ELSEVIER

Cognition 58 (1996) 1–73

COGNITION

Are humans good intuitive statisticians after all? Rethinking some conclusions from the literature on judgment under uncertainty

Leda Cosmides*, John Tooby

Center for Evolutionary Psychology, University of California, Santa Barbara, CA 93106, USA

Received September 1, 1992, final version accepted January 2, 1995

Abstract

Professional probabilists have long argued over what probability means, with, for example, Bayesians arguing that probabilities refer to subjective degrees of confidence and frequentists arguing that probabilities refer to the frequencies of events in the world. Recently, Gigerenzer and his colleagues have argued that these same distinctions are made by untutored subjects, and that, for many domains, the human mind represents probabilistic information as frequencies. We analyze several reasons why, from an ecological and evolutionary perspective, certain classes of problem-solving mechanisms in the human mind should be expected to represent probabilistic information as frequencies. Then, using a problem famous in the "heuristics and biases" literature for eliciting base rate neglect, we show that correct Bayesian reasoning can be elicited in 76% of subjects – indeed, 92% in the most ecologically valid condition – simply by expressing the problem in frequentist terms. This result adds to the growing body of literature showing that frequentist representations cause various cognitive biases to disappear, including overconfidence, the conjunction fallacy, and base-rate neglect. Taken together, these new findings indicate that the conclusion most common in the literature on judgment under uncertainty – that our inductive reasoning mechanisms do not embody a calculus of probability – will have to be re-examined. From an ecological and evolutionary perspective, humans may turn out to be good intuitive statisticians after all.

* Corresponding author; e-mail: tooby@alishaw.ucsb.edu; fax: 805-965-1163.

0010-0277/96/$15.00
SSDI 0010-0277(95)00664-8

1. Introduction

During the early 1800s, when mathematical theories of probability were only a little over a century old, Pierre Laplace argued that probability theory was "only good sense reduced to calculus" (Laplace, 1814/1951, p. 196). When theories of probability conflicted with the intuitions of "reasonable men", mathematicians went back to the drawing board and changed their theories: a clash between probability theory and intuition meant the theory was wrong, not the intuition (Daston, 1980, 1988).

By the 1970s, this view of the relationship between intuition and probability theory had been reversed. In cognitive psychology, we began to view clashes between intuition and probability theory as evidence that our intuitions were wrong, not the theory. According to Kahneman and Tversky, for example, "The presence of an error of judgment is demonstrated by comparing people's responses either with an established fact (e.g., that the two lines are equal in length) or with an accepted rule of arithmetic, logic, or statistics" (1982, p. 123). In a series of seminal papers on inductive reasoning, Kahneman and Tversky used evidence of such clashes to reject the theory that "good sense" embodied a calculus 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 & Tversky, 1973, p. 237)

The view that people's untutored intuitions do not follow a calculus of probability has become the conventional wisdom in psychology today. The literature on human judgment under uncertainty has become a catalog of "cognitive biases" and "normative fallacies", and terms such as "base-rate fallacy", "overconfidence" and "conjunction fallacy" have entered the lexicon of cognitive psychology. In studies of Bayesian reasoning, for example, psychologists argue about whether people ignore base rates entirely or sometimes use them slightly, but not about whether untutored subjects follow the calculus of Bayes' rule. It is presumed that they do not.

As well established as this conclusion appears to be, we think it is premature. For one thing, there is not just one "calculus of probability", but many, and the fact that subjects do not follow one of them does not preclude the possibility that they are following another. There is Bayes's theorem, Neyman–Pearson decision theory, Fisherian null-hypothesis testing, non-additive Baconian probabilities – all of which have different assumptions and are therefore appropriate to different kinds of problems.[1]

Not only are there different statistical theories, but professional probabil-

[1] Indeed, one can even develop specialized normative systems for computing probabilities and making decisions, which are tailored to particular domains or circumstances. Models from evolutionary biology of risk-sensitive foraging are an example (Real & Caraco, 1986).

ists themselves disagree – often violently – about the central issues in their field. Different probabilistis looking at the same problem will frequently give completely different answers to it – that is, they will make contradictory claims about which answer is normatively correct. They even disagree about what probability itself means. For example, Bayesians argue that probability refers to a subjective degree of confidence and, accordingly, because one can express one's confidence that a single event will occur, one can sensibly refer to the probability of a single event. In contrast, frequentists argue that probability is always defined over a specific reference class (such as an infinite – or very large – number of coin tosses), and refers to the relative frequency with which an event (such as "heads") occurs. A frequentist would say that the calculus of probability cannot be used to compute the "probability" of a single event, because a single event cannot have a probability (i.e., a relative frequency).

The conceptual distinction between the probability of a single event and a frequency is fundamental to mathematical theories of probability. Recently, Gigerenzer (1991) has argued that the same conceptual distinction is made by untutored subjects. He has further argued that, for many domains, the human mind represents probabilistic information as frequencies. In essence, Gigerenzer has hypothesized that some of our inductive reasoning mechanisms are good "intuitive statisticians" of the frequentist school.

At first glance, this hypothesis seems implausible, for two reasons. First, it seems to fly in the face of the considerable body of evidence assembled over the last 20 years in the literature on judgment under uncertainty that supports the conclusion that our inductive reasoning procedures do not embody a calculus of probability. But most of these experiments asked subjects to judge the probability of a single event – a question which a frequentist would say has nothing to do with the mathematical theory of probability – and which frequentist mental mechanisms would find hard to interpret. This leaves open the possibility that people are able to make intuitive judgments that accord with a calculus of probability, such as Bayes's rule, as long as they represent probabilities as frequencies.

The second reason this hypothesis seems unlikely on its face is that it seems absurd to claim that untutored subjects are intuitively making a subtle distinction that professional probabilists still argue about in their technical journals. But if you apply David Marr's (1982) approach to the problem of statistical inference and ask what design features you would expect of mechanisms that can operate well under evolutionarily standard and ecologically realistic conditions, the picture changes dramatically. As we will discuss, an architecture designed along frequentist principles is what you would expect given a Marrian functional analysis of statistical inference.

In this article, we will explore what we will call the "frequentist hypothesis" – the hypothesis that some of our inductive reasoning mechanisms do embody aspects of a calculus of probability, but they are designed to take frequency information as input and produce frequencies as output.

We will briefly review evidence from the literature indicating that subjects do conceptually distinguish between frequencies and single-event probabilities, and that there exist cognitive mechanisms that encode frequency information automatically. We will then report a strong test of the frequentist hypothesis: using a problem famous in the "heuristics and biases" literature for eliciting low levels of Bayesian performance, we tested to see whether frequentist representations would elicit from subjects answers that conform to Bayes' rule.

1.1. Subjective probabilities versus objective frequencies: why do Bayesians and frequentists disagree?

To understand why some of our inductive reasoning mechanisms might be designed to be intuitive frequentists, it is important to first understand why Bayesians and frequentists differ in their interpretation of probability theory.

Mathematical theories of probability are theories of inductive reasoning: they specify how to make inferences from data to hypotheses. But the mathematization of inductive inference has not solved Hume's puzzle; there is still no universally agreed upon method of statistical inference (Gigerenzer et al., 1989). There are deep divisions among professional probabilists, not only over how statistical inference should be conducted, but over the very meaning of the word "probability". One of the deepest divisions is between the Bayesians and the frequentists.

In science, what everyone really wants to know is the probability of a hypothesis given data – $p(H|D)$. That is, *given these observations, how likely is this theory to be true?* This is known as an *inverse probability* or a *posterior probability*. The strong appeal of Bayes' theorem arises from the fact that it allows one to calculate this probability: $p(H|D) = p(H)p(D|H)/p(D)$, where $p(D) = p(H)p(D|H) + p(-H)p(D|-H)$. Bayes' theorem also has another advantage: it lets one calculate the probability of a single event, for example, the probability that a particular person, Mrs. X, has breast cancer, given that she has tested positive for it. (This property makes Bayes' theorem popular among economists, who are interested in calculating values such as the probability that an individual consumer will choose one alternative over another.)

Despite these advantages, most professional statisticians are frequentists, not Bayesians. Even the statistical methods for scientific inference that are canonically applied in scientific psychology, such as Fisherian null-hypothesis testing and Neyman–Pearsonian decision theory (known to psychologists as signal detection theory), were derived from the frequentist school. They allow one to calculate the probability of data given a hypothesis – $p(D|H)$ – which is known as a *likelihood*; for example, a level of significance is a likelihood. Unfortunately, they do not allow one to calculate $p(H|D)$, an inverse probability.

Why, then, is Bayes' theorem so rarely used in mathematics and as a

method of scientific inference? R.A. Fisher, one of the architects of modern probability theory, tells us in his foundational book, *The Design of Experiments*, in a section tellingly entitled "The rejection of inverse probability":

> The axiom leads to apparent mathematical contradictions. In explaining these contradictions away, advocates of inverse probability seem forced to regard mathematical probability, not as an objective quantity measured by observed frequencies, but as measuring merely psychological tendencies, theorems respecting which are useless for scientific purposes. (Fisher, 1951, pp. 6–7)

In other words, to calculate the probability of a single event, one must abandon the notion that probability refers to the frequency of an event; one must view probability as an individual's subjective degree of confidence in a hypothesis. A troubling consequence of this interpretation of probability is that, given identical data, $p(H|D)$ can differ from person to person. Yet science strives for intersubjective agreement and consensual methods for arriving at knowledge; to accept Bayes' theorem is to renounce that goal.

Here is why: to use Bayes' theorem, one must first specify one's prior probability that a hypothesis is true – $p(H)$. To get the posterior probability, $p(H|D)$, one revises the prior probability either up or down, depending on $p(D|H)$, which can be determined from experiments. But how does one determine $p(H)$? Some psychologists believe that $p(H)$ should be set according to a particular base rate in the population, and consider it an error when subjects do not do so: for example, Tversky and Kahneman have called base-rate neglect "a sharp violation of Bayes' rule" (1974, p. 1124). But Bayes' theory does not require that the prior probability be set at a base rate. In fact, *Bayes' theory places no constraints whatsoever on how one should set one's prior probability*. Because Bayesian probabilities are subjective, $p(H)$ can vary from person to person, depending on each person's judgment of what is reasonable. And there are many different normative theories specifying what is reasonable (Gigerenzer & Murray, 1987).[2] Consequently, $p(H|D)$, which is a function of $P(H)$, can also differ

[2] For example, consider the "cab problem", in which a subject is told that there are two cab companies in the city, Blue and Green, and is asked how probable it is that a cab that was involved in a hit-and-run accident at night was Blue (e.g., Bar-Hillel, 1980). The subject is told that a witness who is known to be correct 80% of the time (the reliability) identified the cab as Blue, and that 15% of cabs in the city are Blue and 85% are Green (the base rate). But why should the subject use this base rate – "percent of Blue cabs in the city" – as his or her prior probability? There are many other legitimate possibilities (see, for example, Gigerenzer, 1990). For instance, suppose the subject knows based on earlier experience that drivers from small cab companies are less well trained and therefore relatively more likely to cause accidents, or that drivers from small cab companies are more likely to flee the scene of an accident because they are less able to afford higher insurance premiums than drivers from the larger, more well-organized companies, which assume some of their drivers' insurance burden. Any real-world knowledge of this kind *should* be integrated into the subject's prior probability estimate. Consequently, a large number of prior probability estimates are defensible, and a good Bayesian could quite legitimately ignore the "percent of Blue cabs in city" base rate.

from person to person. This means that different people can draw entirely different conclusions from one and the same experiment.

For example, having read extensively about "heuristics and biases" in inductive reasoning, you might think that the hypothesis we are testing in this article is unlikely to be true, and therefore set your prior probability low – perhaps at .2. But another reader might think our hypothesis is reasonably likely to be true, and therefore set her prior probability higher – perhaps at .6. Let us say that the experiments then showed that $p(D|H)$ was .7 and $p(D|-H)$ was .3. What would Bayes' theorem lead each of you to conclude from these experiments? Given these data, you would conclude that the probability of the hypothesis is .37, whereas the other reader would conclude that it is .78. In other words, her confidence that the hypothesis is true would be more than twice what yours would be. If one is seeking a method of inductive reasoning that is both grounded in the world and leads to intersubjective agreement, then Bayes' theorem obviously cannot be it – as Fisher, for instance, recognized.[3]

Suppose one rejects the notion that probability refers to subjective degrees of confidence, as von Mises, Neyman, Pearson, and many other probabilists have. Why can't one be a frequentist and still talk about the probability of a single event?

Accepting the frequentist interpretation of probability as the relative frequency of an event defined over a specified reference class entails the idea that it is meaningless to refer to the probability of a single event. First, a single event either happens or not, and therefore cannot have a "relative frequency". But there is a second, more far-reaching reason. A single event cannot have "a" probability because it belongs to many reference classes, not just one. In fact, the number of reference classes an event belongs to is potentially infinite, depending on the system of categorization applied to the event. Let us illustrate this problem with an updated version of an example provided by Richard von Mises, a twentieth-century pioneer of probability theory, and a thoroughgoing frequentist.

Consider the reference class, "All American women between the ages of 35 and 50", and assume that the attribute we are interested in is the probability that women in this category get breast cancer. Let us say that the relative frequency with which this happens in a year is 4 out of 100. Can we meaningfully "collapse" this figure onto a single individual, says Mrs. X, who is 49, perfectly healthy, and whose mother had breast cancer? Can we say that Mrs. X has a 4% chance of getting breast cancer in the next year? According to von Mises, such reasoning would be "utter nonsense" (1957/1981, p. 18). This is because Mrs. X belongs to an indefinitely large set of

[3] Not if one wants to be consistent, at least. Fisher, for example, seems to have switched back and forth in an inconsistent way between his frequentist view and his interpretation of the level of significance as the degree of confidence that one should have in a hypothesis (see for example, Gigerenzer et al., 1989, pp. 98–106).

different reference classes, and the relative frequency of breast cancer may differ for each of them. For example, the relative frequency of breast cancer for the reference class "women between 45 and 90", to which Mrs. X also belongs, is higher – say 11%. She is also a member of the category "women between 45 and 90 whose mothers had breast cancer", and the relative frequency for this reference class is even higher – say 22%. So what is the "probability" that Mrs. X will get breast cancer? Is it 4%? 11%? 22%? We can devise as many different figures as we can devise reference classes to which Mrs. X belongs.

Can we solve the problem by restricting the reference class as far as possible, by taking into account all of Mrs. X's individual characteristics – for example, "All women who are 49, live in Palo Alto, smoke, do not drink, had a mother with breast cancer, are Silicon Valley executives, had two children before the age of 25 and one after 40, are of Greek descent . . ."? This approach doesn't work either. The more narrowly we define the reference class, the fewer individuals there are in it, until we are left, at the limit, with only one person, Mrs. X. This creates an indeterminacy problem: the fewer the number of individuals in a reference class, the less reliable is any relative frequency derived from that class – the "error term" grows towards the infinitely large. And at the limit of only one person in a reference class, "relative frequency" loses all meaning.

If probability refers to relative frequency, then the question, "What is the probability of a single event?" has no meaning; if one wants to talk about the probability of a single event, then one must accept the interpretation of probability as subjective degrees of confidence. Normatively, it is a package deal: one cannot be a frequentist and accept single-event probabilities (see footnote 3).

Although a frequentist cannot meaningfully speak about the *probability* of a single event, frequentist statistics can be used to make a *decision* about a single case, such as whether Mrs. X, who has tested positive for breast cancer, should start treatment. Indeed, Neyman–Pearson decision theory is a frequentist method that was expressly designed for making such decisions. By making a distinction between *believing* that a hypothesis is true and behaving *as if* it were true, it provides a method for combining information about the relative frequency of events with information about the costs and benefits associated with alternative courses of action (i.e., with a given matrix of payoffs for hits, misses, false alarms, and true rejections). In this system, each individual decision is made in accordance with rules that would yield the best payoff across the expected series of decisions. Assume, for example, that 80% of women who test positive for cancer actually have it. If the costs of not treating a woman with cancer are high and the costs of treating a healthy woman are low, then administering treatment to all women who test positive will save more lives than withholding treatment. Consequently, the physician should act *as if* Mrs. X has cancer, and treat her, even though the physician *believes* that 20% of the women he or she

will be treating are perfectly healthy. Notwithstanding the fact that they are sometimes conflated, making decisions and estimating probabilities are two very different things. Although probability information can serve as input to a decision-making procedure, no decision can be rationally made until it is coupled to information about values.

At this point it is important that we distinguish between the Bayesian *interpretation* of probability, and the *calculus* of Bayes' rule. The calculus of Bayes' rule, which is nothing more than a formula for calculating a conditional probability, is a simple consequence of the elementary axioms of probability theory as laid out, for example, by Kolmogorov (1950). This axiomatic system can be interpreted in a number of ways, and both the subjectivist and the frequentist interpretations of probability are consistent with these axioms. One can use Bayes' rule to calculate the probability of a single event, such as "What is the probability that Mrs. X actually has breast cancer, given that she tested positive for it?", which entails *interpreting* probability as a subjective degree of confidence. But one can also use Bayes' rule to calculate a relative frequency, such as "How many women who test positive for breast cancer actually have it?" In this second case, one interprets probability as a frequency: the inputs to the equation are frequencies, and the output is a frequency. The first case entails a subjectivist interpretation, the second a frequentist interpretation. But in both cases Bayes' rule was used to calculate the relevant probability.

So whether you are a frequentist or a subjectivist, the formula known as "Bayes' rule" is a component of your calculus of probability. It specifies constraints that must be satisfied when prior probabilities and likelihoods are mapped onto posterior probabilities. In practice, these constraints can be realized by many different algorithms – ones that multiply and divide, ones that count category members in a representative population, and so on.

What, then, does it mean to say that a person's *reasoning* is Bayesian? It depends on which professional community you are addressing. When philosophers and mathematicians refer to themselves as Bayesian, with a capital "B", it means they take a subjectivist position on the nature of probability. But the subjectivist/frequentist debate has played virtually no role in experimental psychology. When psychologists (including ourselves) argue about whether people do, or do not, engage in Bayesian reasoning, they are discussing the extent to which our inductive reasoning mechanisms map inputs onto the same outputs that Bayes' rule would, regardless of the actual cognitive procedures employed to accomplish this. That is, psychologists are asking whether humans have inductive reasoning mechanisms that implement the constraints specified by Bayes' rule. They count any answer that satisfies these constraints as correct – whether it is expressed as a frequency or single-event probability, and regardless of the algorithm by which it was computed. (In fact, the design of most of the relevant experiments contains no way of determining the method by which subjects arrive at their answers.) Psychologists use these inclusive criteria because

they are concerned with the *calculus* of Bayes' rule, not its interpretation. They want to know whether humans are "good" intuitive statisticians – that is, whether their inductive reasoning mechanisms embody aspects of a calculus of probability.

That is our concern as well. To remind the reader of this, we will use the term *bayesian reasoning* – with a small "b" – to refer to any cognitive procedure that causes subjects to reliably produce answers that satisfy Bayes' rule, whether that procedure operates on representations of frequencies or single-event probabilities. In this way we can, without contradiction, ask the question, "Do frequentist representations elicit bayesian reasoning?"

1.2. Re-evaluating the "prior probability" of judgmental heuristics

In science, ideas rise and fall not only because of the evidence for them, but also because they seem plausible or compelling – because people assign them a high "prior probability". Quite apart from the evidence for it, the hypothesis that people cannot spontaneously use a calculus of probability and rely instead on judgmental heuristics seems plausible and compelling because of certain arguments that have been made in cognitive psychology. In this section we evaluate these arguments, to see if the "heuristics and biases" hypothesis deserves the high prior probability that it is usually accorded.[4]

One of the goals of cognitive psychology is to discover what information-processing mechanisms are reliably developing features of the human mind – that is, what mechanisms can be thought of as part of human nature. We assume that this goal is shared by those cognitive psychologists who study

[4] The purpose of this section is to examine some of the core principles underlying an influential and persistant school of thought within psychology, and to suggest an alternative analytic framework that we think may lead to new discoveries and a consequent re-evaluation of human inductive competences. It is not intended as an historical exigesis of the ideas of Tversky, Kahneman, or any other individual. Like all creative scientists, their ideas are complex and may change over time. That we quote heavily from Tversky and Kahneman should be seen as a compliment to the power of their ideas, the clarity with which they have expressed them, and the centrality that their arguments have come to have in the intellectual community that studies human statistical inference. Our concern is with the widely accepted logic of the heuristics and biases position, rather than with its creators.

