of possible interpretations to meanings that follow mathematical probability.

The results in Table 12.3 are consistent with the earlier work by Inhelder and Piaget (1969), who showed children a box containing wooden beads, most of them brown, but a few white. They asked the children, "Are there more wooden beads or more brown beads in this box?" By the age of eight, a majority of children responded that there were more wooden beads, indicating that they understand conjunctions (class inclusions). Note that Inhelder and Piaget asked children for frequency judgments, not probability judgments.

Base-Rate Fallacy

Among all cognitive illusions, the "base-rate fallacy" has probably received the most attention. The neglect of base rates seems in direct contradiction to the widespread belief that judgments are unduly affected by stereotypes (Landman & Manis, 1983), and for this and other reasons it has generated a great deal of interesting research on the limiting conditions for the "base-rate fallacy" in attribution and judgment (e.g., Ajzen, 1977; Borgida & Brekke, 1981). For instance, in their review, Borgida and Brekke argue for the pervasiveness of the "base-rate fallacy" in everyday reasoning about social behavior, ask the question "Why are people susceptible to the base-rate fallacy?" (1981, p. 65), and present a list of conditions under which the "fallacy" is somewhat reduced, such as "vividness," "salience," and "causality" of base-rate information.

My analysis is different. Again I first address the normative claims that people's judgments are "fallacies" using two examples that reveal two different aspects of the narrow understanding of good probabilistic reasoning in much of this research.

The first is from Casscells, Schoenberger, and Grayboys (1978, p. 999) and presented by Tversky and Kahneman (1982b, p. 154) to demonstrate the generality of the phenomenon:

If a test to detect a disease whose prevalence is 1/1000 has a false positive rate of 5%, what is the chance that a person found to have a positive result actually has the disease, assuming you know nothing about the person's symptoms or signs?

Sixty students and staff at Harvard Medical School answered this medical diagnosis problem. Almost half of them judged the probability that the person actually had the disease to be 0.95 (modal answer), the average answer was 0.56, and only 18% of participants responded 0.02. The latter was considered to be the correct answer. Note the enormous variability in judgments. Little has been achieved in explaining *how* people make these judgments and *why* the judgments are so strikingly variable.

The Normative Issue But do statistics and probability give one and only one "correct" answer to that problem? The answer is again "no." And for the same

reason, as the reader will already have guessed. As in the case of confidence and conjunction judgments, participants were asked for the probability of a *single event*, that is, that “a person found to have a positive result actually has the disease.” If the mind is an intuitive statistician of the frequentist school, such a question has no necessary connection to probability theory. Furthermore, even for a Bayesian, the medical diagnosis problem has several possible answers. One piece of information necessary for a Bayesian calculation is missing: the test’s long-run frequency of correctly diagnosing persons who have the disease (admittedly a minor problem if we can assume a high “true positive rate”). A more serious difficulty is that the problem does not specify whether or not the person was *randomly* drawn from the population to which the base rate refers. Clinicians, however, know that patients are usually not randomly selected—except in screening and large survey studies—but rather “select” themselves by exhibiting symptoms of the disease. In the absence of random sampling, it is unclear what to do with the base rates specified. The modal response, 0.95, would follow from applying the Bayesian principle of indifference (i.e., same prior probabilities for each hypothesis), whereas the answer 0.02 would follow from using the specified base rates and assuming random sampling. In fact, the range of actual answers corresponds quite well to the range of possible solutions.

How to Make the Cognitive Illusion Disappear The literature overflows with assertions of the generality and robustness of the “base-rate fallacy,” such as: “the base-rate effect appears to be a fairly robust phenomenon that often results from automatic or unintentional cognitive processes” (Landman & Manis, 1983, p. 87); and “many (possibly most) subjects generally ignore base rates completely” (Pollard & Evans, 1983, p. 124; see also Table 12.1). Not only are the *normative* claims often simplistic and, therefore, misleading, but so too are the *robustness* assertions.

What happens if we do something similar as for the “overconfidence bias” and the “conjunction fallacy,” that is, rephrase the medical diagnosis problem in a frequency format? Cosmides and Tooby (1996) did so. They compared the original problem (above) with a frequency format, in which the same information was given:

One out of 1000 Americans has disease X. A test has been developed to detect when a person has disease X. Every time the test is given to a person who has the disease, the test comes out positive. But sometimes the test also comes out positive when it is given to a person who is completely healthy. Specifically, out of every 1000 people who are perfectly healthy, 50 of them test positive for the disease.

Imagine that we have assembled a random sample of 1000 Americans. They were selected by a lottery. Those who conducted the lottery had no information about the health status of any of these people. How many people who test positive for the disease will actually have the disease?
 _____ out of _____

In this frequentist version of the medical diagnosis problem, both the information and the question are phrased in terms of frequencies. (In addition, the two pieces of information missing in the original version [see above] are supplied. In numerous other versions of the medical diagnosis problem, Cosmides and Tooby showed that the striking effect [see Table 12.4] on participants’ reasoning is mainly due to the transition from a single-event problem to a frequency format, and only to a lesser degree to the missing information.) Participants were Stanford University undergraduates.

