GERD GIGERENZER

PSYCHOLOGICAL CHALLENGES FOR NORMATIVE MODELS

Some years ago, at the Center for Advanced Study at Stanford, one of my economist colleagues concluded a discussion on cognitive illusions with the following dictum: 'Look, either reasoning is rational or it's psychological'. In this chapter, I argue against the widespread view that the rational and the psychological are opposed. According to this view, the rational is defined by the laws of probability and logic — that is, by content-free axioms or rules, such as consistency, transitivity, Bayes's theorem, dominance, and invariance. The irrational is left to be explained by the laws of psychology. Here I present examples that are intended to illustrate that defining human rationality independent of psychology is myopic. The 'challenges' in the title of this chapter are not directed against probability theory and logic, or specific versions thereof, but against using these systems as psychologically uninformed, content-free norms. Before I turn to these challenges, I begin with a historical example that illustrates how norms have been revised and made more realistic by the introduction of psychological concepts.

Back to the Blackboard

In the 17th century, a new conception of rationality emerged. This was a modest kind of reasonableness that could handle everyday dilemmas on the basis of uncertain knowledge, in contrast to the traditional rationality of demonstrative certainty [Daston, 1981; 1988]. Those dilemmas were numerous: Believe in God? Invest in an annuity? Accept a Gamble? The mathematical theory of probability was to codify this new brand of rationality, and its primitive concept was rational *expectation*, with expectation defined as *expected value*. Soon, however, it became apparent that minds do not always follow the dictates of the expected value. The *St. Petersburg paradox*, explicated below, marked the celebrated clash between the new theory of expected value and human intuition.

Pierre offers to sell Paul an opportunity to play the following coin-tossing game. If the coin comes up heads on the first toss, Pierre agrees to pay Paul $ 1; if heads does not turn up until the second toss, Paul receives $ 2; if not until the third toss, $ 4; and so on. According to the standard method of calculating expected value, Paul's expectation E—and therefore the *fair price* of the game—is

$$(1) \quad E = 1/2\$1 + 1/4\$2 + 1/8\$4 + 1/16\$8 + \ldots + (1/2)^n\$2^{n-1} + \ldots$$

D.M. Gabbay and Ph. Smets (eds.),
Handbook of Defeasible Reasoning and Uncertainty Management Systems, Vol. 1, 441–467.

Paul's monetary payoffs increase with decreasing probabilities of occurrence: Each of the terms is equal to 50 cents, and the expected value E is infinite. This calculation is straightforward, and there is nothing in the definition of expectation that excludes an infinite value. Nicholas Bernoulli, who first proposed this game in 1713, however, observed that no reasonable person would pay a large amount of money to play the game.

What should be done when the dictates of a norm (E) deviate from human intuitions about the reasonableness of a behaviour? One can either stick to the norm and declare the behaviour irrational, or incorporate the psychological into the norm. In 1738, Nicholas's cousin Daniel Bernoulli published a resolution of the paradox in the annals of the Academy of St. Petersburg (hence the name). Daniel Bernoulli psychologized the norm. He proposed that in situations such as the St. Petersburg gamble, the prospect of winning a certain amount of money, say \$16, means something different for the rich and the poor man. Therefore a theory of reasonableness needs to incorporate personal characteristics such as a person's current wealth, whereas the concept of expected value was based on the impersonal notion of fairness. Bernoulli proposed replacing expected value, which excluded personal circumstances that might prejudice equal rights in legal contexts, with the 'moral' expectation of prudence, defined as the product of the probability of an outcome and what later became known as its *utility*. The utility of money, Bernoulli argued, decreases the more you have.

In modern terminology, let U be the utility of an outcome, w a person's current wealth, and g the sure gain that would yield the same expectation as the St. Petersburg gamble. Then

$$(2)\quad U(w+g) = 1/2U(w+1)+1/4U(w+2)+1/8U(w+4)+\ldots+(1/2)^n U(w+2^{n-1})+\ldots$$

Suppose that $U(x) = ln(x)$, that is, the utility of money for Paul diminishes logarithmically with the amount of money he has. Then, if Paul's current wealth is \$ 50,000, g is about \$ 9. On this psychological assumption, Paul should be willing to pay no more than \$ 9 for the St. Petersburg gamble.[1]

Daniel Bernoulli's revision of *expected value theory* into what is today known as *expected utility theory* exemplifies the Enlightenment attitude toward the relation between the rational and the psychological. By putting psychology into the equations, Bernoulli reunified the rational with the psychological. Expected value theory was a model, not a rigid norm, of rationality. When educated minds reasoned differently from what the theory predicted, this was seen as a problem for the theory, not for the mind, and mathematicians went back to the blackboard to change the equations. Today, as we will see, few researchers respond to such discrepancies by going back to the blackboard and revising their equations. The blame is placed on the mind, not on the model.

[1] Daniel Bernoulli's psychological solution was not the only one, and there exists a large literature on the St. Petersburg Paradox (e.g. [Daston, 1988; Jorland, 1987; Lopes, 1981]).

Separating the mathematical theory of probability from its applications would have seemed foreign to Bernoulli and the Enlightenment probabilists: Their theory was at once a description and a prescription of reasonableness. Along with hydrodynamics and celestial mechanics, the calculus of probability was part of what was then called 'mixed mathematics', a term stemming from Aristotle's explanation of how optics and harmonics mixed the forms of mathematics with the matter of light and sound [Daston, 1992]. Classical probability theory had no existence independent of its subject matter—the beliefs of reasonable men. This is why classical probabilists perceived problems of the St. Petersburg kind as paradoxes—not because there was a mathematical contradiction, but because the mathematical result contradicted good sense.

I invite you in the following pages to look with a Bernoullian eye at some present-day uses of normative models. I proceed by means of examples, each one chosen to illustrate how psychology can be brought to rationality. Some believe that is impure; but don't be misled.

1 CHALLENGE ONE: ALGORITHMS WORK ON INFORMATION THAT NEEDS REPRESENTATION

Probability theory is mute about the representation of the information on which its rules should work. But systems that calculate, machines and minds alike, are sensitive to the representation of numerical information [Marr, 1982]. Computational algorithms work on information, and information needs representation. For instance, my pocket calculator has an algorithm for multiplication. This algorithm is designed for Arabic numbers as input data and would perform badly if I entered binary numbers. Similarly, mental algorithms are designed for particular representations. Consider, for example, how difficult it would be to perform long division with Roman numerals. Arabic, Roman, and binary representations can be mapped onto each other one-to-one and are in this sense mathematically equivalent, but that does not mean they are psychologically equivalent. Physicist Richard Feynman [1967] made this point more generally, explaining that new discoveries can come from different formulations of the same physical law, even if they are mathematically equivalent: 'Psychologically they are different because they are completely unequivalent when you are trying to guess new laws' (p. 53). Let us consider the issue of information representation in research on Bayesian inference.

1.1 The Norm

The question whether humans reason the Bayesian way has been studied in problems with two hypotheses, H and $-H$ (e.g. breast cancer and no breast cancer), and one datum D (e.g. a positive mammogram). Here is one example:

> The probability of breast cancer is 1% for a woman at age forty who participates in routine screening. If a woman has breast cancer, the

probability is 80% that she will have a positive mammogram. If a woman does not have breast cancer, the probability is 9.6% that she will also have a positive mammogram. A woman in this age group had a positive mammogram in a routine screening. What is the probability that she actually has breast cancer? _____%

If one inserts these numbers into Bayes's theorem, the posterior probability $p(H|D)$ is:

$$(3) \quad p(H|D) = \frac{p(H)(D|H)}{p(H)p(D|H) + p(-H)p(D|-H)}$$

$$= \frac{(.01)(.80)}{(.01)(.80) + (.99)(.096)}$$

The result is .078, or 7.8%. In sharp contrast, Eddy [1982] reported that 95 out of 100 physicians estimated the posterior probability p(cancer| positive) to be between 70% and 80%. Psychology undergraduates tend to give the same estimates. Staff at the Harvard Medical School showed not much more insight into a similar problem [Casscells *et al.*, 1978]. In short, very few people have an intuitive understanding of what to do with these probabilities.

Because many people estimated the posterior probability as being close to the hit rate (80%), it has been concluded that mental algorithms generally neglect base-rate information [see Gigerenzer and Murray, 1987, Ch. 5; Koehler, 1996]. Results from these and other studies have been taken as evidence that the human mind does not reason with Bayesian algorithms. Yet this conclusion is not warranted, as the pocket calculator example illustrates. If I feed my pocket calculator binary numbers, and garbage comes out, it does not follow that the calculator has no algorithm for multiplication. Similarly, it would be impossible to detect a Bayesian algorithm in a system by feeding it information in a representation to which it is not tuned. A normative model must therefore specify both the information representation and the algorithm that works on this representation. What are the external representations of information for which cognitive algorithms are designed?