Indeed, among the scores of factors and framing effects that Kahneman and Tversky have discussed over the years as potentially impeding or facilitating "correct" judgments, they have mentioned frequencies (Kahneman & Tversky, 1982). The fact that such a powerful and organizing dimension can nevertheless remain virtually uninvestigated for two decades underscores why we think the shift to a new analytic framework is in order. The assumption of severe processing limitations has forestalled many researchers from seriously considering or investigating a contrasting possibility: that our minds come equipped with very sophisticated intuitive statistical competences that are well-engineered solutions to the problems humans normally encountered in natural environments (Tooby & Cosmides, 1992b), and that ecologically valid input (e.g., frequency formats) may be necessary to activate these competences.

judgment under uncertainty, and that this is why Tversky and Kahneman liken cognitive heuristics to perceptual heuristics and the study of cognitive illusions to the study of perceptual illusions (Tversky & Kahneman, 1974; Kahneman & Tversky, 1982). Vision scientists study perceptual illusions in order to discover perceptual heuristics, which are usually thought of as reliably developing features of the human brain. As Kahneman and Tversky put it, "we use illusions to understand principles of *normal* perception" (1982, p. 123; emphasis ours).

The perceptual analogy extends to the choice of experimental methods in the judgment under uncertainty program: to document heuristics and biases, researchers frequently use the same experimental logic that psychophysicists use to demonstrate that a perceptual heuristic is a design feature of the cognitive architecture. For example, pains have been taken to show that cognitive biases are widespread, not caused by motivational factors, and difficult to eradicate (Nisbett & Ross, 1980). Cognitive biases "seem reliable, systematic, and difficult to eliminate" (Kahneman & Tversky, 1972, p. 431); a judgmental error often "remains attractive although we know it to be an error" (1982, p. 123). The persistence of a perceptual illusion even when one "knows better" is usually taken as evidence that it is a reliably developing feature of the brain. The persistence of cognitive illusions among those who have some expertise in inferential statistics plays a similar role in the heuristics and biases literature (Tversky & Kahneman, 1971; Casscells, Schoenberger, & Graboys, 1978):

> The reliance on heuristics and the prevalence of biases are not restricted to laymen. Experienced researchers are also prone to the same biases – when they think intuitively . . . Although the statistically sophisticated avoid elementary errors, such as the gambler's fallacy, their intuitive judgments are liable to similar fallacies in more intricate and less transparent problems. (Tversky & Kahneman, 1974, p. 1130)

The message is that statistical training may create new, competing cognitive structures, but our heuristics and biases will remain.

The implication of this experimental logic is that judgmental heuristics and biases[5] – like perceptual heuristics and biases – are reliably developing cognitive processes, and not accidents of personal history, such as whether one was schooled in probability and statistics. Indeed, if the scientific community thought that the documentation of normative fallacies revealed nothing more fundamental about the human mind than how one was schooled, then the heuristics and biases program would be of no more interest than a research program documenting that native speakers of English with no schooling in Japanese commit "errors and fallacies" when they try to conjugate Japanese verbs.

Although heuristics and biases are believed to be reliably developing cognitive processes, mechanisms embodying statistical rules are not thought

[5] or the learning mechanisms that induce them.

to be part of our architecture. According to Kahneman and Tversky, "the laws of chance are neither intuitively apparent, nor easy to apply" (1972, p. 431). Pointing out that few people discover statistical rules even though "everyone is exposed, in the normal course of life, to numerous examples from which these rules could have been induced" (Tversky & Kahneman, 1974, p. 1130), they argue that the human mind is not designed to spontaneously learn such rules: "Statistical principles are not learned from everyday experience because the relevant instances are not coded appropriately" (1974, p. 1130).

Evolutionary questions are inescapable in psychology. Any claim that one cognitive process is a reliably developing feature of the human mind whereas another cognitive process is not, raises the question, "Why would that design have been selected for rather than the other one?" In principle, the adaptive problem of judging uncertain events could be solved by judgmental heuristics or by statistical rules. If making accurate judgments under uncertainty is an important adaptive problem, why would natural selection have designed a mind that uses error-prone heuristics rather than an accurate calculus of probability?

Although they do not cast their answer in explicitly evolutionary terms, Tversky and Kahneman answer this question thus: cognitive psychology "is concerned with internal processes, mental limitations, and the way in which the processes are shaped by these limitations" (Kahneman, Slovic, & Tversky, 1982, p. xii). Mental limitations prevent one from correctly solving problems that are too complex. Heuristics are strategies that simplify complex tasks and get the job done well enough – they don't optimize, but they do "satisfice". Heuristics "are highly economical and usually effective"; they "reduce the complex tasks of assessing probabilities and predicting values to simpler judgmental operations", and "make them tractable for the kind of mind that people happen to have" (Tversky & Kahneman, 1974, pp. 1131, 1124; Kahneman et al., 1982, p. xii). This line of reasoning was developed in the 1950s by Herbert Simon (e.g., 1956), who they credit for having helped inspire the heuristics and biases program (Kahneman et al., 1982, pp. xi–xii).

The logic of this position is perfectly respectable. It is obviously true that a computational mechanism cannot solve a problem that is more complex than it can handle. It is also true that natural selection, which is a "better than" principle,[6] does not necessarily produce "optimal" algorithms. But are there good grounds for applying this general line of reasoning to this domain – that is, for thinking that people will judge uncertain events using heuristics rather than a calculus of probability? Bayes' rule, for example, is

[6] An algorithm that is "better than" an existing alternative can be selected for, whether it optimizes or not. Also, "optimality" is not as simple a concept as the heuristics and biases literature might lead one to think; it must always be defined with respect to a predesignated set of constraints (see, for example, Maynard Smith, 1978).

simple enough – a small, hand-held calculator can perform the necessary operations with only a line or two of code. So why should the application of Bayes' rule be difficult for an information-processing system of the size and complexity of the human brain?

There are several problems with the mental limitation argument, some related to the complexity of the presumed mechanisms, others to the complexity of the task.

(1) *Necessity of heuristics.* The visual system must use heuristics because there is no alternative "calculus of vision". Perceptual heuristics are necessary because the only information available in the environment for the system to use are cues that are not perfectly correlated with the size, shape, color, and texture of distal objects. In contrast, there is a calculus of probability, and its rules are simple enough to be implemented in a calculator.

(2) *Mechanism complexity.* Although determining whether and how a statistical rule should be applied to a given domain can be complex, statistical rules themselves are not.[7] Indeed, natural selection has produced computational mechanisms in the visual system that are vastly more complex than those that would be required to apply the calculus of Bayes' rule. There is no "natural limit" on evolved complexity that would prevent the evolution of computational mechanisms that embody Bayes' rule.

(3) *Task complexity.* This justification assumes that some tasks are inherently complex, and some are inherently simple, independent of the nature of the computational mechanisms that solve them. But whether a problem is complex or simple depends, in part,[8] on the design of the computational mechanisms that are available to solve it. To borrow an example from Sperber (1985), recalling a 20-digit number is simple for a digital computer, but difficult for a human being, whereas remembering the gist of the story of Little Red Riding Hood is simple for a human being, but difficult for a digital computer. Seeing objects seems effortless compared to long division because we have mechanisms specially designed for visual perception, but not for long division – not because seeing objects is a

[7] For the purposes of this point, complexity can be defined in any of a number of ways: number of processing steps, amount of information to be integrated or reconstructed, and so on.

[8] It also depends on the quality of the information available in the environment for solving the problem, as when information is degraded or ambiguous due either to permanent properties of the world or because it is coming from an antagonistically co-evolving organism. But judgment under uncertainty researchers are not referring to this kind of task complexity; because subjects are *told* the relevant information, it is usually assumed that the complexity of the task resides in the computations necessary rather than in the mapping from world to rule. Thus, little research has been done on whether "errors" are due to discrepancies between the subject's and the experimenter's judgment about how the problem information should be mapped onto probabilistic concepts.

"simpler task" than long division. (Indeed, machine vision has proved elusive at a time when statistical programs have become ubiquitous on desktop computers.)

(4) *Specifying the limitation.* It is true that previously existing structures can limit what mechanisms can evolve. Such phylogenetic constraints are responsible, for example, for the fact that nerve cells project from the front of the retina rather than the back, causing a blind spot in vision where they meet to form the optic nerve bundle. But the argument that computational procedures embodying statistical rules did not evolve (or cannot be induced[9]) due to "mental limitations" is empirically vacuous unless one specifies what those limitations are. To say that we cannot solve a statistical problem correctly because it is "too complex" or because we have "mental limitations" is merely to restate the fact that we did not evolve the mechanisms for correctly solving it. It begs the question of why we did not.

In short, the argument for judgmental heuristics appears weak when scrutinized from an evolutionary perspective. A calculus of probability exists, it is not inherently complex, it can be instantiated by simple mechanisms, and there are no known phylogenetic constraints that would prevent the evolution of such mechanisms. There is nothing inherently flawed with the line of reasoning used by Simon and others to argue for the possibility of heuristics of reasoning; rather, its surface plausibility breaks down in the particular case of the probability calculus.

Indeed, the superficial plausibility of the mental limitation argument depends on a certain old-fashioned image of the mind: that it has the architecture of an early model, limited-resource general-purpose computer that is incapable of running programs of much complexity. Given this image, one would expect crude rules-of-thumb rather than well-designed mechanisms because a complex set of procedures cannot be run by such a limited system. But we now know that the human mind contains a number of specialized mechanisms of considerable complexity, from color vision, to motor control, to grammar acquisition (see, for example, Shepard, 1992; Bizzi, Mussa-Ivaldi, & Giszter, 1991; and Pinker, 1984, respectively). How, then, can one sustain the argument that a computationally trivial algorithm instantiating Bayes' rule is too complex to be run by our cognitive architecture, but that vision is not?

One could claim, post hoc, that there are neural or developmental constraints of unknown nature that allow the evolution of mechanisms for maintaining color constancy and induction of the past tense, but that preclude the evolution of well-designed mechanisms for statistical

[9] To say that we cannot induce statistical rules "because the relevant instances are not coded appropriately" (Tversky & Kahneman, 1974, p. 1130) simply pushes the problem one step back: one must explain what kept us from evolving a computational system that does code the relevant instances appropriately.

inference – but there is nothing, *a priori*, to make one think this is true. In fact, there are good reasons to think it is false. Behavioral ecologists who study foraging have found evidence of very sophisticated statistical reasoning in organisms with nervous systems that are considerably simpler than our own, such as certain birds and insects (e.g., Real, 1991; Real & Caraco, 1986). Moreover, John Staddon (1988) has argued that in animals from sea snails to humans, the learning mechanisms responsible for habituation, sensitization, classical conditioning and operant conditioning can be formally described as Bayesian inference machines.

This evidence suggests that bird brains and insect minds are capable of performing statistical calculations that some psychologists have assumed human brains are "too limited" to perform. But if a bird brain can embody a calculus of probability, why couldn't a human brain embody one as well? We are not arguing here that humans *must* have well-designed mechanisms for statistical inference – we are merely arguing that the prevailing arguments for why we should *not* have such mechanisms are not substantial. As long as chance has been loose in the world, animals have had to make judgments under uncertainty. If an adaptive problem has endured for a long enough period, and is important enough, then mechanisms of considerable complexity can evolve to solve it. When seen in this light, the hypothesis that humans have inductive reasoning mechanisms that embody a calculus of probability, just like other organisms do, doesn't seem so intrinsically improbable.

1.3. Raising the prior probability of the frequentist hypothesis: Marr and evolution

What, then, should the design of well-engineered reasoning mechanisms be like? Until you answer this Marrian question, you cannot construct experiments that can detect the presence of such designs or recognize their operation (Marr, 1982; Cosmides & Tooby, 1987). In this section we will discuss why, if we do have well-designed mechanisms for statistical reasoning, one might expect them to operate on frequency representations.

Gigerenzer's advocacy of a frequentist approach to inductive reasoning draws much of its motivation from a scrupulous analysis of probability theory and the logic of its application. But the hypothesis that at least some cognitive machinery in the human mind operates on frequentist principles makes sense from a functional, that is, from an evolutionary, point of view as well. By an evolutionary and functional view we simply mean that one should expect a mesh between the design of our cognitive mechanisms, the structure of the adaptive problems they evolved to solve, and the typical environments that they were designed to operate in – that is, the ones that they evolved in. Just as David Marr (1982) studied the reflectant properties of surfaces to understand what kinds of information are available to our visual system, one can ask what kinds of probabilistic information would

have been available to any inductive reasoning mechanisms that we might have evolved. Because cognitive mechanisms evolved to recognize and process information in the form it was regularly presented in the environment of evolutionary adaptedness, to study inductive reasoning one must examine what form such problems regularly took,[10] and what form the information relevant to solving such problems took.

In the modern world, we are awash in numerically expressed statistical information. But our hominid ancestors did not have access to the modern system of socially organized data collection, error checking, and information accumulation which has produced, for the first time in human history, reliable, numerically expressed statistical information about the world beyond individual experience. Reliable numerical statements about single event probabilities were rare or nonexistent in the Pleistocene – a conclusion reinforced by the relative poverty of number terms in modern band-level societies. In our natural environment, the only database available from which one could inductively reason was one's own observations, and possibly those communicated by the handful of other individuals one lived with.

More critically, the "probability" of a single event is intrinsically unobservable. No sense organ can discern that if we go to the north canyon, there is a .25 probability that today's hunt will be successful. Either it will or it won't; that is all one can observe. As useful as a sense organ for detecting single-event probabilities might be, it is theoretically impossible to build one. No organism can evolve cognitive mechanisms designed to reason about, or receive as input, information in a format that did not regularly exist.

What *was* available in the environment in which we evolved was the encountered frequencies of actual events – for example, that we were

[10] Examining the nature of the adaptive problem is just as important as examining the nature of the information available for solving it. This is because selection will not shape decision rules so that they act solely on the basis of what is most likely to be true, but rather on the basis of the weighted consequences of acts given that something is held to be true. Neyman–Pearson decision theory (signal detection theory), for example, is a normative statistical theory that allows one to combine frequency information with cost–benefit information (for an application to the cab problem, see Birnbaum, 1983). Should you walk under a tree that might conceal a predator? Even if such trees have been predator-free 51 (or even 95) times out of 100, an adaptive decision rule should, under many circumstances, cause you to avoid the tree – that is, to act as if the predator were there. The benefits of calories saved via a shortcut, scaled by the probability that there is no predator in the tree, must be weighed against the benefits of avoiding becoming catfood, scaled by the probability that there is a predator in the tree. Because the costs and benefits of hits, misses, false alarms, and correct rejections are often unequal, decision rules that use frequency information may treat as true situations that are unlikely to be true (Tooby & Cosmides, 1990). There is no need to assume that such calculations are consciously accessible, and we take no position on whether the end product of this frequentist process sometimes manifests itself as a consciously experienced subjective degree of certainty.

successful 5 out of the last 20 times we hunted in the north canyon. Our hominid ancestors were immersed in a rich flow of observable frequencies that could be used to improve decision-making, given procedures that could take advantage of them. So if we have adaptations for inductive reasoning, they should take frequency information as input.

Once frequency information has been picked up, why not convert it into a single-event probability? Why not store the encountered frequency – "5 out of the last 20 hunts in the north canyon were successful" – as a single-event probability – "there is a .25 chance that *a* hunt in the north canyon will be successful"? There are advantages to storing and operating on frequentist representations because they preserve important information that would be lost by conversion to a single-event probability. For example:

(1) The number of events that the judgment was based on would be lost in conversion. When the *n* disappears, the index of reliability of the information disappears as well.

(2) One is continuously encountering new information, and having to update one's database. Frequentist representations, which preserve the number and the categorization of events, can be easily updated with each new instance; single-event probabilities cannot. For example, if the next hunt in the north canyon fails, "5 out of 20" can easily be converted to "5 out of 21".

(3) Frequentist representations allow reference classes to be constructed after the fact, allowing one to reorganize one's database flexibly (e.g., Hintzman & Stern, 1978). This allows one to answer a larger array of questions. Assume you have hunted in the north canyon 100 times, and 5 out of the last 20 hunts were successful. But suppose those last 20 times were in summer, and it is now winter. Given that season can modify game distribution, the reference class "hunts during winter one year ago" might be better than the reference class "most recent hunts". Or the criterion for "successful hunt" may need to be changed from "caught a small game animal" to "caught a large game animal", because many more individuals have joined your group and need to be fed. It is computationally trivial to reconstruct a reference class according to new criteria given frequentist representations.

Once a frequency representation has been computed, it can serve as input to decision rules and planning mechanisms. When fed into an appropriate decision rule, a frequency representation can easily produce a subjective degree of confidence – that, for example, a hunt in the north canyon will be successful today. The fact that people routinely report experiencing subjective degrees of confidence does not weaken the claim that the machinery that underlies them operates along frequentist principles. Given these considerations, we can place Gigerenzer's hypothesis that the mind is a good intuitive statistician of the frequentist school into an evolutionary frame-

work, and expand it as follows: during their evolution, humans regularly needed to make decisions whose success could be improved if the probabilistic nature of the world was taken into account. They had access to large amounts of probabilistic information, but primarily or perhaps solely in the form of encountered frequencies. This information constituted a rich resource available to be used to improve decision-making, given procedures that could take advantage of it. Consequently, they evolved mechanisms that took frequencies as input, maintained such information as frequentist representations, and used these frequentist representations as a database for effective inductive reasoning.

A number of predictions follow from this hypothesis:

(1) Inductive reasoning performance will differ depending on whether subjects are asked to judge a frequency or the probability of a single event.

(2) Performance on frequentist versions of problems will be superior to non-frequentist versions.

(3) The more subjects can be mobilized to form a frequentist representation, the better performance will be.

(4) (Strong version) Performance on frequentist problems will satisfy some of the constraints that a calculus of probability specifies, such as Bayes' rule. This would occur because some inductive reasoning mechanisms in our cognitive architecture embody aspects of a calculus of probability.

We are not hypothesizing that every cognitive mechanism involving statistical induction necessarily operates on frequentist principles, only that at least one does, and that this makes frequentist principles an important feature of how humans intuitively engage the statistical dimension of the world. It is possible – we think it likely – that our cognitive architecture includes a constellation of specialized mechanisms whose designs embody non-frequentist principles, alongside or integrated with frequentist designs. Such mechanisms would be deployed when ancestrally valid cues signal the presence of problem-types appropriate to the principles they embody (Tooby & Cosmides, 1990). In short, in contrast to the standard view that the human cognitive architecture does not embody either a calculus of probability or effective statistical competences, we suggest that the human mind may contain a series of well-engineered competences capable of being activated under the right conditions, and that a frequentist competence is prominent among these.

1.4. Do frequentist representations elicit better statistical reasoning?

If people are capable of applying a calculus of probability to frequency representations, then why is the literature on judgment under uncertainty so littered with apparent errors in reasoning?

We should never be surprised to find "errors" in reasoning, even from an "optimally designed" algorithm. This is because it is theoretically impossible to build an "omniscient algorithm": an algorithm that can operate error-free no matter what the format of the information that is fed into it, and no matter what output it is required to produce. So-called "errors" can be produced by the most elegantly designed mechanism, and when they are, one can gain insight into how the mechanism represents information and what kind of output it was designed to produce. The existence of "errors" does not necessarily mean that the mechanisms involved embody a quick-and-dirty rule-of-thumb.

For example, no computational mechanism can correctly process information that it cannot "read"; information can be processed properly only if it is in a format that the mechanism can interpret. A calculator that is designed to correctly multiply numbers presented to it in the format of base 10 will not be able to correctly multiply numbers presented to it in the format of base 2 – it will interpret input such as "10" as ten rather than as two and therefore produce the "wrong" output.

More importantly, no computational mechanism can be expected to correctly produce an answer that it was not designed to produce. For example, choosing food and choosing a spouse both involve "preferences". One can even ask questions about these choices in the same linguistic format: "How much do you like your salad/boyfriend?" But a mechanism that is well designed for choosing nutritious food will not be able to choose the best spouse. Similarly, even though addition and finding a logarithm both involve numbers, a mechanism that is well designed for adding will not be able to find a logarithm.

Suppose people do have reliably developing mechanisms that allow them to apply a calculus of probability, but that these mechanisms are "frequentist": they are designed to accept probabilistic information when it is in the form of a frequency, and to produce a frequency as their output. Let us then suppose that experimental psychologists present subjects with problems that ask for the "probability" of a single event, rather than a frequency, as output, and that present the information necessary to solve the problem in a format that is not obviously a frequency. Subjects' answers to such problems would not appear to have been generated by a calculus of probability, even though they have mechanisms designed to do just that.[11]

[11] In everyday discourse, the word "probability" has meanings other than "relative frequency", such as "weight of evidence". Many situations in real life that ask for the "probability" of a single event are really asking for a weight of evidence judgment: in a court of law, for example, judgments of "probable cause" and "guilt beyond a reasonable doubt" appear to be based on Baconian probabilities and weight of evidence, not on relative frequencies (e.g., Cohen, 1979, 1988). Cohen (1979) has argued that certain answers based on "representativeness" may be normatively correct on a Baconian view of probability. And Gigerenzer (personal communication) has suggested that when subjects are asked for the probability of a single event they may think they are being asked for a weight of evidence judgment and be answering accordingly.

