The Knowledge Engineering Review, Vol. 10:1, 1995, 21–41

Bias in human judgement under uncertainty?

PETER AYTON AND EVA PASCOE

Department of Psychology, City University, Northampton Square, London EC1V 0HB, UK.

Abstract

The claim is frequently made that human judgement and reasoning are vulnerable to cognitive biases. Such biases are assumed to be inherent in that they are attributed to the nature of the mental processes that produce judgement. In this paper, we review the psychological evidence for this claim in the context of the debate concerning human judgemental competence under uncertainty. We consider recent counter-arguments which suggest that the evidence for cognitive biases may be dependent on observations of performance on inappropriate tasks and by comparisons with inappropriate normative standards. We also consider the practical implications for the design of decision support systems.

1 Introduction

Over recent years, designers of expert systems as well as psychologists and management scientists have been developing tools and methods to aid, or even replace, human decision making. Although both communities are focused on similar problems, the tools and general frameworks for analysis tend to differ. Perhaps because of the rather separate disciplinary backgrounds of decision analysts and expert systems designers, two rather distinct and independent approaches have emerged. The decision analytic approach emphasized by management science relies heavily on quantitative estimation of values and calculation using normative decision rules, while expert systems have utilised predominantly qualitative representations of particular knowledge about causality, time and constraints.

A challenge for both approaches is the frequently cited claim that human judgement—even expert judgement—suffers from a number of characteristic biases and deficiencies (cf. Jacob et al., 1986; Ayton, 1992). For expert system designers this raises two particular issues. First, the validity of the general strategy of building systems by emulating experts can be called in to question (cf. Slatter, 1987). If the goal of constructing an expert system is to replicate an expert's decisions, then it would be necessary to model the expert's biases. However, if we wish to make accurate decisions then presumably efforts should be made to eliminate the biases. This in turn prompts a second, and more specific challenge: if human decisions are "plagued" with unconscious biases (cf. Jacob et al., 1986), how might biases in the knowledge base be avoided?

For the decision analytic approach, the problems posed by cognitive biases are no less severe. Subjective expected utility theory, as it is known, specifies that normative decisions can be prescribed on the basis of two independent sorts of information: subjective probabilities attached to the occurrence or non-occurrence of future events, and utilities of subjective values attached to the possible outcomes of the interplay between human actions and events at some time in the future (Edwards, 1954). In the 1960s this psychological decision theory was incorporated into the technique known as decision analysis and was pioneered by Raiffa (1968) and Schlaifer (1969) as a method to improve decision making. However, these business-school-based proponents of decision analysis were more concerned with the computational aspects of the technology which implemented subjective expected utility theory, and less with the subjective assessment of the inputs or probability and value.

The strategy of decision analysis is to decompose what may be very complex decisions involving many considerations into basic components. The decomposition rationale suggests that it prevents information overload of limited capacity human information processing and permits consideration of more factors than unaided intuition would allow. Using a decision tree the components are then evaluated using a normative rule so that the attractiveness of each option can be determined. The appeal of the procedure is that it represents a systematic attempt to analyse all relevant considerations and give them their appropriate weight. However, although computation may proceed normatively, the component inputs to the analysis will always depend—and usually to a large extent—upon judgements. If these judgements are corrupted by cognitive bias, then the end result cannot be held to be ideal.

The first significant evidence for deficiencies in expert judgement was Meehl's (1954) book, which evaluated clinical judgement. Meehl compared the intuitive clinical judgements made by experts (e.g. is this patient schizophrenic?) with those that could be made by a statistical formula using the same information. The statistical decisions were based on a "linear model". A linear model summarizes the relationship between a set of predictor variables and some criterion value—the outcome to be predicted. For example, if predicting the chances of survival from major surgery, relevant predictor variables may be the age, weight and general fitness of the patient. The linear model is constructed in such a way as to maximise the statistical relationship between the predictor variables and the criterion to be predicted. The value of each of the predictor variables is differentially weighted according to the strength of its diagnostic relationship to the criterion, and then all the variables are summed.

In approximately twenty studies which compared clinical decisions with statistical decisions, Meehl found that the statistical model provided more accurate predictions or the two models tied. Over the years since there have been many more studies comparing clinical and statistical judgement in an enormous range of areas of judgement. The superiority of the statistical method over clinical judgement has been replicated in all of these studies. Meehl (1986) commented: "There is no controversy in social science which allows such a large body of qualitatively diverse studies coming out so uniformly in the same direction as this one."

Despite this claim, the effect of the research on the practice of clinical judgement has been limited; according to Dawes (1988) it is "almost zilch". Dawes argues that this is because the findings are a challenge to the self-perceptions of experts. It is difficult for highly trained clinicians to accept that they cannot outperform a procedure which simply adds up the cues in favour of each judgement and picks the one with the highest score. This resistance may well be stiffened by the knowledge that the statistical method will not be perfect. There is evidence that resistance to the use of simple decision rules, which—given present knowledge—cannot be outperformed, increases with expertise and the importance of the decision (cf. Yates et al., 1991). Doctors may find it unacceptable to settle for the given number of errors implied by the statistical approach when they feel that their judgement might do better. Moreover, when statistical decision conflicts with the doctor's decision, the statistical decision may be seen as risky while their own judgement is seen as safe—quite opposite to the conclusion drawn from research.