1.2 Psychologizing the Norm: Ecological Bayesianism

The problem we need to solve has one more unknown than the pocket calculator example. In the latter, we know the input representation and so can make informed guesses about the nature of the algorithm. In the case of human minds, we must also speculate about the external representation of statistical information for which cognitive algorithms are designed. We know some candidate representations, and some facts about them. In the mammography problem, information is represented in single-event probabilities (percentages). We know that probabilities and percentages are very recently invented means of representing information. The notion of

'probability' did not gain prominence in probability theory until the 18th century, a century after the calculus of chance was invented [Gigerenzer *et al.*, 1989]. Percentages became common ways to represent numerical information during the 19th century (mainly for interest and taxes), after the metric system was introduced during the French Revolution, and became common tools for representing uncertainty only in this century. Therefore, it is unlikely that cognitive algorithms were designed for probabilities and percentages, if we think in evolutionary terms. In what representation have humans (and animals) acquired numerical information during most of their history? I assume here that they acquired it in terms of *natural frequencies* as actually experienced in a series of events, rather than probabilities or percentages [Cosmides and Tooby, 1996; Gigerenzer and Hoffrage, 1995]. By 'natural frequencies' I mean absolute frequencies (rather than relatie frequencies), as defined by natural sampling (see the right side of Figure 1; [Gigerenzer and Hoffrage, 1995; Kleiter, 1994]).

For a simple demonstration of the role of representation in reasoning, let us represent the information about the base rate (1%), hit rate (80%), and false alarm rate (9.6%) of the mammography problem in natural frequencies rather than percentages. Imagine 100 women. One has cancer (the base rate) and will possibly test positive (the hit rate). Of the 99 women without cancer, about 10 will also test positive (the false alarm rate). So altogether 11 women will test positive. Question: How many of those who will test positive actually have breast cancer? Now most people easily 'see' the answer: one out of 11.

Why is this? Consider Figure 1. On the left side are probabilities (as in a typical medical text), and the Bayesian algorithm a physician would have to use to compute the posterior probability from a probability representation. On the right side, the same information is represented in terms of natural frequencies. The interesting difference is that the Bayesian algorithm is computationally simpler when information is expressed in natural frequencies than in probabilities or percentages. Furthermore, only two kinds of frequencies need be attended to—'symptom & disease', and 'symptom & no disease'. Base rates (e.g., 10 out of 1,000 in Figure 1) need not be attended to; they are implicit in these two frequencies.

The simple demonstration above used approximate figures; a frequency representation of the mammography problem that is numerically equivalent to the probability representation can be constructed by using a class of 1,000 instead of 100 women, as in the 'natural sampling tree' in Figure 1:

> Ten out of every 1,000 women at age 40 who participate in routine screening have breast cancer. Eight out of these 10 women with breast cancer will get a positive mammogram. Of the 990 women without breast cancer, 95 will also get a positive mammogram.
>
> Here is a new representative sample of women at age 40 who got a positive mammogram in routine screening. How many of these women do you expect actually to have breast cancer? _____ out of _____

Ulrich Hoffrage and I have given 15 problems of this kind (concerning cancer,

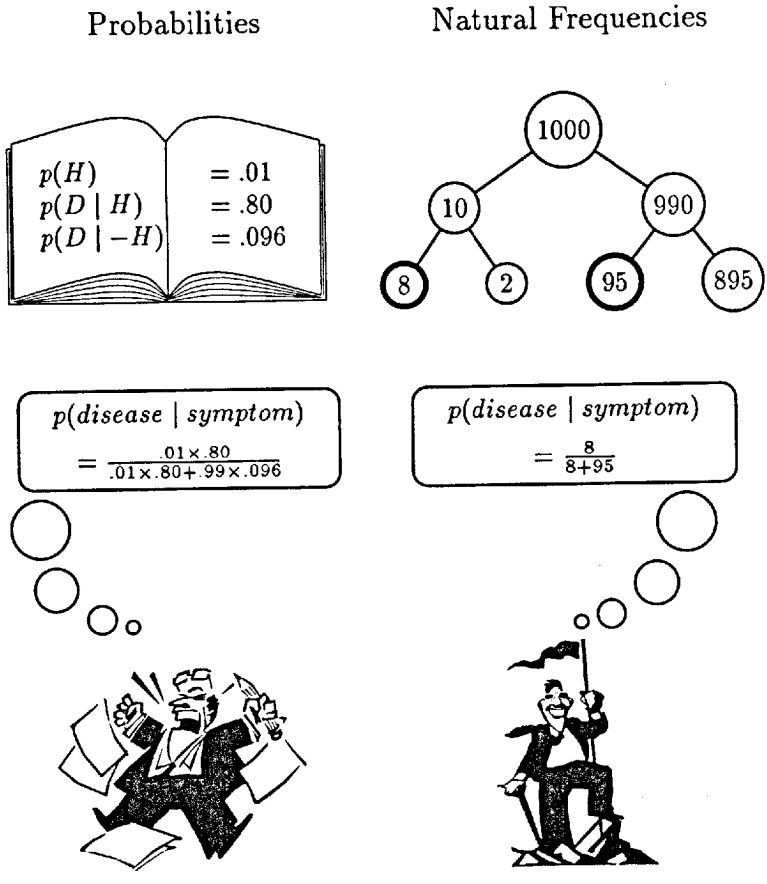

Figure 1. Bayesian inference and information representation (probabilities and natural frequencies). Adapted from 'How to improve Bayesian reasoning without instruction: Frequency formats', by G. Gigerenzer and U. Hoffrage, [1995, p. 689] Copyright 1995 by APA.

HIV, pregnancy, and other everyday matters) to students who had never heard of Bayesian inference, using various frequency and probability representations (but no visual aids such as the tree in Figure 1). When information was represented in terms of natural frequencies, in 46% of cases students found the exact numerical answer and used a Bayesian algorithm, as revealed by protocols. The corresponding value when information was represented in terms of probabilities was only 16% (for details see [Gigerenzer and Hoffrage, 1995]). In a second study, we tested whether natural frequencies improve Bayesian reasoning in physicians, using four medical problems, including mammography. Forty-eight physicians (mean professional experience was 14 years) worked an average of 30 minutes on these problems. Despite the fact that these physicians were experts, the results were similar. When information was represented in a probability format, in only 10% of the cases did the physicians reason the Bayesian way, but when the information was in natural frequencies, Bayesian responses increased to 46% (for details see [Gigerenzer, 1996a; Hoffrage and Gigerenzer, 1996]). These results are consistent with the claim that cognitive algorithms are tuned to natural frequencies (as defined by the tree in Figure 1).[2] The practical consequences are straightforward: Physicians, patients, and students should be taught to transform probabilities into natural frequencies, in which they can 'see' the solutions to diagnostic problems. We have designed such a computerised tutorial program that teaches people how to represent probabilities in natural frequencies. Students using this tutorial scored about twice as high as those who used a traditional program that taught them how to insert probabilities into Bayes's rule. Five weeks later, students who had learned to construct frequency representations still maintained their high level of accuracy, but the others showed the usual steep forgetting curve [Sedlmeier and Gigerenzer, 1996]. It is easier to be a Bayesian when working with frequencies.

To sum up: Bayes's theorem is often used as a norm for rational reasoning, but this rule is mute about the representation of information it is supposed to work on. If evolution has shaped mental algorithms that make inferences about an uncertain world, then it is likely that these algorithms were designed for natural frequencies, as encoded by natural sampling, and not for probabilities and percentages. Comparing human judgment to Bayes's theorem without considering the representation of the numerical information is, according to this argument, like comparing the outputs of a pocket calculator to multiplication tables without considering whether the numbers were entered in Arabic numerals, binary numerals, or in another representation. Challenge One is to come up with normative models for human reasoning that deal with algorithms *and* the input representations on which the

[2]Note that this result applies to the simple type of Bayesian inference with binary hypotheses and data and to one piece of information (e.g. one test result). In situations with multiple pieces of information that are not independent but redundant, however, Bayes's theorem quickly becomes mathematically complex and computationally intractable–at least for an ordinary human mind. In these situations, even frequency representations may not be able to reduce the complexity sufficiently to enable minds to "see' the Bayesian way. In Challenge Six, I will deal with such complex situations and present evidence that simple psychological mechanisms can make inferences as accurate as sophisticated statistical models that use large amounts of knowledge.

algorithms are designed to operate. Rules per se are incomplete normative models for machine and mental computers alike.

2 CHALLENGE TWO: PSYCHOLOGICAL MECHANISMS DETERMINE THE RELEVANT NUMBERS

So far we have linked algorithms to the representation of numerical information but have not thought about the numerical information itself. Let us now put some psychology into the numbers. I illustrate this by summarising Birnbaum's [1983] application of a standard psychological theory, the theory of signal detectability (TSD), to a Bayesian inference problem. TSD is formally equivalent to the Neyman–Pearson theory of hypotheses testing [Gigerenzer and Murray, 1987]. The important point is that TSD can direct our attention to psychological mechanisms which determine the relevant numbers that should be inserted into Bayes's theorem.

