Rationality

Psychological and philosophical perspectives

Edited by K.I. Manktelow and D.E. Over

London and New York

First published in 1993
by Routledge
11 New Fetter Lane, London EC4P 4EE

Simultaneously published in the USA and Canada
by Routledge
29 West 35th Street, New York, NY 10001

Typeset in Times by J&L Composition Ltd, Filey, North Yorkshire.

Printed and bound in Great Britain by
Mackays of Chatham PLC, Chatham, Kent.

British Library Cataloguing in Publication Data
A catalogue record for this book is available from the British Library

Library of Congress Cataloging in Publication Data
Rationality: psychological and philosophical perspectives / edited by K.I. Manktelow and D.E. Over.
p. cm. – (International library of psychology)
Includes bibliographical references and index.
1. Reasoning (Psychology). 2. Cognitive psychology.
3. Logic. 4. Psychology and philosophy.
I. Manktelow, K.I., 1952– . II. Over, D.E., 1946–
III. Series.
BF442.R38 1993
153.4'3–dc20 92–47072
CIP

ISBN 0–415–06955–6

Chapter 11

The bounded rationality of probabilistic mental models

G. Gigerenzer

Imagine you are a subject in a psychological experiment. In front of you is a text problem, and you begin to read:

> 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. Which of two alternatives is more probable?
> (a) Linda is a bank teller.
> (b) Linda is a bank teller and is active in the feminist movement.

Which alternative would you choose? Assume you chose (b), just as most subjects in previous experiments did. The experimenter explains to you that (b) is the conjunction of two events, namely that Linda is a bank teller *and* is active in the feminist movement, whereas (a) is one of the constituents of the conjunction. Because the probability of a conjunction of two events cannot be greater than that of one of its constituents, the correct answer is (a), not (b), the experimenter says. Therefore, your judgement is recorded as an instance of a reasoning error, known as the *conjunction fallacy*. You may be inclined to admit that you have committed a reasoning error. The experimenter now explains that these reasoning errors are like visual illusions: once the error is pointed out, people like you show insight, but this knowledge does not necessarily help. People see the same illusion again, or continue to reason in the same way, despite showing insight. Therefore, in analogy to visual illusions, stable reasoning errors such as the conjunction fallacy have been labelled *cognitive illusions*.

Cognitive illusions, and their explanations, cognitive heuristics, are the stock-in-trade of a research programme known as the heuristics-and-biases programme (for example, Tversky and Kahneman, 1974, 1983). Cognitive illusions 'seem reliable, systematic, and difficult to eliminate' (Kahneman and Tversky, 1972: 431). Stable cognitive illusions are not the first assault on human rationality by psychologists. Sigmund Freud's attack on human rationality is probably the best-known: the unconscious desires and wishes of the Id are a steady source of intrapsychical conflict that may manifest

itself in all kinds of irrational fears, beliefs, and behaviour. But the cognitive-illusion assault is stronger than the psychoanalytic. It does not need to invoke unconscious wishes or desires to overwhelm human rationality. Cognitive illusions are seen as a straightforward consequence of the laws of human reasoning. Humans do not possess the proper mental algorithms.

Paleontologist Stephen J. Gould, referring to the 'Linda problem', puts the message clearly: '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' (Gould, 1992: 469). The purpose of this chapter is to evaluate this claim and to provide an alternative. In the first part, I will draw the reader's attention to the fact that both proponents and opponents of rationality tend to focus on the same single psychological concept: algorithms in the mind. Second, I will extend this focus by conceptual distinctions drawn from philosophy, statistics, and cognitive science, and argue that these distinctions are not just the province of philosophers and statisticians but have quite tangible implications for understanding the cognitive processes in reasoning and for the rationality debate. Third, I demonstrate that these implications are so powerful that they can make apparently stable cognitive illusions disappear. Finally, I will present a model of bounded rationality, the theory of *probabilistic mental models*, as an alternative to traditional explanations in terms of the heuristics-and-biases programme. Using the overconfidence effect as an illustration, I will show that this theory explains both the old data (cognitive illusions), predicts new phenomena, and provides a fresh look at what rational probabilistic reasoning means.

RATIONALITY: WHAT KIND OF MENTAL ALGORITHM?

In his *Movements of Animals*, Aristotle described a *practical syllogism* as one that guides practical rationality:

> For example, when you conceive that every man ought to walk and you yourself are a man, you immediately walk; or if you conceive that on a particular occasion no man ought to walk, and you yourself are a man, you immediately remain at rest.
>
> (Aristotle, 1945: 701a)

The foundation of present-day theories of rationality, however, was laid in the mid-seventeenth century with the classical theory of probability (Daston, 1988). In contrast to syllogisms, probability could deal with degrees of beliefs, weights of evidence, expectations, and other forms of uncertainty that are characteristic of everyday affairs – from weighing the evidence in a law court to calculating insurance premiums. Probability

theory and rational reasoning came to be seen as two sides of the same coin; probability theory is 'nothing more at bottom than good sense reduced to a calculus' (Laplace, 1951/1814: 196). For instance, in his famous treatise of 1854, the mathematician George Boole set out to demonstrate that the laws of logic, probability and algebra can in fact be derived from the laws of human reasoning.

> There is not only a close analogy between the operations of the mind in general reasoning and its operations in the particular science of Algebra, but there is to a considerable extent an exact agreement in the laws by which the two classes of operations are conducted.
>
> (Boole, 1958/1854: 6)

Bärbel Inhelder and Jean Piaget echo this belief a century later: 'Reasoning is nothing more than the propositional calculus itself' (Inhelder and Piaget, 1958: 305).

According to these views, the laws of probability or logic are the algorithms of the mind, and they define rational reasoning as well. According to some critics of these views, the laws of probability are not the algorithms of the mind, but the laws still *define* rationality. Rather, mental algorithms are non-statistical heuristics causing cognitive illusions. Defenders and detractors of human rationality alike have tended to focus on the issue of algorithms. Only their answers differ. Here are some prototypical arguments in the current debate.

Statistical algorithms

For philosophers such as L. Jonathan Cohen, the assumption that human intuition is rational is absolutely indispensable for legitimizing their own profession. If intuition were not rational, this would 'seriously discredit the claims of intuition to provide – other things being equal – dependable foundations for inductive reasoning in analytical philosophy' (Cohen, 1986: 150). Cohen (1983: 511) assumes that statistical algorithms (Baconian and Pascalian probability) are in the mind, but distinguishes between not having a statistical rule and not applying such a rule, that is, between competence and performance. Cohen's interpretation of cognitive illusions parallels J.J. Gibson's interpretation of visual illusions (Gigerenzer, 1991): illusions are attributed to non-realistic experiments using impoverished information, to experimenters acting as conjurors, and to other factors that mask the subjects' competence: 'unless their judgment is clouded at the time by wishful thinking, forgetfulness, inattentiveness, low intelligence, immaturity, senility, or some other competence-inhibiting factor, all subjects reason correctly about probability: none are programmed to commit fallacies or indulge in illusions' (Cohen, 1982: 251). Cohen does not claim, I think, that people carry around the collected

works of Kolmogoroff, Fisher, and Neyman in their heads, and merely need to have their memories jogged, like the slave in Plato's *Meno*. But his claim implies that people do have at least those statistical algorithms in their competence that are sufficient to solve all reasoning problems studied in the heuristics-and-biases literature, including the Linda problem.

The Enlightenment view that human reasoning is in part probability theory does *not* imply that humans make no mistakes in reasoning. Nobody would deny that, even Cohen. According to Boole, for instance, errors in reasoning 'are due to the interference of other laws with those laws of which *right* reasoning is the product' (Boole, 1958/1854: 409). The message of the heuristics-and-biases programme, however, is stronger than reminding us that emotions, desires, and the like make us err in reasoning.

Non-statistical algorithms: heuristics

Proponents of the heuristics-and-biases programme seem to assume that the mind is not built to work by the rules of probability:

> 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.
>
> (Kahneman and Tversky, 1973: 237)

A few more quotations illustrate the claim that the mind lacks statistical algorithms and, therefore, rationality. In a paper on biases in bargaining, Bazerman and Neale say, 'The biases of framing and overconfidence just presented suggest that individuals are generally affected by systematic deviations from rationality' (Bazerman and Neale, 1986: 317). The human mind lacks 'the correct programs for many important judgmental tasks' (Slovic *et al.*, 1976). 'We know that our uneducated intuitions concerning even the simplest statistical phenomena are largely defective' (Piattelli-Palmarini, 1989: 9). The experimental demonstrations have 'bleak implications for human rationality' (Nisbett and Borgida, 1975: 935), and 'For anyone who would wish to view man as a reasonable intuitive statistician, such results are discouraging' (Kahneman and Tversky, 1972/1982: 46).