So *why* are statistical decisions superior? Of course, the statistical approach relies on all the relevant evidence being coded in a quantitative fashion, something which may itself require considerable clinical skill but which would not ordinarily be performed in clinical situations. The statistical model will, moreover, utilize this evidence in an entirely consistent fashion. The statistical model will not be influenced by fatigue and boredom or distracted by spurious factors as human judgement is. A large number of studies have discovered inconsistency in expert medical judgements. Studies reviewed by Schwartz and Griffin (1986) have shown that doctors will show substantial disagreement with each other when interpreting chest X-rays, electrocardiograms and electroencephalograms, as well as more global quantities such as severity of depression. They will also sometimes disagree with their own previous judgements.

The studies using linear models say very little—if anything—about the cognitive processes underlying human reasoning. However, other claimed deficiencies of human decisions depend on

the notion that there are *inherent* biases which emerge that are not due to inattention or fatigue, but are quite characteristic of the mental processes used to make judgements. This claim has been most strongly made with regard to judgements and decisions made under uncertainty (cf. Kahneman et al., 1982). However, the general view that human judgement and reasoning is biased has become the subject of an intense debate (e.g. Cohen, 1981). In the following sections of this paper, we review the empirical evidence for the view that human judgement under uncertainty is biased. The evidence reporting bias in judgement is considered together with recent claims that the cognitive biases are due to an artifact in the design and methodology of the psychological research and inappropriate comparison with standards.

It is worth noting that, although there are difficulties for expert systems designers in modelling uncertain knowledge, most expert systems now have some sort of facility to cope with uncertainty. This is desirable given that experts are often uncertain regarding the assumptions they use or the conclusions they draw. The earliest attempt to deal systematically with uncertainty was MYCIN (Shortliffe & Buchanan, 1976) where conclusions would be qualified with certain factors. However, as the semantics of the resulting numbers was not well-defined (for example, the numbers are not Bayesian probabilities) their meaning, and hence their utility, is somewhat questionable.

Prospector (Duda et al., 1978) has a mechanism based on an approximate form of Bayes' theorem for assigning probabilities to conclusions. However, there are severe difficulties with applying probabilities to uncertain knowledge. The key problems encountered in applying a Bayesian framework are linked to the typical context of decision: for example, in the medical domain, diagnoses can involve several diseases. Traditional Bayesian schemes requiring mutual exclusivity of hypotheses (for example that each patient has only one disease) and conditional independence of evidence are not appropriate as a solution because these assumptions are often unrealistic (Henrion, 1990).[1]

Such difficulties with modelling uncertainty in expert systems might be taken as indicative of the complexities that confront human cognition and for which, on the face of it, we might expect it to be designed. If expert system designers have discovered difficulty in applying simple probability algorithms to drive knowledge-based inferences, we might expect that human cognition would be equipped to deal with uncertainty by methods other than application of probability laws. However, as we shall see, in order to facilitate the analysis of experimental subjects' performance, very often psychological researchers have devised tasks that are very simple in conception precisely so that the responses of subjects can be quite easily compared with what we might expect from simple laws of probability. This, of course, seems perfectly sensible; probability theory is concerned with the likelihoods of events. However, such a step has been held by some to be responsible for misconstruals of the nature and competence of human reasoning. There are those who argue that there is nothing really so very wrong with the human judgements underlying subjective probabilities (e.g. Gigerenzer, 1994) while others claim that there are observable inadequacies (e.g. Kahneman et al., 1982). What can be concluded from these arguments? We now turn to consider some of the evidence concerning the quality of subjective probabilities and consider the role of the standards that have been employed to study subjective probability.

2 Evaluating subjective probabilities

Subjective probabilities represent degrees of belief in the truth of particular propositions. For example, I may be able to say that I feel 50% sure that I have an appointment at the dentist's this afternoon, or 90% sure that the Suez canal is the longest in the world or 99% sure that a person

[1]More recent efforts using Bayesian network techniques (e.g. Pearl, 1988) have made some progress in dealing with the latter of these two difficulties through the modelling of conditional dependencies between separate pieces of evidence. However, modelling the causal links and specifying the conditional probabilities among them in a real knowledge domain is a non-trivial task.

tried in court for an offence is guilty. Such probabilities are subjective in the sense that they reflect an individual's assessment based on his or her knowledge and opinion.[2] People with different knowledge and beliefs will be perfectly entitled to offer different judgements of the likelihoods attached to the same propositions. For example, it is not possible using probability theory to calculate the probability that the Suez canal is the longest in the world. Indeed, unless we are dealing with random sampling from known populations, probability theory does not offer anyone the means to compute *the* probability of uncertain events. It follows that for any particular statement there is no "correct" subjective probability.