2.1 The Norm

The following version of the cab problem is from Tversky and Kahneman [1980, p. 62]:

> A cab was involved in a hit-and-run accident at night. Two cab companies, the Green and the Blue, operate in the city. You are given the following data:
>
> 1. 85% of the cabs in the city are Green and 15% are Blue.
> 2. A witness identified the cab as a Blue cab. The court tested his ability to identify cabs under the appropriate visibility conditions. When presented with a sample of cabs (half of which were Blue and half of which were Green) the witness made correct identifications in 80% of the cases and erred in 20% of the cases.
>
> Question: What is the probability that the cab involved in the accident was Blue rather than Green?

Tversky and Kahneman assumed that the cab problem has one and only one correct answer, which is obtained by inserting the given numbers into Bayes's theorem in the form of Equation 3. Let G and B stand for the two hypotheses ('Cab was Green' and 'Cab was Blue'), and "B" for 'Witness testified that cab was Blue'. Inserting the numbers into Bayes's theorem results in the following probability $p(B|"B")$ that the cab involved in the accident was Blue:

$$(4) \quad p(B|"B") = \frac{p(B)p("B"|B)}{p(B)p("B"|B) + p(G)p("B"|G)}$$

$$= \frac{(.15)(.80)}{(.15)(.80) + (.85)(.20)}$$

The result is .41. Tversky and Kahneman [1980] reported that the modal and median response of several hundred subjects was .80. The median response was identical to the witness's hit rate—much as in the mammography problem. This result has been interpreted as evidence that subjects neglect base rates.

Tversky and Kahneman's use of Bayes's theorem assumes that the content of the problem is merely decorative—for instance, what we know about the cognitive mechanisms of eyewitnesses in visual discrimination tasks is assumed to be not relevant for a normative model. In this content-independent application of Bayes's theorem, there is no need to distinguish between a mammography test and an eyewitness report, except for the numbers. Now, let us have a second look at the norm.

2.2 *Determining the Relevant Numbers: Psychological Assumptions*

Figure 2a illustrates the cab problem from the point of view of the theory of signal detectability (TSD), a theory of sensory discrimination and detection [Birnbaum, 1983]. TSD assumes that each colour, G and B, produces a normal distribution of sensory values on a sensory continuum (although other distributions are possible). The two distributions overlap, which is why errors in identification can occur. There is a decision criterion that balances the probabilities of the two possible errors a witness can make, the probability $p(\text{“}B\text{”}|G)$ of a false alarm, and the probability $p(\text{“}G\text{”}|B)$ of a miss. (The complement of the miss rate is the probability $p(\text{“}B\text{”}|B)$, called the hit rate.) If on some occasion the value on the sensory continuum is to the right of the criterion, the witness says 'Blue'; otherwise, the witness says 'Green'. If it is important to reduce the probability of false alarms, then the criterion is shifted to the right, causing the probability of misses to increase. The reverse follows if the criterion is shifted to the left. As one shifts the criterion from the very left side of the sensory continuum to the very right side, one gets a series of pairs of hit and false alarm rates, one of which is shown in Figure 2a. The distance between the means of the two distributions is known as the sensitivity d' of the witness.

When we look at the cab problem from the point of view of TSD, we notice two key differences between Birnbaum's content-based model and Tversky and Kahneman's content-free model. The first is the decision criterion, which is central to TSD and is absent in the content-free norm. In the content-free approach, the witness is characterised by a single pair of likelihoods (a false alarm and a miss rate), whereas in TSD the witness is characterised by a continuum of such pairs. The second difference is linked to the first: no prior probabilities or base rates are explicit in Neyman–Pearson theory, and consequently, in TSD. However, TSD allows for shifting the decision criterion in response to a shift in base rates, consistent with the empirical finding that the ratio of the hit rate to the false alarm rate varies with the signal probabilities, that is, with the base rates [Birnbaum, 1983; Luce, 1980]. Note that this finding is inconsistent with the independence of base rates and likelihoods

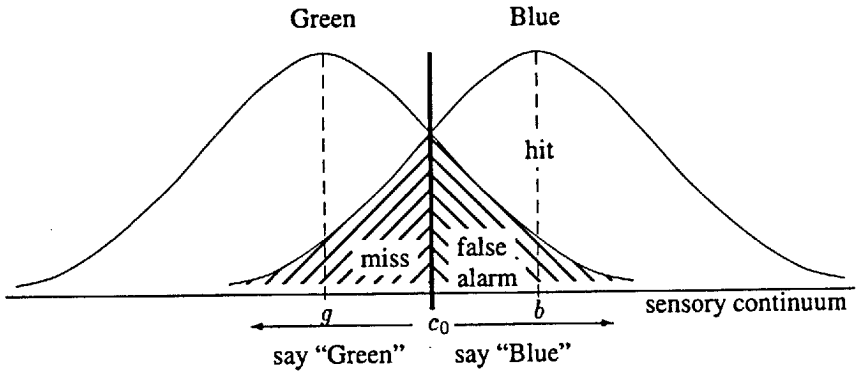

Figure 2a

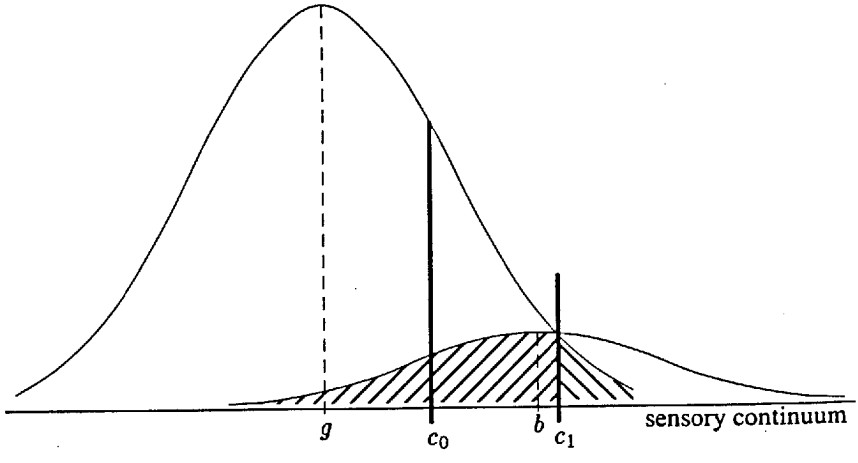

Figure 2b

Figure 2a: The theory of signal detectability (TSD) applied to the cab problem: Representation of the sensory continuum of a witness. From *Cognition as Intuitive Statistics*, p. 168. G. Gigerenzer and D. J. Murray, Erlbaum, Hillsdale, NJ, 1987. Copyright 1987 by Lawrence Erlbaum. Reprinted with permission.
Figure 2b: Location of the criterion c1 that minimises the overall error when the base rates of Blue and Green cabs are different (see text).

assumed in the content-free norm. According to TSD, a change in base rates can be manifested as a change in the criterion (the likelihood ratio). According to the content-free norm, in contrast, a change in base rates does not affect the likelihood ratio.

With this background, we can now address the question: What are the relevant numbers to be inserted into Bayes's theorem? TSD suggests that some of the details in the Cab problem are relevant for finding these numbers—whereas the entire content was irrelevant to the way the "normative" answer of .41 was calculated. Remember that there were two points in time: the night of the accident and the time when the court tested the witness. If the criterion was set at c_0 at the time of the test, where was it set at the critical time of the accident? We are told that the visibility conditions of the test were appropriate; thus we can assume that the sensitivity d' of the witness was similar during the test and on the night of the accident. That is, on the night of the accident, the distance between the means of the two distributions was the same. But where was the criterion? To answer this, we need a psychological theory of criterion shift.

In the absence of further information, we may start with the plausible hypothesis that the witness adjusted his criterion so as to minimize incorrect testimony. During the test, when the base rates of Green and Blue cabs were equal, the criterion (c_0) was at the intersection of the two curves in Figure 2a. Now it becomes clear how crucial it is to know whether or not the witness knew the base rates. Assume that the witness knew the base rates of cabs in the city and attempted to minimize incorrect testimony. This implies that on the night of the accident the criterion was to the right of c_0 because there were many more Green cabs, and the most likely error was to mistake a Green cab as Blue. The criterion that minimises the sum of the overall proportion of errors is called c_1 in Figure 2b. It is defined by a false alarm rate of .03 and a hit rate of .43 [Birnbaum, 1983].[3] From the assumption that on

[3]The false alarm rate $p(\text{"}B\text{"}|G)$ and the hit rate $p(\text{"}B\text{"}|B)$, which minimize the error $.85p(\text{"}B\text{"}|G) + .15p(\text{"}G\text{"}|B)$, are calculated as follows. First, d' is determined from the test situation. Here we know that the base rates of Green and Blue were the same, and we can assume that the two errors $p(\text{"}B\text{"}|G)$ and $p(\text{"}G\text{"}|B)$ were both equal to .20. Assuming that the two distributions are normal distributions with variance 1.0, we can find the difference $(c_0 - g)$ (Figure 2a) using the fact that the cumulative distribution function $\Phi(x)$ of the standard normal distribution takes the value .8 at $x = (c_0 - g)$. Thus, $(c_0 - g) = .84$. From the symmetry of the test situation, we can conclude that the difference $(b - c_0)$ is also equal to .84. Because $d' = b - g = (b - c_0) + (c_0 - g)$, d' is equal to 1.68.