Cognitive illusions are explained by non-statistical algorithms, known as cognitive heuristics. For instance, the standard explanation for the conjunction fallacy in the Linda problem is that the mind assesses the probability by calculating the similarity between the description of Linda and each of the alternatives, and chooses that alternative with the highest similarity. Because the description of Linda was constructed to be representative of an active feminist and the conjunction contains the

term 'feminist', people judge the conjunction more probable – so the explanation goes. Judging probability by similarity has been termed the *representativeness heuristic*. This heuristic was only vaguely defined when first proposed in the early 1970s, and it still is. It has not yet been linked to any of many existing theories of similarity, nor has it been spelled out how exactly similarity or representativeness is calculated.

Statistical and non-statistical heuristics

So far we have two research programmes. Cohen assumes that statistical algorithms are in the competence of humans, and one should explain cognitive illusion by identifying performance-inhibiting factors. Tversky and Kahneman assume that mental algorithms are non-statistical heuristics, which cause stable cognitive illusions. Proponents of a third position do not want to be forced to choose between statistical and non-statistical algorithms, but want to have them both. Fong and Nisbett (1991: 35) argue that people possess both rudimentary but abstract intuitive versions of statistical principles, such as the law of large numbers, *and* non-statistical heuristics such as representativeness. The basis for these conclusions are the results of training studies. For instance, the experimenters first teach the subject the law of large numbers or some other statistical principle, and subsequently also explain how to apply this principle to a real-world domain such as sports problems. Subjects are then tested on similar problems from the same or other domains. The typical result is that more subjects reason statistically, but transfer to domains not trained in is often low. Evans (1984) has proposed a similar interpretation of deductive reasoning, assuming both a mental logic and non-logical heuristics.

To summarize: I have briefly sketched three positions in the present debate on the rationality of probability judgement. My point is that the discussion between these three positions focuses on the kind of mental algorithm – is it probability, heuristics, or both? I now invite you to look beyond algorithms, to different questions and new kinds of experiments. Let me start with three ideas and distinctions.

THERE IS MORE THAN MENTAL ALGORITHMS

The distinction between algorithms and information representation

Information needs representation. In order to communicate information, it has to be represented in some symbol system (Marr, 1982). Take numerical information. This information can be represented by the Arabic numeral system, by the binary system, by Roman numbers, or other systems. These different representations can be mapped in a one-to-one way, and are in this sense equivalent representations. But they are not

necessarily equivalent for an algorithm. Pocket calculators, for instance, generally work on the Arabic base-10 system, whereas general-purpose computers work on the base-2 system. The numerals 100000 and 32 are representations of the number thirty-two in the binary and Arabic system, respectively. The algorithms of my pocket calculator will perform badly with the first kind of representation but work well on the latter.

The human mind finds itself in an analogous situation. The algorithms most western people have stored in their minds – such as how to add, subtract, or multiply – work well on arabic numerals. But contemplate for a moment division in Roman numerals, without transforming them first into Arabic numerals.

There is more to the distinction between an algorithm and a representation of information. Not only are algorithms tuned to particular representations, but different representations make explicit different features of the same information. For instance, one can quickly see whether a number is a power of 10 in an Arabic numeral representation, whereas to see whether that number is a power of 2 is more difficult. The converse holds with binary numbers. Finally, algorithms are tailored to given representations. Some representations allow for simpler and faster algorithms than others. Binary representation, for instance, is better suited to electronic techniques than Arabic representation. Arabic numerals, on the other hand, are better suited to multiplication and elaborate mathematical algorithms than Roman numerals – possibly one of the reasons for the superior development of mathematics in the early Arabic cultures as opposed to Roman culture.

The distinction between algorithms and information representation is central to David Marr's (1982) analysis of visual information processing systems. From vision to reasoning, I argue, understanding of cognitive processes needs to take account of both algorithms *and* information representation. I now connect this distinction with another conceptual distinction prominent in philosophy and probability theory.

The distinction between subjective degrees of belief and objective frequencies

The classical probabilists of the Enlightenment slid with breathtaking ease and little justification from one sense of probability to another: from objective frequencies to physical symmetry (today referred to as 'propensity') to subjective degrees of belief. Lorraine Daston (1988) has argued that this ease was a consequence of the associationist psychology of these days, of the belief, advanced *inter alia* by John Locke and David Hartley, that the matching of objective frequencies to subjective belief was rational. Only when associationist psychology shifted its emphasis to illusions and distortions introduced by passion and prejudice, did the

gap between objective and subjective probabilities become evident. Philosophers and mathematicians now drew a bold line between the first two objective meanings on the one hand and subjective probabilities on the other. The unity of belief and frequency crumbled in the first half of the nineteenth century. After the fall of the classical interpretation of probability the frequency interpretation emerged as the dominant view in the nineteenth and twentieth centuries.

For proponents of the frequency view such as Richard von Mises (1957/1928) and Jerzy Neyman (1977), probability theory is about frequencies, and does *not* deal with degrees of belief in single events. In the subjective ('Bayesian') interpretation that re-emerged in this century, however, degrees of belief are what probability means. Others wanted to have it both ways, or have proposed alternative interpretations of probability. The question, What is probability about? is still with us.[1]

My intention here is not to take sides in this debate, but to liken the conceptual distinction between single-event probabilities and frequencies to the concept of information representation. This leads us to distinguish two kinds of representations: frequency information or single-event probabilities. Finer distinctions can be made, but this will suffice for a start.

Monitoring of event frequencies

The third idea is an evolutionary speculation that links with the above distinctions. Bumblebees, birds, rats, and ants all seem to be good intuitive statisticians, highly sensitive to changes in frequency distributions in their environments, as recent research in foraging behaviour indicates (Gallistel, 1990; Real and Caraco, 1986). From sea snails to humans, as John Staddon (1988) argued, the learning mechanisms responsible for habituation, sensitization, and classical and operant conditioning can be described in terms of statistical inference machines.

Assume that some capacity or algorithm for statistical reasoning has been built up through evolution by natural selection. What information representation would such an algorithm be tuned to? Certainly not percentages and single-event probabilities (as in the typical experiments on human reasoning), since these took millenia of literacy and numeracy to evolve as tools for communication. Rather, in an illiterate world, the input representation would be *frequencies* of events, sequentially encoded, such as 3 out of 20 (as opposed to 15 per cent or $p = 0.15$). Such a representation is couched in terms of discrete cases. Moreover, frequencies such as 3 out of 20 contain *more* information than percentages such as 15 per cent. These frequencies contain information about the sample size (here: 20), which allows one to compute the *ambiguity* or precision of the estimate.

The notion that the mind infers the structure of the world through

monitoring event frequencies is an old one. Locke and Hartley assumed that the mind is a kind of counting machine that automatically registered frequencies of past events, an assumption that is now called *automatic frequency processing* (Hasher and Zacks, 1979). David Hume thought the mind was very sensitive to small differences in frequency: 'When the chances or experiments on one side amount to ten thousand, and on the other to ten thousand and one, the judgement gives the preference to the latter, upon account of the superiority' (Hume, 1975/1739: 141).

Now we can put these three ideas together. First, to analyse probabilistic reasoning, information representation and algorithms have to be distinguished. Second, there are (at least) two kinds of representations, frequencies and single-event probabilities. Finally, if evolution has selected some kind of algorithm in the mind, then it will be tuned to frequencies as representation.

In the next section I will show that these ideas, still rather general, are powerful enough to make several apparently stable cognitive illusions disappear.

HOW TO MAKE COGNITIVE ILLUSIONS DISAPPEAR

Cognitive illusions have become a hard currency in many debates. When Stephen Stich argued against Donald Davidson's philosophy of language and Daniel Dennett's philosophy of mind, he pointed out that these two systems are inconsistent with the psychologists' 'evidence for extensive irrationality in human inference' (Stich, 1990: 11). When I discuss with colleagues the actual evidence underlying such claims, the *conjunction fallacy* is often thrown in as *the* truly convincing and replicable demonstration of irrational reasoning.

So let us first see what the distinction between algorithm and information representation, and between frequency and single-event format, does to this cognitive illusion.