At first it might be supposed that this would pose an insurmountable problem for anyone interested in evaluating judgements of subjective probability, for if any value is as "correct" as any other who can say what a good judgement is? However there are ways by which *sets* of subjective probabilities are constrained by axioms of probability theory and also external correspondence to facts about the world. For example the additivity axiom states that the probabilities for a mutually exclusive and exhaustive set of events (e.g. the possible winner of a horse race) must add to one. Although, insofar as probability theory is concerned, an individual is entitled to believe anything he likes about the chances of each horse winning, the total of all the probabilities he estimates for each horse winning must add to one.[3] If the subjective probabilities produced by an individual conform to an axiom of probability theory then they are said to be coherent with respect to that axiom. The state of the world can also be referred to in order to measure how well *calibrated* a set of subjective probabilities are (cf. McClelland & Bolger, 1994). If the statements concern some verifiable aspect of the world then they can be checked for external correspondence (cf. Yates, 1982). For example, if I claim to be 70% sure about each of a whole set of statements being true then, to be well calibrated, 70% of them must in fact be true.

The relationship between the coherence and calibration properties of subjective probabilities is similar to that between reliability and validity in psychometric test design (Ayton & Wright, 1987). For example, a personality test designed to measure say, intelligence is said to be reliable if on different occasions it gives the same assessment of a group of individuals. Reliability is of interest as it is a pre-requisite for validity; a test that proved to be unreliable couldn't systematically be providing a valid measure of intelligence. However, as many critics of intelligence tests have pointed out, of itself, reliability is no guarantee that the test really is measuring intelligence—it might be measuring some other stable trait. The validity of a test has to be determined by comparing the test results with some other measure of intelligence. In a similar way coherence and calibration of subjective probabilities are interdependent. Subjective probabilities that are incoherent with respect to an axiom of probability theory cannot systematically be well-calibrated; they therefore cannot be taken as a guide to the relative likelihood of the events that they describe. However, subjective probabilities that are shown to be coherent are not necessarily well-calibrated.

Why should the quality of subjective probabilities be of any concern? In the classic decision analytic framework (see von Winterfeldt & Edwards, 1986; Goodwin & Wright, 1991), numerical probabilities are ascribed to all the different events identified in a decision tree. The best alternative is selected by combining the probabilities and the utilities corresponding to the possible outcomes associated with each of the possible alternatives. *Subjective* probabilities are thus one of the prime numerical inputs into decision analysis (Raiffa, 1968), cross-impact analysis (Dalkey, 1972), fault-tree analysis (Fischhoff, Slovic & Litchtenstein, 1978) and many other

[2]Savage (1954), introducing Bayesian ideas to the newly emerging field of decision analysis, proposed that these probabilities should be understood as a property of the person and for this reason suggested that they be known as personal probabilities. The expression has never really caught on in general usage which is perhaps a pity; for some, the term "subjective' automatically evokes negative connotations that the word "personal" would avoid.

[3]For the sake of simplicity, our example neglects those outcomes such as dead heats or races where no horses finished which could nonetheless, in principle, be included.

management technologies. This is because actuarial statistics concerning relative frequencies of the pertinent future events will often be unavailable or may be believed to be inappropriate for current circumstances. A decision-maker may realise for example that there have been changes in the world which have some causal impact on the events being judged. Such changes would invalidate statistical methods for calculating probabilities, for example by using regression or time-series methods based on averaging techniques. And, of course, one might take the decision analytic framework as a psychological model of (unaided) human choices. Perhaps people intuitively attempt some sort of expected utility analysis in order to make their choices; plainly, subjective expectations in this context would be crucial. For example, our decision about whether or not to go on a picnic might be strongly influenced by our estimate of the likelihood of good weather.

3 Biases in probability judgement?

3.1 Conservatism bias

So how good are judgmental probabilities? One early benchmark used for comparison was Bayes' theorem. Bayes' theorem defines mathematically how probabilities should be combined and can be used as a normative theory of the way in which subjective probabilities representing degrees of belief attached to the truth of hypotheses should be revised in the light of new information. In the 1960s, Ward Edwards and his colleagues conducted a number of studies using the book-bag and poker-chip paradigm. A typical experiment might involve two opaque bags. Each bag contained one hundred coloured poker-chips in different but stated proportions of red and blue. One contains 70 red chips and 30 blue while the second contains 30 red chips and 70 blue. The experimenter chooses one bag at random and draws a series of chips from it. After each draw, the poker-chip is replaced and the bag well shaken before the next chip is drawn. The subject's task is to say how confident they are—in probability terms—that the chosen bag is bag 1 or bag 2.

A crucial aspect of the logic of these studies is that the experimenter is able to say what the correct subjective probabilities should be for the subjects by the simple expedient of calculating them using Bayes' theorem. All of the information required as inputs to Bayes' theorem is explicit and unambiguous. Ironically enough, though, this meant that the *subjectivity* of probability was not a part of the studies in the sense that experimenters assumed that they could objectively compute the correct answer—which they would be able to assume should be the same for all the subjects faced with the same evidence.