Now for the situation on the night of the accident where the base rates of Blue and Green are different, we use this value of d' to find the value of the criterion c_1, which determines the minimum value of the error $.85p(\text{"}B\text{"}|G) + .15p(\text{"}G\text{"}|B)$. Notice that $p(\text{"}B\text{"}|G) = 1 - \Phi(c_1 - g)$, and $p(\text{"}G\text{"}|B) = \Phi(c_1 - b)$, where Φ is the area under the cumulative normal distribution. So, the error to be minimised can be recast as $.85(1 - \Phi c_1 - g)) + .15\Phi(c_1 - b)$. We use standard techniques of differential calculus to minimize this expression. Its derivative is $\frac{1}{\sqrt{2\pi}}\{-.85\exp\frac{1}{2}(c_1 - g)^2 + .15\exp\frac{1}{2}(c_1 - b)^2\}$, which is zero only at $c_1 = g + 1.87$. This point corresponds both to the minimal value of the error and the intersection of the two curves in Figure 2b. With this value of c_1, the false alarm rate is equal to $p(\text{"}B\text{"}|G) = 1 - \Phi(c_1 - g) = 1 - \Phi(1.87) = .03$, and $p(\text{"}G\text{"}|B) = \Phi(c1 - b) = \Phi(g + 1.87 - b) = \Phi(1.87 - d') = \Phi(1.87 - 1.68) = \Phi(.19) = .57$. Therefore the hit rate $p(\text{"}B\text{"}|B)$ is $1 - .57 = .43$.

Note that Birnbaum [1983] has reported a slightly different value, which seems to be based on a cal-

the night of the accident the witness set his criterion at c_1 rather than c_0, these two values are the relevant numbers to be inserted into Bayes's theorem:

$$(5) \quad p(B|\text{“}B\text{”}) = (.15)(.43)/[(.15)(.43) + (.85)(.03)]$$

The result is .72. Note that this result could be mistaken as an instance of base-rate neglect because it is again close to the hit rate (80% in the text of the cab problem). Ironically, the value of .72 was computed based on the assumption that the witness knows and uses the base rates. The TSD analysis of the cab problem illustrates how psychological assumptions (e.g. the criterion setting) and mathematical assumptions (e.g. the identical normal distributions of the perceptual processes) go hand in hand in a content-sensitive normative model. Good statistical reasoning cannot be reduced to the mechanical insertion of numbers into a formula, an insight that the intellectual parents of TSD, Jerzy Neyman and Egon S. Pearson, emphasised repeatedly [Gigerenzer, 1993].

If the witness adjusts the criterion in a different way than minimising incorrect testimony, this will lead to a different posterior probability. Birnbaum [1983] has studied various such psychological strategies a witness might use. I should mention that the criterion shift is not limited to situations in which the base rates at the time of the accident and at the test differ. Even if the base rates were identical, the witness who testified 'Blue' knows that he can be accused of making one and only one error, that is, of saying 'Blue' although the cab was Green (a false alarm). The other possible error, a miss (mistaking a blue cab for a green one) is excluded, because he testified that the cab was blue. If he wants to protect himself from being accused of erroneous testimony, he may shift the criterion far to the right so that the probability of a false alarm is minimised. Shifting the criterion to the right also increases the posterior probability (this can be inferred from Figure 2a).

Challenge Two is to build normative models from psychological assumptions, rather than to insert numbers into a formula, purified of the content of the situation. The fundamental normative role of the assumptions a person makes is not peculiar to the cab problem; for instance, it is crucial for the normative evaluation of the three-doors problem [Falk, 1992], the four-cards problem [Gigerenzer and Hug, 1992; Oaksford and Chater, 1994], gambler's fallacy, and 'conservatism' in information processing [Cohen, 1982].

To emphasise psychology is emphatically not to say that 'anything goes'. The contrast I wish to draw is not between a norm that is created mechanically in a content-free way and no norms at all. My point is that psychological assumptions (the semantics and pragmatics of the situation) are indispensable for constructing a sensible norm. The particular assumptions that are made about a situation determine the choice among possible candidates for a normative model. A consequence is that claims of the kind 'this is the only correct answer' need to be based on fleshing out the psychological assumptions [Levi, 1983]. The cab problem is of particular interest here because it illustrates how two different statistical approaches, the

culation error.

Neyman-Pearson theory of hypotheses testing (which is formally equivalent with TSD) and (the content-free application of) Bayes's theorem, can highlight different aspects of the problem as important, such as the decision criterion of the witness.

Challenge One adds to Challenge Two. When the information in the cab problem is represented in natural frequencies as opposed to probabilities or percentages, people can 'see' the numerical answer much more easily, whatever numbers they chose as relevant [Gigerenzer and Hoffrage, 1995].

3 CHALLENGE THREE: THE INDETERMINACY OF CONSISTENCY

The take-home message so far is that modeling rational judgment involves (a) assumptions about the information representation for which cognitive algorithms are designed, and (b) assumptions about psychological mechanisms that determine which numbers (prior probabilities, likelihoods) enter an algorithm. Let us now extend our discussion of the role of content in defining sound reasoning and turn to internal consistency of choice. Internal consistency is often seen as *the* requirement of rational choice in decision theory, behavioural economics, and game theory. Challenge Three is to define consistency in terms of something external to the choice behaviour, such as social objectives and values, rather than in terms of content-independent formulations (axioms). Only then can we decide whether a behaviour is actually consistent or not.

3.1 The Norm: Property Alpha

One basic condition of internal consistency of choice is known as 'Property Alpha', also called the 'Chernoff condition' and 'independence of irrelevant alternatives' [Sen, 1993]. The symbols S and T denote two nonempty sets of alternatives, and $x(S)$ that alternative x is chosen from the set S.

Property Alpha:

(6) $x(S)$ and $x \varepsilon T \subseteq S \Rightarrow x(T)$.

Property Alpha demands that if x is chosen from S, and x belongs to a subset T of S, then x must be chosen from T as well. For instance, assume you won a free subscription to any weekly magazine in the world (S) of your choice. You choose the *Economist* (x). Now you learn that you can actually only chose a weekly magazine published in English (T). You still chose the *Economist*. The following two choices would be inconsistent in that they violate Property Alpha:

1. x is chosen given the options $\{x, y\}$
2. y is chosen given the options $\{x, y, z\}$

Property Alpha is violated because x is chosen when the two alternatives $\{x, y\}$ are offered, but y is chosen when z is added to the menu. (Choosing x is interpreted here as a rejection of y, not as a choice that results from mere indifference.) It may indeed appear odd and irrational that someone who chooses x and rejects y when offered the choice set $\{x, y\}$ would choose y and reject x when offered the set $\{x, y, z\}$. Such violations are known as preference reversals. For illustration, here is a story told about the Columbia University philosopher Sidney Morgenbesser. Sidney went to the donut store on 116th Street. 'Would you like a plain or a glazed donut?' the waitress asked. 'I'll have a plain donut', responded Sidney. 'Oh, I forgot, we also have a jelly donut', the waitress added. 'In this case,' Sidney replied, 'I'll take the glazed donut'. My philosopher friends laugh at this story: Sidney has violated Property Alpha.

3.2 *Psychologizing the Norm: Making Consistency Work*

Sen [1993] has launched a forceful attack on internal consistency as defined by Property Alpha and similar principles, and what follows is based on his ideas and examples. Property Alpha formulates consistency exclusively in terms of the internal consistency of choice behaviour with respect to sets of alternatives. No reference is made to anything external to choice, for instance, to intentional states such as a person's social objectives, values, and motivations. This exclusion of everything psychological beyond behaviour is in line with Samuelson's [1938] program of freeing theories of behaviour from any traces of utility and from the priority of the notion of 'preference'.

But consider Property Alpha in the context of social politics at a dinner party. Everyone makes his or her way through the main course and conversation. Finally, a fruit basket is passed around for dessert. When the basket reaches Mr. Polite, there is only one apple left. Mr. Polite has the choice of taking nothing (x) or taking the apple (y). Mr. Polite loves apples, but because there is only one left he decides to behave decently and take nothing (x), because this would deprive the next person from having a choice. If the basket had contained another piece of fruit (z), he could have chosen y over x without violating standards of good behaviour. Choosing x over y from the choice set $\{x, y\}$ and choosing y over x from the choice set $\{x, y, z\}$ violates Property Alpha, even though there is nothing irrational about Mr. Polite's behaviour given his values regarding social interaction. If he had not held to such values of politeness in company, or had chosen to dine alone, then Property Alpha would not have been violated. It is social values that determine what the perceived alternatives in the choice set are: For the selfish person it is *apple* versus *nothing* in both choice sets, but for Mr. Polite it is last *apple* versus *nothing* in the first set. Property Alpha tells us little about consistency unless one looks beyond choice behaviour to a person's intentions and values.