Conjunction fallacy

Tversky and Kahneman (1983) reported that 85 per cent of 142 undergraduates indicated that the conjunction 'Linda is a bank teller and is active in the feminist movement' (T&F) is more probable than 'Linda is a bank teller' (T). They and others have shown that this judgement is replicable and stable – not only with statistically naïve undergraduates but with 'highly sophisticated respondents' such as doctoral students in the decision science programme of the Stanford Business School who had taken advanced courses in probability, statistics, and decision theory (Tversky and Kahneman, 1983: 298).

The conjunction fallacy and the conclusion that the mind is not

programmed by the laws of probability but by non-statistical heuristics (albeit only very loosely defined ones) has become the accepted wisdom in much of cognitive and social psychology, philosophy of mind, and beyond. The conjunction fallacy has been proposed as the cause of various kinds of human misfortune, such as US security policy, where 'the conjunction fallacy . . . lends . . . plausibility to highly detailed nuclear war-fighting scenarios' (Kanwisher, 1989: 671).

Stephen J. Gould, explaining the Linda problem to his audience, writes:

> Tversky and Kahneman's 'studies have provided our finest insight into "natural reasoning" and its curious departure from logical truth . . . 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".'
>
> (Gould, 1992: 469)

Gould should have trusted his homunculus. In what follows, I will discuss the claim that the judgement called 'conjunction fallacy' is an error in probabilistic reasoning. I will argue that this claim is not tenable, and Gould's homunculus will be vindicated. Thereafter I will show what the distinction between algorithm and information representation can do to the conjunction fallacy.

Cognitive illusion illusory?

Is the conjunction fallacy a violation of probability theory? Has a person who chooses T&F violated probability theory? The answer is no, if the person is a frequentist such as Richard von Mises or Jerzy Neyman; yes, if he or she is a subjectivist such as Bruno de Finetti; and open otherwise.

The mathematician Richard von Mises, one of the founders of the frequency interpretation, used the following example to make his point:

> We can say nothing about the probability of death of an individual even if we know his condition of life and health in detail. The phrase 'probability of death', when it refers to a single person, has no meaning at all for us. This is one of the most important consequences of our definition of probability.
>
> (von Mises, 1957/1928: 11)

In this frequentist view, one cannot speak of a probability unless a reference class has been defined. The relative frequency of an event such as death is only defined with respect to a reference class, such as 'all male pub-owners fifty-years old living in Bavaria'. Relative frequencies may vary from reference class (pub-owners) to reference class (HIV-positives).

Since a single person is always a member of many reference classes, no unique relative frequency can be assigned to a single person. As the frequentist statistician G.A. Barnard put it, if one wants to evaluate subjective probabilities of single events, one 'should concentrate on the works of Freud and perhaps Jung rather than Fisher and Neyman' (Barnard, 1979: 171). Thus, for a strict frequentist, the laws of probability are about frequencies and not about single events such as whether Linda is a bank teller. Therefore, in this view, no judgement about single events can violate probability theory.

From the frequency point of view, the laws of probability are mute on the Linda problem, and what has been called a conjunction fallacy *is not* an error in probabilistic reasoning – probability theory simply doesn't apply to such cases. Seen from the Bayesian point of view, the conjunction fallacy *is* an error. Note that the experimental subjects were neither told that the Linda problem is meant to be a Bayesian probability textbook problem, nor did the experimenters try to persuade and commit their subjects to the Bayesian view.

How shall we evaluate this situation? The frequency view has been dominant since the nineteenth century, and teaching in statistics departments today as well as in undergraduate psychology courses is still predominantly frequentist in philosophy. Therefore, we cannot expect psychology undergraduates to carry around a Bayesian superego in their minds. One should be careful not to evaluate reasoning against some norm, unless subjects have been committed to that particular norm. Thus, choosing T&F in the Linda problem is *not* a reasoning error. What has been labelled the 'conjunction fallacy' here does *not* violate the laws of probability. It only looks so from *one* interpretation of probability.

How to make the conjunction fallacy disappear

We apply now the distinction between single-event and frequency information representation to the Linda problem. We just change the format from single event to a frequency representation (see italicized passage), leaving everything else as it was.

> 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.
> *There are 100 people who fit the description above. How many of them are:*
> *bank tellers?*
> *bank tellers and active in the feminist movement?*

Subjects are now asked for frequency judgements rather than for the probability of a single event. If one focuses on mental algorithms, this

change appears irrelevant. If the mind solves the Linda problem by using a representativeness heuristic, changes in 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 T & F than of T. Subjects therefore should still exhibit the conjunction fallacy. Similarly, if one assumes with Cohen that the laws of probability are in the mind, but that subjects have been misled by the experimenter into bad performance, changes in representation should not matter either. For instance, subjects may have been misled by assuming that the description of Linda is of any relevance to the solution, whereas it is completely irrelevant to finding the solution. This irrelevancy argument is not altered by the frequency format.[2]

However, if there is some statistical algorithm in the mind *that is tuned to frequencies* as information representation, then something striking should happen to this stable cognitive illusion. Violations of the conjunction rule should largely disappear.

The experimental evidence available confirms this prediction. Klaus Fiedler (1988) reported that the number of conjunction violations in the Linda problem dropped from 91 per cent in the original, single-event representation to 22 per cent in the frequency representation ($n = 44$). The same result was found, when he replaced 'There are 100 people' by some odd number such as 'There are 168 people'. The drop in the number of conjunction violations here was from 83 per cent to 17 per cent ($n = 23$). Fiedler reported similar results for other standard problems from which the conjunction fallacy has been inferred as a stable cognitive illusion. Tversky and Kahneman (1983: 308–9) reported similar phenomena.

To summarize: The debate between Cohen and Tversky and Kahneman has centred on the question of algorithm. I have argued that in order to understand probabilistic reasoning, one should distinguish between algorithms and information representation. The philosophical and statistical distinction between single events and frequencies clarifies that judgements hitherto labelled instances of the 'conjunction fallacy' cannot be properly called reasoning errors in the sense of violations of the laws of probability. The conceptual distinction between single event and frequency representations is sufficiently powerful to make this allegedly stable cognitive illusion disappear. The conjunction fallacy is not the only cognitive illusion that is subject to this argument.

Base-rate fallacy

Casscells *et al.* (1978) presented sixty staff and students at Harvard Medical School with the following problem:

> 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?

If one inserts these numbers into Bayes' theorem, the posterior probability that the person actually has the disease is 0.02, or 2 per cent (assuming that the test correctly diagnoses every person who has the disease – a piece of information that is missing). However, almost half of the sixty staff and students at Harvard Medical School estimated this probability as 0.95, or 95 per cent, not 2 per cent. Only eleven participants answered 2 per cent. Note the variability in the judgements of physicians about the probability of the disease! The modal answer of 0.95 was taken as an instance of the *base-rate fallacy*. This term signifies that the base rate of the disease (1/1000) is neglected, and the judgement is based only (or mainly) on the characteristics of the test (the false-positive rate). Tversky and Kahneman (1982) used the results of this study to illustrate the generality and stability of the base-rate fallacy, a cognitive illusion that has been widely discussed and given much prominence. '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' (Kahneman and Tversky, 1973: 243). Little is known about how the participants made these judgements, and why these were so variable. It just seems that students and staff did not get effective training in statistical reasoning at Harvard Medical School.

How to make the base-rate fallacy disappear

I will now apply the same argument to the Harvard Medical School problem as I did to the Linda problem. Assume there is some kind of algorithm for statistical reasoning that works on frequency representations. Therefore, if we change the information representation in the Harvard Medical School problem from percentages and single-event probabilities to frequencies, then the base-rate fallacy should also disappear. As a consequence, the large variability in judgements should disappear. This is a testable prediction.

When I made this prediction during luncheon discussions at the Center for Advanced Study in the Behavioral Sciences, two of the other fellows, Leda Cosmides and John Tooby, got up from the table and went down the hill to Stanford University, where they tested the prediction with 425 undergraduate subjects (Cosmides and Tooby, 1991). They constructed a dozen or so versions of the medical problem as controls; of chief interest here is the frequency version:

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 version, the representation of the input information is changed from percentages, such as 5 per cent, to frequencies such as '50 out of 1000'. The representation of the output information is changed from a single-event probability ('What is the probability that a person . . .?') to a frequency judgement ('How many people . . . ?'). This made the proportion of Bayesian answers skyrocket from 12 per cent (in a replication using the original representation) to 76 per cent (and to 92 per cent, if subjects were instructed to visualize frequencies in a graphical display).