If the question was rephrased in natural frequencies, as shown above, then the Bayesian answer of 0.02—that is, the answer “one out of 50 (or 51)” —was given by 76% of the participants. The “base-rate fallacy” disappeared. By comparison, the original single-event version elicited only 12% Bayesian answers in Cosmides and Tooby’s study. Chapter 6 provides an explanation for this effect.

Cosmides and Tooby identified one condition in which almost every participant found the Bayesian answer of 0.02. Participants received the frequentist version of the medical diagnosis problem (except that it reported a random sample of “100 Americans” instead of “1000 Americans”), and in addition a page with 100 squares (10×10). Each of these squares represented one American. Before the frequentist question “How many people who test positive . . .” was put, participants were asked to (i) circle the number of people who will have the disease and (ii) fill in squares to represent people who will test positive. After that, 23 out of 25 participants came up with the Bayesian answer (see frequency format, pictorial, in Table 12.4).

All three examples point in the same direction: The mind acts as if it were a frequentist; it distinguishes between single events and frequencies in the long run—just as probabilists and statisticians do. Despite the fact that researchers in the “heuristics and biases” program routinely ignore this distinction fundamental to probability theory when they claim to have identified “errors,” it

Table 12.4 How to make the “base-rate fallacy” disappear: The medical diagnosis problem

Medical diagnosis problem	<i>N</i>	Bayesian answers (%)
Original single-event version (Casscells, Schoenberger, & Grayboys, 1978)	60	18
Single-event version, replication (Cosmides & Tooby, 1996)	25	12
Frequency format (Cosmides & Tooby, 1996)	50	76
Frequency format, pictorial (Cosmides & Tooby, 1996)	25	92

would be foolish to label these judgments “fallacies.” These results not only point to a truly new understanding of judgment under uncertainty, but they also seem to be relevant for teaching statistical reasoning.

Selected versus Random Sampling: More on the Base-Rate Fallacy

Another conceptual distinction routinely used by probabilists and statisticians is that between random sampling and selected sampling. Again, little attention has been given to that distinction when intuitive statistical reasoning is investigated. The original medical diagnosis problem is silent about whether the patient was randomly selected from the population. That this crucial information is missing is not atypical. For instance, in the “Tom W.” problem (Kahneman & Tversky, 1973), *no* information is given about how the personality sketch of Tom W. was selected, whether randomly or not. The same holds for the personality sketches of Gary W. and Barbara T. in Ajzen’s (1977) base-rate studies.

But the issue is not necessarily resolved simply by asserting random sampling verbally in the problem. Consider the following famous demonstration of base-rate neglect in which random sampling is actually mentioned. A group of students had to solve the engineer–lawyer problem (Kahneman & Tversky, 1973, pp. 241–242):

A panel of psychologists have interviewed and administered personality tests to 30 engineers and 70 lawyers, all successful in their respective fields. On the basis of this information, thumbnail descriptions of the 30 engineers and 70 lawyers have been written. You will find on your forms five descriptions, chosen at random from the 100 available descriptions. For each description, please indicate your probability that the person described is an engineer, on a scale from 0 to 100.

Two of these thumbnail descriptions were:

Jack is a 45-year-old man. He is married and has four children. He is generally conservative, careful, and ambitious. He shows no interest in political and social issues and spends most of his free time on his many hobbies which include home carpentry, sailing, and mathematical puzzles. The probability that Jack is one of the 30 engineers in the sample of 100 is _____%.

Dick is a 30-year-old man. He is married with no children. A man of high ability and high motivation, he promises to be quite successful in his field. He is well liked by his colleagues.

A second group of students received the same instructions and the same descriptions, but were told that the base rates were 70 engineers and 30 lawyers (as opposed to 30 engineers and 70 lawyers). Kahneman and Tversky found that the mean response in both groups of students was for the most part the same, and concluded that base rates were largely ignored. Their explana-

tion was that participants use a representativeness heuristic, that is, they judge the probability by the similarity (representativeness) between a description and their stereotype of an engineer. Kahneman and Tversky believed that their participants were violating “one of the basic principles of statistical prediction,” the integration of prior probability with specific evidence by Bayes’s rule. The result was given much weight: “The failure to appreciate the relevance of prior probability in the presence of specific evidence is perhaps one of the most significant departures of intuition from the normative theory of prediction” (p. 243).²

The Normative Issue The phrase “the normative theory of prediction,” or probability, is standard rhetoric in the “heuristics and biases” program. But what is this normative theory? Certainly it is not frequentist—to name, for example, only the most popular theory of probability. So let us infer what the authors mean by “the normative theory” from what they want their participants to do. This seems to be simply to apply a formula—Bayes’s rule—to the engineer–lawyer problem. But there is more to *good* probabilistic reasoning than applying formulas mechanically. There are assumptions to be checked (see Mueser, Cowan, & Mueser, 1999). Is the structure of the problem the same as the structure of the statistical model underlying the formula?