One way to see if this is the case is to compare performance on tests that ask for the probability of a single event to similar tasks that ask for the answer as a frequency. "Errors" that are reliably elicited by the single-event task should disappear on the frequency task.

That seems to be just what happens (for review, see Gigerenzer, 1991). For example, Klaus Fiedler (1988) showed that the "conjunction fallacy" virtually disappears when subjects are asked for frequencies rather than single-event probabilities (see Table 1). Whereas 70–80% of subjects commit the conjunction fallacy when asked for the probability of single events, 70–80% of subjects *do not* commit the conjunction fallacy when asked for relative frequencies (a finding for which there is also evidence in Tversky & Kahneman's (1983) original study).

The same manipulation can also cause the "overconfidence bias" to disappear. "Overconfidence" is usually defined as a discrepancy between one's degree of belief (confidence) in a single event and the relative frequency with which events of that class occur. But such a discrepancy is not a violation of frequentist theories of probability. When one compares subjects' judged frequencies with actual frequencies, as Gigerenzer, Hoffrage, and Kleinbolting (1991) did, "overconfidence" disappears; subjects' judgments turn out to be quite accurate. According to their *probabilistic mental model* theory, one's confidence in a single answer is an estimate of the ecological validity of the cues used in providing that answer, not an

Table 1
Single-event and frequency versions of Fiedler's (1988) conjunction problems

Single-event version	Frequency version
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.	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. To how many out of 100 people who are like Linda do the following statements apply?
Please rank order the following statements with respect to their probability:	
Linda is a bank teller	Linda is a bank teller
Linda is a bank teller and active in the feminist movement	Linda is a bank teller and active in the feminist movement

[a] For both versions, several other statements had to be judged as well (e.g., "Linda is a psychiatric social worker"), but the crucial comparison is between the two statements listed above. For any two categories, A and B, instances of A should be judged more frequent than instances of A&B. In the above case, there are more bank tellers than there are bank tellers who are feminists, because the category "bank tellers" includes both feminists and non-feminists.

estimate of the long-run relative frequency of correct answers. Indeed, by assuming that people accurately encode and store frequency information from their environment, Gigerenzer et al.'s theory allowed them to predictably elicit well-calibrated performance, overestimation, or underestimation, depending on whether the questions asked were a random sample from the subjects' reference class, selected to be difficult, or selected to be easy.

This result fits well with the literature on automatic frequency encoding. One would expect an organism that relies on frequency information in making judgments under uncertainty to be constantly picking up such information from the environment in a way that does not interfere with the organism's ongoing activities. Hasher, Zacks, and colleagues have found evidence for just such a mechanism. People encode frequency information very accurately, and they appear to do so automatically. Their performance on frequency discrimination tasks is unaffected by the kinds of factors that affect more "effortful" processes, such as free recall. For example, frequency performance is not hindered by competing task demands, it is not affected by the amount or appropriateness of practice, and it is not affected by the accuracy of the subject's test expectations. Moreover, there are no stable individual differences in performance, and second-graders do just as well as adults: just what one would expect of a reliably developing, automatic mechanism (See Alba, Chromiak, Hasher, & Attig, 1980; Attig & Hasher, 1980; Hasher & Chromiak, 1977; Hasher & Zacks, 1979; Hintzman & Stern, 1978; Zacks, Hasher, & Sanft, 1982).

There is even preliminary evidence to suggest that asking for frequencies rather than single-event probabilities causes base rate neglect to disappear and subjects to act like good bayesians. In Kahneman and Tversky's (1973) "Tom W." problem, subjects were given a personality description of a graduate student, "Tom W.", and asked to judge which field of study he was most likely to be in – that is, the probability of a single event. Their subjects ignored base rates, and appeared to assign probabilities based on the similarity of Tom W.'s personality to their stereotypes of the various fields (performance which is consistent both with a Baconian and a "weight of evidence" interpretation of probability; see footnote 11). A frequentist version of an analogous problem was administered by McCauley and Stitt (1978) to subjects who were untutored in statistics. Instead of asking for the probability that an individual with certain personality traits was a member of a category, as in the Tom W. problem, they asked their subjects to estimate the frequencies of various personality traits and categories. For example, subjects were asked to estimate "the percent of Germans who are efficient" (p(trait|German)), "the percent of all the world's people who are efficient" (p(trait)), "the percent of efficient people who are German" (p(German|trait)), and "the percent of the world's people who are German" (p(German)). McCauley and Stitt, who were interested in stereotyping, wanted to see whether subjects' judgments of p(trait|German) followed a bayesian logic. They found that base rate neglect disappeared: the judged inverse probability – p(trait|German) – was correlated both with the base

rate (p(trait)) and with the likelihood (p(German|trait)), even though judged base rate and judged likelihood were not correlated with each other. This means that base rates were exercising an effect independent of likelihood. In a stronger test, McCauley and Stitt used Bayes' theorem to calculate p(trait|German) from their subjects' estimates of p(trait), p(German|trait) and p(German). These calculated values were highly correlated with the subjects' directly judged estimates of p(trait|German) ($r = .91$). It is difficult to see how this striking internal consistency among estimates could be achieved unless their subjects were somehow applying a bayesian logic of statistical prediction.

The mounting evidence that people are good "intuitive statisticians" when they are given frequencies as input and asked for frequencies as output, suggests that the issue of whether our inductive reasoning mechanisms embody a calculus of probability should be reopened. To this end, we conducted a particularly strong test of Gigerenzer's hypothesis that our inductive reasoning mechanisms were designed to operate on and to output frequency representations (henceforth called the "frequentist hypothesis"). We conducted a series of experiments to see whether casting a single-event probability problem in frequentist terms would elicit bayesian reasoning. If it does, then the conclusion that our inductive reasoning mechanisms do not embody a calculus of probability – that they consist of nothing more than a few quick-and-dirty rules-of-thumb – will have to be re-examined.

1.5. *The medical diagnosis 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 that you know nothing about the person's symptoms or signs? ____%

The above reasoning problem is called the medical diagnosis problem, and it was designed to assess whether people engage in bayesian reasoning. It is famous in the literature on judgment under uncertainty for eliciting base rate neglect even from technically educated subjects. Casscells et al. (1978) asked a group of faculty, staff and fourth-year students at Harvard Medical School to solve this problem. Only 18% of them answered "2%", which is the correct bayesian answer under most interpretations of the problem.[12] Forty-five percent of them answered "95%". Because "95%" is inconsistent with a population base rate for the disease of 1 in 1000, Casscells et al. concluded that their subjects were violating Bayes' theorem

[12] "2%" is the correct answer only if one assumes that the true positive rate is 100% (this information was not provided in the original problem), that the population base rate is the appropriate prior probability, and that the individual tested was randomly drawn from the population. If the subject believes that any of these assumptions are false, then, according to Bayes' theorem, other answers would be correct. See Experiments 5 and 6 below.

by ignoring the base rate. The usual explanation for base rate neglect is the operation of a representativeness heuristic, but this cannot account for base rate neglect in the medical diagnosis problem (Tversky & Kahneman, 1982, p. 154). Accordingly, Tversky and Kahneman use the results of Casscells et al.'s study to make the point that judgmental biases are widespread and difficult to eradicate:

> Evidently, even highly educated respondents often fail to appreciate the significance of outcome base rate in relatively simple formal problems . . . The strictures of Meehl and Rosen (1955) regarding the failure to appreciate base rates are not limited to clinical psychologists; they apply to physicians and other people as well. (Tversky & Kahneman, 1982, p. 154)

Physicians are taught statistics so that they will know how to evaluate diagnostic test results of the kind presented in the medical diagnosis problem. If even they fail to use a calculus of probability, then it seems compelling to argue that the human mind does not embody one.

We wanted to stand this experimental logic on its head. We chose the medical diagnosis problem for our experiments precisely because it had elicited such low levels of correct bayesian reasoning *even* from statistically educated subjects. We wanted to see what would happen if the same problem were posed in frequentist terms – that is, if the problem information was presented as frequencies and the answer was asked for as a frequency. Could a frequentist version of the medical diagnosis problem elicit correct bayesian reasoning "even" from undergraduates, most of whom have had little or no training in statistics? That would be strong evidence for the hypothesis that we do have mechanisms that embody some aspects of a calculus of probability, but that frequency representations are their natural format.

The remainder of this article is divided into three parts. In Part I we show that very high levels of bayesian reasoning are elicited by frequentist versions of the medical diagnosis problem. In Part II, we show that simply clarifying non-frequentist versions of the problem does *not* produce these high levels of bayesian reasoning. In Part III, we successively eliminate various elements of the frequentist problem to determine which are critical for producing high levels of bayesian reasoning, and show that the crucial elements are (1) asking for the answer as a frequency rather than as a single-event probability, and (2) presenting the problem information as frequencies.

PART I. CAN FREQUENTIST VERSIONS OF THE MEDICAL DIAGNOSIS PROBLEM ELICIT CORRECT BAYESIAN REASONING?

In Part I, we describe four experiments that were designed to test whether frequentist versions of the medical diagnosis problem could elicit correct

bayesian reasoning from subjects. The frequentist problems in this section have all the characteristics that a good frequentist problem should have: (1) they present the problem information as frequencies; (2) they ask for the answer as a frequency rather than as a single-event probability; (3) they specify the true positive rate; (4) they define the notion of a false positive rate rather than assuming that subjects already know what this term means; (5) they make the random sampling assumption explicit; and (6) they specify the size of the random sample, thus giving the subject a concrete reference class to think in terms of. If we have well-designed mechanisms for bayesian reasoning that operate on and produce frequency representations – and can do this using verbal input – then these problems should elicit substantial levels of bayesian performance.

The subjects and procedure were identical for all of the experiments. There were 25 subjects in each condition, all of them students at Stanford University. Their average age was 19.6 years, and they were paid volunteers recruited by advertisement. Each subject was given a booklet that consisted of an instruction page followed by one medical diagnosis problem. The instructions were minimal: they merely asked the subject to read the problem carefully before answering any questions. Although most subjects finished in less than 10 min, they were allowed to take all the time they needed.

EXPERIMENT 1

The purpose of Experiment 1 was (1) to see if we could replicate Casscells et al.'s results on the original version of the medical diagnosis problem, and (2) to see whether we could create a frequentist version of the problem that would elicit a higher percentage of bayesian responses than that elicited by the original version.

2. Materials

Experiment 1 had three conditions.

Condition 1

Condition 1 was an exact replication of the problem that Casscells et al. administered to physicians and fourth-year medical students at Harvard Medical School teaching hospitals:

> 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 that you known nothing about the person's symptoms or signs? ____%

Condition 2

The problem tested in Condition 2 was as follows:

1 out of every 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 (i.e., the "true positive" rate is 100%). 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 (i.e., the "false positive" rate is 5%).

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.

Given the information above:

on average,

How many people who test positive for the disease will *actually* have the disease? ____ out of ____

Condition 2 differs from Condition 1 – Casscells et al.'s original version – in several respects: (1) it gives the true positive rate, which Casscells et al. had omitted entirely; (2) the base rate and false positive rate are given as frequencies, as well as percentages; (3) the text defines what the term "false positive" means; (4) it specifies that the question refers to a random sample (it is inappropriate to use base rates as a prior probability if a sample was not randomly drawn); (5) it specifies the size of the random sample ("1000 Americans"), giving a concrete reference class to think in terms of; and (6) it asks for the answer as a frequency ("How many people who . . ."), rather than as a single-event probability ("What is the chance that *a* person . . .").

Condition 3

The first and second paragraphs of Condition 3 were identical to those for Condition 2; the remainder of the problem read as follows:

Given the information above:

on average,

(1) How many of these 1000 people will have the disease? ____

(2) How many of the 1000 people will have the disease AND test positive for it? ____

(3) How many of the 1000 people will be healthy AND test positive for the disease? ____

(4) How many of the 1000 people will test positive for the disease, whether they have the disease or not? ____

(5) How many people who test positive for the disease will *actually* have the disease? ____ out of ____

Questions 1–4 were asked so that we could ascertain whether subjects understood the information given in the problem. In addition, we wanted to see whether answering these questions would boost performance. Answering these probe questions correctly would make all the information necessary for solving the problem explicit in the subject's mind. If the human mind is equipped to do bayesian reasoning, but subjects sometimes fail to correctly extract the necessary information from the problem, then this problem should boost performance. By the same token, a small number of "2%" responses on this problem would be strong evidence that the mind is not naturally equipped to do bayesian reasoning.

3. Results

The results are pictured in Figs. 1 and 2. Fig. 1 shows the results for all subjects; Fig. 2 shows the frequency with which subjects' judgments fell into the three categories that were both most common and of greatest theoretical interest: "2%", the correct bayesian answer on most interpretations of the problem; "1/1000", which reflects base rate conservatism (here, taking into account nothing *but* the base rate); and "95%", which was the modal response in the Casscells et al. study and has been taken to reflect base rate neglect. (Eighty-four percent of subjects' responses fell into one of these three categories.) Inspection of Figs. 1 and 2 shows that the pattern of results obtained for the original, non-frequentist version of the problem in Condition 1 literally reversed itself in the two frequentist versions tested in Conditions 2 and 3. The two frequentist versions elicited a preponderance of correct bayesian answers, and caused base rate neglect to vanish. More specifically:

3.1. Did Casscells et al.'s original study replicate on Stanford students?

Condition 1 replicated the results of Casscells et al. very nicely: 12% of Stanford students, as compared to 18% of Casscells et al.'s medical school students and staff, gave the bayesian answer of "2%". In addition, "95%" was the modal answer for Stanford students, just as it was for Casscells et al.'s subjects: 56% of Stanford students versus 45% of Casscells et al.'s subjects gave this response. Another 12% of our subjects were base rate "conservatives": rather than neglecting the base rate, these subjects paid attention to *nothing but* the base rate, answering "0.1%" (i.e., 1/1000). (We have no figures for comparison from Casscells et al., who reported

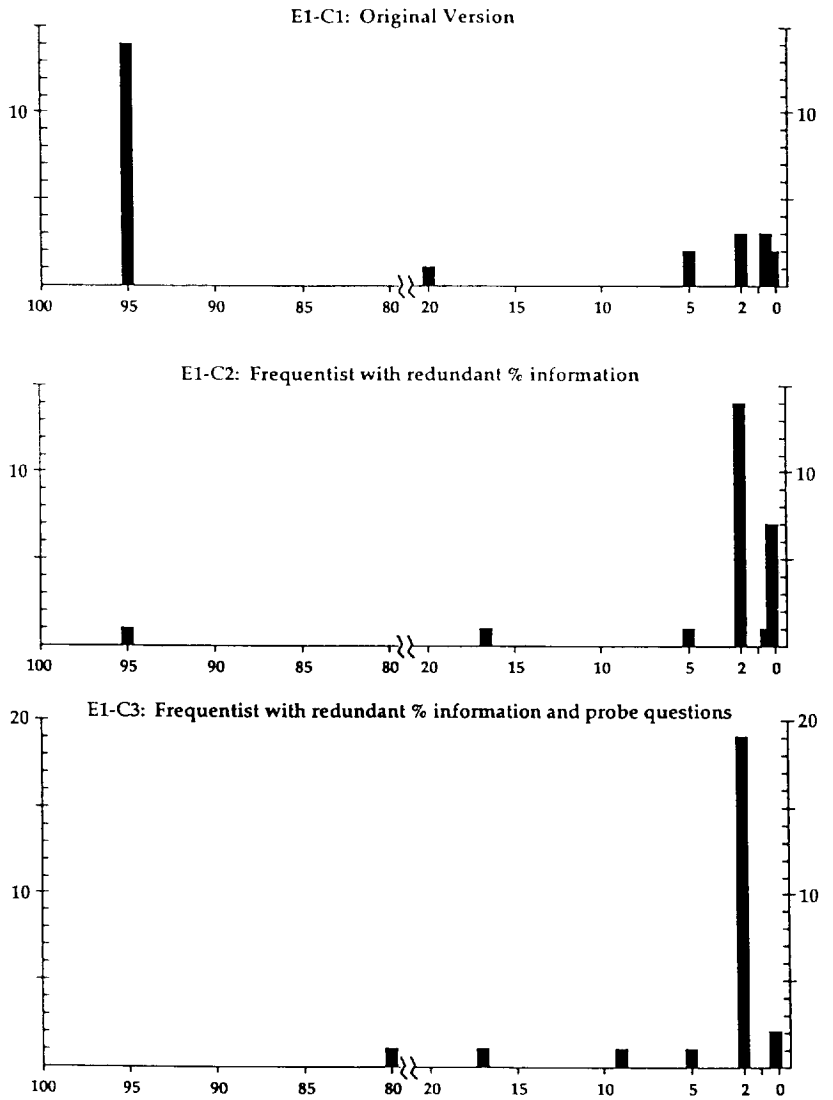

Fig. 1. Results of Experiment 1. The y-axis represents how many subjects gave a particular answer; the x-axis represents which answer, from "0%" to "100%", was given.

percents only for those subjects who responded "2%" or "95%", which accounted for only 63% of the subjects they tested.)

3.2. *Did the frequentist versions of the problem boost performance?*

There is a striking difference between the pattern of results obtained for Condition 1 and that obtained for Conditions 2 and 3: the pattern of results for the non-frequentist version of the problem *reverses itself* for the two

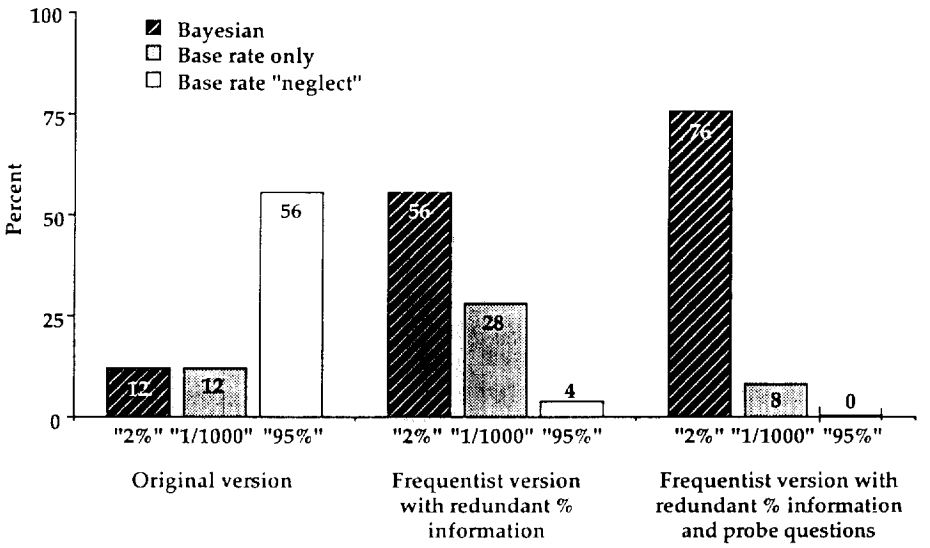

Fig. 2. Percentage of subjects who answered either "2%", "1/1000", or "95%" in Experiment 1; "2%" is the correct bayesian answer under most interpretations of the problem. The distribution of responses for the two frequentist problems is almost the reverse of the distribution for the original version, which is non-frequentist.

frequentist versions. The modal response for the two frequentist versions of the problem was "2%" – the correct bayesian answer. The correct bayesian answer was elicited from 56% of subjects in condition 2, and 76% in condition 3. This is, of course, a significantly higher percentage of bayesian answers than the 12% elicited in the non-frequentist condition 1 (56% vs. 12%; $Z = 3.28$, phi = .46, $p = .0005$).

The second most common response for the frequentist versions was "0.1%", which was given by 28% of subjects in condition 2 and 8% in Condition 3. Far from neglecting the base rate, these subjects were weighting it too heavily – they were base rate conservatives.

Base rate neglect virtually disappeared in the frequentist versions: only one subject in Condition 2, and no subjects in Condition 3 answered "95%", which was the modal response for the non-frequentist version tested both in Condition 1 and by Casscells et al. Indeed, only one subject in each of the frequentist conditions gave an answer greater than "20%".

The extent to which base rate neglect vanished in the frequentist versions of the problem can be seen by considering how many subjects gave answers of "2%" *or lower* in the three conditions. Rather than neglecting base rates, such subjects are either giving the correct bayesian answer or weighting base rates too heavily. In the non-frequentist Condition 1 this figure was only 32%, but in the frequentist Conditions 2 and 3 it leapt to 88% and 84%, respectively.