If only the representation of the input information was changed into frequencies, but not that of the output information, or vice versa, the effect of the change in information representation was halved. All other changes, such as adding the missing information about the false-negative rate and the explicit information about random sampling, had little effect on the judgements, as the control versions showed.

We have the same result as for the Linda problem. Judgements labelled 'base-rate fallacy' largely disappear in the Harvard Medical School problem when we change the information representation from single events to frequencies. The effect is about as strong as in the Linda problem.

Results in the same direction have been obtained on other reasoning problems when information representation was only partially changed into a frequency format, such as using sequential monitoring of frequency information and random sampling from a collective (e.g. Borgida and Brekke, 1981; Gigerenzer *et al.*, 1988; McCauley and Stitt, 1978).[3]

It is also instructive that some researchers tend to change their own information representation when they turn away from the subject and explain the correct solution to the reader. An early example is Hammerton, who used single-event probabilities to communicate information to his subjects:

> 1. A device has been invented for screening a population for a disease known as psylicrapitis. 2. The device is a very good one, but not perfect. 3. If someone is a sufferer, there is a 90% chance that he will be recorded positively. 4. If he is *not* a sufferer, there is still a 1% chance that he will be recorded positively. 5. Roughly 1% of the population has the disease. 6. Mr. Smith has been tested, and the result is positive. The chance that he is in fact a sufferer is: ____.
>
> (Hammerton, 1973: 252)

When the author explains the correct answer to his readers, he switches, without comment, into a frequency representation:

> Out of every 100 persons tested, we expect 1 to have the disease; and the device is nearly certain to say that he has. Also, out of that 100, we expect the machine to say that 1 healthy person has the disease. Thus, in the long run, out of every 100 persons tested, we expect 2 positive results, one of which will be correct and the other incorrect. Therefore the odds on any positive result being valid are roughly even.
>
> (ibid: 252)

The frequency format is easily digested by Hammerton's readers. However, Hammerton's subjects not surprisingly failed on the single-event representation. Their median response was not one-to-one (i.e. 50 per cent), but 85 per cent.

Thus far, we have seen how to make two cognitive illusions, the conjunction fallacy in the Linda problem and the base-rate fallacy in the Harvard Medical School problem, largely disappear. I will now turn to a third prominent illusion.

Overconfidence bias

Confidence in one's 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%

Imagine you are an experimental subject: your task is to choose one of the two alternatives. Possibly you chose Islamabad, as most subjects in previous studies did. (If your choice was indeed Islamabad, you agree with the majority of subjects but are, regrettably, wrong.) Then you are asked to rate your confidence that your answer 'Islamabad' is correct. Fifty per cent confident means guessing, 100 per cent confident means that you are absolutely sure that Islamabad is the larger city. After many subjects answer many questions, the experimenter counts how many answers in each of the confidence categories were actually correct.

The major finding of some two decades of research is the following: in all the cases where subjects said, 'I am 100 per cent confident that my answer is correct', the relative frequency of correct answers was only about 80 per cent; in all the cases where subjects said, 'I am 90 per cent confident' the relative frequency of correct answers was only about 75 per cent; when subjects said 'I am 80 per cent confident' the relative frequency of correct

answers was only about 65 per cent, and so on (Lichtenstein *et al.*, 1982). Values for confidence were systematically *higher* than relative frequencies. This systematic discrepancy has been interpreted as an error in reasoning and has been named 'overconfidence bias'. Quantitatively, overconfidence bias is defined as the difference between mean confidence and mean percentage correct.

Consistent with the general research strategy of the heuristics-and-biases programme, the explanandum is a *discrepancy* (overconfidence bias) between a confidence judgement and a norm (frequency), not the confidence judgements by themselves (there are some exceptions, e.g. May (1987)). Little, however, has been achieved in explaining this discrepancy. A common proposal is to explain 'biases' by other, deeper mental flaws. For instance, Koriat *et al.* (1980) propose that the overconfidence bias is caused by a 'confirmation bias'. Here is their explanation. After one alternative is chosen (e.g. Islamabad), the mind searches for further information that confirms the answer given, but not for information that could falsify it. This selective information search artificially increases confidence. The key idea is that the mind is not a Popperian. Other deficiencies in cognition and motivation have been suggested as explanations: Fischhoff, Edwards, and others proposed that subjects are insensitive to item difficulty (von Winterfeldt and Edwards, 1986: 128). Dawes suggested the tendency of humans in the western world to overestimate their intellectual powers, which 'has been reinforced by our realization that we have developed a technology capable of destroying ourselves' (Dawes, 1980: 328). Others have proposed motivational reasons such as 'fear of invalidity' or 'illusion of validity'. Note that in all these explanatory attempts the experimental phenomenon is seen as a 'cognitive illusion', that is, an error in probabilistic reasoning, that has to be explained by some deeper flaw in our mental or motivational programming.

Similar to the conjunction fallacy, overconfidence bias has been suggested as an explanation for human disasters of many kinds, including deadly accidents in industry (Spettell and Liebert, 1986), errors in the legal process (Saks and Kidd, 1980), and systematic deviations from rationality in negotiation and management (Bazerman and Neale, 1986).

Checking the normative yardstick

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 average 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 in the long run? From the point of view of the frequency interpretation, the answer is 'no', for the reasons already discussed. Probability theory is restricted to frequencies; it does not apply

to single-event judgements like confidences. Therefore, no statement about confidences can violate the laws of probability. Even for Bayesians, however, the answer is not 'yes', as it was with the conjunction fallacy. The issue here is not internal consistency or coherence, but the relation between subjective probability and external (objective) frequencies, which is a more complicated issue and depends on conditions such as exchangeability (for a discussion related to overconfidence see Kadane and Lichtenstein, 1982).

To summarize: a discrepancy between confidence in single events and relative frequencies in the long run is not an 'error' in the sense of a violation of probability theory, contrary to the claims in the heuristics-and-biases literature. It only looks that way from the perspective of a narrow interpretation of probability theory that blurs the fundamental distinction between single events and frequencies.

How to make overconfidence bias disappear

Many experiments have demonstrated the stability of the overconfidence phenomenon despite various 'debiasing methods' (Fischhoff, 1982). In our own experiments, we have also confirmed the stability despite *those* methods (Gigerenzer *et al.*, 1991). We warned subjects, prior to the experiment, of overconfidence, or gave them monetary incentives – this did not decrease overconfidence. We tried it with a bottle of French champagne as an incentive – to no avail. To quote von Winterfeldt and Edwards (1986: 539): 'Overconfidence is a reliable, reproducible finding.' And they conclude, with a tone of regret 'Can anything be done? Not much' (Edwards and von Winterfeldt, 1986: 656). Let's see.

I will now apply to the overconfidence bias the same argument as before to the conjunction fallacy and base-rate fallacy. Assume an experiment in which you present subjects with fifty general-knowledge questions of the Hyderabad–Islamabad type and ask them for confidence judgements, as usual. Here is where this experiment diverges from earlier work. You also ask the same subjects about judgements of the frequency of correct answers: 'How many of these fifty questions do you think you have answered correctly?' Assume your subjects' mean confidence judgements are, just like in earlier studies, systematically higher than their relative frequency of correct answers. That is, you replicate the earlier findings and get a typical overconfidence bias of about 15 per cent. What do you guess how your subjects' frequency judgements will compare with the true frequency of correct answers?

If confidence in one's knowledge were truly biased due to confirmation bias, wishful thinking, or other deficits in cognition, motivation, or personality, then the difference between a single-event and a frequency representation should not matter. Overestimation should remain stable, just as it does with warnings and French champagne.

Table 11.1 Overestimation disappears in judgements of frequency

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

Ulrich Hoffrage, Heinz Kleinbölting and I have performed this and related experiments (for details see Gigerenzer *et al.*, 1991). Table 11.1 shows the results of two experiments with 80 and 97 subjects, respectively. Only averages are shown here, because individual results conformed well to averages. In both experiments, the stable discrepancy between mean confidence and the true relative frequency of correct answers could be replicated. This is necessary for control, but no surprise. Overconfidence bias, expressed in percentage (by multiplying the difference by the factor 100) was 13.8 per cent and 15.4 per cent, respectively. What about the frequency judgements? When we compared subjects' estimated frequencies with their true frequencies, overestimation *disappeared*. In both experiments subjects showed a tendency towards underestimation. In Table 11.1, the differences between estimated and true frequencies are also expressed in percentages, for comparison. For instance, in Experiment 1, the average estimated frequency of correct answers (in a series of 50 questions) was 1.2 lower than the true frequency of correct answers, which corresponds to −2.4 in 100 questions, or −2.4 per cent. Negative signs denote underestimation, positive signs overestimation.