The fact that the experimenter assumes he is able to calculate what the subjective probabilities *should* be for all the subjects was absolutely necessary if one was to be able to judge judgement. However, it is also an indication of the artificiality of the task—and is at the root of the difficulties that were to emerge with interpreting the subjects' behaviour. The experiments conducted with this procedure produced a good deal of evidence that human judgement under these conditions is not well described by Bayes' theorem. Although subjects' opinion revisions were proportional to the values calculated from Bayes' rule, they did not revise their opinions sufficiently in the light of the evidence, a phenomenon that was labelled *conservatism*. The clear suggestion was that human judgement was to this extent poor, although there was some debate as to the precise reason for this. It might be due to a failure to understand the impact of the evidence, or to an inability to aggregate the assessments according to Bayes' theorem. Aside from any theoretical interest in these possibilities, there were practical implications of this debate. If people are good at assessing probabilities but poor at combining them (as Edwards, 1968 suggested), then perhaps they could be helped; a relatively simple remedy would be to design a support system that took the human assessments and combined them using Bayes' theorem. However, if they were poor at assessing the component probabilities then there wouldn't be much point in devising systems to help them aggregate these. "Garbage in garbage out" used to be a popular aphorism for summarizing this sort of predicament.

3.2 The disappearance of conservatism bias

Before any firm conclusions were reached as to the cause of conservatism, however, the research exploring the phenomenon rather fizzled out (see Fischhoff & Beyth-Marom, 1973). The reasons for this seem to be twofold. One cause, considered in the next section, was the emergence of the heuristics and biases research and, in particular, the discovery of what Kahneman and Tversky (1973) called base-rate neglect. Before this development occurred, however, there was growing disquiet as to the validity of this sort of study as a model for judgement in the real world.

A number of studies had shown that there was considerable variability in the amount of conservatism manifested according to various quite subtle differences in the task set to subjects. For example, the diagnosticity of the data seemed an important variable. Imagine, instead of our two bags with a 70/30 split in the proportions of blue and red poker-chips, the bags contained 49 red and 51 blue or 49 blue and 51 red chips. Clearly, two consecutive draws of a blue chip would not be very diagnostic as to which of the two bags we were sampling from. Experiments have shown that the more diagnostic the information the less optimal is the subject. When the information is very weakly diagnostic, as in our example, human probability revision can be too extreme (Phillips & Edwards, 1966).

Another factor is the way in which the information is presented. Presenting the information about draws sequentially or all at once is irrelevant according to Bayes' theorem but Peterson et al. (1965) found that presenting the information one item at a time, with revisions after each item, were less conservative than those subjects who were given all the information in one go. Pitz et al. (1967) described an "inertia effect", where subjects tended not to revise their probabilities downward once the initial sequence of information had favoured one of the hypotheses under evaluation.

DuCharme and Peterson (1968) attempted to investigate probability revisions in a situation they considered nearer to real life than the standard paradigm. They argued that the fact that the information was restricted to one or two different possibilities (red chip or blue chip) meant that there were very few possible revisions that could be made. In the real world, information leading to revision of opinion doesn't have discrete values but may more fairly be described as varying along a continuum. In an experimental study, DuCharme and Peterson used a hypothesis test consisting of the population of male heights and the population of female heights. The subjects' task was to decide which population was being sampled from on the basis of the information given by randomly sampling heights from one of the populations. Using this task, DuCharme and Peterson found conservatism greatly reduced to half the level found in the more artificial tasks. They concluded that this was due to their subjects' greater familiarity with the data generating process underlying their task.

The argument concerning the validity of the conclusions from the book-bag and poker-chip paradigm was taken further by Winkler and Murphy (1973). Their paper, entitled "Experiments in the laboratory and the real world", argued that the standard task differed in several crucial aspects from the real world. First, the bits of evidence that are presented to the subjects are conditionally independent. That is, two or more pieces of information have an identical implication for the posterior probability to be credited to the hypotheses, regardless of the order in which they are produced. Knowing one piece of information does not change the likelihood of the other, producing one red chip from the urn and then replacing it does not affect the likelihood of drawing another red chip. However, in real world probability revision this assumption often does not make sense. Suppose I see a football supporter wandering along the street waving a blue scarf. This may well cause me to revise my opinions about which team may be visiting my home team; I would now be more confident that it would be a team that wears blue. However, the sight of another supporter also wearing a blue scarf will hardly change my views very much more—having seen the first blue scarf, the sight of another blue scarf is very much more likely.

For another example, consider a problem posed by medical diagnosis. Loss of appetite is a symptom which, used in conjunction with other symptoms, can be useful for identifying the cause

of certain illnesses. However, if I know that a patient is nauseous I know that they are more likely (than in the absence of nausea) to experience loss of appetite. These two pieces of information therefore are not conditionally independent, and so, when making my diagnosis, I should not rely on the loss of appetite symptom as much as I might, in the absence of nausea, to diagnose diseases indicated by loss of appetite. Winkler and Murphy argued that in many real world situations, lack of conditional independence of the information would render much of it redundant. In the standard tasks, subjects may have been treating the data as if it was conditionally dependent and so one possible explanation for conservatism is that the subjects are behaving much as they do in more familiar situations involving redundant information sources.