Sidney Morgenbesser's reversal of preference looks irrational. However, consider the following. I grew up in Bavaria and I love roasted pork with potato dumplings. In a restaurant in Illinois I once had a choice between roasted pork

and steak, and I chose the steak over roasted pork (from bitter experience). But when the waiter added 'It's not on the menu, but we also have blood-and- liver sausages with sauerkraut', then I switched and chose roasted pork over steak. The third alternative, although I did not choose it, indicated by its very existence that this restaurant's cook might really know how to make Bavarian roasted pork with potato dumplings. Again, choosing x over y from the choice set $\{x, y\}$ and choosing y over x from the choice set $\{x, y, z\}$ violates Property Alpha, even though there is nothing irrational about this behaviour. The mere emergence of a new alternative may carry information about the previous alternatives.

To summarize the argument: Consistency, as defined by Property Alpha, deals only with choice behaviour. However, consistency in observed choice can be a poor indication of consistency, as the examples illustrate (for more see [Gigerenzer, 1996b; Sen, 1993]). Only once a person's social values, objectives, and expectations are known can axioms such as Property Alpha capture consistency. Challenge Three is to develop concepts of consistency that are not merely syntactical and leave out semantics and pragmatics, but start from psychological entities such as a persons expectations and social values.

4 CHALLENGE FOUR: SEMANTIC INFERENCES

Challenges Two and Three emphasised the role of content in building normative models. So far I have dealt with content that did not look relevant from the point of view of content-free normative models. In this view, the relevant information is assumed to be reducible to those words in a problem description that sound similar to concepts in logic and probability theory, such as 'AND', 'OR', 'probable', and 'likely'. In this section, I deal with these key terms. Unlike logic and probability theory, natural languages are polysemous, and the meaning of these terms must be inferred from the content in which they occur. Challenge Four is to analyze the semantic inferences people make about the meaning of terms, and to judge the soundness of a person's reasoning on the basis of these inferences, rather than assuming that natural language terms map one-to-one into similar-sounding concepts in probability theory and logic.

4.1 The Norm

Consider the following, known as the Linda problem [Tversky and Kahneman, 1983]:

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 anti-nuclear demonstrations.

Which of these two alternatives is more probable?

- Linda is a bank teller. (T)

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

In numerous experiments, a majority of subjects (often 80% to 90%) chose $T\&F$ as more probable. Tversky and Kahneman [1983] argued that this choice is an error in reasoning: $T\&F$ is the conjunction of two propositions, namely that Linda is a bank teller (T) and that she is active in the feminist movement (F), whereas T is one of the conjuncts. The mathematical probability of a conjunction cannot be greater than that of one of its conjuncts—this rule has been referred to as the conjunction rule:

(7) $p(T\&F) \leq p(T)$

Tversky and Kahneman argued that because of this rule, the correct answer to the problem is T. They therefore concluded that the majority of their subjects, who chose $T\&F$, had committed a reasoning error they called the 'conjunction fallacy'. The explanation of the phenomenon was that people do not reason by the laws of probability, and instead use similarity to judge probability, a strategy termed the 'representativeness heuristic'. The Linda problem has been used by other researchers to draw hefty conclusions about human rationality, for instance, that 'our minds are not built (for whatever reason) to work by the rules of probability' [Gould, 1992].

4.2 *Psychologizing the Norm: Semantic Inferences*

This use of the conjunction rule as a normative model for the Linda problem (and other similar problems) assumes that (a) all that counts for rational reasoning are the English terms 'and' (in 'T and F') and 'more probable', and (b) these natural language terms can be mapped in a one-to-one fashion into logic and probability theory: The English 'and' is assumed to be immediately translatable onto the logical 'AND', and the English 'probable' onto mathematical probability. Everything else, including Linda's description and the content of the two propositions, is considered irrelevant for sound reasoning.

The critical point here is that this one-to-one mapping from natural language to logic or probability theory cannot capture sound reasoning. All natural languages embody polysemy. For example, connectives such as 'and', 'or', and 'if' have several meanings, such as the inclusive and exclusive meanings of 'or'. Consider the proposition 'Joan and Jim married and they had a baby' versus 'Joan and Jim had a baby and they married'. If one mapped the 'and' in these two sentences onto the logical 'AND', one would miss the difference between the information communicated in the first versus the second sentence. Not mapping the natural language 'and' onto the logical operator allows us immediately to understand that the 'and' in these sentences indicates temporal order, and thus unlike logical AND, is not commutative. The cognitive mechanism that infers the meaning of terms such as 'and' from the content of a sentence is a most impressive feature of the human mind; no

computer program exists yet that can make these inferences. I refer to this mapping process as 'semantic inference'.

The polysemy of the English term 'and' makes it important to find out which meaning of 'and' a person infers when reading the proposition 'Linda is a bank teller and active in the feminist movement'. There is a good reason to interpret 'and' as something other than logical AND: The logical meaning would render the description of Linda and the content of the two alternatives irrelevant, thus violating the conversational maxim of relevance, that is, the assumption that what the experimenter tells you is relevant to the task [Adler, 1991; Grice, 1975]. Polysemy also holds for the English term 'probable'. The Oxford English Dictionary lists a wide range of legitimate meanings for 'probable', including 'plausible', 'having an appearance of truth', 'that may in view of present evidence be reasonably expected to happen', and 'likely', among others. Most of these meanings cannot be mapped onto the mathematical concept of probability. For instance, if 'Which is more probable?' is understood as 'Which makes a more plausible story?' or 'Which is supported by evidence?' then $T\&F$ seems to be the better alternative. There exist a number of studies indicating that people indeed draw semantic inferences that lie outside logic and probability theory, such as the '$T \rightarrow T\&$ not F' implicature (that is, to infer that 'Linda is a bank teller' means 'Linda is a bank teller and not active in the feminist movement') and the '$T\&F \rightarrow F$ given T' implicature (that is, to infer that 'Linda is a bank teller and active in the feminist movement' means 'Linda is a feminist given she is a bank teller'); see e.g. [Dulany and Hilton, 1991; Hertwig and Gigerenzer, 1997; Tversky and Kahneman, 1983]. These implicatures render the description of Linda relevant to the problem.

There are experiments that indicate that the reason why many people chose $T\&F$ as more probable is their outstanding ability to perform semantic inferences rather than their alleged failure to reason according to the laws of probability. One experiment simply couches the problem in terms of frequency rather than subjective degree of belief, replacing the ambiguous term 'probable' by the less ambiguous term 'how many' [Hertwig and Gigerenzer, 1997]. This version of the problem informs the subject that there are 200 women who fit Linda's description. Subjects are then asked

How many of the 200 women are bank tellers? _____ of 200

How many of the 200 women are active in the feminist movement? _____ of 200

How many of the 200 women are bank tellers and are active in the feminist movement? _____ of 200

In this and similar 'frequency versions' of the Linda problem, violations of the conjunction rule dropped to 0%–20% (from 80%–90% in the original 'probability version'). Substituting 'how many' for the ambiguous term 'probable' is, to the best of my knowledge, the strongest and most consistent way to reduce the conjunction fallacy (see [Fiedler, 1988; Hertwig, 1995] and [Tversky and Kahneman,

1983, p. 309] for similar results). Most subjects reason according to the conjunction rule when linguistic ambiguity is resolved.

To summarize: The use of the conjunction rule as a content-free norm for correct thinking overlooks the capacity of the human mind to make semantic inferences. Challenge Four is to model these impressive semantic inferences, rather than to assume as normative a one-to-one mapping of natural language terms into probability theory and logic.

5 CHALLENGE FIVE: DOUBLE STANDARDS

Normative models are sometimes used by convention rather than by reflection. Unreflective use of norms can lead to double standards. For example, the same researcher sometimes uses two mutually inconsistent norms, one prescribing what is rational inference for subjects and another prescribing what is rational inference for the researcher. Each norm is used mechanically, without consideration of content.