To summarize: I have argued that the discrepancy between mean confidence and relative frequency of correct answers, known as 'overconfidence bias', is not an error in probabilistic reasoning. It only looks that way from a narrow normative perspective, in which the distinction between single-event confidence and frequencies is blurred. If we ask our subjects about frequencies instead of single-event confidences we can make this stable phenomenon disappear.

It is easy to see how my argument, illustrated here by three prominent examples, can be extended to and tested for other cognitive illusions. The philosophical distinction between single-event probabilities and frequencies teaches us that the irrationality claim, at least as based on these examples, is premature. The normative yardstick does not stand up to closer examination. The distinctions between algorithm and information representation, and between single event and frequencies, combined with the notion of the mind as a frequency-monitoring device, teaches us how to make apparently stable cognitive illusions disappear. This is of course

good news for those who would like to believe in human rationality, or for those biologically minded people who wonder how a species so bad at statistical reasoning could have survived so long, and also for those unfortunate souls charged with teaching undergraduate statistics.

Earlier explanations of reasoning in terms of a general representativeness heuristic or a general confirmation bias cannot account for these striking results. We have to look for a fresh understanding of cognitive processes that explains both the old and new facts. What follows is a brief introduction into the theory of probabilistic mental models (Gigerenzer *et al.*, 1991). The theory explains both the old facts (the robust overconfidence and hard–easy effects of the last two decades) the new facts (the disappearance of overconfidence) and makes several other novel predictions.[4]

PROBABILISTIC MENTAL MODELS

I will illustrate the theory of probabilistic mental models (for short, PMM theory) by the following problem:

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

Assume that subjects do not know the answer, but have to make an inference under uncertainty. How is that inference made?

Before I start outlining the theory, a general remark on explanatory strategy is helpful. Our explanandum is confidence and choice, and not overconfidence bias. That is, we attempt to explain judgement, not the deviation of judgement from some controversial norm. As a consequence, we do not need to invoke deeper-level biases (such as confirmation biases) or error-prone heuristics as explanations. This contrasts with the heuristics-and-biases programme. Nor do we invoke explanations that assume perfect knowledge and unlimited computational and attentional capacities, as in traditional rational-choice theories. Instead, PMM theory postulates cognitive mechanisms that work well given limited knowledge, limited attention, and limited computational capacities. In these respects, PMM theory is a model of 'bounded rationality' (Simon, 1955).

PMM theory assumes that a frame of inference is constructed to solve a particular problem such as the Heidelberg–Bonn problem. This frame of inference is called a *PMM*. A PMM generalizes the two alternatives, Heidelberg and Bonn, to a reference class, such as 'all cities in Germany'. And it generalizes the target variable, number of inhabitants, to a network of probability cues that co-vary with the target. Thus, a PMM consists of

Table 11.2 Probability cues for solving tasks of the Heidelberg–Bonn type. Examples given are for the reference class 'cities in Germany'

Probability cues
1 Soccer-team cue (one city's soccer team plays in the soccer 'Bundesliga', the other city's team does not).
2 State capital cue (one city is a state capital, the other city is not).
3 Industrial cue (one city is located in the 'Ruhrgebiet', the other in rural Bavaria).
4 Licence-plate cue (the letter code that identifies a city on a licence plate is shorter for one city than for the other).
5 Familiarity cue (one had heard of one city, but not of the other).
6 Capital cue (one city is a capital, the other city is not).

a reference class (that contains the two alternatives), a target variable, and probability cues.

Table 11.2 shows examples of probability cues for population size in the reference class 'German cities'. Take the soccer-team cue. Large cities are likely to have a team playing in the Soccer Bundesliga, in which the eighteen best teams compete. The *ecological validity* of this cue can be determined by checking all pairs in which one city has a team in the Bundesliga but the other does not. For instance, one finds that in 91 per cent of these cases the city with the Bundesliga team has more inhabitants (calculated for 1988/89, for what then were West German cities with more than 100,000 inhabitants). Thus, the ecological validity of the soccer cue is 0.91 in this reference class. Note that it is defined as a relative frequency, not as a Pearson correlation as in Brunswik's (1955) framework. Ecological validity is defined on the environment, whereas *cue validity* is the corresponding concept in a subject's PMM. I will call a PMM well-adapted if the cue validities correspond well to the ecological validities.

Note, however, that the soccer team cue cannot be *activated* for the Heidelberg–Bonn problem: neither city has a team in the Bundesliga; so the cue does not differentiate. In fact, only the last cue in this list can be activated, and this *capital cue* does not have a particularly high cue validity – because it is well known that Bonn is not exactly London or Paris. (The low cue validity may change soon, however, because Bonn's days as capital are numbered.)

PMM theory assumes when activation rates are low or time pressure occurs, as is typical for studies of general knowledge, that the *first* cue that can be activated determines choice (here: Bonn) and that *confidence equals cue validity* (Table 11.3). This algorithm is 'satisficing' (Simon, 1982) in the sense that it produces good, but not necessarily optimal, performance. The algorithm is a variant of bounded rationality (Simon, 1955) in so far as it is designed to work on limited knowledge and on the first cue activated. The latter avoids computationally complex integrations of many cues.

Table 11.3 PMM algorithm for choice and confidence

Task:	Choose the correct alternative, *a* or *b*, and give confidence judgment.
Algorithm:	
Step 1:	Generalize *a* and *b* to a reference class *R*, where $a, b \in R$.
Step 2:	Generate cue C_i highest in cue validity.
Step 3:	Generate values of *a* and *b* for cue C_i. If one or both values are unknown, go back to step 2 and generate the cue next highest in cue validity.
Step 4:	Test whether values of *a* and *b* differ, i.e., whether C_i can be activated. If yes, denote this as aC_ib. If not, go back to step 2.
Step 5:	Choose *a* if $p(a\|aC_ib;R) > p(b\|aC_ib;R)$. (For example, let aC_ib stand for '*a* has a soccer team in the Bundesliga but *b* does not'. Then $p(a\|aC_ib;R)$ is the probability that *a* has the larger population given aC_ib, for all $a,b \in R$. This probability is the cue validity, and *R* is the reference class.)
Step 6:	Confidence = $p(a\|aC_ib;R)$. (The confidence that the choice *a* is correct is equal to the cue validity of the activated cue C_i.)

Source: Gigerenzer *et al.* (1991).
Note: Knowledge of cues can be limited, i.e. only a subset of all ecological valid cues may be available from memory (step 2). Knowledge of values can be limited, too. Cues have binary values (yes/no; see Table 11.2), but knowledge is tertiary (yes/no/unknown; see step 3).

Table 11.4 Probabilistic mental models for single-event (confidence) and frequency tasks

PMM	*Confidence task*	*Frequency task*
Target variable	Number of inhabitants	Number of correct answers
Reference class	Cities in Germany	Similar sets of general-knowledge questions in similar testing situations
Probability cues	E.g. soccer team cue, state capital cue	E.g. base rates of previous performance

Now consider a frequency task. Subjects answer several hundred questions of the Heidelberg–Bonn type. After each group of fifty questions they are asked: 'How many of the last fifty questions do you think you have answered correctly?' The point is that according to PMM theory, confidence and frequency judgements are based on different cognitive processes, because different PMMs have to be constructed (Table 11.4).

The target variable in the confidence task is number of inhabitants, whereas in the frequency task it is number of correct answers. As a consequence, reference class and probability cues are different, too. A soccer cue, for example, no longer helps. A task that involves judgements of frequencies of correct answers has a different reference class: sets of general knowledge questions in similar testing situations. And base rates of earlier performance in such testing situations are an example of a probability cue for frequency judgements. Note that both single-event

confidence and frequency judgements are explained by reference to experienced frequencies. However, these experienced frequencies relate to different reference classes, which are in turn part of different PMMs.

PMM theory can be quantitatively simulated; for the present purpose, however, qualitative predictions are sufficient. In the following sections, I will derive several novel predictions from PMM theory, some of them being counterintuitive and therefore surprising. First, however, we will see how PMM theory explains the stable overconfidence bias.