A second difference between experimental environments and reality was that in most experiments the data generators (the book-bags) are "stationary". The contents of the bags are fixed, but in reality our hypotheses are not always constant; indeed, the evidence may cause us to change the set of hypotheses under consideration. A third difference is that in reality the information may be somewhat unreliable, and therefore less diagnostic than the perfectly reliable colours of the poker-chips. In support of this argument, Yousseff and Peterson (1973) found that, when laboratory tasks included unreliable data, probability revision was less conservative. A fourth difference is that the typical experiments have offered very diagnostic evidence—clearly favouring one hypothesis—whereas in reality the evidence may very often be weakly diagnostic. Again, the result of generalizing from experience may be the appearance of conservatism. Recall that Phillips and Edwards (1966) found that probability revision can be too extreme with very weakly diagnostic evidence. In summary, the arguments considered by Winkler and Murphy led them to conclude that "conservatism may be an artifact caused by dissimilarities between the laboratory and the real world."[4]

An early supposition from this line of experimentation was that subjective probabilities were inappropriate by virtue of the observed discrepancies with the probabilities derived from the Bayesian normative standard. However, over the decade of research that we have just described, a curious reversal of this conclusion was arrived at; now the normative standard is considered inappropriate, and thereby subjective probabilities may be appropriate. As we shall see, the nature of contemporary criticisms of more recent research, which documents suboptimality in other aspects of probabilistic judgement, suggests that this cycle is about to repeat itself: once again, it is being claimed that inappropriate presumption lies behind the conclusion that poor judgement is responsible for the experimentally observed disparities between judgement and the normative standard (e.g. Gigerenzer 1994: Beach & Braun, 1994). The notion that judgement is valid and that poorly chosen experimental tasks have led to a misconceived view of poor human capability has strong advocates.

3.3 Base-rate neglect

In reporting studies of peoples' intuitions of random sampling, Tversky and Kahneman (1971) commented, in a footnote, that their respondents ". . . can hardly be described as conservative. Rather . . . they tend to extract more certainty from the data than the data, in fact, contain." Their discovery of base-rate neglect—the antithesis of conservatism—seems to have been the final nail in the coffin for the hypothesis that we are all conservative Bayesians. In Kahneman and Tversky's (1973) experiments, demonstrating neglect of base-rates, subjects were found to ignore information concerning the prior probabilities of the hypotheses. For example, in one study subjects

[4]Edwards and his colleagues actually conducted many of their experiments not using book-bags at all but a display consisting of 48 numbered locations each containing a push button and a red and green light. On pushing the button, one of the lights was illuminated. Subjects are told that this is equivalent to sampling a chip of the corresponding colour from a book-bag. The program was carefully prepared so that the sample would be representative of the book-bag being sampled. It is possible that this 'artificial' book-bag might induce subjects to respond in ways that they might not if they could see a real book-bag being sampled (cf. Winefield, 1966; Gigerenzer et al., 1988).

were presented with this brief personal description of an individual called Jack and told that the description was drawn at random from those of seventy engineers and thirty social scientists.

> *Jack is a 45-year old man. He is married and has four children. He is generally conservative, careful and ambitious. He shows no interest in political and social issues and spends most of his free time on his many hobbies which include home carpentry, sailing and mathematical puzzles.*

Half of the subjects were told that the description had been drawn from a sample of seventy engineers and thirty lawyers, while the other half were told that the description was drawn from a sample of thirty engineers and seventy lawyers. Both groups were asked to estimate the probability that Jack was an engineer (or a lawyer). The mean estimates of the two groups of subjects were only very slightly different (50% vs. 55%). On the basis of this result and others, Kahneman and Tversky concluded that prior probabilities are largely ignored when individuating information was made available.

Although subjects used the base-rates when told to suppose that they had no information whatsoever about the individual (a "null description"), a description designed to be totally uninformative with regard to the profession of an individual called Dick produced complete neglect of the base-rates.

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

When confronted with this description, subjects in both base rate groups gave median estimates of 50%. Kahneman and Tversky concluded that when no specific evidence is given, the base-rates were properly utilized; but when worthless information is given base-rates were neglected.

Kahneman and Tversky's theoretical approach to judgement assumes that, because human information processing capacity is limited, people do not judge under uncertainty using computationally ideal strategies. Instead they use mental *heuristics*—rules of thumb—to judge uncertainties (Kahneman et al., 1982). Although within artificial intelligence research the term *heuristic* appears not to have any negative connotation, the association of heuristics with bias within cognitive psychology has led to a certain denigration of heuristic reasoning. It is perhaps rather telling that, although heuristics are avidly sought out and seized upon by computer scientists to produce artificial intelligence (e.g. Lenat, 1982), within cognitive psychology heuristics are more often invoked to account for real stupidity.

One such heuristic is *representativeness*. This heuristic determines how likely it is that an event is a member of a category according to how similar or typical the event is to the category. For example, people may judge the likelihood that a given individual is employed as a librarian by the extent to which the individual resembles a typical librarian. This may seem a reasonable strategy, but it neglects consideration of the relative prevalence of librarians. When base-rates of different categories vary, judgements may be correspondingly biased.

Another heuristic used for probabilistic judgement is *availability*. This heuristic is invoked when people estimate likelihood or relative frequency by the ease with which instances can be brought to mind. Instances of frequent events are typically easier to recall than instances of less frequent events, so availability will often be a valid cue for estimates of likelihood. However, availability is affected by factors other than likelihood. For example, recent events and emotionally salient events are more easy to recollect. It is a common experience that the perceived riskiness of air travel rises in the immediate wake of an air disaster. Judgements made on the basis of availability then are vulnerable to bias.