The two frequentist versions of the problem, Conditions 2 and 3, differed

in that Condition 3 asked a series of questions about the information presented in the problem. Asking these questions increased the number of bayesian answers by 20 percentage points (76% vs. 56%: $Z = 1.49$, phi = .21, $p = .07$), and decreased the number of "1/1000" answers by 20 percentage points (28% vs. 8%: $Z = 1.84$, phi = .26, $p = .03$). Thus, asking these questions seemed to clarify the problem for subjects. In Condition 3, all of the 21 subjects who answered either "2%" or "1/1000" answered the probe questions in ways indicating that they understood the information presented in the problem. For two of the four subjects who gave a different answer, the probe questions and the answer given indicated correct bayesian reasoning on the assumption that the prevalence of the disease in the population was other than 1/1000 (one subject answered as if it were 1/100, the other as if it were 5/1000).

The fact that all of the base rate conservatives in Condition 3 seemed to understand the information presented in the problem suggests to us that they may have misread the last question as a conjunction rather than as a conditional probability. If one were not careful, it would be easy to mistake "How many people who test positive for the disease will actually have the disease?" (a conditional probability) for "How many people test positive for the disease and actually have the disease?" (a conjunction). One out of 1000 is, of course, the correct answer to the latter question.

4. Discussion for Experiment 1

The results of Experiment 1 show that a simple rewording of the medical diagnosis problem can elicit correct bayesian reasoning from the majority of subjects and eliminate base rate neglect. But what accounts for this dramatic difference in performance? The frequentist versions of the problem in Conditions 2 and 3 differ from the original version tested in Condition 1 in six ways, only three of which are relevant to the frequentist hypothesis. Later, in Parts II and III, we will independently manipulate these variables to see which ones account for this effect.

In the remainder of this article we will want, in certain cases, to compare performance across experiments. We will therefore refer to results by their experiment and condition numbers. Thus, E1–C3 will refer to Experiment 1, Condition 3; E2–C2 will refer to Experiment 2, Condition 2, and so on. For easy reference, Table 2 provides a complete list of all the experiments and conditions reported in this article.

EXPERIMENTS 2–4

Experiments 2–4 were designed first to see whether the startling results of Experiment 1 would replicate, and second to see how high we could push

Table 2
List of experiments

Part I	
Experiment 1	
Condition 1 (E1–C1)	Non-frequentist. Original version of Casscells et al. problem
Condition 2 (E1–C2)	Frequentist version with redundant percent information
Condition 3 (E1–C3)	Frequentist version with redundant percent information and probe questions
Experiment 2	
Condition 1 (E2–C1)	Frequentist problem
Condition 2 (E2–C2)	Frequentist problem with probe questions
Experiment 3	
Condition 1 (E3–C1)	Frequentist version with redundant percent information (replication of E1–C2)
Condition 2 (E3–C2)	Frequentist problem (replication of E2–C1)
Experiment 4	
Condition 1 (E4–C1)	Active pictorial (frequentist)
Condition 2 (E4–C2)	Passive pictorial (frequentist)
Part II	
Experiment 5 (E5)	Non-frequentist. Clarified version of original Casscells et al. problem. True positive rate specified; meaning of false positive rate clarified
Experiment 6	
Condition 1 (E6–C1)	Non-frequentist. Like E5, but with random sampling assumption made explicit (population size not explicitly enumerated)
Condition 2 (E6–C2)	Non-frequentist. Like original Casscells et al. version tested in E1–C1, but with a 40%, rather than a 5%, false positive rate
Part III	
Experiment 7	
Condition 1 (E7–C1)	Like E5, but answer asked for as a frequency (percent information; random sampling not mentioned)
Condition 2 (E7–C2)	Like E2–C1, but answer asked for as a single-event probability (frequency information; random sampling explicit; explicitly enumerated population)
Condition 3 (E7–C3)	Like E6–C1, but answer asked for as a frequency (percent information; random sampling explicit; population size not explicitly enumerated)
Experiment 8	
Condition 1 (E8–C1)	Like E7–C3, but with an explicitly enumerated population (percent information; random sampling explicit; frequency answer)
Condition 2 (E8–C2)	Like E2–C1, but without an explicitly enumerated population

bayesian performance. Is 56% correct as high as performance can get without asking the subject "leading questions"? Does 76% correct represent a ceiling for a problem that does ask leading questions – that is, ones that

make the information necessary to solve the problem explicit in the subject's mind?

EXPERIMENT 2

Experiment 2 allowed us to address two questions: (1) was the high level of bayesian performance elicited by the frequentist problems in Experiment 1 a fluke? and (2) can correct bayesian performance be pushed even higher by presenting information *only* as frequencies?

In the frequentist problems of Experiment 1, the true positive rate and the false positive rate were presented in two ways: as frequencies *and* as percents. Although percents are, technically, frequencies normalized on a population of 100, this information is implicit in the definition, not explicit in the notation. It is easy to forget that percents are implicit frequencies for two reasons: (1) by using school-taught algorithms that allow one to plug percents directly into formulas, one can solve problems by symbol manipulation in a way that bypasses the formation of a frequentist representation of the problem; and (2) in natural language, percents are sometimes used to express degrees of effort or confidence, rather than frequencies (e.g., "The football team was playing at only 70%"; "I'm 90% sure the party is today"). (See Part III below for more discussion of this point.)

If there are inductive reasoning mechanisms that are designed to process frequency information, then the more explicitly frequentist the representation, the better performance should be. It occurred to us that presenting the information in a percent format might actually confuse some subjects and *impair* performance. We tested this hypothesis in Experiment 2. If such redundant information impairs performance, this should be most evident in problems that do *not* ask "leading" probe questions. This is because the clarity achieved by asking probe questions may be sufficient to override a negative effect of percent format information.

5. Materials

Experiment 2 had two conditions.

Condition 1 (E2–C1)

Condition 1 was virtually identical to the frequentist problem tested in Experiment 1, Condition 2 (E1–C2). The only difference between these two problems was that the parenthetical expression that gave the redundant percentage information in Experiment 1 – "(i.e., the 'true positive' rate is 100%)" and "(i.e., the 'false positive' rate is 5%)" – were deleted from the text of the problem tested in Experiment 2. No probe questions were asked in this condition.

Condition 2 (E2–C2)

Condition 2 was virtually identical to the frequentist problem tested in Experiment 1, Condition 3 (E1–C3) – the problem that asked the four probe questions. Again, the only difference between these two problems was that the parenthetical expressions that gave the redundant percent information in Experiment 1 were deleted from the text of the problem tested in Experiment 2.

6. Results of Experiment 2

6.1. Did the high level of bayesian performance found in Experiment 1 replicate in Experiment 2?

Yes. The correct bayesian response of "2%" was elicited from 72% of subjects in both Condition 1 and Condition 2 of Experiment 2. This high level of bayesian performance was elicited by Condition 1 in spite of the fact that it asked no "leading" probe questions.

6.2. Does redundant percentage information depress bayesian performance?

This question can be answered by comparing performance on E2–C1 with performance on E1–C2. Neither of these problems ask probe questions. The only difference between them is that E2–C1 presents the information only as frequencies, whereas E1–C2 presents it as both a frequency and a percent. The level of bayesian performance was 16 percentage points higher when the information was presented only as a frequency – 72% correct for E2–C1 versus 56% correct for E1–C2 ($Z = 1.18$, phi = .17, $p = .12$).

Our sample size is not large enough for a 16 percentage point difference to show up as significant at the conventional .05 level. Hence, to see if this effect is real, we attempted to replicate it in Experiment 3.

6.3. Does redundant percentage information depress bayesian performance when probe questions are asked?

To answer this question we need to compare the results of E2–C2 to those of E1–C3. Both of these problems ask the four probe questions, but the former presents the information only as frequencies, whereas the latter presents it redundantly, as both a frequency and a percent. E2–C2 elicited the correct bayesian answer, "2%", from 72% of subjects tested. This is virtually identical to the 76% correct found for the matching problem tested in E1–C3. This suggests that the clarity achieved by having subjects answer

probe question is sufficient to override any negative effect of presenting information in a percent format.

EXPERIMENT 3

We were so intrigued by the fact that bayesian performance was *better* when the information was presented only as a frequency that we wanted to see if this effect would replicate.

7. Materials

Experiment 3 had two conditions.

Condition 1 (E3–C1)

Condition 1 was identical to the frequentist problem tested in E1–C2. In other words, the false positive rate and true positive rate were each presented redundantly, both as a frequency and as a percent. No probe questions were asked.

Condition 2 (E3–C2)

Condition 2 was identical to the frequentist problem tested in E2–C1. In other words, the false positive rate and true positive rate were each presented only as a frequency. No probe questions were asked.

8. Results

Eighty percent of subjects in the frequency only condition and 64% of subjects in the frequency and percent condition gave the correct bayesian answer. Thus, the 16 percentage point difference found previously between these conditions (E2–C1 vs. E1–C2) replicated exactly in Experiment 3 ($Z = 1.25$, phi = .18, $p = .1$). This supports our earlier hypothesis that the 16 percentage point difference is real, but too small an effect to show up as significant in a between group design with $n = 25$ per group. Indeed, when we increase our sample size to $n = 50$ per group by combining the results of the two frequency only conditions (E2–C1 and E3–C1) on the one hand, and the two frequency and percent conditions (E1–C2 and E3–C2) on the other, one can see that providing the redundant percentage information

does significantly decrease performance (frequency only: 76%; frequency and percent: 60%; $Z = 1.71$, phi = .17, $p = .04$).

9. Discussion for Experiments 2 and 3

First, Experiments 2 and 3 provide four new independent replications of the main effect found in Experiment 1, namely, that wording the medical diagnosis problem in frequentist terms elicits correct bayesian performance from the majority of subjects tested. Second, by presenting the problem information as frequencies only, we were able to push bayesian performance up to 72% and 80%, *even in the absence of any "leading" probe questions.* Indeed, presenting the information only as frequencies elicited bayesian performance that was just as high as that elicited by problems that do ask probe questions (76% and 72%, in E1–C3 and E2–C2, respectively). This suggests that purely frequentist representations make the problem so clear that probe questions are superfluous. Third, when it is presented in the form of a percent, redundant problem information actually seems to impair performance. (We will see additional evidence of the negative effect of presenting problem information as a percent in Part III.) Fourth, the presence of probe questions that make the subject explicitly represent the information necessary to solve the problem seems to be sufficient to override the negative effect of redundant percent information.

Thus, even in the absence of any probe questions, 76% of subjects gave the correct bayesian answer when presented with a purely frequentist version of the medical diagnosis problem (E2–C1, E3–C2, $n = 50$). This is dramatically better performance than the 12% elicited by the original Casscells et al. problem tested in Experiment 1; the effect size, phi, for this comparison is 0.61 (76% vs. 12%: $Z = 5.25$, $p = .0000001$).

EXPERIMENT 4

Can bayesian performance be pushed even higher than the average of 76% correct found for the pure frequentist problems tested in Experiments 2 and 3, or does 76% represent some sort of ceiling? The assumption we started out with was that there would be inductive reasoning mechanisms that represent information as frequencies because in our natural environment that is what we would have been encountering: a series of real, discrete, countable events. If true, then the highest levels of bayesian performance should be elicited when subjects are *required* to represent the information in the problem as numbers of discrete, countable individuals. This is what we tried to do in Experiment 4.

10. Materials

Experiment 4 had two conditions.

Condition 1 (E4–C1)

Condition 1 was our "active" pictorial condition; in this condition we forced our subjects to actively construct a concrete, visual frequentist representation of the information in the problem. The text of condition 1 read as follows:

> 1 out of every 100 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 100 people who are perfectly healthy, 5 of them test positive for the disease.
>
> Imagine that we have assembled a random sample of 100 Americans. They were selected by a lottery. Those who conducted the lottery had no information about the health status of any of these people. The 100 squares pictured below represent this random sample of 100 Americans. Each square represents one person.
>
> Using these squares, we would like you to depict the information given above. To indicate that a person actually has the disease, *circle* the square representing that person. To indicate that a person has tested positive for the disease, *fill in* the square representing that person.
>
> Given the information above:
>
> on average,
>
> (1) *Circle* the number of people who will have the disease.
> (2) *Fill in* squares to represent the people who will test positive for the disease.
> (3) How many people who test positive for the disease will *actually* have the disease? ____ out of ____

Following the text of the problem on the same page were line drawings of 100 squares, arranged as 10 rows and 10 columns. (For this problem, we used a base rate of 1/100 rather than 1/1000, simply because we could not fit 1000 squares on one sheet of paper.) Thus the correct bayesian response for this problem is either 1 out of 6 or 1 out of 5, depending on whether the subject estimates 5% of 99 – which is 4.95 – to be 5 or 4. Either is defensible – 5 if one is rounding by standard methods, and 4 if one decides there is no such thing as .95 of a person, or rounds by truncation of the

decimal places, or loosely estimates without computing, noting that 5% of 99 is less than 5 but not by much.

It should now be clear why we call this an "active" pictorial condition: the subject is forced to actively construct a concrete, visual frequentist representation of the information in the problem. To correctly solve the problem, the subject should circle one person to represent the one out of 100 who has the disease. The square representing this person should then be filled in, because the problem states that everyone who has the disease tests positive for it. Then, either four or five of the other squares should be filled in, to represent the 5% of healthy people who test positive. Once one has done this, solving the problem is trivial, and requires no formalisms whatsoever: to answer the question "How many people who test positive for the disease will actually have the disease?" one simply counts up the number of squares that are circled and filled in – 1 – and the number of squares that are filled in – either 5 + 1 or 4 + 1. The answer, then, is either 1 out of 6 or 1 out of 5.

Condition 2 (E4–C2)

Condition 2 was our "passive" pictorial condition. Here, we did not require that the subject actively construct a frequentist representation. But we did represent the information pictorially as well as verbally. There was nothing, however, to prevent the subject from ignoring the pictorial information. The text of the passive pictorial condition read as follows:

> 1 out of every 100 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 100 people who are perfectly healthy, 5 of them test positive for the disease.
>
> We have represented this information below pictorially. The 100 squares represent a random sample of 100 Americans. Each square represents one person. If a person actually has the disease, we *circled* the square representing that person. If a person tested positive for the disease, we *filled in* the square representing that person.
>
> Given this information:
>
> on average,
>
> How many people who test positive for the disease will *actually* have the disease? ____ out of ____

An array of 100 line drawings of squares like the one described for the active pictorial condition followed the text. One of the squares was circled and filled in; five of the other squares were filled in. These circled and filled in squares were scattered throughout the array.

11. Results

The "active" pictorial condition – the condition in which we forced our subjects to actively construct a frequentist representation of the information in the problem – elicited the correct bayesian response from 92% of subjects tested.[13] And although the remaining two subjects answered "1 out of 100", they had filled in and circled the boxes correctly. This suggests that they were reasoning correctly but misread the final question as a conjunction rather than as a conditional probability – an easy error to make, as discussed above. We suspect that this condition represents a ceiling on bayesian performance!

The passive pictorial condition elicited the correct bayesian answer from 76% of subjects tested. This is the same level of performance elicited by the two frequentist problems that had no probe questions (E2–C1 and E3–C2 averaged to 76%). Thus there was a 16 percentage point difference between the active and passive pictorial problems, as well as between the active pictorial problem and the two other frequentist problems. Naturally, a real 16 point difference is too small to count as significant given a comparison of two groups of size $n = 25$ (although the effect size, phi, for this comparison is .22). But when we increase the power of the test by comparing the results of the active pictorial problem (92%, $n = 25$) to those of the passive one plus the other two comparable frequentist problems (57/75 = 76%, $n = 75$), the difference is significant ($Z = 1.73$, phi = .17, $p = .04$).

The data on rounding from the active pictorial condition, the "frequency only" probe question condition (E2–C2), and the "frequency and percent" probe question condition (E1–C3) provide indirect evidence that there are inductive reasoning mechanisms that represent probabilities as frequencies of discrete, countable entities. For these three problems, subjects must state what they consider 5% of 99 or of 999 to be. In computing this figure, some subjects rounded by truncation whereas others simply rounded up. We suggested earlier that rounding by truncation is reasonable if one believes that it is not sensible to talk about .95 of a person. If frequentist representations encourage one to represent the problem in terms of discrete

[13] As is the usual practice in the literature, this figure includes two subjects who seem to have forgotten that the person with the disease will test positive for it; that is, for these two subjects, the "5" in "1 out of 5" reflects only people who are *healthy* and test positive. We doubt that these two subjects actually believe that diseased people do not test positive, first, because it would be exceedingly strange to believe that a test designed to detect the presence of a disease would never yield a positive result for a person who has that disease, and second, because no subjects in the other two conditions that included probe questions (E1–C3, E2–C2) held this belief. If one wanted to adopt a stricter scoring procedure, however, the total would be 84% rather than 92%. We prefer the "final answer only" scoring criterion because (1) that is the standard dependent measure in the heuristics and biases literature (i.e., points are neither given nor subtracted depending on the subject's reasoning process – which the experimenter rarely knows), and (2) it allows us to directly compare the results of our probe question conditions to those conditions that lack probe questions.

individuals, then we might expect a more pronounced tendency to round by truncation than for percent representations, which may encourage one to think in terms of continuous distributions.[14] This did, in fact, seem to happen: in E1–C3, the "frequency and percentage" probe question condition, only one subject rounded by truncation, as opposed to nine subjects in each of the "frequency only" probe question conditions (36% vs. 4% ($n = 25$): $Z = 2.83$, phi = .40, $p = .0023$). We have already seen that presenting problem information as percents can depress bayesian performance (more evidence on this point will be presented in Part III). This, together with the data on rounding by truncation, supports not only the frequentist hypothesis, but also a related claim by Johnson-Laird (1983): that logical problems are easier to solve using representations of discrete, countable individuals than using representations that map finite sets of individuals into infinite sets of points, such as Venn diagrams or Euler circles (see General Discussion, below).

12. Discussion for Experiment 4

The passive pictorial condition (E4–C2) elicited the correct bayesian response from 76% of subjects tested, which is the same level of performance as that found for frequentist versions of the problem that do not include a pictorial depiction of the problem information. We assume that this is because the mere presence of the pictorial information does not mean that subjects will use it. But when subjects are *required* to represent the information in the problem as numbers of discrete, countable individuals – that is, when they are required to construct a frequentist representation, as they were in the active pictorial condition – 92% of them gave the correct bayesian answer.

SUMMARY OF PART I

The original, non-frequentist version of the medical diagnosis problem, which presents the problem information as percents and asks for the answer as a single-event probability, elicited bayesian performance from only 12% of subjects. But by simply translating this problem into frequentist terms, we were able to elicit correct bayesian reasoning from 76% of our subjects. By requiring them to create a concrete, visual frequentist representation of the problem, we were able to push their performance to 92% correct. Fig. 3 summarizes these results.

[14] For the reasons discussed above. Because a percent can be plugged directly into a formula, one can bypass the formation of a frequentist representation and treat it as a continuous variable, creating concepts such as "4.95 persons".

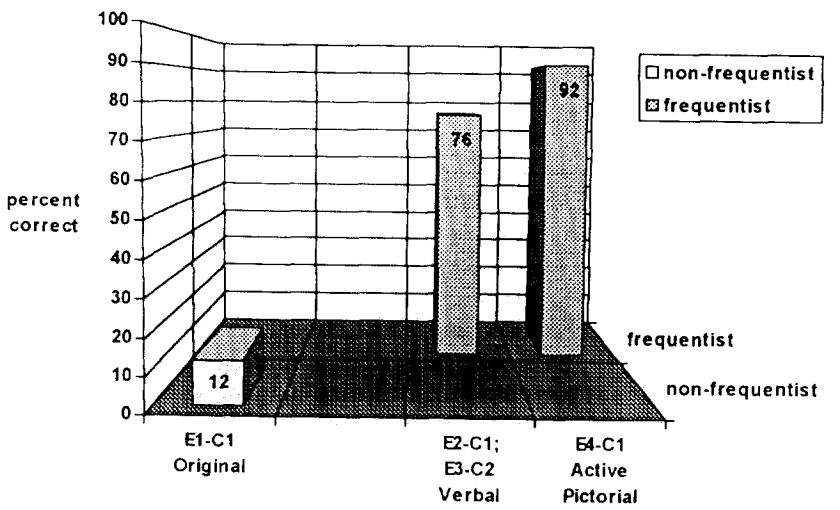

Fig. 3. Summary of the main results of Part I, comparing the percent of correct bayesian answers for the original, non-frequentist version to the percent correct for the two frequentist versions. In the active pictorial condition, which elicited the highest levels of bayesian performance, subjects were required to form a frequentist representation.

PART II. CAN NON-FREQUENTIST VERSIONS OF THE MEDICAL DIAGNOSIS PROBLEM ELICIT HIGH LEVELS OF BAYESIAN REASONING?