Explaining old facts: overconfidence bias

PMM theory explains the stable overconfidence effect in the following way. Assume that subjects' PMMs are, on the average, well adapted. This means that although subjects' knowledge about some domain (such as about German urban centres) may be limited, it should not be systematically biased. This implies that cue validities roughly correspond to ecological validities, but it does not imply that subjects know all the relevant cues. If the general-knowledge questions were a representative sample from the knowledge domain, zero overconfidence would be expected. For instance, if the soccer cue has an ecological validity of about 0.9, and the cue validity matches this, it follows from PMM theory that confidence would be around 0.9 in those cases where the soccer cue can be activated. From the definition of the ecological validity it follows that the relative frequency of correct answers would be 0.9, too. However, general-knowledge questions typically are not representative samples from some domain of knowledge, but are selected to be difficult or even misleading. The Hyderabad–Islamabad question is an example for a misleading question. Here, a usually valid cue, the capital cue (Islamabad is a capital, Hyderabad is not), leads to a wrong choice: Hyderabad has a much larger population.

Selecting difficult and misleading questions decreases the number of correct answers, and 'overconfidence bias' results as a consequence of selection, not of some deficient mental heuristic. To the best of my knowledge, all previous studies that have demonstrated overconfidence in general knowledge have used selected questions: this explains the stability of the phenomenon against warning, monetary incentives, and French champagne. Here are several novel predictions.

Novel predictions

Prediction 1. Confidence and representative sampling

Assume (1) well-adapted PMMs as above, and (2) use a *representative* sample of questions from some knowledge domain. Then, PMM theory

predicts that overconfidence will disappear. We have tested this prediction using random samples from the reference class 'all cities in Germany with more than 100,000 inhabitants' (Gigerenzer *et al.*, 1991). In Experiment 1, 'overconfidence bias' decreased from 13.8 per cent in a set of selected questions to 0.9 per cent in a representative sample; in Experiment 2 this decrease replicated from 15.4 per cent to 2.8 per cent. Juslin (in press, a) independently confirmed this novel prediction using random samples from several other domains of knowledge.

Prediction 2. Frequency judgements and selected sampling

Recall that PMM theory implies that frequency judgements such as 'How many of the last fifty questions do you think you got right?' are solved by a PMM with a different reference class (e.g. other general-knowledge tests). Assume (1) that the PMMs for a frequency task are well adapted and (2) use a set of questions that are representative for *this* reference class. Because the typical sets of general-knowledge questions used in earlier research are representative for this reference class, frequency judgements should be accurate. We have tested and confirmed this novel prediction (see Table 11.1).

The crucial point is that confidence and frequency judgements refer to different kinds of reference classes. The same set of questions can be representative with respect to one reference class, and at the same time selected with respect to the other class. Thus a set of fifty general-knowledge questions of the city-type may be representative for the reference class 'general-knowledge questions', but not for the reference class 'cities in Germany' (because city pairs have been selected for being difficult or misleading). Asking for a confidence judgement summons up a PMM based on the reference class 'cities in Germany'; asking for a frequency judgement summons up a PMM based on the reference class 'sets of general-knowledge questions'.

Prediction 3. Underestimation in frequency judgements

We use here the situation of prediction 1 to deduce a condition in which frequency judgements *underestimate* the true frequency of correct answers. If a PMM for frequency judgement is well adapted to its reference class (i.e. sets of selected items), but the actual set of questions is not selected, then we expect frequency judgements to be underestimations of true frequencies. We have tested and confirmed this novel prediction (Gigerenzer *et al.*, 1991). In Experiment 1, the difference between estimated and true frequency of correct answers decreased from −2.4 per cent (set of selected items, see Table 11.1) to −11.8 per cent; and from −4.2 per cent (see Table 11.1) to −9.3 per cent in Experiment 2.

Further novel predictions can be derived from quantitative simulations of PMM theory. Here is one last example. The prediction is about percentage correct, that is, about correct choice rather than about confidence or frequency judgements, on which we have focused so far.

Prediction 4. When little knowledge is as good as good knowledge

Recall that in the experiments just reported, our subjects were German, and they were answering questions about German cities. Their mean percentage of correct answers varied between 70 per cent and 75 per cent (for representative samples of cities). Assume we take a new sample of German students who are just as good as the earlier ones – they are familiar with German cities and know the relevant probability cues. We do the same kind of experiment; the only difference is that we give them questions about an environment which is highly unfamiliar to them: cities in the USA. More precisely, we take the 75 largest cities in the USA, draw a random sample of 100 pairs, and give these 100 questions to our German subjects. What would you predict?

All theories of overconfidence I am aware of are mute on the issue of percentage correct. All the people I have asked so far concluded that percentage correct will be much lower when subjects answer these 100 questions about foreign cities. From our simulations with PMM theory, we derived a quite different and surprising prediction: subjects will do just as well with American as with German cities. That is, their percentage correct will be the same for German and US cities. I will deduce this prediction here by a simplified calculation.

We take the 75 largest cities in the USA. Assume that our German subjects have not even heard of half of these, such as Mesa, Mobile, and Shreveport, and that they know nothing about the other half, except that they have heard of these cities. Thus, their PMM is poor; the only probability cue it can generate is the familiarity cue, that is whether one has heard of a city or not. This familiarity cue is of high cue validity, but it plays almost no role in judgements about German cities, because most of our subjects have heard of all these German cities. Thus, it can rarely be activated. The point is that for judgements about US cities, the familiarity cue has a high activation rate. To be precise, if half of the US cities are familiar, the activation rate is 50.7 per cent.[5] What is the validity of the cue? Pretests have shown that the cue validity is around 0.90.[6] Thus, we have about 50 per cent of questions where the familiarity cue can be activated, and 50 per cent where it cannot (because either the names of both cities are known or both unknown). For the first group, we expect 90 per cent correct – given a cue validity of 0.90 – that is, in absolute terms, 45 per cent correct answers. In the other group, we expect by mere guessing an additional 25 per cent correct answers, that is, altogether, 70

per cent correct answers. This value is counterintuitively large. Note that this value is in the range of the percentage correct for German cities (70–75 per cent), although it has been calculated on the assumption of no specific knowledge. Any such knowledge (e.g. that New York is larger than Boston) will add on to this estimate.

Thus, PMM theory makes a counterintuitive prediction: in the situation described, German subjects will get about the same percentage correct in judgements about unfamiliar US cities as in judgements about German cities.

Horst Kilcher, Ulrich Hoffrage, and I conducted an experiment. Fifty-six subjects each answered 200 questions of the Heidelberg–Bonn type, 100 being a random sample of city pairs from the 75 largest US cities, the other 100 being a random sample from the 75 largest German cities. Half of the subjects got the questions about the German cities first, the other half those about US cities. Consistent with our earlier experiments, mean percentage correct was 75.6 per cent for German cities. But what was the percentage correct for judgements about US cities?

Table 11.5 shows that mean percentage correct for US cities was 76 per cent, that is, about the same as for the German cities about which our subjects had considerably more knowledge. This result follows from PMM theory. Here, we have an interesting situation, where quite limited knowledge (but not *no* knowledge) produces the same good performance (percentage correct) as quite good knowledge.

To summarize my second part: I have briefly presented PMM theory, which specifies the cognitive processes underlying choice, confidence, and frequency judgements. The theory implies conditions under which overconfidence appears: either a PMM for a task is not properly adapted to a corresponding environment (for example, cue validities do not correspond to ecological validities), or the set of objects used is not a representative sample from the corresponding environment, but is selected for difficulty. In our experiments, overconfidence disappeared when random samples instead of selected samples were used, which is consistent with the latter explanation. Thus, the source of overconfidence seems to be in the relation between the sample of objects used in the task and the reference class in a corresponding environment. Overconfidence does not seem to be located in the relation between PMMs and corresponding environments (that is, in a low correspondence between cue validities and ecological validities).

Table 11.5 Mean percentage of correct answers

	US	*German*
Mean percentage correct	76.0 (SE_m=0.7)	75.6 (SE_m=0.9)
Mean confidence	72.3 (SE_m=1.0)	79.5 (SE_m=0.8)

PMM theory specifies conditions under which the 'robust' overconfidence effect of the last fifteen years appears, disappears, and even inverts. One can no longer speak of a general overconfidence bias, in the sense that it relates to deficient processes of cognition or motivation. I have not dealt here with how the theory explains the second robust finding in the literature – the hard–easy effect (that is, overconfidence increases with item difficulty). I will simply mention that the theory also provides an explanation for the hard–easy effect on the same principles, and specifies conditions under which it disappears or even inverts. Juslin (in press, b) has tested and confirmed a prediction from PMM theory that specifies conditions that make the hard–easy effect disappear (Gigerenzer *et al.*, 1991: 512). Simulations with PMM theory have led us to explain several anomalies in the literature, and to integrate earlier explanatory attempts into a comprehensive theoretical framework. For instance, Koriat's and colleagues' (1980) results testing the confirmation bias explanation can be fully integrated into PMM theory (Gigerenzer *et al.*, 1991). PMM theory seems to be the first theory in this field that offers a coherent explanation not only of the effects previously reported in the literature on judgement under uncertainty, but also for the new results we have obtained in our experiments.