5.1 Fisherian Norms for Me, Bayesian Norms for You

R. A. Fisher's *The Design of Experiments* [1935] is possibly the single most influential book on experimental methodology in the social and biological sciences. Fisher disapproved of the routine application of Bayes's theorem. In the introduction, Fisher congratulates the Reverend Thomas Bayes for being so critical of this theorem as to withhold its publication (Bayes's 1763 treatise was published only after Bayes died). Fisher thought that the preconditions for applying Bayes's theorem, such as an objective prior distribution over the set of possible hypotheses, rarely hold, and that routine applications of the theorem would lead to unacceptable subjectivism wherein the strength of evidence would be just a matter of taste. In his book, Fisher successfully sold researchers his method of null hypothesis testing instead. By the 1950s, null hypothesis testing, also known as significance testing, became institutionalised in many social, biological, and medical fields as the sine qua non of scientific inference.[4]

In the 1960s, Ward Edwards and colleagues proposed that researchers (a) abandon null hypothesis testing and turn Bayesian instead [Edwards, Lindman and Savage, 1963], and (b) study whether the untutored mind reasons by Bayesian principles [Edwards, 1968]. The first proposal fell stillborn from the press while the second became a raging success. Researchers began to test whether the mind draws inferences according to Bayes's theorem (as described in Challenges One and Two) at the same time that they continued to use significance testing. Researchers had been taught to use significance testing (which promised

[4] What was institutionalised was actually a mishmash between Fisher's null hypothesis testing and some concepts of a theory that Fisher deeply disliked, namely Neyman and Pearson's theory of hypotheses testing. Textbooks and curricula are generally silent about the fact that they teach a hybrid creature that would have been rejected by both camps (Gigerenzer, [1987; 1993] and [Gigerenzer *et al.*, 1989, Chs. 3 and 6]).

objectivity) mechanically, and Bayes's theorem smelled of subjectivity [Gigerenzer, 1987]. Thus those who went only half-way with Ward Edwards unwittingly committed themselves to a double standard. Ordinary people who do not make inferences according to Bayes's theorem are branded irrational, but the researchers who brand them do not apply the same standard to their own inductive inferences. They use significance testing, not Bayes's theorem, to infer whether people are Bayesians or not.

5.2 *Beyond Double Standards*

Challenge Five is to construct normative models (for the reasoning of researchers and their experimental subjects alike) in a thoughtful rather than a mechanical way. One may end up having to tailor different normative models for different situations, but not mechanically use one norm for experimenters and another for subjects. I and others have traced how the mindless use of null hypothesis testing in psychology (and many social and medical sciences) became institutionalised in textbooks, curricula, and editorial practices (e.g. [Gigerenzer, 1993], [Gigerenzer and Murray, 1987, Ch. 1] and [Gigerenzer *et al.*, 1989, Chs 3 and 6]). Remember that Hume's problem of inductive inference has not yet been solved; there is no single method that works in all situations, and we need to teach students and researchers what the methods are and how to choose between them. In my opinion, inferential statistics—significance testing, Neyman–Pearson testing, Bayesian statistics, and so on—are rarely needed in research. What is needed is good descriptive statistics, knowledge of the data (e.g. look at the scatter diagram instead of just at the correlation coefficient), adequate representations of the data, and the formulation of precise alternative hypotheses instead of a single null hypothesis. There is hope on the horizon that after four decades, the reign of the null hypothesis testing ritual is at last in decline. For instance, Geoffrey Loftus, editor of *Memory & Cognition*, seems to be the first editor of a major psychology journal in the United States to speak out and explicitly discourage researchers from mechanically submitting p, F, and t-values for no good reason (Loftus, [1991; 1993]). He asked researchers instead to provide good descriptive statistics and to think about the representation of information, for example, to provide figures with error bars instead of p values.

When one thinks of statistical models as models of some situation rather than mechanically applicable tools, then double standards can be avoided. To come up with reasonable normative models of inference, one must try to model the situation instead of imposing content-independent standards onto the reasoning of either subjects or experimenters.

6 CHALLENGE SIX: THE POWER OF SIMPLE PSYCHOLOGICAL MECHANISMS

In the first four challenges, I argued that syntactical rules or axioms are not sufficient to define rational behaviour, unless they take psychological mechanisms into account. These were standard rules and axioms, which are taken by many to define rationality in a content-free way. In this section, I invite you to look beyond standard rules and axioms to the power of simple, 'satisficing' psychological mechanisms that violate classical assumptions of rationality. Challenge Six is to design simple satisficing mechanisms that work well under real-world constraints of limited time and knowledge: mechanisms that are fast and frugal but nevertheless about as accurate as computationally 'expensive' statistical models that satisfy classical norms.

6.1 The Norm

Imagine that you have to infer which of two alternatives, a or b, has a higher value on some criterion, and there are 10 predictors of the criterion with different validities. One method that is used to make such an inference is multiple regression, which computes the beta weights for each of the predictors, computes the value of each alternative on the criterion, and chooses the alternative that scores higher. This amounts to formulating the following multiple regression equation:

$$(8) \quad y_a = x_{a1}\beta_1 + x_{a2}\beta_2 + x_{a3}\beta_3 + \ldots + x_{a10}\beta_{10},$$

where y_a is the value of a on the criterion, x_{a1} is the value of alternative a on predictor 1, β_1 is the optimal beta weight for predictor 1, and so on.

Note that we are now dealing with a much more complex situation than in Challenges One and Two: There are many pieces of information (the predictors) rather than only one (e.g., positive mammogram), and these may be partially redundant. The two norms dealt with here are more general than the multiple regression model. The first norm is that sound inference implies *complete search*, that is, taking account of all pieces of information available, and the second requires *complete integration*, that is, integrating all pieces of information in a some reasonable way [Gigerenzer and Goldstein, 1996]. These norms hold for multiple regression as well as for Bayesian inference and neural networks, all of which look up and integrate all available information.

6.2 Beyond Complete Search and Integration: Take The Best

Humans often need to make inferences about aspects of their environment under constraints of limited time, limited knowledge, and limited computational resources. The linear multiple regression is in conflict with all three. When one is driving fast and the road suddenly forks, one does not have the time to think about all the reasons that would favor going right or left, nor the knowledge and computational

aides to calculate all the beta weights, multiply these with the predictor values, and calculate sums. Similarly, a doctor in an emergency room who has to make a decision whether a heart attack patient should be treated as a high risk or a low risk case does not know the values of the patient on all relevant predictors, nor can she always take the time to measure these. In these and many other situations, humans have to rely on fast and frugal psychological mechanisms rather than on multiple regression. In Herbert Simon's terms, humans 'satisfice' rather than 'optimize'. Consider the following demographic problem:

Which city has more inhabitants:

(a) Bremen

(b) Bielefeld

Assume you do not know the answer, but have to make an inference. There are many predictors (cues) that signal larger population, such as whether or not a city has a soccer team in the major German league ('Bundesliga'), whether or not it is a state capital, has a university, and so on. Thus, according to the norms of complete search and information integration, one should search in memory for all predictors, estimate the values of the two cities on those predictors, estimate the weights for each predictor, multiply these with the estimated values, sum the products up and choose the city with the higher value. Such a model assumes that the mind is a supercomputer like Laplacean Demon with almost unlimited time and knowledge. What is the alternative?

My students and I have developed a family of satisficing algorithms [Gigerenzer *et al.*, 1991; Gigerenzer and Goldstein, 1996], one of which I will describe here. It is based on psychological mechanisms that a mind can utilize given limited time and knowledge. One of these simple mechanisms, the 'recognition heuristic', says that if one has heard of city a but not of city b, then the search for further information can be stopped, and the inference that a is the larger can be made. Thus, in the example problem, if you have never heard of Bielefeld, you infer that Bremen has more inhabitants. The recognition heuristic can be invoked when there is a correlation between recognition and the criterion. For instance, advertisement companies (e.g. Benetton) exploit recognition by making sure that consumers recognize the brand name while providing no information about the product itself [Goldstein and Gigerenzer, 1996].

If recognition cannot be used as a cue, that is, if someone has heard of both cities, a second mechanism is invoked: one-reason decision making. For instance, if the fact is retrieved from memory that Bremen has a soccer team in the major league but Bielefeld does not (or one does not know), then 'one-reason decision making' makes the inference that Bremen is the larger city. No further information is sought. What we call the Take The Best algorithm (because of its motto 'take the best and ignore the rest') is based on just these two psychological principles and the assumption of a subjective ranking of the predictors in terms of their validities. The flow chart of Take The Best is shown in Figure 3. For simplicity, only binary cues are

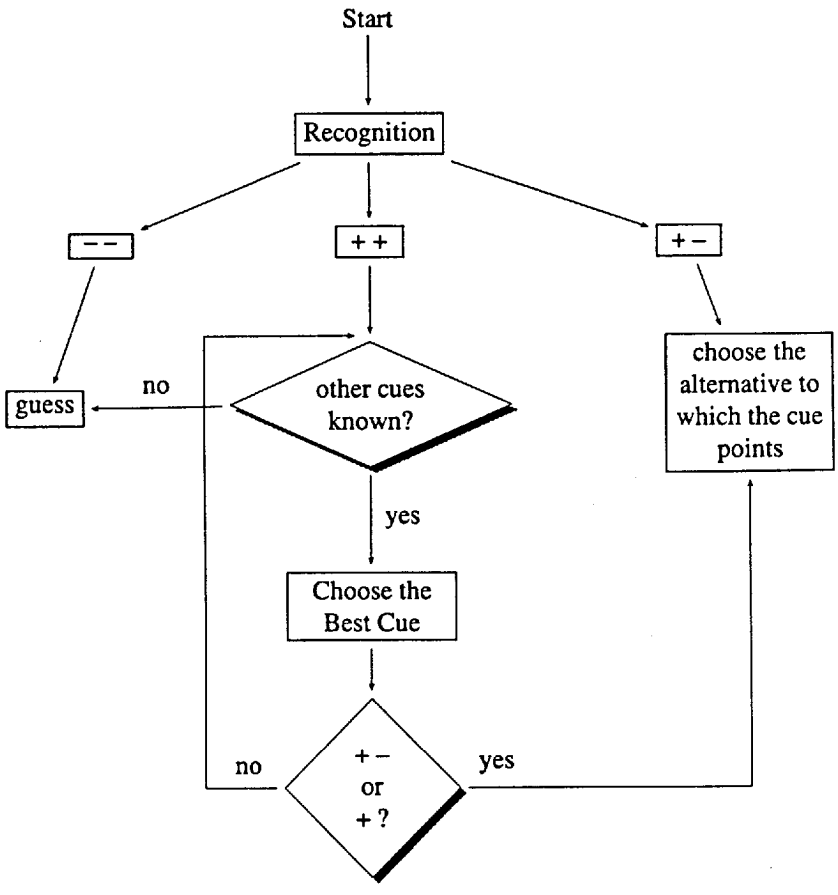

Figure 3. Flow diagram of a satisficing algorithm: Take The Best.