Because strong claims have been made that people are not good at bayesian reasoning, the results of Part I would be interesting even if they turned out to be unrelated to the frequentist hypothesis. But because we are interest in *how* probabilistic information is represented and processed by the human mind, we would like to go one step further and ask what it is about these problems that elicits such high levels of bayesian performance.

If the mind contains inductive reasoning mechanisms that can apply a calculus of probability when it represents probabilities as frequencies, then the high levels of bayesian performance elicited in Part I should be due to frequentist aspects of the problem, such as asking for the answer as a frequency and presenting the problem information as a frequency. If one can create non-frequentist versions of the problem that elicit high levels of bayesian performance, then we would not be justified in concluding from the previous experiments that frequentist representations elicit bayesian reasoning.

In Part II we ask the following questions: (1) Can simply clarifying the original Casscells et al. problem elicit high levels of bayesian performance? (2) Are subjects who were given the original, non-frequentist Casscells et al. problem actually good Bayesians who simply believed that the random

sampling assumption has been violated? and (3) If a violation of random sampling does not account for their distribution of responses, what does?

EXPERIMENT 5

Earlier, we noted that the original Casscells et al. problem is somewhat ambiguous because (1) it does not specify the true positive rate, and (2) subjects might not know what the term "false positive rate" means. Did the frequentist problems in Part I elicit high levels of bayesian performance because they were frequentist, or simply because they clarified these ambiguities in the original, non-frequentist version of the medical diagnosis problem? To address this question, we tested subjects on a "cleaned up" version of the Casscells et al. problem: one that specifies the true positive rate and that clearly defines the notion of a false positive rate.

13. Materials

The text of the problem tested in Experiment 5 was as follows:

> The prevalence of disease X is 1/1000. 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, 5% of all people who are perfectly healthy test positive for the disease.
>
> What is the chance that a person found to have a positive result actually has the disease, assuming that you know nothing about the person's symptoms or signs? ____%

Note that this problem specifies the true positive rate and explains the false positive concept. (Like the frequentist problems tested in Experiments 2–4, the terms "true positive rate" and "false positive rate" were not used.) We tried to stay as close as possible to the wording of the original Casscells et al. problem while still providing this information.

14. Results

Thirty-six percent of subjects in Experiment 5 gave the correct bayesian answer, "2%". This figure is significantly higher than the 12% elicited by the original, ambiguous version tested in E1–C1 (36% vs. 12%: $Z = 1.99$, phi = .28, $p = .023$). Nevertheless, this level of bayesian performance is nowhere near the average of 76% correct elicited by the two comparable frequentist versions tested in Part I (E3–C2 vs. E5: 80% vs. 36%: $Z = 3.15$,

phi = .45, $p = .0008$; E2–C1 vs. E5: 72% vs. 36%: $Z = 2.55$, phi = .36, $p = .0054$).

One can get some insight into how these subjects were solving the problem by looking at their calculations. Although they were not asked to show their work, some subjects wrote their calculations on the test sheet. This was true of seven of the nine subjects who had produced the correct answer. All seven had analyzed the problem in frequentist terms, that is, by assuming a population of fixed size and figuring out how many individuals would fall into each category. For example, one subject had written, "51 positive/1000 total, 1 real/51 pos"; another, "1000 people; 1 tests positive is positive; 999 healthy, 5% test positive; ~50 test positive are negative/~51 positive; 1/51 have disease." This was also true of the original version of the problem tested in E1–C1: two out of the three subjects who gave the correct bayesian answer to that problem wrote out a frequentist analysis. (Naturally, many subjects wrote out this kind of calculation for the frequentist versions of the problem.) In contrast, the only subject in Experiment 5 whose calculations indicated that he had tried to use Bayes' formula came up with "99.905%" rather than "2%"!

This suggests that many people who correctly solve non-frequentist versions of the medical diagnosis problem do so by translating it into frequentist terms. Indeed, formulas are of no use unless one can correctly map the concepts in the problem onto the terms of the formula, which may be particularly difficult for probability problems that are not expressed in frequentist terms.

Why was "95%" the modal response for the original Casscells et al. problem? It could be because subjects were taking "false positive rate" to refer to $p(\text{healthy}|\text{positive})$ – what a patient who tests positive wants to know – rather than $p(\text{positive}|\text{healthy})$ – what a scientist who constructs diagnostic tests wants to know. If so, then clarifying the meaning of this term in Experiment 5 should reduce the number of subjects who made this response. That is just what happened. Whereas 56% of subjects answered "95%" on the original version (E1–C1), only 32% of subjects did on the clarified version tested in Experiment 5 (56% vs. 32%: $Z = 1.71$, phi = .24, $p = .044$). Actually, 28% is a more appropriate figure than 32%, because one subject answered "95.2%", the precision of which indicates that he was using the Bayesian principle of indifference, rather than interpreting "false positive rate" as an inverse probability. As we discuss in Experiment 6, the Bayesian principle of indifference can be appropriate to use when one does not know whether a sample was randomly drawn.

EXPERIMENT 6

To apply Bayes' theorem – or any other statistical theory – correctly, one must first make sure that the assumptions of the problem match the

assumptions of the theory. For example, it is inappropriate to use a base rate as one's prior probability if one does not believe that the sample one is reasoning about was randomly drawn.

How does one decide whether the assumptions of a problem match those of a statistical theory? Because factors such as random sampling and independence of events differ from domain to domain, Gigerenzer and his colleagues have argued that one should draw on one's knowledge of the problem domain (Gigerenzer & Murray, 1987). This appears to be just what subjects do. In an elegant series of experiments, Gigerenzer, Hell, and Blank (1988) showed that subjects take base rates into account only when they are convinced that the random sampling assumption has been met. When their real-world knowledge of a domain confirms the random sampling assumption, subjects will take base rates into account. But when their real-world knowledge of a domain contradicts the random sampling assumption, they will take base rates into account only if the experimenter can convince them that the assumption has, in fact, been met – by, for example, asking subjects to reach into urns and draw the sample themselves (Gigerenzer et al., 1988). These results indicate that, in the past, subjects have sometimes been categorized as performing erroneously when they were intuitively (and presumably nonconsciously) more sophisticated than the formal framework that experimenters were using to judge them.

The physicians and fourth-year medical students tested by Casscells et al. had a great deal of real-world knowledge about the conditions under which patients are given diagnostic tests – most of which would have contradicted the random sampling assumption. Under normal circumstances, clinicians give diagnostic tests only to patients who are already exhibiting some symptoms of a disease. This is a highly select group of people – not a random sample from the general population. Yet a base rate of "1 out of 1000" applies to the population as a whole, not to this select group. Gigerenzer and Murray (1987, p. 166) point out that if these physicians had assumed that the person being tested did not represent a random draw from the general population, then setting their prior probability at 1/1000 would have been a normative error.

If these physicians had been told what disease was being tested for, they could have set their prior probability based on the frequency with which their patients have that disease. But in the original Casscells et al. problem the disease is unspecified. One way of dealing with this lack of information within the framework of Bayes' theorem is to adopt the "principle of indifference", and set one's prior probability at 50%. On this assumption, the correct bayesian answer is "95.2%" – and "95%" was the modal response for Casscells et al.'s subjects. This raises the possibility that Casscells et al.'s Harvard Medical School students and staff were engaging in *correct* bayesian reasoning (Gigerenzer & Murray, 1987, p. 166). We will call this hypothesis the "indifference" hypothesis.

Alternatively, "95%" might have been the modal answer in both Casscells

et al.'s experiment and our replication of it because subjects thought that a "false positive rate" is an inverse probability (i.e., p(healthy|positive)), rather than a likelihood (i.e., p(positive|healthy)). In other words, they may think that a 5% false positive rate means that (i) out of every 100 people who test positive for the disease, 5 test falsely so – that is, are actually healthy (an inverse probability), rather than (ii) out of every 100 healthy people, 5 will test positive (a likelihood). If one takes a false positive rate to be an inverse probability, then "95%" is the correct answer, because the p(disease|positive) = 1 – p(healthy|positive) = 1 – .05 = .95. We will call this hypothesis the "inverse probability" hypothesis.

Although we have no way of finding out what Casscells et al.'s physicians and medical students were thinking, these two competing hypotheses can be tested on our subject population. In Experiment 6 we do this in two ways: (1) by making the random sampling assumption explicit and (2) by increasing the false positive rate.

If the indifference hypothesis is correct, then making the random sampling assumption explicit should elicit the answer "2%" from more subjects than the problem tested in Experiment 5, where no statement about random sampling was made. In contrast, if the inverse probability hypothesis is correct, then the distribution of responses should be the same as that for Experiment 5.

Increasing the false positive rate tests between these two hypotheses in a different way. With a false positive rate of 5%, the indifference hypothesis and the inverse probability hypothesis yield the same answer – "95%". But for a false positive rate of 40%, a good Bayesian who was applying the principle of indifference would answer "71%",[15] whereas a person who assumed that the false positive rate was an inverse probability would answer "60%".

15. Materials

Experiment 6 had two conditions.

Condition 1 (E6–C1)

The text of the problem tested in Condition 1 was very similar to the text of the problem tested in Experiment 5, the non-frequentist version in which the meaning of false positive rate was clarified:

[15] In applying the indifference principle, one assumes that p(disease) = p(healthy). Out of a population of 200 people, 100 would have the disease, and all of these would test positive. One hundred would be healthy, but because the false positive rate is 40%, 40 of these healthy people would also test positive. The total number of people who test positive would therefore be 140. Thus 100 people would have the disease and test positive for it, out of a total of 140 who test positive for it: 100/140 = 71%.

The prevalence of disease X among Americans is 1/1000. 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, 5% of all people who are perfectly healthy test positive for the disease.

Imagine that we have given this test to a random sample of Americans. They were selected by a lottery. Those who conducted the lottery had no information about the health status of any of these people.

What is the chance that a person found to have a positive result actually has the disease? ____%

The second paragraph tells the subject that the random sampling assumption has been met. Gigerenzer et al. (1988) found that interjecting the single word "random" was insufficient to convince subjects that the random sampling assumption had been met when their experience dictated otherwise, so we used this longer and more explicit statement in Condition 1. Condition 2 is important, however, because we cannot be absolutely sure that even a long verbal statement is sufficient to counteract a lifetime of experience.

Condition 2 (E6–C2)

The text of Condition 2 read as follows:

If a test to detect a disease whose prevalence is 1/1000 has a false positive rate of 40%, what is the chance that a person found to have a positive result actually has the disease, assuming that you know nothing about the person's symptoms or signs? ____%

It is identical to the original Casscells et al. problem and to the problem tested in E1–C1, except that the false positive rate is 40% rather than 5%. We modeled this problem on the original Casscells et al. problem rather than on the clarified version tested in Experiment 5 because (1) the meaning of "false positive" is ambiguous only in the original version, and (2) the original elicited the highest proportion of "95%" responses ("95%" is the response we are trying to explain).

16. Results

We will compare the results of E6–C1 to those of Experiment 5 because the only difference between these two problems is that E6–C1 made the random sampling assumption explicit. If the indifference hypothesis were true, then the number of subjects answering "95%" in E6–C1 should drop, and the number answering "2%" should rise. Although only 16% of

subjects answered "95%" in this condition, this is not significantly different from the 32% who gave this answer in Experiment 5 ($Z = 1.32$, phi = .19, $p = .09$). Similarly, the number of subjects answering "2%" did not rise, even by the most liberal scoring criterion. Twenty-eight percent of subjects in Condition 1 answered "2%", as compared to 36% in Experiment 5. (By an unusually liberal scoring criterion that mitigates against the indifference hypothesis,[16] the figure for E6–C1 was 48% – which is still not significantly different from the results obtained in Experiment 5.) Thus the results of Condition 1 do not support the Bayesian indifference hypothesis.

Nor do the results of Condition 2 support the Bayesian indifference hypothesis (see Fig. 4). If the indifference hypothesis were correct, then there should be a large number of subjects who answer "71%". In contrast, if the inverse probability hypothesis were correct, as many subjects should answer "60%" in Condition 2 as answered "95%" in E1–C1 – the original Casscells et al. problem. Only one subject in this condition answered "71%" – the Bayesian indifference answer. In contrast, 60% of them answered "60%" – which is almost identical to the 56% of subjects who answered "95%" in E1–C1. These results clearly support the inverse probability hypothesis over the indifference hypothesis.

Indeed, the response profiles for the two original versions tested (E6–C2 and E1–C1) were virtually identical, despite the large difference in given false positive rates (40% vs. 5%). For E6–C2, the correct bayesian response if one accepts the 1/1000 base rate is 0.25%. One subject in this condition

[16] If we were to use the same (standard) scoring criterion for this experiment that we did for all of the others, the results would look even better for the frequentist hypothesis that is the major thrust of this article: only 28% would be scored as answering correctly on this clarified non-frequentist version, and the various frequentist versions discussed elsewhere in this article would look even better in comparison. However, to be conservative, we analyzed the results as follows: there were five subjects in E6–C1 who answered ".02". It seems likely that three of these subjects had simply not seen the percent sign next to the answer blank, because in their calculations these three had written that the answer was 1/50 – which is 2%. It is impossible to determine whether the other two subjects who answered ".02" also missed the percent sign, or whether they simply solved the problem incorrectly. A liberal scoring criterion that gave all five of these subjects the benefit of the doubt would bring the total to 48%. For analyses later in the paper, we preferred to take a middle course, and count as correct the three subjects who clearly understood that 1/50 was the correct answer, bringing the total to 40% correct. This apparent confusion did not arise in any of the other problems that asked for the answer as a percent – errors were not variants of "2%" that merely differed by a misplaced decimal point in one direction or the other. Therefore, the low proportion of Bayesian responses obtained in these conditions cannot be accounted for by the assumption that some subjects simply did not see the percent sign. Nor can they be accounted for by the hypothesis that subjects cannot transform a number into a percent. If subjects could not transform percents to frequencies (or vice versa), then errors consisting of misplaced decimal points would be the rule rather than the exception. Moreover, performance would not have been so high (64% correct) on E7-C1, which asked for the answer as a frequency, but which presented the problem information as a percent: to give the correct answer, these subjects had to be able to understand what that percent information meant.

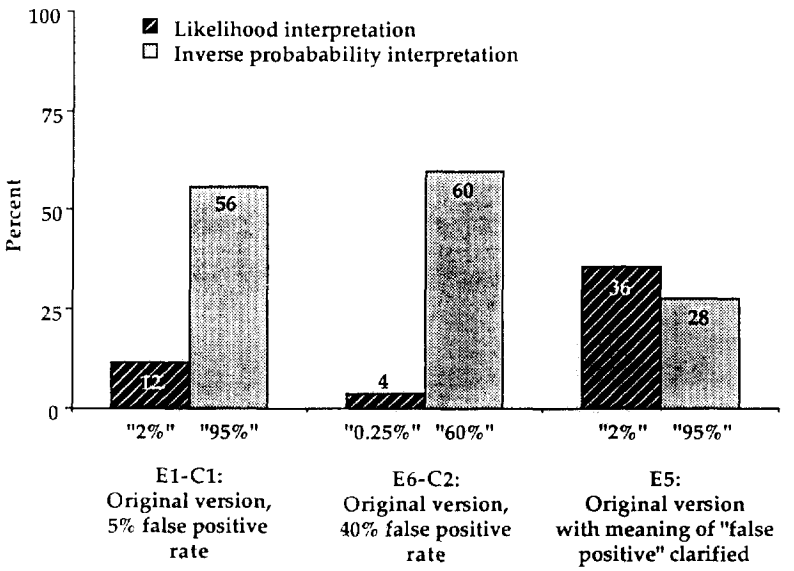

Fig. 4. Percentage of subjects who gave the correct answer assuming they interpreted a false positive rate to be a likelihood, versus the percentage who gave the correct answer assuming they interpreted it to be an inverse probability. The results support the inverse probability hypothesis over the Bayesian indifference hypothesis.

did a correct frequentist analysis, rounding 1/400 to 0.2%; another, who applied Bayes' rule, answered 0.5% due to an arithmetic error (he thought $0.4 \times 999 = 200$, rather than 400). Whether we count only the first individual as correct or both of them, the number of subjects in E6–C2 who gave the correct bayesian answer given a prior probability of 1/1000 was very similar to the matching condition in Experiment 1 (E6–C2: 1 or 2/25; E1–C1: 3/25).

DISCUSSION FOR PART II

Can one elicit the very high levels of bayesian performance found in Part I merely by eliminating ambiguities in the original, non-frequentist Casscells et al. problem, or is there something special about representing a problem in frequentist terms?

In Part II we eliminated the hypothesis that merely clarifying terms in the original Casscells et al. problem is sufficient to produce these high levels of bayesian performance. Experiment 5 tested an amended version of the original problem, which included the true positive rate and clarified the notion of a false positive rate. This elicited correct bayesian performance from only 36% of subjects tested – a far cry from the 76% correct elicited by the two comparable frequentist versions in Part I, or the 92% correct

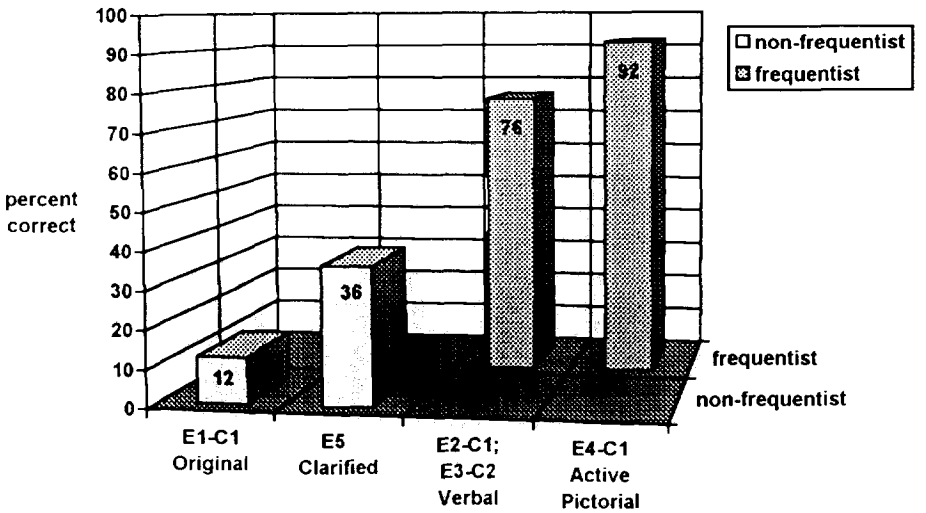

Fig. 5. Although clarifying a non-frequentist problem boosts performance slightly, it is not sufficient to elicit the high levels of bayesian performance that frequentist problems do.

elicited by the active-pictorial condition, which forced subjects to construct a concrete, visual frequentist representation (see Fig. 5). Moreover, at least 7 out of the 9 subjects who did provide the correct bayesian answer for this non-frequentist problem did so by translating the problem into frequentist terms. The low levels of bayesian performance on the original Casscells et al. problem are not a mere artifact of ambiguities in the wording of the problem.

Did the original Casscells et al. problem elicit low levels of "2%" responses and high levels of "95%" responses because the subjects were good Bayesians who believed that the random sampling assumption had been violated and therefore applied the principle of indifference? (If so, then "95%" is the *correct* answer.) We eliminated this hypothesis as well. Making the random sampling assumption explicit in E6–C1 elicited the same distribution of responses as the comparable problem tested in Experiment 5 – the number of "2%" responses was no higher, the level of "95%" responses no lower. Because verbal statements about random sampling are not always sufficient to cause a subject to ignore their real-world knowledge of a domain (Gigerenzer et al., 1988), we also tested this hypothesis in a different way. If people were applying the Bayesian principle of indifference, then a version of the original Casscells et al. problem with a false positive rate of 40% should elicit the answer "71%" from as many subjects in E6–C2 as answered "95%" in E1–C1 – that is, 56% of them. Instead, only one subject in E6–C2 answered "71%" – that is, 4% of subjects tested in that condition.

If subjects are not answering "95%" in the original problem because they

are using the Bayesian principle of indifference, then why is this the modal response? The data of Part II support the hypothesis that subjects gave this answer because they were assuming that a false positive rate is an inverse probability rather than a likelihood. The inverse probability hypothesis and the indifference hypothesis predict the same response – "95%" – when the false positive rate is set at 5%, but they predict different responses when the false positive rate is 40% – inverse probability predicts "60%", whereas indifference predicts "71%". Sixty percent of subjects tested in E6–C2 answered "60%" and only one answered "71%". The number of subjects who gave the inverse probability response in E6–C2 was almost identical to the number who gave the comparable response ("95%") in the original Casscells et al. problem. The hypothesis that subjects were assuming that the false positive rate was an inverse probability in the original problem (E1–C1) is also supported by the results of Experiment 5. In Experiment 5 we made it clear that 5% was a likelihood rather than an inverse probability. This caused the number of subjects answering "95%" to drop from a high of 56% in E1–C1 to 28% in Experiment 5.