CONCLUSIONS

Since the Enlightenment, probability theory has been seen as the codification of human rationality. Consequently, recent experiments suggesting that human reasoning systematically violates the laws of probability have been widely cited as evidence for human irrationality. Here are the arguments of this chapter.

1 I have argued that the cognitive illusions I have dealt with are not genuine illusions, contrary to the assertions in the heuristics-and-biases literature. They only look like errors from a narrow normative view about what is right and wrong in reasoning, a view that blurs the philosophical distinction between single-event probabilities and frequencies.
2 I have linked this philosophical distinction with Marr's (1982) distinction between algorithms and information representation, and with the evolutionary idea that the mind's algorithms are tuned to frequency information. This framework suggests how to make apparently stable cognitive illusions disappear. I have demonstrated this using three cognitive illusions, widely cited as evidence for human irrationality. The new facts cannot be accounted by the old explanations invoking heuristics such as representativeness.
3 I introduced the theory of probabilistic mental models (PMM theory) as

an alternative explanation of intuitive reasoning, using confidence in one's knowledge as an example. The theory explains both old and new facts. PMM theory postulates a mental algorithm that processes frequency information from the environment. This algorithm works well given only limited knowledge, limited attention, and limited computational capacities, and is a variant of bounded rationality. The theory describes reasoning and performance in terms of relations between a PMM, an environment, and an experimental task. Focusing on mental algorithms alone, whether they seem to be good or bad ones, turns out to be too narrow for understanding the mind, and also, for discussing rationality.

ACKNOWLEDGEMENT

This chapter is based on a lecture delivered at Harvard University, 2 October 1991. I wrote this chapter under a fellowship at the Center for Interdisciplinary Research, University of Bielefeld, Germany, and with the support of the *Fonds zur Förderung der wissenschaftlichen Forschung* (*P 8842-MED*), Austria. I am grateful to Lorraine Daston, Ralph Hertwig and Ulrich Hoffrage for many helpful comments.

NOTES

1 The debate between the frequentists and Bayesians was particularly lively before the 1970s. Today, both sides know each other's arguments well and the vital debate has turned into sterile, well-rehearsed argument. The two sides seem to have quit listening. As Glenn Shafer (1989) complained, statistics departments no longer provide a forum for the debate, and the main divisions over the meaning of probability now follow disciplinary lines: frequentists dominate statistics and experimental social sciences, Bayesians predominate in artificial intelligence and theoretical economics. 'Conceptually and institutionally, probability has been balkanized' (Shafer, 1989: 15).

2 The attentive reader will have noticed that the frequency version of the Linda problem asks for a quantitative judgement, whereas the single-case version asks for a comparative judgement. The latter, however, is an accidental feature of our choice of example. Single-case versions asking for quantitative judgements ('What is the probability that Linda is . . . ?') are known to give about the same amount of conjunction errors as comparative judgements (Tversky and Kahneman, 1983).

3 In some studies widely cited as demonstrating base-rate neglect, subjects were not informed about random sampling. In the 'Tom W.' Problem (Kahneman and Tversky, 1973), the crucial information about how the personality sketch of Tom W. was selected, whether randomly or not, is missing. The same holds for the Gary W. and Barbara T. problems that Ajzen (1977) used. Several studies have demonstrated that it can make a difference to subjects' reasoning when they learn about random sampling (e.g. Ginossar and Trope, 1987; Grether, 1980; Hansen and Donoghue, 1977; Wells and Harvey, 1977), or, even better, when they can actually watch random sampling. For instance, the neglect of base

rates in the engineer–lawyer problem (Kahneman and Tversky, 1973) largely disappears when subjects themselves *do* the random sampling (Gigerenzer *et al.*, 1988). For general critical discussions of the evidence see Berkeley and Humphreys (1982), Gigerenzer and Murray (1987, Ch. 5), Lopes (1991), Lopes and Oden (1991), Macdonald (1986), and Scholz (1987).

4 Other proposals have been made in the literature to explain the old facts, that is, the cognitive illusions. I cannot discuss these here, but only mention a few: the role of conversational principles in the experimenter–subject interaction (Adler, 1991); the evolutionary idea that there are domain-specific reasoning mechanisms (e.g. cheating detection) that reflect our inherited social intelligence rather than a domain-general logic (e.g. Cosmides, 1989; Gigerenzer and Hug, 1992), and the idea that category judgements such as in probability revision problems and in the Linda problem can be modelled by connectionist architectures (e.g. Gluck and Bower, 1988).

5 There are 75 cities, 38 are familiar, 37 not (or 37 familiar, 38 not, which leads to the same result). If two familiar cities are compared, or two unfamiliar ones, the familiarity cue cannot be activated; it can only be activated if one city is familiar but the other is not. The number of such familiar–unfamiliar pairs is 38×37, and the number of all possible pairs is 75×74/2. Thus, the activation rate is 38×37 divided by 75×74/2, which is 38/75 or 50.7 per cent. The activation rate can be determined in this way for each individual separately depending on the number of familiar and unfamiliar cities. For instance, if not one-half, but only one-third of the cities were familiar, the activation rate would change slightly, from 50.7 per cent to 45 per cent.

6 The cue validity of the familiarity cue can be calculated for each individual from the set of familiar–unfamiliar pairs. The relative frequency in which the familiar city actually has the larger population is the cue validity.

REFERENCES

Adler, J.E. (1991) 'An optimist's pessimism: conversation and conjunction', *Posnan Studies in the Philosophy of the Sciences and Humanities* 21: 251–82.

Ajzen, I. (1977) 'Intuitive theories of events and the effects of base-rate information on prediction', *Journal of Personality and Social Psychology* 35: 303–14.

Aristotle (1945) *Movements of Animals*, trans. E.S. Foster, Cambridge, MA: Harvard University Press.

Barnard, G.A. (1979) 'Discussion of the paper by Professors Lindley and Tversky and Dr. Brown', *Journal of the Royal Statistical Society*, (A), 142: 171–2.

Bazerman, M.H. and Neale, M.A. (1986) 'Heuristics in negotiation', in H.R. Arkes and K.R. Hammond (eds) *Judgment and Decision making: An Interdisciplinary Reader*, Cambridge: Cambridge University Press, pp. 311–21.

Berkeley, D. and Humphreys, P. (1982) 'Structuring decision problems and the "bias heuristic"', *Acta Psychologica* 50: 201–52.

Boole, G. (1958) *An Investigation of the Laws of Thought on which are Founded the Mathematical Theories of Logic and Probabilities*, New York: Dover (original work published in 1854).

Borgida, E. and Brekke, N. (1981) 'The base rate fallacy in attribution and prediction', in J.H. Harvey, W. Ickes, and R.F. Kidd (eds) *New Directions in Attribution Research*, vol. 3, Hillsdale, NJ: Erlbaum, pp. 63–95.

Brunswik, E. (1955) 'Representative design and probabilistic theory in a functional psychology', *Psychological Review* 62: 193–217.

Casscells, W., Schoenberger, A., and Grayboys, T. (1978) 'Interpretation by physicians of clinical laboratory results', *New England Journal of Medicine* 299: 999–1000.

Cohen, L.J. (1982) 'Are people programmed to commit fallacies? Further thoughts about the interpretation of experimental data on probability judgment', *Journal for the Theory of Social Behavior* 12: 251–74.

—— (1983) 'The controversy about irrationality', *Behavioral and Brain Sciences* 6: 510–17.

—— (1986) *The Dialogue of Reason*, Oxford: Clarendon Press.

Cosmides, L. (1989) 'The logic of social exchange: has natural selection shaped how humans reason? Studies with the Wason selection task', *Cognition* 31: 187–276.

Cosmides, L. and Tooby, J. (1991) 'Are humans good intuitive statisticians after all? Rethinking some conclusions from the literature on judgment under uncertainty', manuscript submitted for publication.

Daston, L. (1988) *Classical Probability in the Enlightenment*, Princeton, NJ: Princeton University Press.

Dawes, R.M. (1980) 'Confidence in intellectual judgments vs. confidence in perceptual judgments', in E.D. Lantermann and H. Feger (eds) *Similarity and Choice: Papers in Honor of Clyde Coombs*, Bern, Switzerland: Huber, pp. 327–45.