considered, and the values +, −, and ? signify that an object has a positive, negative, or unknown value on a cue. The key features of Take The Best are (a) limited search, that is, it stops with the first cue (including recognition) that discriminates between two objects, and (b) no integration, that is, choice is made on the basis of only one cue (but this cue may be different from one pair of cities to the next). These two features violate the norms of exhaustive search and integration.

Well, Mr. Optimal says, people may use Take The Best or similar satisficing algorithms, given the constraints of limited time and knowledge, but this is certainly a stupid, quick-and-dirty algorithm. Look, Mr. Satisficing responds, studies on the 'flat maximum' have shown that in real-world environments, the 'optimal' beta weights (in Equation 8) may not lead to better predictions than unit weights (+1 or −1), suggesting that the world can be predicted as well with simpler algorithms [Dawes, 1979; Lovie and Lovie, 1986]. Well, Mr. Optimal replies, unit-

weight models use simpler weights, but outside of that they do not violate the classical norms of rationality: They look up all information available and integrate it, whereas Take The Best does neither. Thus, it will make lousy inferences. How do you know? asks Mr. Satisficing; let us empirically determine, in a real-world environment, how much better algorithms that obey the two norms actually perform, compared to the satisficing Take The Best algorithm.

To answer this question, Daniel Goldstein and I conducted a competition between five integration algorithms (including multiple regression and unit-weight linear models) and Take The Best. The algorithms inferred which of a pair of cities was the larger one, as in the Bielefeld-Bremen example, and the criteria used were the speed and the accuracy of the inferences [Gigerenzer and Goldstein, 1996]. Inferences were made about all (pairs of) cities in Germany (after reunification) with more than 100,000 inhabitants (there were 83 cities). We simulated subjects with varying degrees of limited knowledge, ranging from those who did not recognize a single German city, and consequently did not know any information (values on predictors for population) about these cities, to subjects who recognised all 83 cities and knew all values of these cities on all predictors (cues). There were 10 predictors, including the aforementioned soccer team, state capital, and university cues. The simulation included 84 (number of cities recognised, from 0 to 83) times 6 (proportion of cue values known for the cities recognised: 0%, 10%, 20%, 50%, 75%, 100%) types of subjects, and within each type we used 500 individual subjects that differed randomly in the particular cities and cue values known. Each of these $84 \times 6 \times 500$ simulated subjects drew inferences about all pairs of cities (as in the Bielefeld–Bremen example) using six algorithms (one at a time). The algorithms were five integration algorithms that looked up all information (including multiple regression) and Take The Best algorithm. In the competition we measured the speed (proportion of cue values searched before making an inference), and accuracy (proportion of correct inferences). When simulated subjects made their inference with Take The Best, they searched on average for only 30% of the information that the integration algorithms used, thus outperforming all other contestants in speed. After all, speed and computational simplicity is what this algorithm is designed for. But how accurate were the inferences that Take The Best drew? The striking result was that Take The Best matched or outperformed all competitors in accuracy, multiple regression included (for details see [Gigerenzer and Goldstein, 1996]).

This result is an existence proof that cognitive mechanisms capable of successful performance in a real-world environment do not need to satisfy classical norms of rational inference: Exhaustive information search and integration may be sufficient but not necessary for a mind capable of sound reasoning. There is independent evidence that 'one-reason decision making', as demonstrated by Take The Best, can classify heart attack patients into high and low risk groups as well as or better than standard statistical integration models with many valid predictors [Breiman *et al.*, 1993]. One important question concerning norms turns out to be ecological: What are the structures of real-world environments that can be exploited by simple cog-

nitive mechanisms, and how can we talk about these structures? We have recently identified several of these structures [Gigerenzer, 1997; Martignon and Hoffrage, in press].

The result of this competition defeats the widespread view that only 'rational' algorithms—ones that search and integrate all information—can be accurate. But does this result generalize? It does. In a sample of 20 real-world environments, the fast and frugal Take The Best made on the average *more* accurate predications (using cross-validation) than multiple regression [Czerlinski *et al.*, in press]. Models of inference do not have to forsake accuracy for simplicity, or rationality for psychological plausibility. Challenge Six is to design psychologically plausible models of sound inference that can operate under constraints of limited time and knowledge. Reasoning can be rational and psychological.

7 A PSYCHOLOGICAL APPROACH TO NORMS

I started out with the opposition between the rational and the psychological: Rational judgment is defined by the laws of probability and logic, and only by these. Psychology does not come in until things go wrong, that is, when people's judgements deviate from the laws of probability and logic. In contrast, I argued that psychological principles are indispensable for defining and evaluating what sound judgment is. Axioms and rules from probability theory and logic are, by themselves, indeterminate. In particular, I discussed the role of the representation of numbers, the role of content for inferring what the relevant numbers are, the role of a person's social values, motives, and expectations in defining and evaluating norms for sound judgment, and the power of fast and frugal algorithms.

ACKNOWLEDGEMENTS

I am grateful to Phil Blythe, Valerie Chase, Jean Czerlinski, Berna Eden, Dan Goldstein, Ralph Hertwig, Ulrich Hoffrage, Alejandro Lopes, Laura Martignon, Geoffrey Miller, Anita Todd, and Peter Todd for their critical comments on earlier versions of this chapter. Arnold Davidson told me the donut story.

Max Planck Institute for Human Development, Berlin

REFERENCES

[Adler, 1991] J. Adler. An optimist's pessimism: conversation and conjunction. In *Studies on L. Jonathan Cohen's Philosophy of Science*, E. Eells & T. Maruszewski, eds. pp. 251–282. Rodopi, Amsterdam-Atlanta, GA, 1991

[Birnbaum, 1983] M. H. Birnbaum. Base rates in Bayesian inference: signal detection analysis of the cab problem. *American Journal of Psychology*, **96**, 85–94, 1983.

[Breiman *et al.*, 1993] L. Breiman, J. H. Friedman, R. A. Olshen and C. J. Stone. *Classification and Regression Trees*. Chapman and Hall, New York, 1993.

[Casscells *et al.*, 1978] W. Casscells, A. Schoenberger and T. Grayboys. Interpretation by physicians of clinical laboratory results. *New England Journal of Medicine*, **299**, 999–1000, 1978.

[Cohen, 1982] L. J. Cohen. Are people programmed to commit fallacies? Further thoughts about the interpretation of experimental data on probability judgment. *Journal of the Theory of Social Behavior*, **12**, 251–274, 1982.

[Cosmides and Tooby, 1996] L. Cosmides and J. Tooby. Are humans good intuitive statisticians after all? Rethinking some conclusions from the literature on judgment under uncertainty. *Cognition*, **58**, 1–73, 1996.

[Czerlinski *et al.*, in press] J.Czerlinski, D. G. Goldstein and G. Gigerenzer. When it pays to be a lazy thinker: A simulation study. In *Simple Heuristics that Make Us Smart*, G. Gigerenzer and P. M. Todd, eds. Oxford University Press, New York, in press.

[Daston, 1981] L. Daston. Mathematics and the moral sciences: the rise and fall of the probability of judgments, 1785–1840. In *Epistemological and Social Problems of the Sciences in the Early Nineteenth Century*, H. N. Jahnke and M. Otte, eds. pp. 287–309. D. Reidel Publishing Company, Dordrecht, Holland, 1981.

[Daston, 1988] L. Daston. *Classical probability in the Enlightenment*. Princeton University Press, Princeton, NJ, 1988.

[Daston, 1992] L. Daston. The doctrine of chances without chance: determinism, mathematical probability, and quantification in the seventeenth century. In *The Invention of Physical Science*, M. J. Nye *et al.*, eds. pp. 27–50. Kluwer Academic Publishers, 1992.

[Dawes, 1979] R. M. Dawes. The robust beauty of improper linear models. *American Psychologist*, **34**, 571–582, 1979.