We would like to emphasize that if one believes that a false positive rate is an inverse probability, then "95%" is the correct answer – not a normative error. Our results accord well with those of Eddy (1982) who, in an analysis of the medical literature and an informal survey of physicians, found that many physicians interpret likelihoods, such as true and false positive rates, as inverse probabilities.

The frequentist problems tested in Part I differed from the original, non-frequentist Casscells et al. problem in a number of ways: (1) they were less ambiguous; (2) they made the random sampling assumption explicit; and (3) they expressed the problem in frequentist terms. The experiments of Part II eliminated the hypothesis that the dramatic effects obtained in Part I were caused by the first two factors, either singly or in combination; neither stating a problem clearly, nor making the random sampling assumption explicit is sufficient to elicit high levels of bayesian reasoning from a problem that is not expressed in frequentist terms. This leaves factor (3) – the fact that the problems in Part I that elicited high levels of bayesian performance were expressed in frequentist terms. In Part III, we further test the hypothesis that frequentist representations afford correct bayesian reasoning by seeing whether systematically subtracting various frequentist elements lowers bayesian performance.

PART III. DO FREQUENTIST REPRESENTATIONS OF THE MEDICAL DIAGNOSIS PROBLEM ELICIT CORRECT BAYESIAN REASONING?

In Part III we further investigate the hypothesis that frequentist representations elicit correct bayesian reasoning by (1) adding frequentist representations to problems that lack them, and (2) subtracting frequentist

representations from problems that have them. We will address three questions: (1) Does asking for the answer as a frequency rather than as a single-event probability improve bayesian performance, all else equal? (2) Does asking the subject to answer the problem with respect to an explicitly enumerated population improve performance? and (3) Does presenting the problem information as frequencies, rather than as percents, improve performance?

EXPERIMENT 7

For a died-in-the-wool frequentist, a probability can refer only to a relative frequency defined with respect to a specified reference class; it cannot, in principle, refer to a single event. In other words, it is meaningful to ask "How many people who test positive for the disease will actually have the disease?", but it is not meaningful to ask "What is the chance that *a person* who tests positive for the disease actually has it?" The first question asks for a frequency, the second for the probability of a single event.

Does the untutored human mind also distinguish between frequencies and single-event probabilities? If there are mechanisms that can apply a calculus of probability when they represent probabilities as frequencies defined over a reference class, then asking for the answer to the medical diagnosis problem as a frequency should elicit higher performance than asking for the answer as a single-event probability. We tested this hypothesis in Experiment 7.

17. Materials

Experiment 7 had three conditions.

Condition 1 (E7–C1)

First, we wanted to see whether asking for the answer as a frequency could improve performance on a problem that was otherwise non-frequentist. Thus, Condition 1 was designed as a focused comparison for the clarified version of the original Casscells et al. problem tested in Experiment 5.

E7–C1 and E5 were identical, except for one thing: whereas Experiment 5 asked for the answer as a single-event probability, this condition asked for the answer as a frequency. In all other respects, the two problems are exactly the same: both present the problem information as percentages, neither makes the random sampling assumption explicit, and neither provides an explicitly enumerated population for the subject to think in terms of.

Thus, where the text of Experiment 5 read: "What is the chance that a person found to have a positive result actually has the disease, assuming that you know nothing about the person's symptoms or signs? ____%", the text of Experiment 7, Condition 1 read:

> Assume you know nothing about any person's symptoms or signs.
> Given the information above:
> on average,
> How many people who test positive for the disease will *actually* have the disease? ____ out of ____

If frequentist representations activate bayesian reasoning mechanisms, then this condition should elicit a higher percentage of bayesian responses than the 36% elicited by Experiment 5.

Condition 2 (E7–C2)

The purpose of Condition 2 was to see whether asking for the answer as a single-event probability could lower performance on a problem that was otherwise frequentist. Thus, Condition 2 was designed as a focused comparison for the two pure frequentist problems tested in E2–C1 and E3–C2, for which the average level of bayesian performance was 76% ($n = 50$). Both the frequentist problems from Part I, and the problem tested here, presented the problem information as frequencies, made the random sampling assumption explicit, and gave subjects an explicitly enumerated population of 1000 to think in terms of (as in "Imagine that we have assembled a random sample of 1000 Americans"). These problems differed in only one respect: whereas the experiments reported in Part I asked for the answer as a frequency, this condition asked for the answer as a single-event probability. Thus, where the text of the Part I problems read: "How many people who test positive for the disease will *actually* have the disease? ____ out of ____", the text of this condition read: "What is the chance that a person found to have a positive result actually has the disease? ____%"

If the frequentist hypothesis is correct, then this condition should elicit fewer bayesian responses than the 76% elicited by E2–C1 and E3–C2 in Part I.

Condition 3 (E7–C3)

Condition 3 was designed as a focused comparison for the non-frequentist problem tested in Experiment 6, Condition 1 (E6–C1). Both problems present the problem information as percentages; both make the random sampling assumption explicit; neither provide an explicitly enumerated population for the subject to think in terms of. They differ in only one way:

whereas E6–C1 asked for the answer as a single-event probability, this condition asked for the answer as a frequency. Thus, where the text of E6–C1 read: What is the chance that a person found to have a positive result actually has the disease? ____%", the text of Experiment 7, condition 3 read:

Given the information above:
on average,
How many people who test positive for the disease will *actually* have the disease? ____ out of ____

If the frequentist hypothesis is correct, then this condition should elicit more bayesian responses than the 40% (see footnote 16) elicited by E6–C1.

18. Results

All three predictions were confirmed (see Fig. 6). All else equal, problems that asked for the answer as a frequency elicited the correct

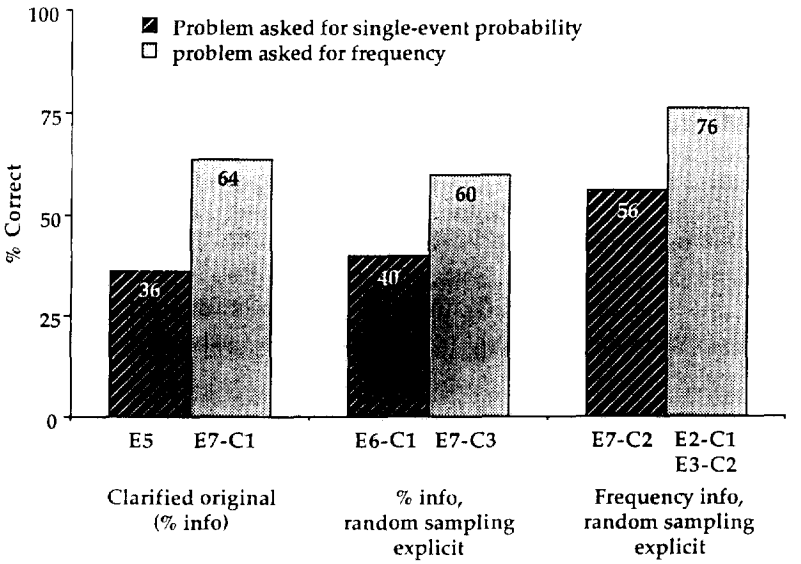

Fig. 6. All else equal, asking for the answer as a frequency enhances performance over asking for the answer as a single-event probability. Each pair of bars represents two problems that are identical, except that the first asked for the answer as a single-event probability whereas the second asked for the answer as a frequency. The first two pairs represent conditions in which the problem information was presented as a percent (hence overall performance is lower); the last pair represents conditions in which the problem information was presented as a frequency.

bayesian answer ("2%") from more subjects than ones that asked for the answer as a single-event probability.

For example, although E7–C1 and E5 were otherwise identical, 64% of subjects responded correctly when the answer was asked for as a frequency (E7–C1), whereas only 36% of subjects did so when the answer was asked for as a single-event probability (E5). This is a 28 percentage point rise in performance ($Z = 2.00$, phi = .28, $p = .023$). This is all the more impressive given that the answer format was the only frequentist element in the problem: the information was presented as percents, subjects were not given an explicitly enumerated population to think in terms of, and the random sampling assumption was not made explicit.

E7–C2 showed that asking for the answer as a single-event probability can depress performance even when all other elements in a problem are frequentist. The two, pure frequentist problems of Part 1 (E2–C1 and E3–C2) elicited the correct bayesian answer from 76% of subjects tested ($n = 50$). E7–C2 was identical to these problems, except for the fact that it asked for the answer as a single-event probability. Yet it elicited the correct response from only 56% of subjects. Here, asking for the answer as a single-event probability decreased bayesian performance by 20 percentage points ($Z = 1.77$, phi = .20, $p = .04$). An analysis of the errors showed that this decrement in bayesian performance was not the result of subjects not seeing the percent sign next to the answer blank, or by any inability of subjects to convert a number into a percent.

Condition 3 tested a problem that was a hybrid between those tested in Conditions 1 and 2. Although the problem information was presented as percents, as in Condition 1, the random sampling assumption was made explicit, as in Condition 2. Here the focused comparison was between Condition 3, which asked for the answer as a frequency, and E6–C1, which was identical except that it asked for the answer as a single-event probability. Whereas 40% of subjects gave the correct bayesian answer in E6–C1, where they were asked for a single-event probability, 60% of subjects gave the correct bayesian answer in Condition 3, where they were asked for a frequency. This is a 20 percentage point difference, just as in the Condition 2 focused comparison. Although this difference is not significant at the .05 level given only $n = 25$ per group, we note that the effect size, phi, is identical to that found in the Condition 2 comparison ($Z = 1.41$, phi = .20, $p = .08$).

By combining the results of these three focused comparisons, we can estimate the amount by which asking for the answer as a frequency, as compared to a single-event probability, increases bayesian performance. On average, the three problems that asked for the answer as a frequency elicited the correct bayesian response from 69% of subjects tested, whereas the three that asked for the answer as a single-event probability elicited this response from 44% of subjects tested. This produces an effect size, phi, of .25 ($Z = 3.32$, $p = .0005$).

Can the decrement in performance for problems that ask for the answer as a single-event probability be accounted for either by arithmetic errors in converting a number into a percent, or by the failure to see the percent sign next to the answer blank? No. No one gave an answer like ".02" or ".2" in either E7–C2, or E5, or E1–C1, nor did anyone give a functionally equivalent answer in E6–C2 (the 40% false positive rate problem) – yet all of these problems asked for the answer as a single-event probability. The only single-event probability problem where it appeared that some subjects may have failed to see the percent sign was E6–C1, and for this problem we used a more liberal scoring criterion to reflect this fact (see footnote 16).

Thus, the results of Experiment 7 show that asking for the answer as a frequency rather than as a single-event probability enhances bayesian performance by 20–28 percentage points.

EXPERIMENT 8

If frequencies are the natural format for some of our inductive reasoning mechanisms, then a very straightforward way of solving probability problems is to imagine a population of fixed size and then simply count up how many individuals fall into each category – in this case, how many have the disease and test positive for it, and how many test positive for the disease whether they have it or not. For example, if one imagines a population of 1000 people (randomly drawn), then 1 should have the disease, and this person should test positive for it, as well as 50 people who are healthy. Thus one person will have the disease and test positive for it out of 51 who test positive for it.

Anything that encourages subjects to represent the problem in this way should enhance bayesian performance. We thought that asking subjects to solve the problem with respect to a population of fixed size might encourage the formation of this kind of representation, and therefore improve performance. Hence, in Experiment 8, the question we addressed was, "Does asking the subject to answer the problem with respect to an explicitly enumerated population improve performance?"

19. Materials

Experiment 8 had two conditions.

Condition 1 (E8–C1)

Condition 1 was designed as a focused comparison to the problem tested in E7–C3. Both present the problem information as percents, both make the random sampling assumption explicit, and both ask for the answer as a

frequency. The only difference between the two is that this condition provides the subject with an explicitly enumerated population whereas E7–C3 did not. Thus, where E7–C3 read: "Imagine that we have given this test to a random sample of Americans", this condition read: "Imagine that we have assembled a random sample of 1000 Americans." If providing the subject with an explicitly enumerated population improves performance, then this condition should elicit more bayesian responses than the 60% elicited by E7–C3.

Condition 2 (E8–C2)

Condition 2 was designed as a focused comparison for the pure frequentist problems tested in Part I (E2–C1 and E3–C2). Both present the problem information as frequencies, both make the random sampling assumption explicit, and both ask for the answer as a frequency. The only difference between the two is that the frequentist problems from Part I provided the subject with an explicitly enumerated population whereas this condition did not. Thus, where the Part I problems read: "Imagine that we have assembled a random sample of 1000 Americans", this condition read: "Imagine that we have given this test to a random sample of Americans." If providing the subject with an explicitly enumerated population improves performance, then this condition should elicit fewer bayesian responses than the 76% ($n = 50$) elicited by the pure frequentist problems of Part I.

20. Results

Providing the subject with an explicitly enumerated population does not seem to make any difference. Condition 1, which provided an explicitly enumerated population, elicited bayesian performance from 52% of subjects tested, whereas the figure for E7–C3 was 60%. Similarly, whereas the pure frequentist problems of Part I, which provided an explicitly enumerated population, elicited bayesian performance from 76% of subjects tested, the figure for Condition 2, which lacked an explicitly enumerated population, was 68%. This is not significantly lower than 76% ($Z = 0.74$, phi = .09, $p = .23$).

This should not be terribly surprising. Being asked to imagine a random sample of people of unspecified size is not too different from being asked to imagine a random sample of specified size – both encourage the subject to represent the problem information as frequencies defined with respect to a population. But it seems that even more minimal cues are sufficient to elicit a frequentist representation of the problem. The problem tested in E7–C1 was identical to those tested in E8–C1 and E7–C3, insofar as they all presented the problem information as percents and asked for the answer as a frequency. The only difference is that E8–C1 and E7–C3 ask the subject

to imagine a random sample of people (explicitly enumerated in E8–C1, but not in E7–C3), whereas the problem tested in E7–C1 does not. Even though it does not ask subjects to imagine a population at all, E7–C1 elicited the correct bayesian response from 64% of subjects tested, a number that does not differ significantly from the average of 56% correct ($n = 50$) found for E7–C3 and E8–C1, which did ask subjects to imagine a population (64% vs. 56%: $Z = 0.66$, phi = .08, $p = .25$).

Considering performance on E6–C1 also supports the conclusion that providing a population to think in terms of, whether explicitly enumerated or not, does not affect subjects' tendency to represent the problem in frequentist terms. By making the random sampling assumption explicit, this condition provided a population to think in terms of, but it presented the problem information as percents and asked for a single-event probability answer. This condition elicited the same (low) level of bayesian performance as E5 did, even though it asked subjects to imagine a population and E5 did not (40% vs. 36%; see analysis for Experiment 6).

Taken together, these results indicate that asking the subject to imagine a population, whether explicitly enumerated or not, does not affect the level of bayesian performance one way or the other. If other elements in the problem cause the subject to construct a frequentist representation, the level of bayesian performance will be high; if not, it will be low, regardless of whether the subject is explicitly asked to imagine a population.[17]

20.1. Does presenting the problem information as a percent rather than a frequency decrease bayesian performance?

This question may, at first glance, appear to be a strange one: after all, isn't a percent simply a frequency out of a population of 100? Technically, this is true. But people with even junior high school mathematics skills have been taught how to plug percents directly into formulas, thus allowing them to use percents mathematically by simply manipulating symbols. When used in this way, it is easy to forget that a percent is a frequency because cranking through a formula does not require one to construct a frequentist representation of what that percent means. And, as we saw from the data on rounding in Part I, percent representations seemed to encourage subjects to represent the problem information in terms of continuous distributions, whereas frequentist representations seemed to encourage them to represent the problem in terms of discrete, countable individuals.

[17] In keeping with the frequentist hypothesis, we used round population numbers large enough that subjects would not have to think about fractions of individuals. It is possible, however, that requiring subjects to answer the question with respect to a population of, say 87, might worsen performance, for two related reasons: (1) given a base rate of 1/1000, they would have to think about fractions of people; and (2) arithmetic errors are more likely when one has to convert "50 out of 1000" to 5% and then compute 5% of 87, than when one can work directly with "50 out of 1000".

Furthermore, raw frequency information is more informative than percent information because it contains information about sample size. Because percents are normalized on a population of 100, they contain no information about the sample size the figure was derived from, and hence no information about how reliable that figure is. A mechanism that was designed to represent probabilities in the world in frequentist terms should therefore act on actual frequencies rather than automatically normalizing them into a percent: "5 out of 100" should seem a more "brittle", less trustworthy number than "50 out of 1000". Percentage representations are most useful when one wants to compare frequencies between groups (as we wanted to in analyzing the data for these experiments). They allow one to think in terms of frequencies out of a normalized population of 100.

We wanted to see whether presenting the problem information as a frequency rather than a percent would enhance bayesian performance. Therefore, in some of our problems, the false positive rate was presented as a frequency, whereas in others it was presented as a percent. We have been calling the first "frequency information" problems, and the second "percent information" problems. So, for example, a frequency information problem would present the false positive rate by saying, "Specifically, out of every 1000 people who are perfectly healthy, 50 of them test positive for the disease", whereas a matching percent information problem would say "Specifically, 5% of all people who are perfectly healthy test positive for the disease." For the frequency information problems we chose a base population large enough that the smallest category in the problem would still have at least one whole individual in it; thus, for problems where the base rate was 1/1000, we said "out of every 1000 people who are perfectly healthy, 50 of them test positive for the disease", and for problems where the base rate was 1/100 (the pictorial problems), we said "out of every 100 people who are perfectly healthy, 5 of them test positive for the disease". We thought this would encourage our subjects to choose a reference class large enough that they would not have to think in terms of fractions of individuals. If we have mechanisms that represent probabilities as frequencies defined over a reference class of discrete, countable entities, then this should improve performance.

For the same reason, we never presented the base rate, 1 out of 1000, as a percent. Although we thought presenting it as .001 or 0.1% would certainly have a negative effect on performance, we thought this would be stacking the deck too much in favor of the frequentist hypothesis for two reasons, one legitimate and the other not. From the point of view of the frequentist hypothesis, the legitimate reason is that 0.1% represents one-tenth of a person out of a population of 100. If there are mechanisms that represent frequencies in terms of discrete, countable entities, it should be difficult to think about a tenth of a person, and therefore the level of bayesian performance should decrease. On the other hand, a decimal percent is more likely to engender arithmetic errors when one is converting it to a frequency

than is a whole number percent. This kind of error is theoretically uninteresting, because we wanted to test our subjects' ability to reason probabilistically, not their ability to do arithmetic. As a compromise solution, we presented the base rate in a way that emphasized that it represents a frequency in the frequency information problems, but de-emphasized this in the percent information problems. Thus, the frequency information problems read: "1 out of every 1000 Americans has disease X", whereas the percent information problems read (following Casscells et al.): "The prevalence of disease X is 1/1000." The true positive rate was presented the same way in every problem, by saying "Every time the test is given to a person who has the disease, the test comes out positive."

Experiments 1–8 allow a number of focused comparisons. Because we have already found that making the random sampling assumption explicit or providing an explicitly enumerated population have no effect on the level of bayesian performance, we will ignore these factors. The first comparison is between the six problems that asked for the answer as a frequency, but differed in whether they presented the problem information as a frequency or a percent (see Table 3). Three problems that asked for the answer as a frequency also presented the problem information as a frequency: E2–C1, E3–C2, and E8–C2. These problems elicited a bayesian answer from 72%, 80%, and 68% of subjects, respectively, yielding an average value of 73.3% ($n = 75$). Three other problems asked for the answer as a frequency, but presented the problem information as a percent: E8–C1, E7–C3, and E7–C1, which elicited a bayesian answer from 52%, 60%, and 64% of subjects, respectively, yielding an average value of 58.7% ($n = 75$). This means there was, on average, a 14.6 percentage point difference between the frequency information problems and the percent information problems, given that the answer is asked for as a frequency. This difference is significant ($Z = 1.90$, phi = .15, $p = .03$). We note that 58.7% for these "percent information/frequency answer" problems is almost identical to the 60% found in Part I for the "frequency *and* percent information/frequency answer" problems (E1–C2: 56%; E3–C1: 64%). So providing redundant percent information appears to decrease performance to the same level found for problems that present the information only as a percent. Furthermore, the 14.6 point difference found here compares very favorably with the 16 point difference found in Part I between the frequency information problems and the frequency *and* percent information problems. The effect sizes were also very similar: phi = .15 for this comparison and .17 for the comparison from Part I.