One important structural assumption is random sampling. If the descriptions of Jack, Dick, and the others were not randomly sampled but selected, the base rates of engineers and lawyers specified were indeed irrelevant. In fact, the descriptions were made up and *not* randomly sampled from a population with the base rates specified—although the participants were told the contrary. Whether the single word “random” in the instruction is enough to commit participants to this crucial structural assumption is a problem in itself—particularly because we cannot assume that people are familiar with situations in which profession guessing is about randomly drawn people. For instance, both in the United States and in Germany there is a popular TV program in which a panel of experts guesses the profession of a candidate, who answers only “yes” or “no” to their questions. Here, the experts would perform badly if they started out with the known base rates of professions, say in the United States, and revised them according to Bayes’s rule. The candidates were selected, not randomly drawn.

2. The terms “prior probabilities” and “base rates” are frequently used interchangeably in the psychological literature. But these concepts are not identical. It is the prior probabilities that are fed into Bayes’s rule, and these priors may be informed by base rates. Base rates are just one piece of information among several that a person can consider relevant for making up her prior probabilities. Equating prior probabilities with *one* particular kind of base-rate information would be a narrow understanding of Bayesian reasoning. Such reasoning might be defensible in those situations in which one knows very little, but not in real-life situations in which one can base judgments on rich knowledge.

Random Sampling Increases Use of Base Rates One way to understand participants' judgments is to assume that the engineer-lawyer problem activates earlier knowledge associated with profession guessing, which can be used as an inferential framework—a "mental model"—to solve the problem.³ But, as I have argued, we cannot expect random sampling to be part of this mental model. If my analysis is correct, then base-rate use can be increased if we take care to commit the participants to the crucial property of random sampling—that is, break apart their mental models and insert the new structural assumption. In contrast, if the true explanation is that participants rely on the representativeness heuristic, then the participants should continue to neglect base rates.

There is a simple method of making people aware of random sampling in the engineer-lawyer problem, which we used in a replication of the original study (Gigerenzer, Hell, & Blank, 1988). The participants themselves drew each description (blindly) out of an urn, unfolded the description, and gave their probability judgments. There was no need to tell them about random sampling because they did it themselves. This condition increased the use of base rates. Participants' judgments were closer to Bayesian predictions than to base-rate neglect. When we used, for comparison, the original study's version of the crucial assumption—as a one-word assertion—neglect of base rates appeared again (although less intensely than in Kahneman and Tversky's study).

Worthless Specific Evidence and the Base-Rate Fallacy The description of "Dick" (see above) is a particularly interesting case. It was constructed to be totally uninformative for distinguishing engineers from lawyers. Kahneman and Tversky (1973) reported that the median probabilities were the same (0.50) in both base-rate groups. That people neglect base rates even when only "worthless specific evidence" is given has been taken by the authors to demonstrate the "strength" (p. 242) of the representativeness heuristic. This striking result led to many a speculation:

The fact that the base rate is ignored even when the individuating information is useless (for example, the target is 'ambitious' or 'well liked') suggests that the preference for specific-level evidence is so great that the base rate or high-level default information is not even retrieved once the subject tries to make a prediction on the basis of the specific information. (Holland et al., 1986, p. 218)

Such statements need to be corrected. First, if the crucial structural assumption of random sampling is made clear, base rates are no longer ignored in participants' judgments about the "uninformative" description of Dick, just as for "informative" descriptions such as Jack's. Second, and equally striking,

3. I use the term "mental model" in a sense that goes beyond Johnson-Laird's (1983). As in the theory of probabilistic mental models (Chapter 7), a mental model is an inferential framework that generalizes the specific task to a reference class (and probability cues defined on it) that a person *knows* from his or her environment.

I show that even with Kahneman and Tversky's original "verbal assertion method," that is, a one-word assertion of random sampling, there is in fact no support for the claim that judgments about an uninformative description are guided by a general representativeness heuristic—contrary to assertions in the literature.

Table 12.5 lists all studies of the uninformative description "Dick" that I am aware of—all replications of Kahneman and Tversky's (1973) verbal assertion method. The two base-rate groups were always 30% and 70% engineers. According to Kahneman and Tversky's argument, the difference between the two base-rate groups should approach the difference between the two base rates, that is, 40% (or somewhat less, if the description of Dick was not perceived as totally uninformative by the participants). The last column shows their result mentioned above, a zero difference, which we (Gigerenzer, Hell, & Blank, 1988) could closely replicate. Table 12.5 also shows, however, that several studies found substantial mean differences up to 37%, which comes very close to the actual difference between base-rate groups.