The *anchor and adjust* heuristic is used when people make estimates by starting from an initial value that is adjusted to yield a final value. The claim is that adjustment is typically insufficient. For instance, one experimental task required subjects to estimate various quantities stated in percentages (e.g. the percentage of African countries in the UN). Subjects communicated their answers by using a spinner wheel showing numbers between 0 and 100. For each question the wheel was spun

and then subjects were first asked whether the true answer was above or below this arbitrary value. They then gave their estimate of the actual value. Estimates were found to be considerably influenced by the initial (entirely random) starting point.

The phenomenon of base-rate neglect was attributed to the operation of the representativeness heuristic. Subjects were not judging by any kind of statistical reasoning, but were judging the probability of the profession by the extent to which the descriptions were similar to the stereotype of the profession. Kahneman and Tversky argued that people did not engage in statistical reasoning as such but instead invoked heuristics such as representativeness and availability to judge uncertainties. Their 1972 paper on representativeness argued that the Bayesian approach to the analysis and modelling of subjective probability did not capture the essential determinants of the judgement process. This was because "In his evaluation of evidence man is apparently not a conservative Bayesian: he is not Bayesian at all".

The literature reporting investigations of expert judgement provides several instances of poor judgement or faulty reasoning—some of which have serious potential consequences. Eddy (1982) reports alarming evidence of fundamental errors in the probabilistic reasoning employed by physicians to make diagnoses of breast cancer on the basis of X-rays. X-rays can give an indication as to whether or not a lesion is malignant. X-rays are used as a basis for making decisions as to whether or not surgery, involving the removal of tissue for further examination, is merited. However, the indication from the X-ray test is not perfectly reliable; some malignant lesions will be incorrectly classified as benign and some benign lesions will be classified as malignant. Consequently, the task of diagnosis can be viewed as a process of statistical inference.

Eddy set a sample of physicians the task of estimating the likelihood that a patient had cancer given that, prior to the X-ray, their examination of the patient indicated a 99% probability that the lesion was benign but that the X-ray test was positive and had indicated it was malignant. They were told that research into the accuracy of the test showed that 79.2% of malignant lesions were correctly diagnosed and 90.4% of benign lesions were correctly diagnosed by the test.

As 90.4% of the patients with benign lesions will be correctly diagnosed, 9.6% will not be, and thereby show positive on the test, along with 79.2% of those that do not develop cancer. But of course, there are many more people who do not develop cancer than those who do. Consequently, the population of people who show positive on the test consists of a small proportion (9.6%) of the vast majority of people who do not develop the disease (99%), plus a large proportion (79.2%) of the tiny minority of the people who do develop the disease (1%).

Bayes' theorem can be used to combine the information about the reliability of the test with the physicians prior probability judgement to assess the correct likelihood that the patient has cancer. Substituting the values into the formula tells us that the likelihood of cancer in the light of this information [p(cancer/positive)] is just under 8%. However, most of the physicians misinterpreted the information about the reliability of the test and estimated the likelihood of cancer to be about 75%. When asked about their reasoning the physicians report that they assumed that the probability of cancer given a positive test result [p(cancer/positive)] is equal to the probability of a result of a positive X-ray in a patient with cancer [p(positive/cancer)]. They can therefore be said to have used a representativeness heuristic in that they judged the likelihood of cancer in patients with a positive test in terms of how typical (or representative) they were of patients with cancer. It is clear that they failed to properly consider the impact on the outcome of the very low incidence of the disease (base-rate) together with the tendency of the test to (falsely) show positive test results.

3.4 The disappearance of base-rate neglect

Later research established that base rates might be attended to more (though usually not sufficiently) if they were perceived as relevant (Bar-Hillel, 1980) had a causal role (Kahneman & Tversky, 1982b) or were "vivid" rather than "pallid" (Nisbett & Ross, 1980). However, Gigerenzer et al. (1988) have argued that the real reason for variations in base-rate neglect is nothing to do with any of these factors *per se*, but because the different tasks may to varying degrees encourage the

subject to represent the problem as a Bayesian revision problem. They claim that there are few inferences in real life that correspond to Bayesian revision where a known base-rate is revised on the basis of new information. Just because the experimenter assumes that he has defined a Bayesian revision problem does not imply that the subject will see it the same way. In particular, the subjects may not take the base-rate asserted by the experimenter as their subjective prior probability. In Kahneman and Tversky's original experiments the descriptions were not of course actually randomly sampled (as the subjects were told) but especially selected to be "representative" of the professions. To the extent that the subjects suspected that this was the case, then they would be entitled to ignore the offered base-rates.

In an experiment, Gigerenzer et al. (1988) found that when they let the subjects experience the sampling themselves base-rate neglect disappeared. In the experiment their subjects could examine the ten pieces of paper, each marked lawyer or engineer (according to the base rates). Subjects then drew one of the pieces of paper from an urn and it was unfolded so they could read a description of an individual without being able to see the mark defining it as being of a lawyer or engineer. In these circumstances, subjects clearly used the base-rates in a proper fashion—including for the "uninformative" description Dick.[5] However, in a replication of the verbal presentation where base-rates were asserted, rather than sampled, Kahneman and Tversky's base-rate neglect was replicated.