We can also compare performance for the three problems that asked for the answer as a single-event probability, but differed in whether they presented the problem information as a frequency or a percent. One problem, E7–C2, asked for the answer as a single-event probability, but gave the problem information as a frequency. This elicited the bayesian answer from 56% of subjects tested. Two problems, E6–C1 and E5, asked

Table 3
Does asking for answers as frequencies contribute to performance independently of presenting information as frequencies?

	Information as:		
Answer as	Frequency	Percent	
Frequency	73.3% ($n = 75$)	58.7% ($n = 75$)	(66%)
Single-event probability	56% ($n = 25$)	38% ($n = 50$)	(47%)
	(65%)	(48%)	

for the answer as a single-event probability, and gave the problem information as a percent. These elicited the bayesian answer from 38% of subjects tested (E6–C1: 40%; E5: 36%). Thus, presenting the information as a frequency rather than a percent resulted in an 18 percentage point advantage in this comparison. This number is even larger than the 14.6 point advantage found above for the "frequency answer" problems, but because this comparison involves half as many subjects it is not significant at the .05 level ($Z = 1.48$, phi = .17, $p = .07$). The effect size for this comparison, .17, is very similar to the effect size of .15 for the previous one.

If we combine all these problems, regardless of the form the answer was given in, then 69% ($n = 100$) of subjects in the frequency information conditions gave the correct bayesian answer, compared to 50.4% ($n = 125$) for the percent information conditions ($Z = 2.81$, phi = .19, $p = .0025$).

These comparisons show that presenting the problem information as frequencies does in fact elicit higher levels of bayesian reasoning than presenting it as percents.

20.2 Is the advantage obtained by presenting the problem information as a frequency rather than a percent independent of the advantage obtained by asking for the answer as a frequency rather than a single-event probability?

We can answer this question by doing an analysis of variance on the nine problems described in the section above. Table 3 shows the breakdown of results.

An analysis of variance confirms what is already obvious from Table 3 and from the previous analyses: asking for the answer as a frequency as opposed to a single-event probability contributes to performance ($F(1, 221) = 7.25$, $r = .18$, $p = .01$) and presenting the information as a frequency as opposed to a percent contributes to performance ($F(1, 221) = 5.36$, $r = .15$, $p = .025$), and they do so independently of one another ($F(1, 221) = .056$ for the interaction). Moreover, the effect of asking for the answer as a frequency is a bit bigger than the effect of presenting the information as a frequency: $r = .18$ versus .15.

GENERAL DISCUSSION

Although the original, non-frequentist version of Casscells et al.'s medical diagnosis problem elicited the correct bayesian answer of "2%" from only 12% of subjects tested, pure frequentist versions of the same problem elicited very high levels of bayesian performance: an average of 76% correct for purely verbal frequentist problems and 92% correct for a problem that requires subjects to construct a concrete, visual frequentist representation. This was shown in Part I. In Parts II and III, we tried to discover what accounts for the very high levels of bayesian performance elicited by these frequentist problems.

In Part II, we showed that manipulating aspects of the medical diagnosis problem that are unrelated to the issue of frequentist representations – such as clarifying the meaning of "false positive rate" and making it clear that the sample was randomly drawn – cannot produce these dramatic effects. Clarifying the meaning of "false positive rate" and providing the true positive rate for a non-frequentist problem increased bayesian performance slightly, from 12% to 36%, but this is nowhere near the average of 76% correct elicited by the pure frequentist problems, let alone the high of 92% correct for the more ecologically valid problem that required subjects to construct a concrete, visual frequentist representation. Also, in Part II we asked why the original Casscells et al. problem elicited a high level of "95%" answers both from our subjects and from Casscells et al.'s physicians and medical students. One possibility was that subjects understood that a false positive rate is a likelihood but, because they did not believe the sample was randomly drawn, they were applying the Bayesian principle of indifference. The other possibility was that they believed that a false positive rate is an inverse probability rather than a likelihood. Our results showed that the latter hypothesis was correct: in the absence of information to the contrary, many subjects interpret a false positive rate to be an inverse probability.

In Part III, we tackled the question of the causal efficacy of frequentist representations directly by systematically adding and subtracting elements that we thought would induce subjects to construct a frequentist representation of the problem and seeing how this affected performance. We found that (1) asking for the answer as a frequency rather than as a single-event probability improves bayesian performance, (2) although requiring subjects to actively construct a concrete, visual representation of frequencies does enhance bayesian performance, verbal instructions to merely imagine a population (whether explicitly enumerated or not) does not, and (3) presenting the problem information as frequencies, rather than as percents, improves bayesian performance. Asking for the answer as a frequency produces the largest effect, followed closely by presenting the problem information as frequencies.

Verbally instructing subjects to merely imagine a population is insufficient

to enhance bayesian performance when the problem information is in percents and the answer is asked for as a single-event probability; moreover, such instructions seem to be superfluous when both answer and information are asked for and presented as frequencies. But this conclusion applies only to verbal instructions to imagine. The active pictorial condition tested in Part I showed that when subjects are instructed to construct a concrete *visual* representation of a population that depicts the relevant frequencies, bayesian performance is, in fact, enhanced to near perfect levels.

In short, all the predictions of the frequentist hypothesis were confirmed: (1) inductive reasoning performance differed depending on whether subjects were asked to judge a frequency or the probability of a single event; (2) performance on frequentist versions of problems was superior to non-frequentist versions; (3) the more subjects could be mobilized to form a frequentist representation, the better their performance was; and (4) performance on frequentist problems satisfied those constraints of a calculus of probability that we tested for (i.e., Bayes' rule). Taken together, the results of Parts I–III support the hypothesis that frequentist representations activate mechanisms that produce bayesian reasoning, and that this is what accounts for the very high levels of bayesian performance elicited by the pure frequentist problems that we tested.

21. Representations, algorithms and computational theories

When we say that people have mechanisms that produce good bayesian reasoning, what exactly does that mean? More generally, what does it mean to say that the mind "embodies" aspects of a calculus of probability?

According to David Marr:

> [There are] different levels at which an information-processing device must be understood before one can be said to have understood it completely. At one extreme, the top level, is the abstract computational theory of the device, in which the performance of the device is characterized as a mapping from one kind of information to another, the abstract properties of this mapping are defined precisely, and its appropriateness and adequacy for the task at hand are demonstrated. In the center is the choice of representation for the input and output and the algorithm to be used to transform one into the other. And at the other extreme are the details of how the algorithm and representation are realized physically – the detailed computer architecture, so to speak. (Marr, 1982, pp. 24–25)

The hypotheses we tested about bayesian reasoning concern Marr's first two levels of explanation. Bayes' theorem is part of a computational theory – that is, a normative theory – of how inductive reasoning should be conducted in certain domains. It is an abstract specification of how information about prior probabilities and likelihoods should be mapped onto posterior probabilities.

Bayes' theorem is *not* a hypothesis about the representations and

algorithms that effect this transformation in the human mind. In principle, many different algorithms and representational systems can produce performance that accords with Bayes' theorem. When we tested the hypothesis that the mind embodies aspects of a calculus of probability, such as Bayes' rule, we were testing for the presence of mechanisms that implement a particular computational theory.

The frequentist hypothesis tested in this article is at Marr's second level: it is a hypothesis about the kind of input and output representations used by an algorithm that accomplishes the transformation specified by Bayes' theorem. In saying that frequentist representations afford bayesian reasoning, we are proposing that there exists at least one algorithm that maps frequentist representations of prior probabilities and likelihoods onto a frequentist representation of a posterior probability in a way that satisfies the constraints of Bayes' theorem.

We have made no strong claims about the exact nature of the algorithm that effects this transformation on the contents we tested. Indeed, different algorithms may accomplish bayesian reasoning for different kinds of adaptive problems (see section 23 below). The algorithm could, for example, have subroutines that multiply $p(\mathrm{H})$ by $p(\mathrm{D}|\mathrm{H})$ and then divide by $p(\mathrm{D})$, but be unable to do this unless the information is in a frequentist format. Alternatively, the algorithm might involve no arithmetic operations beyond a counting subroutine. If one represents the base rate and likelihood information as numbers of discrete, countable individuals – that is, if one forms a frequentist representation – then multiplication and division become unnecessary. If one understands what categories of information the problem is asking about – in this case, "people who have the disease and test positive for it" and "people who test positive whether they have the disease or not" – then the only remaining step is to count up how many individuals fall into each category. Indeed, to judge from their side calculations, this appears to be just how our successful subjects were proceeding, a conclusion that is also supported by our subjects' uniformly high performance on the problem in which they were required to construct a visual representation of countable individuals. If a counting subroutine is involved, then the most interesting part of the algorithm would be that which allows one to map a complex set of relationships among different categories of information. These set manipulation procedures may require representations of discrete individuals to operate properly. On this account, both the set manipulation procedures and the counting subroutine would require frequentist representations. Together, these procedures would produce reasoning performance that conforms to Bayes' rule; they would thereby "embody" that aspect of a calculus of probability. (If this account is true, then certain aspects of inductive and deductive reasoning may be accomplished via some of the same mechanisms. For example, Johnson-Laird's mental model theory for syllogistic reasoning also requires a representational format of

discrete, countable individuals and set manipulation procedures for combining them.[18])

To discover what our intuitive statistical competences are – that is, to discover what aspects of the human cognitive architecture reliably develop in the absence of explicit instruction – basic anthropology can be informative. It can tell one what kinds of information would, and would not, have been reliably present in the environments in which these competences evolved. People everywhere are exposed to the actual frequencies of real events, and we seem to have unconscious mechanisms that can keep track of these frequencies, just as many nonhuman animals do (Staddon, 1988; Gallistel, 1990). In our experiments, however, subjects were exposed not to actual events, but to linguistic propositions about numbers – such as "50 out of every 1000 Americans who are perfectly healthy test positive for the disease." Is it possible, then, that our intuitive statistical procedures were designed to take linguistic propositions about numbers as input and produce linguistic propositions about numbers as output? Evidence from the ethnographic record quickly eliminates this hypothesis: The number lexicon is extremely limited or virtually non-existent in many languages (especially for band-level societies). Linguistically transmitted numerical propositions were not a regular part of the environment in which we evolved; one would not expect humans to have evolved cognitive mechanisms designed to reason about, or accept as input, information in a form that did not exist. We may have mechanisms that allow linguistic information to be translated into a format that our intuitive statistical procedures are capable of reading, but these statistical procedures were surely not designed to take this kind of input directly.

What, then, are we to make of our subjects' performance in these experiments? They have been given a number lexicon by their culture and taught how to perform arithmetic operations on this symbol system in school. Without a number lexicon at least, they could not have even

[18] In his mental model theory, Johnson-Laird (1983) has suggested that people also solve syllogistic reasoning problems by representing category information in the form of discrete individuals. Moreover, he has claimed that syllogistic problems in which the category information is represented as discrete countable individuals are easier to solve than problems using representations that map finite sets of individuals into infinite and continuous sets of points, as in a Venn diagram. This is similar to our claim about bayesian problems. Indeed, both syllogistic and bayesian problems require one to understand the overlapping relationships among different categories of information – for example, people who have a disease, people who test positive for it, and people who are healthy. It may be that the same set manipulation procedures underlie both kinds of reasoning, and that these procedures require representations of discrete individuals to map the relationships among categories properly. On this view, the distinction between inductive and deductive reasoning would begin to dissolve at the mechanism level (although not at the computational theory level).

understood the problems as presented. Yet the balance of what our subjects needed to successfully solve these problems has not been culturally given: Most of them have not had explicit instruction in bayesian reasoning. In fact, those few who tried to use Bayes' formula performed poorly.

Therefore, one cannot interpret performance on these problems as purely the result of culturally acquired skills, or as purely the result of reliably developing species-typical statistical competences. Insead, we interpret these results as what happens when some simple culture skills (e.g., a number lexicon and some basic arithmetic) tie into and activate an underlying system of inductive reasoning mechanisms. This underlying system must be supplying all that is necessary for solving the problem that was not explicitly taught. The problems we administered therefore provide a window onto the nature of this underlying system, even though they include elements that could not have been understood without an explicitly taught, culturally specific number lexicon. Had our subjects been unable to understand linguistic propositions about frequencies, then we might have observed good bayesian performance only for problems in which the information was presented as the encountered frequencies of real events.

22. Does the mind embody aspects of a calculus of probability?

In 1972, Kahneman and Tversky drew the following conclusion from their research, which is still widely accepted: "In his evaluation of evidence, man is apparently not a conservative Bayesian: he is not a Bayesian at all" (p. 450). It now appears that this conclusion was premature. Frequentist problems elicit bayesian reasoning. This finding adds to the growing body of literature that shows that many cognitive biases in statistical reasoning disappear, and good statistical reasoning reappears, when problems are posed in frequentist terms. The conjunction fallacy disappears, the overconfidence bias disappears, base rate neglect disappears, and good bayesian reasoning emerges.

Frequentist mechanisms could not elicit bayesian reasoning unless our minds contained mechanisms that embody at least some aspects of a calculus of probability. This means that the more general conclusion of the literature on judgment under uncertainty – that the human mind does not embody a calculus of probability, but has instead only crude rules-of-thumb – must also be re-examined. This conclusion was based largely on subjects' responses to single-event probability problems. But if those inductive reasoning procedures that do embody a calculus of probability take frequency representations as input and produce frequency representations as output, then single-event probability problems cannot, in principle, reveal the nature of these mechanisms. It would therefore be illuminating to

restate the classic single-event problems in frequentist terms. This would allow us to discover which aspects of a calculus of probability our inductive reasoning mechanisms do embody, and which aspects they do not.

Furthermore, when reasoning performance deviates from normative predictions based on mathematical theories of probability, we should not automatically conclude that the mechanisms involved are poorly designed. Instead, we should investigate the possibility that our experimental protocol activated domain-specific reasoning procedures that are well designed for solving adaptive problems that were faced by our hominid ancestors, but which do not rely on ontogenetically observed frequencies. Rather than having only one or a few inductive reasoning mechanisms, the mind might include many different ones, each appropriate to a different kind of decision-making problem.

23. Statistical inference in a multimodular mind

In this article, we are emphatically not taking sides on the question of whether a frequentist interpretation of probability is intrinsically "better" or "truer" *for scientists* than a Bayesian subjective confidence interpretation, or whether, for example, single-event probabilities are in fact an incoherent notion. We are instead arguing that certain cognitive mechanisms are frequentist in design – that is, have elements and procedures that embody the theories of statistical inference used by many frequentists – because such designs solve certain adaptive problems with special efficiency. This will not be true, however, for *all* adaptive problems. Therefore, we are also not arguing that all cognitive mechanisms that guide decisions under uncertainty are frequentist. On the contrary, in a multimodular mind, the design of some mechanisms will probably echo the structure of other mathematical theories of probability, because the adaptive problems they address are better solved by designs that embody these other approaches.

Modular, domain-specific, or content-specific approaches ultimately derive their rationale from considerations that are either implicitly or explicitly functional and evolutionary (Cosmides & Tooby, 1987; Marr, 1982; Symons, 1987; Tooby & Cosmides, 1992a). This is because many families of important adaptive problem can be solved more efficiently by cognitive mechanisms that were specially tailored to meet the particular task demands characteristic of that problem-type (e.g., vision (Marr, 1982; Ramachadran, 1990); language acquisition (Pinker & Bloom, 1990); the perception and representation of object motion (Shepard, 1984; Freyd, 1987); the representation of biomechanical motion (Shiffrar & Freyd, 1990); cooperation (Cosmides, 1989; Cosmides & Tooby, 1989; Gigerenzer & Hug, 1992)). The trade-off between generality of application and efficiency in performance

leads to different compromises in different cases, with generality of application by no means always winning out.

Most adaptive problems that animals face – foraging for food, avoiding predators, finding mates, predicting the behavior of conspecifics – require them to make decisions under conditions of uncertainty about the state of the world. Decisions made under uncertainty will be successful to the extent that they turn out to correspond to the actual state of the world. Applying a calculus of probability to subjective degrees of belief that are unlinked to the relative frequencies of real events or the actual structure of a specific situation will not allow one to find real food, avoid real predators, or predict the behavior of real conspecifics. Why, then, might some mechanisms process ontogenetically observed frequencies whereas other mechanisms might not?

24. Frequencies past and present

24.1. Architectures shaped by statistical relationships that endure across generations

The nature of a psychological design should reflect the nature of the information that is available to be processed by that design. For any specific organism there are two primary sources of information: observation (ontogeny) and the evolutionary process (phylogeny). One might expect frequentist mechanisms to be activated by domains in which event frequencies are observable, are relevant to the problem, and are the sole, the primary, or the best source of information available for solving the problem. Obviously, local, recent observed frequencies will usually be the very best predictor of local events and relationships in the near future. But when the relevant frequencies are not observable during one's lifetime, or when a sufficient database has not or cannot be accumulated ontogenetically, mechanisms that rely only on ontogenetically observed frequencies will not generate good decisions. When actual observation during the lifespan of the individual is impossible, other sources of information must be used.

The world has a rich structure of recurring covariant relationships, most of which cannot be directly observed by the limited and local perceptual systems of an individual organism. But frequencies of events in the past can be "observed" by natural selection: Over many generations, cognitive systems whose designs better reflect the recurrent structure of the world tend to outreproduce and replace designs that do not. Natural selection can therefore create designs that "assume" that certain distributions and relationships are true, even though they cannot be observed during the lifetime of an individual (e.g., Staddon, 1988; Tooby & Cosmides, 1990). For example, the recurrent statistical relationship between ingesting toxic

plant secondary compounds and spontaneous abortions during the first trimester is virtually impossible to observe; consequently, women have mechanisms that "assume" this relationship, producing the first trimester food aversions, nausea, and vomiting known as "pregnancy sickness" (Profet, 1988, 1992). Similarly, the statistical relationship between helping in accordance with Hamilton's rule and fitness cannot be induced because it cannot, in principle, be observed during one's lifetime; the rule can be followed only if one has phylogenetically given decision rules that embody it (Cosmides & Tooby, 1987, 1994).

Even when organisms can, in principle, observe the relevant frequencies during their individual lifetimes, they must sometimes make decisions that have large fitness consequences before they have had the opportunity to develop a data base that is large enough to be reliable. This is particularly true of domains in which covariant events happen infrequently. For example, few people have actually observed a lethal snake bite. But over evolutionary time there has been a recurrent statistical relationship between the presence of snakes and the occurrence of potentially lethal bites. This has selected for domain-specific mechanisms that regulate judgment under uncertainty about snakes which "assume" that they pose a danger (Marks, 1987; Cook, Hodes, & Lang, 1986; Cook, Mineka, Wolkenstein, & Laitsch, 1985). Certainly, American children are not pure frequentists about sources of danger, nor are they simple culture absorbers. According to Maurer (1965) "they do not . . . fear the things they have been taught to be careful about. . . . The strange truth is that they fear an unrealistic source of danger in our urban civilization: wild animals" (p. 265). Indeed, almost all the 5- and 6-year-olds in Maurer's study of Chicago schoolchildren mentioned wild animals (most frequently snakes, lions, and tigers) in response to the question "What are the things to be afraid of?" Similarly, rats are not frequentists when it comes to pairing the taste of food with an electric shock, and virtually no experienced frequency of such pairings will produce the correct inference (Garcia & Koelling, 1966).

Lastly, there might be cases in which ontogenetically observable frequencies are ignored by decision-making mechanisms even when a large database of relevant frequencies can, in principle, be observed during an individual's lifetime. For domains in which there are statistical relationships that have endured over long periods of evolutionary time, one might expect to find domain-specific inductive reasoning mechanisms that embody information about these ancestral frequencies. In such cases, a cognitive system that assumes these relationships may be more efficient than one that must induce them *de novo* each generation (e.g., Shepard, 1984, 1987).

For example, an expanding body of literature in the field of cognitive development suggests that very small children have domain-specific inductive reasoning mechanisms that guide how they make inferences and thereby acquire knowledge in a given domain (e.g., Carey & Gelman, 1991; Gelman

& Markman, 1986; Hirschfeld & Gelman, 1994; Keil, 1989; Leslie, 1987; Spelke, 1988, 1990). In this view, the principles that govern induction in one domain may differ profoundly from those that govern induction in another domain. So, for example, considerable evidence has accumulated over the past few years that very young children reliably develop cognitive processes that cause them to reason differently about people's mental states and behavior than they do about physical objects or about the plant and animal world (e.g., Astington, Harris, & Olson, 1988; Atran, 1990; Baron-Cohen, 1994; Carey, 1985; Keil, 1989; Leslie, 1987, 1988; Spelke, 1988, 1990). Domain-specific inference procedures of this kind can supply a computational basis for decisions under uncertainty that is unrelated to frequencies that have been encoded by an individual during that individual's lifetime. For example, suppose you are looking for your dog. Initially, you think it likely that the dog is hiding under your bed, because this has often been true in the past (i.e., the prior probability that the dog is under the bed is high). Then you remember that because you sold the bedframe yesterday, your mattress is now flush to the floor. But if the mattress is flush to the floor, then the dog cannot be under it. Here, the revision of your initial prior probability is not based on new information about relative frequencies; it is based on cognitive mechanisms designed for reasoning about solid objects, which place constraints on the world of possibilities, such as that two physical objects cannot occupy the same space at the same time (Spelke 1988, 1990).