Seen together, the studies seem to be as inconsistent as it is possible to be: Every result between zero difference (base-rate neglect) and the actual base-rate difference has been obtained. This clearly contradicts the rhetoric of robustness and generality of the base-rate fallacy, such as: "Regardless of what kind of information is presented, subjects pay virtually no attention to the base rate in guessing the profession of the target" (Holland et al., 1986, p. 217). And it contradicts the explanation of the so-called fallacy: the proposed general representativeness heuristic.

Table 12.5 How to make the "base-rate fallacy" disappear: The uninformative description "Dick" in the engineer-lawyer problem

Study	No. of descriptions	"Dick" encountered first (relative frequency)	Mean difference between base rate groups ^a
Gigerenzer, Hell, & Blank (1988) ^b	6	1/6	1.2
Kahneman & Tversky (1973) ^c	5	1/5	0.0
Wells & Harvey (1978)	2	1/2	18.0
Ginossar & Trope (1987) ^d	1	1	24.0
Ginossar & Trope (1980) ^e	1	1	31.0
Gigerenzer, Hell, & Blank (1988) ^b	1	1	37.0

a. Entries are $\{p_{70}(E|D) - p_{30}(E|D)\} \times 100$, where $p_n(E|D)$ is the mean probability judgment that "Dick" is an engineer, given the description and the 70% base rate.

b. Order of descriptions systematically varied.

c. Medians (no means reported).

d. Three descriptions were used, but "Dick" was always encountered first.

e. Separate analysis for all participants who encountered "Dick" first.

How to explain these apparently inconsistent results? Table 12.5 gives us a clue. There is a striking correlation between the *number* of descriptions each participant read and judged and the mean difference between base-rate groups. The key variable seems to be the relative frequency with which participants encountered "Dick" first, which is a direct function of the number of descriptions. In all studies in which only Dick was used (i.e., the number of descriptions was 1), or in which a separate analysis was performed for all participants who encountered Dick first, there is a strong base-rate effect. If Dick and one informative description (Jack) were used, as in Wells and Harvey (1978), then the base-rate effect is in between, because of averaging across participants who encountered Dick either before or after the informative description. Thus Table 12.5 supports the following conclusions. (1) Contrary to claims in the literature, participants did make use of the base rates if only uninformative information ("Dick") was presented. (2) The neglect of base rates occurred only in a specific condition, that is, when the participants had encountered one or more *informative* descriptions before they judged "Dick"—in other words, when "Dick" occurred in the second, third, fourth, or later position. (3) The more descriptions a participant encountered, the less often "Dick" was in the first position, and—because of averaging across positions (such as in Kahneman and Tversky's study)—the smaller the difference was between base-rate groups.

This result should *not* occur if the intuitive statistician operates with a representativeness heuristic. Again, an explanatory framework using mental models based on knowledge about a particular domain (here, profession guessing) seems to be superior. If an informative description is encountered first, a mental model is activated that contains probability cues for professions, such as hobbies and political attitudes. Once the mental model is activated, the mind uses it as an inferential framework for similar-looking problems, that is, when "Dick" is encountered as the second or subsequent problem. Carrying over mental models to similar problems is analogous to perceptual judgment. We watch the first few steps and then proceed on the hypothesis that the rest are like the first (Gregory, 1974). This practice can sometimes make us stumble, but this kind of uncertain and "risky" inference is what makes our perceptual apparatus superior to any computer available today.

If the uninformative description is encountered first, however, then such a mental model is not activated, because its probability cues would not discriminate, and participants fall back on the only information available, the base rates (which are, as I argued above, not part of the mental model of profession guessing).

To summarize: (1) There is little justification for calling participants' judgments in the engineer-lawyer problem an "error" in probabilistic reasoning, because (aside from the frequentist argument) participants were not committed to random sampling. (2) If one lets the participants do the random drawing, base-rate neglect disappears. (3) That participants are sensitive to the distinction between random and selected (nonrandom) drawings shows again that the framework of so-called "heuristics and biases" is much too narrow for under-

standing judgments under uncertainty (for similar results see Ginossar & Trope, 1987; Grether, 1980; Hansen & Donoghue, 1977; Wells & Harvey, 1977; but see Nisbett & Borgida, 1975).

Note that the critical variable here is the content of a problem. There seems to be a class of contents for which participants know from their environment that base rates are relevant (as do birds; see Caraco, Martindale, & Whittam, 1980) or that random sampling is common (though they need not represent these concepts explicitly), whereas in other contents this is not the case. Profession guessing seems to belong to the latter category. In contrast, predictions of sports results, such as those of soccer games, seem to belong to the former. For instance, we found that participants revised information about the previous performance of soccer teams (base rates) in light of new information (half-time results) in a way that is indistinguishable from Bayesian statistics (Gigerenzer, Hell, & Blank, 1988). Here verbal assertion of random drawing was sufficient—there was no need for strong measures to break apart mental models.