Edwards, W. and von Winterfeldt, D. (1986) 'On cognitive illusions and their implications', in H.R. Arkes and K.R. Hammond (eds) *Judgment and Decision Making*, Cambridge: Cambridge University Press, pp. 642–79.

Evans, J. St B.T. (1984) 'Heuristic and analytic processes in reasoning', *British Journal of Psychology* 75: 451–68.

Fiedler, K. (1988) 'The dependence of the conjunction fallacy on subtle linguistic factors', *Psychological Research* 50: 123–9.

Fischhoff, B. (1982) 'Debiasing', in D. Kahneman, P. Slovic, and A. Tversky (eds) *Judgment under Uncertainty: Heuristics and Biases*, Cambridge: Cambridge University Press, pp. 422–44.

Fong, G.T. and Nisbett, R.E. (1991) 'Immediate and delayed transfer of training effects in statistical reasoning', *Journal of Experimental Psychology: General* 120: 34–45.

Gallistel, C.R. (1990) *The Organization of Learning*, Cambridge, MA: MIT Press.

Gigerenzer, G. (1991) 'On cognitive illusions and rationality', *Poznan Studies in the Philosophy of the Sciences and the Humanities* 21: 225–49.

Gigerenzer, G. and Hug, K. (1992) 'Domain-specific reasoning: social contracts, cheating, and perspective change', *Cognition* 43: 127–71.

Gigerenzer, G. and Murray, D.J. (1987) *Cognition as Intuitive Statistics*, Hillsdale, NJ: Erlbaum.

Gigerenzer, G., Hell, W., and Blank, H. (1988) 'Presentation and content: the use of base rates as a continuous variable', *Journal of Experimental Psychology: Human Perception and Performance* 14: 513–25.

Gigerenzer, G., Hoffrage, U., and Kleinbölting, H. (1991) 'Probabilistic mental models: a Brunswikian theory of confidence', *Psychological Review* 98: 506–28.

Ginossar, Z. and Trope, Y. (1987) 'Problem solving in judgment under uncertainty', *Journal of Personality and Social Psychology* 52: 464–74.

Gluck, M.A. and Bower, G.H. (1988) 'Evaluating an adaptive network model of human learning', *Journal of Memory and Language* 27: 166–95.

Gould, S.J. (1992) *Bully for Brontosaurus. Further Reflections in Natural History*, Harmondsworth: Penguin.

Grether, D.M. (1980) 'Bayes rule as a descriptive model: the representativeness heuristic', *The Quarterly Journal of Economics* 95: 537–57.
Hammerton, M. (1973) 'A case of radical probability estimation', *Journal of Experimental Psychology*, 101: 252–4.
Hansen, R.D. and Donoghue, J.M. (1977) 'The power of consensus: information derived from one's own and others' behavior', *Journal of Personality and Social Psychology* 35: 294–302.
Hasher, L. and Zacks, R.T. (1979) 'Automatic and effortful processes in memory', *Journal of Experimental Psychology: General* 108: 356–88.
Hume, D. (1975) *A Treatise of Human Nature*, Oxford: Clarendon Press (original work published 1739).
Inhelder, B. and Piaget, J. (1958) *Growth of Logical Thinking: From Childhood to Adolescence*, New York: Basic Books.
Juslin, P. (in press, a) 'The overconfidence phenomenon as a consequence of informal experimenter-guided selection of almanac items', *Organizational Behavior and Human Decision Processes*.
—— (in press, b) 'An explanation of the hard–easy effect in studies of realism of confidence in one's general knowledge', *European Journal of Cognitive Psychology*.
Kadane, J.B. and Lichtenstein, S. (1982) *A Subjectivist View of Calibration*, Report No. 82–86, Eugene, OR: Decision Research.
Kahneman, D. and Tversky, A. (1972) 'Subjective probability: a judgment of representativeness', *Cognitive Psychology* 3, 430–54. Reprinted in D. Kahneman *et al.* (1982) (eds) *Judgment under Uncertainty: Heuristics and Biases*, Cambridge: Cambridge University Press, pp. 32–47.
—— (1973) 'On the psychology of prediction', *Psychological Review* 80:237–51.
Kanwisher, N. (1989) 'Cognitive heuristics and American security policy', *Journal of Conflict Resolution* 33: 652–75.
Koriat, A., Lichtenstein, S., and Fischhoff, B. (1980) 'Reasons for confidence', *Journal of Experimental Psychology: Human Learning and Memory* 6: 107–18.
Laplace, P.S. (1951) *A Philosophical Essay on Probabilities*, New York: Dover (original work published 1814).
Lichtenstein, S., Fischhoff, B., and Phillips, L.D. (1982) 'Calibration of probabilities: the state of the art to 1980', in D. Kahneman, P. Slovic, and A. Tversky (eds) *Judgment under Uncertainty: Heuristics and Biases*, Cambridge: Cambridge University Press, pp. 306–34.
Lopes, L. (1991) 'The rhetoric of irrationality', *Theory & Psychology* 1: 65–82.
Lopes, L. and Oden, G.C. (1991) 'The rationality of intelligence', *Posnan Studies in the Philosophy of the Sciences and the Humanities* 21: 199–223.
McCauley, C. and Stitt, C.L. (1978) 'An individual and quantitative measure of stereotypes', *Journal of Personality and Social Psychology* 36: 929–40.
Macdonald, R.R. (1986) 'Credible conceptions and implausible probabilities', *British Journal of Mathematical and Statistical Psychology* 39: 15–27.
Marr, D. (1982) *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*, San Francisco: Freeman.
May, R.S. (1987) *Realismus von subjektiven Wahrscheinlichkeiten*, Frankfurt/Main: Lang.
Neyman, J. (1977) 'Frequentist probability and frequentist statistics', *Synthese* 36: 97–131.
Nisbett, R.E. and Borgida, E. (1975) 'Attribution and the psychology of prediction', *Journal of Personality and Social Psychology* 32: 932–43.
Piattelli-Palmarini, M. (1989) 'Evolution, selection and cognition: From "learning" to parameter setting in biology and in the study of language', *Cognition* 31: 1–44.

Real, L. and Caraco, T. (1986) 'Risk and foraging in stochastic environments: theory and evidence', *Annual Review of Ecology and Systematics* 17: 371–90.
Saks, M. and Kidd, R.F. (1980) 'Human information processing and adjudication: trial by heuristics', *Law and Society Review* 15: 123–60.
Scholz, R.W. (1987) *Cognitive Strategies in Stochastic Thinking*, Dordrecht, Holland: Reidel.
Shafer, G. (1989) *The Unity and Diversity of Probability*, inaugural lecture, 20 November 1989, University of Kansas.
Simon, H.A. (1955) 'A behavioral model of rational choice', *Quarterly Journal of Economics* 69: 99–118.
—— (1982) *Models of Bounded Rationality*, 2 vols. Cambridge, MA: MIT Press.
Slovic, P., Fischhoff, B., and Lichtenstein, S. (1976) 'Cognitive processes and societal risk taking', in J.S. Carroll and J.W. Payne (eds) *Cognition and Social Behavior*, Hillsdale, NJ: Erlbaum.
Spettell, C.M. and Liebert, R.M. (1986) 'Training for safety in automated person-machine systems', *American Psychologist* 41:545–50.
Staddon, J.E.R. (1988) 'Learning as inference', in R.C. Bolles and M.D. Beecher (eds) *Evolution and Learning*, Hillsdale, NJ: Erlbaum.
Stich, S.P. (1990) *The Fragmentation of Reason*, Cambridge, MA: MIT Press.
Tversky, A. and Kahneman, D. (1974) 'Judgment under uncertainty: heuristics and biases', *Science* 185: 1124–31.
—— (1982) 'Evidential impact of base rates', in D. Kahneman, P. Slovic, and A. Tversky (eds) *Judgment under Uncertainty: Heuristics and Biases*, Cambridge: Cambridge University Press.
—— (1983) 'Extensional versus intuitive reasoning: the conjunction fallacy in probability judgement', *Psychological Review* 90: 293–315.
von Mises, R. (1957) *Probability, Statistics, and Truth*, London: Allen & Unwin (original work published in 1928).
von Winterfeldt, D. and Edwards, W. (1986) *Decision Analysis and Behavioral Research*, Cambridge: Cambridge University Press.
Wells, G.L. and Harvey, J.H. (1977) 'Do people use consensus information in making causal attributions?', *Journal of Personality and Social Psychology* 35: 279–93.