Cosmides and Tooby (in press) report similar experiments to those performed by Eddy (described above), replacing all reference to probabilities with descriptions of the frequencies of the events. Thus, a probability of 5% would be referred to as "50 out of 1000". In addition, rather than estimate the probability that a particular patient had the disease, subjects were asked how many, of a sample of 1000 who tested positive, would have the disease. Under these circumstances the vast majority (92%) of answers given by subjects are in accord with answers given by Bayes' theorem.

If people had a difficulty in reasoning with uncertainty *per se*—rather than merely with single event probabilities–then the transformation into a relative frequency format would not make any difference. However, the limited number of experiments that have been reported using relative frequency to represent statistical information (see Gigerenzer, 1994) show that people appear to reason quite satisfactorily, and certainly far more competently than might be assumed from the large number of studies that have tested abilities to reason with probabilities.

3.5 The gambler's fallacy

One of the most widely known of the phenomena in subjective probability judgement—the gambler's fallacy—also "disappears" in similar contexts. The gambler's fallacy acquired its name after Dostoyevsky's (1866/1966) observation that players of roulette falsely assume that, after a given number has come up, it is much less likely to occur next time. This phenomenon is also known as negative recency. In studies of the ability to generate or recognize random sequences, the typical finding reported in the psychological literature over the past thirty years has been that subjects show a degree of negative recency. That is, subjects appear to believe that alternations of the elements of a random sequence are more likely than they really are; by the same token, subjects under-estimate the degree of repetition that there is. Ayton, et al. (1989, 1991) have discussed the possible reasons for this observation. They noted that there are very few occasions when people

[5]In their experiment two, Gigerenzer, Hell and Blank (1988) found Bayesian conservatism in their subjects. This apparent re-appearance of conservatism occurred in subjects asked to predict the final scores of football matches given the half-time scores. Subjects were unable to do so optimally; they underestimated the extent to which the half-time score predicted the result and tended to give too much emphasis to their estimate of the prior strength of the teams. By their own argument, this is a task with which the subject would be very familiar, and yet they did not indulge in the correct amount of opinion revision. It is not clear whether such a result constitutes evidence for cognitive bias.

would encounter an assuredly random process, and that consequently the apparent bias might actually be a function of generalisation from experience of encounters with non-random sequences. There is also evidence that the effect is extremely sensitive to subtle variations in the instructions given to subjects. A study by Winefield (1966) showed that if subjects had to guess the suit of a card drawn from the deck the usual measure of negative recency disappeared *if they could see the card being placed back in the deck and the deck being well shuffled.* If this was not the case then they continued to display negative recency. The experiment seems very similar to rationale and findings to that of Gigerenzer's et al. (1988) test of the robustness of base rate neglect.

Recently, a number of authors have reported that negative recency varies quite considerably with the task and that tasks can be devised where the subjects perform quite well. A number of authors have found that using an instructional set that avoids reference to probabilistic concepts of chance or randomness will improve performance. For example, Finke (1984) found that subjects asked to produce responses that are as unpredictable as possible produce responses which more closely approached the frequency of repetitions expected by chance than do those subjects asked to produce responses as randomly as possible. If, in a competitive game, subjects are motivated to produce unpredictable sequences of binary responses they can do so in a way that satisfies standard tests of randomness more successfully (Rappoport & Budescu, 1992). Kareev (1992) reports data that also show variability of performance according to task and concludes that people have a basically correct notion of randomness, with apparently non-random behaviour being the result of attempts by a capacity limited information processing system to optimise performance given its interpretation of the standard tasks. In the words of the title of Kareev's paper, maybe human randomness is not so bad after all.

There is also evidence of some confusion about the normative standards used to judge human conceptions of randomness. Curiously the same basic concept (representativeness) is often used both to define the basis for objective "tests for randomness" and to explain why it is that subjects deviate from this standard. The statistical tests commonly assume that a random sequence should contain a *representative* sample of all the possible configurations (e.g. all the possible pairs of outcomes resulting from tossing a fair coin: HT: HH: TH: TT). However, subjects are typically berated when they can be assumed to be applying the same heuristic. Indeed, negative recency has often been assumed to result from the application of representativeness. Subjects have commonly been judged to be suffering from the misconception that even small samples of random output should contain a (representative) number of the basic elements, and should also show disorder and absence of any obvious "patterns". In the light of this contradiction it is perhaps not altogether surprising that the instructions given to experimental subjects by experimental psychologists therefore have sometimes appeared a little confusing, and possibly, in terms of the negative recency hypothesis, rather self fulfilling. Subjects have sometimes been explicitly told that they should produce responses which appear "jumbled" or that don't contain any patterns. Aside from being a potential source of the inappropriate "gambler fallacy", such instructions suggest that the psychologists have been a little uncertain as to quite what the objectives are for their subjects with this task.

3.6 Overconfidence bias

In the 1970s and 1980s, a considerable amount of evidence was marshalled for the view that people suffer from an overconfidence bias. Typical laboratory studies of calibration ask subjects to answer questions such as:

"Which is world's longest canal? (a) Panama
(b) Suez"

Subjects are required to indicate the answer that they think is correct and then state how confident they are on a probability scale ranging from 50% to 100% (as one of the answers is always correct 50% is the probability of guessing correctly). To be well calibrated, assessed probability should

be equal to the percentage correct over a number of assessments of equal probability. For example, if you assign a probability of 70% to each of ten predictions then you should get seven of those predictions correct. For a full review of this aspect of probabilistic judgement see McClelland and Bolger (1994).