Heuristics

The concept of a "heuristic" has various meanings and a long history—from Descartes's 21 heuristic rules for the direction of the mind to Duncker's heuristic methods that guide the stepwise reformulation of a problem until it is solved (Groner, Groner, & Bischof, 1983). The cognitive revolution has reintroduced the concept of a heuristic into psychology, in particular in the work of Herbert Simon (1957). Because of limited information-processing abilities, Simon argued, humans have to construct simplified models of the world. Heuristics are a product of these: They are shortcuts that can produce efficient decisions. Simon understood heuristics such as satisficing (i.e., selecting the first option available that meets an aspiration level) as adaptive strategies in a complex environment, in which alternatives for action are not given but must be sought out.

In the 1970s, Kahneman and Tversky borrowed the term "heuristic" from artificial intelligence to explain "errors" in probabilistic reasoning: "People rely on a limited number of heuristic principles which reduce the complex tasks of assessing probabilities and predicting values to simpler judgmental operations. In general, these heuristics are quite useful, but sometimes they lead to severe and systematic errors" (Tversky & Kahneman, 1974, p. 1124). Although they repeatedly asserted that these heuristics are useful, almost all of their work focused on how they lead to "errors." The three heuristics proposed in the early 1970s—representativeness, availability, and anchoring and adjustment—were a first, promising step to connect the rather atheoretical Bayesian research of the 1960s with cognitive theory. But in the 30 years of "heuristics and biases" research since then, a lack of theoretical progress is possibly the most striking result. The absence of a general theory or even of specific models of underlying cognitive processes has been repeatedly criti-

cized (e.g., Jungermann, 1983; Wallsten, 1983), but to no avail. Why is this? I believe that particular features of the use of the term "heuristic" have led to the present conceptual dead end, and more research in a cul-de-sac will not help. In my opinion, these features are the following.

The Function of Heuristics

In artificial intelligence research one hopes that heuristics can make computers smart; in the "heuristics and biases" program one hopes that heuristics can tell why humans are not smart. The fundamental problem with the latter is that most "errors" in probabilistic reasoning that one wants to explain by heuristics are in fact not errors, as I have argued above. Thus heuristics are meant to explain what does not exist. Rather than explaining a *deviation* between human judgment and allegedly "correct" probabilistic reasoning, future research has to get rid of simplistic norms that evaluate human judgment instead of explaining it.

Simon, and earlier Egon Brunswik, has emphasized that cognitive functions are adaptations to a given environment and that we have to study the structure of environments in order to infer the constraints they impose on reasoning. Heuristics such as representativeness have little to say about how the mind adapts to the structure of a given environment.

Redescription as a Substitute for Theorizing

Several of the explanations using heuristics are hardly more than redescrptions of the phenomena reported. Take, for instance, the explanation of base-rate neglect in the engineer-lawyer problem (and similar base-rate problems) by the representativeness heuristic. Representativeness here means the perceived similarity between a personality description and the participants' stereotype of an engineer. In the vocabulary of Bayes's rule, this similarity is a likelihood: that is, the probability of a description given that the person is an engineer. Now we can see that Bayes's rule, in particular its concepts of base rates (prior probabilities) and likelihoods, provides the vocabulary for both the phenomenon and its purported explanation. The phenomenon is neglect of base rates and use of likelihoods. The "explanation" is that participants use representativeness (likelihoods) and do not use base rates. What is called a representativeness heuristic here is nothing more than a redescription of the phenomenon (Gigerenzer & Murray, 1987, pp. 153-155).

Heuristics Are Largely Undefined Concepts

Representativeness means similarity. Although there are numerous specific models of similarity (including Tversky, 1977), the relationship between the representativeness heuristic and these models has never been worked out. Fiedler (1983), for instance, has analyzed the theoretical weakness of explaining estimated frequencies of events by the availability heuristic. All three heu-

ristics, representativeness, availability, and anchoring and adjustment, are largely undefined concepts and can post hoc be used to explain almost everything. After all, what is similar to what (representativeness), what comes into your mind (availability), and what comes first (anchoring) have long been known to be important principles of the mind.

More Undefined Concepts, Less Theory

Instead of giving up the program of explaining *deviations* of human judgment from simplistic norms by means of redescription and largely undefined heuristics, the last 25 years have witnessed the effort to keep that program going and to add further undefined concepts such as "causal base rates" and "vividness" to account for contradictory results (for an analysis of the "causal base rate" concept, see Gigerenzer & Murray, 1987, pp. 157-162). Heuristics such as representativeness are by now riddled with exceptions, but all this tinkering has not given us much purchase in understanding judgment under uncertainty.