Decision-making architectures of this kind may be considered "non-frequentist" in a number of different senses:

(1) The weighting or "prior probability" assigned to a hypothesis need not be derived from ontogenetically encountered frequencies of the same event; it can be set phylogenetically (Staddon, 1988).

(2) The revision of a prior probability need not be based on data on the relative frequency of the event in question.

(3) The output of the mechanism may be a subjective degree of confidence, rather than a frequency.[19]

(4) The algorithms involved may compute Baconian, rather than Pascalian, probabilities (Cohen 1979, 1989). Pascalian probabilities are computed over repetitions of equivalent events; the more often an outcome occurs, the higher its Pascalian probability. Baconian probabilities increase as one varies potential causes, eliminating confounds that could explain the

[19] Of course, this can also be true of a frequentist mechanism; even though it might initially output a frequency, and perhaps even store the information as such, other mechanisms may make that frequentist output consciously accessible in the form of a subjective degree of confidence.

data. Baconian "eliminative" induction is particularly important in the evaluation of causal hypotheses.[20]

Of course, some inductive reasoning mechanisms may be hybrids, having some frequentist and some non-frequentist elements (see, for example, Gelman, 1990a, 1990b).

Phylogenetically supplied information is useful as a guide to decision-making because much of the structure of the world does endure over time: the statistical and structural properties of certain kinds of events do recur across generations. This recurrent structure of the world makes possible the evolution of complex domain-specific inference engines that can exploit these enduring probabilistic relationships to improve decisions under uncertainty, either to supplement ontogenetically observed frequencies or to operate in their absence.

These domain-specific reasoning mechanisms might violate normative theories that draw their standards from context-free mathematical analyses. That is because domain-specific mechanisms can draw on information about the present situation based on the statistical structure of *ancestral* environments; in other words, they have a "crib sheet" with added hints, clues or even answers, which a content-free mechanism does not (Tooby & Cosmides, 1990, 1992a). Mathematically derived computational theories might be more appropriate for precisely those domains for which we have *not* evolved domain-specific inference engines. But what kinds of domains would fall into this category?

24.2. Architectures that process ontogenetically experienced frequencies

Natural selection is a slow process compared to the length of an individual lifespan. Covariant relationships must endure for a great many generations before natural selection will modify cognitive adaptations to take them into account. But for those aspects of the world that change rapidly compared to

[20] In evaluating the frequentist hypothesis, for example, you (the reader) used both Baconian and Pascalian probabilities. Although it was important to show that good bayesian performance with the pure frequentist problems was replicable – that is, that the hypothesis has a high Pascalian probability – doing the same experiment over and over again would not have convinced you that the frequentist hypothesis is true, because the performance could have been caused by non-frequentist aspects of the problem. So we tried to raise the *Baconian* probability of the frequentist hypothesis, by showing that (1) good bayesian performance cannot be produced by manipulating non-frequentist aspects of the problem (such as clarifying the meaning of "false positive rate"), and (2) it can be enhanced by amplifying frequentist variables (as in the active pictorial condition). If nothing had replicated, you would not have believed the hypothesis (low Pascalian probability), nor would you have believed it if we had not been able to isolate frequentist representations as the causal variable (low Baconian probability).

generation time, phylogenetically supplied information will be unreliable or absent. In such cases, observed frequencies over the lifespan will be the best available predictors. Architectures that can pick up and use this frequency information would be able to solve such problems. Hybrid designs that use domain-specific knowledge – for example, a phylogenetically supplied prior probability – which is then revised based on observed frequencies can be expected where both phylogenetically and ontogenetically given information can jointly improve decision making. Staddon (1988) has argued that many learning mechanisms in nonhuman animals have hybrid designs of this kind, and Gelman (1990a, 1990b) has proposed similar designs for certain human learning mechanisms.

Given the foregoing, inference mechanisms with architectures that process ontogenetically encoded frequency information should have evolved to analyze domains in which information decays rapidly across ontogeny (or at least across generations) and that tend to have large "*n*s" over the lifespan. A large part of the hunter-gatherer's world was comprised of rapidly decaying reference class interrelationships – the changing spatial and temporal distributions of game, of plant foods, of predators, of daily weather, and so on. Phylogeny can supply nothing useful to predict those relationships that rapidly decay. In contrast, recent local sampling of frequencies can provide a reliable basis for prediction in such cases (see, for example, Real, 1991, on foraging algorithms in bumblebees). Precisely because they are content-free, mathematically derived computational theories might be more appropriate for such relationships; where relationships are not stable over time, domain-specific processes cannot improve prediction. We are exposed to an ocean of rapidly decaying relationships in the modern world as well: the architecture of buildings in different parts of a city; the proportions of various dog breeds in the neighborhood; the proportions of various brands of microcomputer in the office complex; changing fashions in running shoes; the popularity of names over time, and so on. Frequency encoding mechanisms seem to pick up such information automatically. For instance, when asked to guess whether "Charles", "Ruth", "Jennifer" or "Jason" are over or under 40, people do so in a way that corresponds well with the actual frequency of names given to people of various generations (Karas & Eisenacher, reported in Brown, 1986, pp. 588–589). One expects architectures that embody content-free normative theories to process just those kinds of information that our domain-specific inductive architectures do not.

Conclusions

If the body of results indicating well-calibrated statistical performance continues to grow, then a new analytic framework may be required to

organize and explain the literature on judgment under uncertainty (see, for example, Gigerenzer, 1991; Gigerenzer & Murray, 1987; Tooby & Cosmides, 1992b). Highly organized, well-calibrated performance cannot occur in the absence of well-designed cognitive mechanisms, so any new analytic framework must admit and explain the existence of such mechanisms. The evolutionary-functional framework proposed by Marr is particularly promising: one looks for a mesh between the nature of the adaptive problem to be solved and the design of the algorithms and representations that evolved to solve it. Mathematics and evolutionary biology provide a broad assortment of alternative normative theories of statistical inference, appropriate to different kinds of adaptive problem. These can help one discover what cognitive processes govern inductive reasoning in various domains, and why they have the functional design they do. By locating functionality in its evolutionary and ecological context, performance that had previously looked erratic and erroneous may begin to look orderly and sophisticated.

It may be time to return to a more Laplacian view, and grant human intuition a little more respect than it has recently been receiving. The evolved mechanisms that undergird our intuitions have been subjected to millions of years of field testing against a very rich and complexly structured environment. With only a few hundred years of normative theorizing under our belts, there may still be aspects of real-world statistical problems that have escaped us. Of course, no system will be completely error-free, even under natural conditions. But when intuition and probability theory appear to clash, it would seem both logical and prudent to at least consider the possibility that there may be a sophisticated logic to the intuition. We may discover that humans are good intuitive statisticians after all.

Acknowledgements

We warmly thank John Cotton, Martin Daly, Lorraine Daston, Jennifer Freyd, Gerd Gigerenzer, Wolfgang Hell, Gernot Kleiter, Duncan Luce, Dick Nisbett, Steve Pinker, Catrin Rode, Paul Romer, Peter Sedlmeier, Roger Shepard, Dan Sperber, Amos Tversky, Margo Wilson and an anonymous reviewer for enlightening discussions of or comments on the issues addressed in this paper. This chapter was prepared, in part, while the authors were Fellows at the Center for Advanced Study in the Behavioral Sciences. We are grateful for the Center's support, as well as that provided by the Gordon P. Getty Trust, the Harry Frank Guggenheim Foundation, and NSF Grant BNS87-00864 to the Center. NSF Grant BNS91-57449 to John Tooby and the McDonnell Foundation provided valuable support while we finished writing this paper, as did Peter Weingart and the Zentrum fur interdisziplinaire Forschung at the University of Bielefeld, Germany.

References

Alba, J.W., Chromiak, W., Hasher, L., & Attig, M.S. (1980). Automatic encoding of category size information. *Journal of Experimental Psychology: Human Learning and Memory, 6*, 370–378.

Astington, J.W., Harris, P.L., & Olson, D.R. (eds.) (1988). *Developing theories of mind*. Cambridge, UK: Cambridge University Press.

Atran, S. (1990). *The cognitive foundations of natural history*. New York: Cambridge University Press.

Attig, M., & Hasher, L. (1980). The processing of frequency of occurrence information by adults. *Journal of Gerontology, 35*, 66–69.

Bar-Hillel, M. (1980). The base-rate fallacy in probability judgments. *Acta Psychologica, 44*, 211–233.

Baron-Cohen, S. (1994). *Mindblindness on autism and theory of mind: An essay*. Cambridge, MA: MIT Press.

Birnbaum, M.H. (1983). Base rates in Bayesian inference: signal detection analysis of the cab problem. *American Journal of Psychology, 96*, 8594.

Bizzi, E., Mussa-Ivaldi, F., & Giszter, S. (1991). Computations underlying the execution of movement: a biological perspective. *Science, 253*, 287–291.

Brown, R. (1986). *Social psychology* (2nd ed.). New York: Free Press.

Carey, S. (1985). *Conceptual change in childhood*. Cambridge, MA: MIT Press.

Carey, S., & Gelman, R. (Eds.) (1991). *The epigenesis of mind: Essays on biology and cognition*. Hillsdale, NJ: Erlbaum.

Casscells, W., Schoenberger, A., & Graboys, T.B. (1978). Interpretation by physicians of clinical laboratory results. *New England Journal of Medicine, 299*, 999–1001.

Cohen, L.J. (1979). On the psychology of prediction: whose is the fallacy? *Cognition, 7*, 385–407.

Cohen, L.J. (1988). The role of evidential weight in criminal proof. In P. Tillers & E.D. Green (Eds.), *Probability and inference in the law of evidence*. Amsterdam: Elsevier.

Cohen, L.J. (1989). *An introduction to the philosophy of induction and probability*. Oxford: Oxford University Press.

Cook, E.W., III, Hodes, R.L., & Lang, P.J. (1986). Preparedness and phobia: effects of stimulus content on human visceral learning. *Journal of Abnormal Psychology, 95*, 195–207.

Cook, M., Mineka, S., Wolkenstein, B., & Laitsch, K. (1985). Observational conditioning of snake fear in unrelated rhesus monkeys. *Journal of Abnormal Psychology, 94*, 591–610.

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., & Tooby, J. (1987). From evolution to behavior: evolutionary psychology as the missing link. In J. Dupre (Ed.), *The latest on the best: Essays on evolution and optimality*. Cambridge, MA: MIT Press.

Cosmides, L., & Tooby, J. (1989). Evolutionary psychology and the generation of culture, Part II. Case study: A computational theory of social exchange. *Ethology and Sociobiology, 10*, 51–97.

Cosmides, L., & Tooby, J. (1994). Origins of domain specificity: the evolution of functional organization. In L. Hirschfeld & S. Gelman (Eds.), *Mapping the mind: Domain-specificity in cognition and culture*. New York: Cambridge University Press.

Daston, L. (1980). Probabilistic expectation and rationality in classical probability theory. *Historia Mathematica, 7*, 234–260.

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

Eddy, D.M. (1982). Probabilistic reasoning in clinical medicine: problems and opportunities. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), *Judgment under uncertainty: Heuristics and biases*. Cambridge, UK: Cambridge University Press.

Fiedler, K. (1988). The dependence of the conjunction fallacy on subtle linguistic factors. *Psychological Research, 50,* 123–129.
Fisher, R.A. (1951). *The design of experiments* (6th ed.). New York: Hafner.
Freyd, J.J. (1987). Dynamic mental representations. *Psychological Review, 94,* 427–438.
Gallistel, C.R. (1990). *The organization of learning.* Cambridge, MA: MIT Press.
Garcia, J., & Koelling, R.A. (1966). Relations of cue to consequence in avoidance learning. *Psychonomic Science, 4,* 123–124.
Gelman, R. (1990a). Structural constraints on cognitive development: introduction to a special issue of *Cognitive Science. Cognitive Science, 14,* 3–9.
Gelman, R. (1990b). First principles organize attention to and learning about relevant data: number and the animate–inaminate distinction as examples. *Cognitive Science, 14,* 79–106.
Gelman, S., & Markman, E. (1986). Categories and induction in young children. *Cognition, 23,* 183–208.
Gigerenzer, G. (1990). Strong AI and the problem of "second order" algorithms. *Behavioral and Brain Sciences, 13,* 663–664.
Gigerenzer, G. (1991). How to make cognitive illusions disappear: beyond heuristics and biases. *European Review of Social Psychology, 2,* 83–115.
Gigerenzer, G., Hell, W., & Blank, H. (1988). Presentation and content: the use of base rates as a continuous variable. *Journal of Experimental Psychology: Human Perception and Performance, 14,* 513–525.
Gigerenzer, G., Hoffrage, U., & Kleinbolting, H. (1991). Probabilistic mental models: a Brunswikean theory of confidence. *Psychological Review, 98,* 506–528.
Gigerenzer, G., & Hug, K. (1992). Domain-specific reasoning: social contracts, cheating and perspective change. *Cognition, 43,* 127–171
Gigerenzer, G., & Murray, D. (1987). *Cognition as intuitive statistics.* Hillsdale, NJ: Erlbaum.
Gigerenzer, G., Swijtink, Z., Porter, T., Daston, L., Beatty, J., & Kruger, L. (1989). *The empire of chance: How probability changed science and everyday life.* Cambridge, UK: Cambridge University Press.
Hasher, L., & Chromiak, W. (1977). The processing of frequency information: an automatic mechanism? *Journal of Verbal Learning and Verbal Behavior, 16,* 173–184.
Hasher, L., & Zacks, R.T. (1979). Automatic and effortful processes in memory. *Journal of Experimental Psychology: General, 108,* 356–388.
Hintzman, D.L., & Stern, L.D. (1978). Contextual variability and memory for frequency. *Journal of Experimental Psychology: Human Learning and Memory, 4,* 539–549.
Hirschfeld, L., & Gelman, S. (Eds.) (1994). *Mapping the mind: Domain specificity in cognition and culture.* New York: Cambridge University Press.
Johnson-Laird, P.N. (1983). *Mental models: Towards a cognitive science of language, inference and consciousness.* Cambridge, MA: Harvard University Press.
Kahneman, D., Slovic, P., & Tversky, A. (Eds.) (1982). *Judgment under uncertainty: Heuristics and biases.* Cambridge, UK: Cambridge University Press.
Kahneman, D., & Tversky, A. (1972). Subjective probability: a judgment of representativeness. *Cognitive Psychology, 3,* 430–454.
Kahneman, D., & Tversky, A. (1973). On the psychology of prediction. *Psychological Review, 80,* 237–251.
Kahneman, D., & Tversky, A. (1982). On the study of statistical intuitions. *Cognition, 11,* 123–141.
Keil, F.C. (1989). *Concepts, kinds, and cognitive development.* Cambridge, MA: MIT Press.
Kolmogorov, A. (1950). *Foundations of probability theory* (N. Morrison, Trans.). New York: Chelsea.
Laplace, P.S. (1814/1951). *A philosophical essay on probabilities.* New York: Dover.
Leslie, A.M. (1987). Pretense and representation: the origins of "theory of mind". *Psychological Review, 94,* 412–426.
Leslie, A.M. (1988). The necessity of illusion: perception and thought in infancy. In L. Weiskrantz (Ed.), *Thought without language* (pp. 185–210). Oxford: Clarendon Press.

Marks, I.M. (1987). *Fears, phobias, and rituals*. New York: Oxford University Press.

Marr, D. (1982). *Vision: A computational investigation into the human representation and processing of visual information*. San Francisco: Freeman.

Maurer, A. (1965). What children fear. *Journal of Genetic Psychology, 106*, 265–277.

Maynard Smith, J. (1978). Optimization theory in evolution. *Annual Review of Ecology and Systematics, 9*, 31–56.

McCauley, C., & Stitt, C.L. (1978). An individual and quantitative measure of stereotypes. *Journal of Personality and Social Psychology, 36*, 929–940.

Meehl, P., & Rosen, A. (1955). Antecedent probability and the efficiency of psychometric signs, patterns, or cutting scores. *Psychological Bulletin, 52*, 194–216.

Nisbett, R.E., & Ross, L. (1980). *Human inference: Strategies and shortcomings of social judgment*. Englewood Cliffs, NJ: Prentice-Hall.

Pinker, S. (1984). *Language learnability and language development*. Cambridge, MA: Harvard University Press.

Pinker, S., & Bloom, P. (1990). Natural language and natural selection. *Behavioral and Brain Science, 13*, 707–784.

Profet, M. (1988). The evolution of pregnancy sickness as protection to the embryo against Pleistocene teratogens. *Evolutionary Theory, 8*, 177–190.

Profet, M. (1992). Pregnancy sickness as adaptation: A deterrent to maternal ingestion of teratogens. In J. Barkow, L. Cosmides, & J. Tooby (Eds.), *The adapted mind: Evolutionary psychology and the generation of culture*. New York: Oxford University Press.

Ramachadran, V.S. (1990). Visual perception in people and machines. In A. Blake & T. Troscianko (Eds.), *AI and the eye*. New York: Wiley.

Real, L.A. (1991). Animal choice behavior and the evolution of cognitive architecture. *Science, 253*, 980–986.

Real, L., & Caraco, T. (1986). Risk and foraging in stochastic environments: theory and evidence. *Annual Review of Ecology and Systematics, 17*, 371–390.

Shepard, R.N. (1984). Ecological constraints on internal representation: resonant kinematics of perceiving, imagining, thinking, and dreaming. *Psychological Review, 91*, 417–447.

Shepard, R.N. (1987). Evolution of a mesh between principles of the mind and regularities of the world. In J. Dupre (Ed.), *The latest on the best: Essays on evolution and optimality*. Cambridge, MA: MIT Press.

Shepard, R.N. (1992). The three-dimensionality of color: an evolutionary accommodation to an enduring property of the world? In J. Barkow, L. Cosmides, & J. Tooby (Eds.), *The adapted mind: Evolutionary psychology and the generation of culture*. New York: Oxford University Press.

Shiffrar, M., & Freyd, J.J. (1990). Apparent motion of the human body. *Psychological Science, 1*, 257–264.

Simon, H.A. (1956). Rational choice and the structure of the environment. *Psychological Review, 63*, 129–138.

Spelke, E.S. (1988). The origins of physical knowledge. In L. Weiskrantz (Ed.), *Thought without language* (pp. 168–184). Oxford: Clarendon Press.

Spelke, E.S. (1990). Principles of object perception. *Cognitive Science, 14*, 29–56.

Sperber, D. (1985). Anthropology and psychology: towards an epidemiology of representations. *Man (N.S.), 20*, 73–89.

Staddon, J.E.R. (1988). Learning as inference. In R.C. Bolles & M.D. Beecher (Eds.), *Evolution and learning*. Hillsdale, NJ: Erlbaum.

Symons, D. (1987). If we're all Darwinians, what's the fuss about? In C.B. Crawford, M.F. Smith, & D.L. Krebs (Eds.), *Sociobiology and psychology* (pp. 121–146). Hillsdale, NJ: Erlbaum.

Tooby, J., & Cosmides, L. (1990). The past explains the present: emotional adaptations and the structure of ancestral environments. *Ethology and Sociobiology, 11*, 375–424.

Tooby, J., & Cosmides, L. (1992a). The psychological foundations of culture. In J. Barkow, L. Cosmides, & J. Tooby (Eds.), *The adapted mind: Evolutionary psychology and the generation of culture*. New York: Oxford University Press.

Tooby, J., & Cosmides, L. (1992b). *Ecological rationality and the multimodular mind* (Tech. Rep. No. 92-1). Santa Barbara: University of California, Center for Evolutionary Psychology.

Tversky, A., & Kahneman, D. (1971). The belief in the "law of small numbers". *Psychological Bulletin, 76*, 105–110.

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: heuristics and biases. *Science, 185*, 1124–1131.

Tversky, A., & Kahneman, D. (1982). Evidential impact of base rates. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), *Judgment under uncertainty: Heuristics and biases*. Cambridge, UK: Cambridge University Press.

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

von Mises, R. (1957/1981). *Probability, statistics and truth* (2nd rev. English ed.). New York: Dover.

Zacks, R.T., Hasher, L., & Sanft, H. (1982). *Journal of Experimental Psychology: Learning, Memory, and Cognition, 8*, 106–116.