[Dulany and Hilton, 1991] D. E. Dulany and D. J. Hilton. Conversational implicature, conscious representation, and the conjunction fallacy. *Social Cognition*, **9**, 85–110, 1991.

[Eddy, 1982] D. M. Eddy. Probabilistic reasoning in clinical medicine: problems and opportunities. In *Judgment under Uncertainty: Heuristics and Biases* , D. Kahneman, P. Slovic and A. Tversky, eds. pp. 249–267. Cambridge University Press, Cambridge, 1982.

[Edwards, 1968] W. Edwards. Conservatism in human information processing. In *Formal Representation of Human Judgment*, B. Kleinmuntz, ed. pp. 17–52. Wiley, New York, 1968.

[Edwards, Lindman and Savage, 1963] W. Edwards, H. Lindman and L. J. Savage. Bayesian statistical inference for psychological research. *Psychological Review*, **70**, 193–242, 1963.

[Falk, 1992] R. Falk. A closer look at the probabilities of the notorious three prisoners. *Cognition*, **43**, 197–223, 1992.

[Fiedler, 1988] K. Fiedler. The dependence of the conjunction fallacy on subtle linguistic factors. *Psychological Research*, **50**, 123–129, 1988.

[Fisher, 1935] R. A. Fisher. *The Design of Experiments*. Oliver and Boyd, Edinburgh, 1935.

[Feynman, 1967] R. Feynman. *The Character of Physical Law*. MIT Press, Cambridge, MA, 1967.

[Gigerenzer, 1987] G. Gigerenzer. Probabilistic thinking and the fight against subjectivity. In *The Probabilistic Revolution, Vol. 2. Ideas in the Sciences*, L. Krüger, G. Gigerenzer and M. S. Morgan, eds. pp. 11–33. MIT Press, Cambridge, MA, 1987.

[Gigerenzer and Hug, 1992] G. Gigerenzer and K. Hug. Domain-specific reasoning: social contracts, cheating and perspective change. *Cognition*, **42**, 127–171, 1992.

[Gigerenzer, 1993] G. Gigerenzer. The superego, the ego, and the id in statistical reasoning. In *A Handbook for Data Analysis in the Behavioral Sciences: Methodological Issues*, G. Keren and C. Lewis, eds. pp. 313–339. Erlbaum, Hillsdale, NJ, 1993.

[Gigerenzer, 1994] G. Gigerenzer. Why the distinction between single-event probabilities and frequencies is relevant for psychology (and vice versa). In *Subjective Probability*, G. Wright and P. Ayton, eds. pp. 129–161. Wiley, New York, 1994.

[Gigerenzer, 1996a] G. Gigerenzer. The psychology of good judgment: Frequency formats and simple algorithms. *Journal of Medical Decision Making*, **16**, 273–280, 1996.

[Gigerenzer, 1996b] G. Gigerenzer. Rationality: Why social context matters. In *Interactive Minds: Life-span Perspectives on the Social Foundation of Cognition*, P. Baltes and U. M. Staudinger, eds. pp. 319–346. Cambridge University Press, Cambridge, 1996.

[Gigerenzer, 1997] G. Gigerenzer. Bounded rationality: Models of satisficing inference. *Swiss Journal of Economics and Statistics*, **133**, 1997.

[Gigerenzer and Goldstein, 1996] G. Gigerenzer and D. G. Goldstein. Reasoning the fast and frugal way: Models for bounded rationality. *Psychological Review*, **103**, 650–669, 1996.

[Gigerenzer and Hoffrage, 1995] G. Gigerenzer and U. Hoffrage. How to improve Bayesian reasoning without instruction: Frequency formats. *Psychological Review*, **102**, 684–704, 1995.

[Gigerenzer *et al.*, 1991] G. Gigerenzer, U. Hoffrage and H. Kleinbölting. Probabilistic mental models: a Brunswikian theory of confidence. *Psychological Review*, **98**, 506–528, 1991.

[Gigerenzer and Murray, 1987] G. Gigerenzer and D. J. Murray. *Cognition as Intuitive Statistics*. Erlbaum, Hillsdale, NJ, 1987.

[Gigerenzer *et al.*, 1989] G. Gigerenzer, Z. Swijtink, T. Porter, L. Daston, J. Beatty and L. Krüger. *The Empire of Chance: How Probability Changed Science and Everyday Life*. Cambridge University Press, Cambridge, 1989.

[Goldstein and Gigerenzer, 1996] D. G. Goldstein and G. Gigerenzer. Recognition: how to exploit a lack of knowledge. Unpublished manuscript, 1996.

[Gould, 1992] S. J. Gould. *Bully for Brontosaurus: Further Reflections in Natural History*. Penguin Books, New York, 1992.

[Grice, 1975] H. P. Grice. Logic and conversation. In *Syntax and Semantics, III: Speech Acts*, P. Cole and J. L. Morgan, eds. pp. 41–58. Academic Press, New York, 1975.

[Hertwig, 1995] R. Hertwig. *Why Dr. Gould's Homunculus doesn't Think Like Dr. Gould: The Conjunction Fallacy Reconsidered*. Hartung-Gorre Verlag, Konstanz. Doctoral dissertation, Universigät Konstanz, Germany, 1995.

[Hertwig and Gigerenzer, 1997] R. Hertwig and G. Gigerenzer The 'conjunction fallacy' revisited: How intelligent inferences look like reasoning errors. Manuscript. Max Planck Institute for Psychological Research, Munich, 1996.

[Hoffrage and Gigerenzer, 1996] U. Hoffrage and G. Gigerenzer. The impact of information representation on Bayesian reasoning. In *Proceedings of the Eighteenth Annual Conference of the Cognitive Science Society*, pp. 126–130. Erlbaum, Mahwah, NJ, 1996.

[Jorland, 1987] G. Jorland. The Saint Petersburg Paradox 1713-1937. In *The Probabilistic Revolution, Vol. 1. Ideas in the Sciences*, L. Krüger, L. Daston and M. Heidelberger, eds. pp. 157–190. The MIT Press, Cambridge, MA, 1987.

[Kleiter, 1994] G. D. Kleiter. Natural sampling: rationality without base rates. In *Contributions to Mathematical Psychology, Psychometrics, and Methodology*, G. H. Fischer and D. Laming, eds. pp. 375–388. Springer, New York, 1994.

[Koehler, 1996] J. J. Koehler. The base rate fallacy reconsidered: descriptive, normative, and methodological challenges. *Behavioral and Brain Sciences*, **19**, 1–54, 1996.

[Levi, 1983] I. Levi. Who commits the base rate fallacy? *Behavioral and Brain Sciences*, **6**, 502–506, 1983.

[Loftus, 1991] G. R. Loftus. On the tyranny of hypothesis testing in the social sciences. *Contemporary Psychology*, **36**, 102–104, 1991.

[Loftus, 1993] G. R. Loftus. Editorial comment. *Memory and Cognition*, **21**, 1–3, 1993.

[Lopes, 1981] L. L. Lopes. Decision making in the short run. *Journal of Experimental Psychology: Human Learning and Memory*, **7**, 377–385, 1981.

[Lovie and Lovie, 1986] A. D. Lovie and P. Lovie. The flat maximum effect and linear scoring models for prediction. *Journal of Forecasting*, **5**, 159–168, 1986.

[Luce, 1980] R. D. Luce. Comments on the chapters by MacCrimmon, Stanbury and Wehrung, and Schum. In *Cognitive Processes in Choice and Decision Making*, T. S. Wallsten, ed. Erlbaum, Hillsdale, NJ, 1980.

[Marr, 1982] D. Marr. *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*. Freeman, San Francisco, 1982.

[Martignon and Hoffrage, in press] L.Martignon and U. Hoffrage. Environments where there is no simplicty/accuracy tradeoff. In *Simple Heuristics that Make Us Smart*, G. Gigerenzer and P. M. Todd, eds. Oxford University Press, New York, in press.

[Oaksford and Chater, 1994] M. Oaksford and N. Chater. A rational analysis of the selection task as optimal data selection. *Psychological Review*, **101**, 608–631, 1994.

[Samuelson, 1938] P. A. Samuelson. A note on the pure theory of consumers' behavior. *Economica*, **5**, 61–71, 1938.

[Sedlmeier and Gigerenzer, 1996] P. Sedlmeier and G. Gigerenzer. Teaching Bayesian reasoning in less than two hours. Manuscript submitted for publication, 1996.

[Sen, 1993] A. Sen. Internal consistency of choice. *Econometrica*, **61**, 495–521, 1993.

[Tversky and Kahneman, 1980] A. Tversky and D. Kahneman. Causal schemata in judgments under uncertainty. In *Progress in Social Psychology*, Vol. 1. M. Fishbein, ed. pp. 49–72. Erlbaum, Hillsdale, NJ, 1980.

[Tversky and Kahneman, 1983] A. Tversky and D. Kahneman. Extensional versus intuitive reasoning: the conjunction fallacy in probability judgment. *Psychological Review*, **90**, 293–315, 1983.