Beyond Heuristics and Biases

I have argued that what have been widely accepted to be the "normative principles of statistical prediction" (e.g., Ajzen, 1977, p. 304), against which human judgment has been evaluated as "fallacious," are a caricature of the present state of probability theory and statistics. I have shown that several so-called fallacies are in fact not violations of probability theory. Conceptual distinctions routinely used by probabilists and statisticians were just as routinely ignored in the normative claims of "fallacies." Most strikingly, in the experimental research reviewed, "fallacies" and "cognitive illusions" tend to disappear if we pay attention to these fundamental distinctions. I am certainly not the first to criticize the notion of "robust fallacies." The only novelty in my research is that the variables that bring "cognitive illusions" under experimental control are those important from the viewpoint of probability and statistics (as opposed to, say, whether participants were given more or less "vivid" or "causally relevant" information).

Together, these results point to several ways to develop an understanding of judgment under uncertainty that goes beyond the narrow notion of a "bias" and the largely undefined notion of a "heuristic."

Use Different Statistical Models as Competing Explanatory Models

The existence of different statistical models of inference is a rich resource for developing theories about intuitive inference. This resource has been rarely touched, possibly because of the misleading normative view that statistics speaks with one voice.

For instance, despite the quantity of empirical data that has been gathered on the cab problem, the lack of a theory of the cognitive processes involved in solving it is possibly the most striking result. Tversky and Kahneman claimed that the cab problem has one "correct answer" (1980, p. 62). They attempted to explain the extent to which people's judgments deviated from that "norm" by largely undefined terms such as "causal base rates." But statistics gives several interesting answers to the cab problem, rather than just one "correct" answer (e.g., Birnbaum, 1983; Gigerenzer, 1998c; Levi, 1983). If progress is to be made and people's cognitive processes are to be understood, one should no longer try to explain the *difference* between people's judgments and Tversky and Kahneman's "normative" Bayesian calculations. People's *judgments* have to be explained. Statistical theories can provide highly interesting models of these judgments. The only theoretically rich account of the cognitive processes involved in solving the cab problem (or similar "eyewitness testimony" problems) was in fact derived from a frequentist framework: Birnbaum (1983) combined Neyman-Pearson theory with psychological models of judgments such as range-frequency theory.

Future research should use competing statistical theories as competing explanatory models, rather than pretending that statistics speaks with one voice (see also Cohen, 1982; Wallendael & Hastie, 1990).

Explore the Metaphor of the Mind as a Frequentist

I reported earlier the striking effect of participants' judging frequencies rather than probabilities for single events. These results suggest that the mind distinguishes between frequencies and other meanings of probability, just as a statistician of the frequentist school does. Because "cognitive illusions" tend to disappear in frequency judgments, it is tempting to think of the intuitive statistics of the mind as frequentist statistics.

Processing of frequencies seems to be fairly automatic, like encoding of time and space (e.g., Hasher & Zacks, 1979)—whereas probabilities are in evolutionary terms recent tools of the mind that seem to be processed less automatically. The theory of probabilistic mental models (Chapter 7) seems to be the first frequentist theory of confidence judgments that integrates Brunswik's frequency-learning view with the notion of mental models. The general theoretical point is that both single-case and frequency judgments are explained by learned frequencies (the probability cues), albeit by frequencies that relate to different reference classes and different networks of cues—in short, to different mental models.

Intuitive Statisticians Need to Check the Structure of the Environment

Good judgment under uncertainty is more than mechanically applying a formula, such as Bayes's rule, to a real-world problem. The intuitive statistician, like his professional counterpart, must first check the structure of the environ-

ment (or of a problem) in order to decide whether to apply a statistical algorithm at all, and if so, which (see Gigerenzer & Murray, 1987, pp. 162–174). There is no good (applied) probabilistic reasoning that ignores the structure of the environment and mechanically uses only *one* (usually mathematically convenient) algorithm. I illustrate this point with a thought experiment by Nisbett and Ross (1980, p. 15), which I have shortened and slightly changed here (in respects unimportant to my argument).

(i) You wish to buy a new car. Today you must choose between two alternatives: to purchase either a Volvo or a Saab. You use only one criterion for that choice, the car's life expectancy. You have information from *Consumer Reports* that in a sample of several hundred cars the Volvo has the better record. Just yesterday a neighbor told you that his new Volvo broke down. Which car do you buy?

Nisbett and Ross comment that after the neighbor's information "the number of Volvo-owners has increased from several hundred to several hundred and one" and that the Volvo's record "perhaps should be changed by an iota" (p. 15). The moral of their thought experiment is that good probabilistic reasoning is applying an algorithm (here, updating of base rates) to the world. There is some truth to this message of resisting the temptation of the vivid and personal, but that is only half the story. Good intuitive statistics is more than calm calculation; first and foremost, the structure of the environment has to be examined. I will now vary the content of Nisbett and Ross's thought experiment to make the point intuitively immediate. Here is the same problem, but with a different content (Gigerenzer, 1990):

(ii) You live in a jungle. Today you must choose between two alternatives: to let your child swim in the river, or to let it climb trees instead. You use only one criterion for that choice, your child's life expectancy. You have information that in the last 100 years there was only one accident in the river, in which a child was eaten by a crocodile, whereas a dozen children have been killed by falling from trees. Just yesterday your neighbor told you that her child was eaten by a crocodile. Where do you send your child?

If good probabilistic reasoning means applying the same algorithm again and again, the neighbor's testimony should make no difference. The base rates would be updated by the testimony from one to two cases in 100 years, and by this reasoning one would send the child into the river. The mind of a parent, however, might use the new information to *reject* the updating algorithm instead of *inserting* the new information into the algorithm. A parent may suspect that the small river world has changed—crocodiles may now inhabit the river.

Why do we have different intuitions for the Volvo and the crocodile problems? In the Volvo problem, the prospective buyer may assume that the Volvo world is stable and that the important event (good or bad Volvo) can be considered as an independent random drawing from the same reference class. In the crocodile problem, the parents may assume that the river world has

changed and that the important event (being eaten or not) can no longer be considered as an independent random drawing from the same reference class. Updating "old" base rates may be fatal for the child.

The question of whether some part of the world is stable enough to use statistics has been posed by probabilists and statisticians since the inception of probability theory in the mid-seventeenth century—and the answers have varied and will vary, as is well documented by the history of insurance (Daston, 1987). Like the underwriter, the layperson has to check structural assumptions before entering into calculations. For instance, the following structural assumptions are all relevant for the successful application of Bayes's rule: independence of successive drawings, random sampling, an exhaustive and mutually exclusive set of hypotheses, and independence between prior probabilities and likelihoods.

How can the intuitive statistician judge whether these assumptions hold? One possibility is that the mind generalizes the specific content to a broader mental model that uses implicit domain-dependent knowledge about these structural assumptions. If so, then the *content* of problems is of central importance for understanding judgment—it embodies implicit knowledge about the structure of an environment.

The Surplus Structure of the Environment

Analyzing the environment (problem) using structural properties of a given statistical model is one way to understand its structure. But natural environments often have *surplus* structure, that is, a structure that goes beyond prior probabilities and likelihoods (the Bayesian structure) or entailment and contradiction (the structure of binary propositional logic). Surplus structure includes space and time (Björkman, 1984), cheating options, perspective, and social contracts (Cosmides, 1989), among others. Surplus structure is the reason that the notion of "structural isomorphs" has only limited value.

The idea of studying inductive reasoning using structural isomorphs (i.e., use a particular statistical model or formal principle and construct problems that all have the same formal structure but different contents) is implicit in much research on reasoning; it postulates that if two problems have different contents, but the same *formal* structure (say, Bayesian probability-revision structure), then judgments should be the *same*. But the structure of natural environments is usually richer than what Bayes's rule has to offer, and two structural isomorphs may differ on relevant surplus structure. If we understand reasoning as an adaptation to the environment, then it should be sensitive to surplus structure.

One way to deal with this is to devise theories that combine statistical theory and psychological principles—just as the most distinguished statisticians of this century, R. A. Fisher, J. Neyman, and E. S. Pearson, emphasized that good statistical reasoning always consists of mathematics *and* personal judgment. Birnbaum (1983) gave several examples of how the Neyman–Pearson theory can be combined with psychological principles to give a theoretically

rich account of intuitive inference. Developing such integrated models is a challenging task for future research on judgments under uncertainty.

The Social Context of Judgment and Decision

Judgment under uncertainty occurs in a social environment in which there are other "players" who make a person's response more or less rational. Here is an anecdote to illustrate this point.

A small town in Wales has a village idiot. He once was offered the choice between a pound and a shilling, and he took the shilling. People came from everywhere to witness this phenomenon. They repeatedly offered him a choice between a pound and a shilling. He always took the shilling.

Seen as a single choice (and by all monotone utility functions), this choice would seem irrational. Seen in its social context, in which a surprising choice increases the probability of getting to choose again, this behavior looks different.

The following are several aspects of the social context of judgment and decision that have been explored recently. First, human judgment seems to be domain specific rather than guided by some general mental logic. In particular, reasoning about social contracts seems to have its own laws. The striking changes of judgment depending on people's perspective and cheating options in a social contract were shown by Cosmides (1989) and Gigerenzer and Hug (1992). Second, the role of conversational principles in social interactions, such as that participants assume the experimenter's contribution will be cooperative (Adler, 1984; Grice, 1975), has sometimes been acknowledged by, but never been integrated into, the judgment under uncertainty literature. Third, humans share knowledge and decisions, and sharing imposes constraints on information processing and judgment as postulated by shareability theory (Freyd, 1983). Fourth, research on group decision making and judgments negotiated by two or more people is still largely disconnected from "individualistic" social cognition research (see Scholz, 1983).

Conclusion

A key metaphor for understanding inductive reasoning is probability theory. Since its origins in the mid-seventeenth century and throughout the Enlightenment, probability theory was viewed as a mathematical codification of rationality. In Pierre Laplace's famous phrase: probability theory is "only good sense reduced to calculus" (1814/1951, p. 196). When there was a striking discrepancy between the judgment of reasonable people and what probability theory dictated—as with the famous St. Petersburg paradox—then the mathematicians went back to the blackboard and changed the equations (Daston, 1980).

Those good old days have gone, although the eighteenth-century link between probability and rationality is back in vogue in cognitive and social psychology. If, in studies on social cognition, researchers find a discrepancy between human judgment and what probability theory seems to dictate, the blame is now put on the human mind, not on the statistical model.

I have used classical demonstrations of overconfidence bias, conjunction fallacy, and base-rate neglect to show that what have been called “errors” in probabilistic reasoning are in fact *not* violations of probability theory. They only look so from a narrow understanding of good probabilistic reasoning that ignores conceptual distinctions fundamental to probability and statistics. These so-called cognitive illusions largely disappear when one pays attention to these conceptual distinctions. The intuitive statistician seems to be highly sensitive to them—a result unexpected from the view that “mental illusions should be considered the rule rather than the exception” (see Table 12.1).

Why do cognitive illusions largely disappear? The examples in this chapter have illustrated three reasons:

1. *Polysemy: not all probabilities are mathematical probabilities.* Asking a frequency as opposed to a probability question can reduce the multiple meanings (polysemy) of the English terms “probable” and “likely.” Frequency questions clarify that the question is actually about mathematical probability and not about one of the other legitimate meanings (see the Oxford English Dictionary), which are often suggested by the cover story of a problem. Reducing polysemy seems to be the major reason the conjunction fallacy in the Linda problem largely disappears (Hertwig & Gigerenzer, 1999).
2. *A mathematical probability refers to a reference class, which may differ depending on the task.* Asking a frequency (as opposed to a probability) question can systematically cue different reference classes (and therefore, different probabilistic mental models). Changing reference classes seems to be the reason overconfidence bias appears in probability judgments and disappears in frequency judgments (Chapter 7).
3. *Natural frequencies facilitate Bayesian reasoning.* When information is represented in natural frequencies rather than in conditional probabilities (or relative frequencies), Bayesian computations become simpler. Using natural frequencies is a powerful tool to reduce people’s mental confusion and foster Bayesian reasoning (Chapter 6).

This is not to say that frequencies always improve judgment. For instance, the theory of probabilistic mental models specifies conditions under which frequency judgments systematically underestimate actual frequencies, and Chapter 6 explains why natural frequencies but not other kinds of frequencies facilitate Bayesian reasoning. The question is not whether or not, or how often, “cognitive illusions” disappear, but *why*. We need precise models of heuristics that make surprising (and falsifiable) predictions, not vague terms that, post hoc, explain everything and nothing. Future progress will be in understanding, not debunking, human thinking.

The Superego, the Ego, and the Id in Statistical Reasoning

Piaget worked out his logical theory of cognitive development, Köhler the Gestalt laws of perception, Pavlov the principles of classical conditioning, Skinner those of operant conditioning, and Bartlett his theory of remembering and schemata—all without rejecting null hypotheses. But by the time I took my first course in psychology at the University of Munich in 1969, null hypothesis tests were presented as *the* indispensable tool, as the *sine qua non* of scientific research. Post–World War II German psychology mimicked a revolution of research practice that had occurred between 1940 and 1955 in American psychology.

What I learned in my courses and textbooks about the logic of scientific inference was not without a touch of moralizing, a scientific version of the Ten Commandments: Thou shalt not draw inferences from a nonsignificant result. Thou shalt always specify the level of significance before the experiment; those who specify it afterward (by rounding up obtained *p* values) are cheating. Thou shalt always design thy experiments so that thou canst perform significance testing.

The Inference Revolution

What happened between the time of Piaget, Köhler, Pavlov, Skinner, and Bartlett and the time I was trained? In Kendall’s (1942) words, statisticians “have already overrun every branch of science with a rapidity of conquest rivalled only by Attila, Mohammed, and the Colorado beetle” (p. 69).

What has been termed the *probabilistic revolution in science* (Gigerenzer et al., 1989) reveals how profoundly our understanding of nature changed when concepts such as chance and probability were introduced as fundamental theoretical concepts. The work of Mendel in genetics, that of Maxwell and Boltzmann on statistical mechanics, and the quantum mechanics of Schrödinger and Heisenberg that built indeterminism into its very model of nature are key examples of that revolution in thought.