Overconfidence of judgements made under uncertainty is commonly found in calibration studies (see Lichtenstein et al., 1982), and has been recorded in the judgements of experts. For example, evidence that physicians overconfidently diagnose is provided by Christensen-Szalanski and Bushyhead (1981), who explored the validity of the probabilities given by physicians to diagnoses of pneumonia. They found that the probabilities were poorly calibrated and very overconfident; the proportion of patients who turned out to have pneumonia was far less than the probability statements implied. These authors had previously established that the physicians' estimates of the probability of a patient having pneumonia was significantly correlated with their decision to give a patient a chest X-ray and to assign a pneumonia diagnosis.

Wagenaar and Keren (1986) found overconfidence in lawyers' attempts to anticipate the outcome of court trials in which they represented one side. As they point out it is inconceivable that the lawyers do not pay attention to the outcomes of trials in which they have participated, so why don't they learn to make well-calibrated judgements? Nonetheless, it is possible that the circumstances in which the lawyers, and other experts, make their judgements, and the circumstances in which they receive feedback, combine to impede the proper monitoring of feedback necessary for the development of well calibrated judgements. A consideration of the reports of well calibrated experts supports this notion; they all appear to be cases where some explicit unambiguous quantification of uncertainty is initially made and the outcome feedback is prompt and unambiguous.

The most commonly cited example of well calibrated judgements are weather forecasters' estimates of the likelihood of precipitation (Murphy & Winkler, 1984) but there are a few other cases. Keren (1987) found highly experienced tournament bridge players (but not experienced non-tournament players) made well calibrated forecasts of the likelihood that a contract, reached during the bidding phase, would be made, and Phillips (1987) reports well calibrated forecasts of horse races by bookmakers. In each of these three cases, the judgements made by the experts are precise numerical statements and the outcome feedback is unambiguous and received promptly, and so can be easily compared with the initial forecast. Under these circumstances, the experts are unlikely to be insensitive to the experience of being surprised; there is very little scope for neglecting, or denying, any mismatch between forecast and outcome.

It seems quite likely that many experts operate in conditions where such an analysis is difficult or impossible. Sometimes the outcomes will be in the remote future or may be attributable to interventions (perhaps by the expert) designed to change the outcomes. Thus doctors may take actions which change the outcomes implied by diagnosis, and thereby remove feedback which might confirm or refute the diagnosis. In any case, as doctors do not typically record precise numerical indices of their uncertainty, they may well forget what their original judgements were by the time outcome information is available.

3.7 Overconfidence-disappears

The overconfidence phenomenon has often been explained as a characteristic of human information processing. Some researchers have implicated the operation of the anchor and adjust heuristic (Keren, 1991; Ferrell & McGooey, 1986); ignorance of processing limitations (Pitz, 1974); motivation (Milburn, 1983); cognitive optimism (Dawes, 1980); and response scale effects (Poulton, 1989).

However, "ecological" theorists (cf. McClelland & Bolger, 1994) claim that overconfidence is essentially an artifact due to the use of artificial experimental tasks and the non-representative sampling of stimulus materials. Gigerenzer et al. (1991) and Juslin (1994) argue that individuals are well adapted to their environments and do not make biased judgements. Overconfidence is

observed because the typical general knowledge quiz used in most experiments contains a disproportionate number of misleading items.

These authors have found that when knowledge items are randomly (representatively) sampled the overconfidence phenomenon disappears. For example, Gigerenzer et al. (1991) presented their subjects with items generated with random pairs of the German cities with more than 100,000 inhabitants and asked them to select the biggest and indicate their confidence they had done so correctly. With this randomly sampled set of items there was no overconfidence.

Moreover, with conventional general knowledge quizzes subjects *are* aware of how well they are likely to perform overall. Gigerenzer et al. found that subjects are really quite accurate at indicating *the proportion of items* that they have correctly answered. Such quizzes are representative of general knowledge quizzes experienced in the past. Thus, even when they appear overconfident with their answers to the individual items, subjects are not overconfident about their performance on the same item as a set. As a corollary, for randomly sampled items, subjects' judgements of overall proportion correct are underestimates.

4 Is human judgement under uncertainty probabilistic, heuristic or frequentistic?

As we noted, Gigerenzer (1994) discusses several cases where, with frequentist representations, putative fallacies of probabilistic judgement disappear and attributes this to the idea that subjects are better equipped to process information concerning frequencies of events than they are single event probabilities. An evolutionary speculation supports the argument: in the course of their evolutionary development, humans (and perhaps other creatures) have acquired the means to effectively represent and manipulate information about frequencies of events—but not their probabilities. After all, probability theory has, relative to the evolutionary scheme, only very recently emerged as a means by which to represent and communicate information. According to Hacking (1975) what we now recognize as the mathematical calculus of probability theory was only formalized in the seventeenth century. It would not be altogether surprising then if human cognition did not naturally compute probabilities—but instead used stored frequencies of events. Consequently, requesting subjects to give their responses to problems in the form of probabilities may be asking them to understand and speak a rather foreign language. Just as it would be unrealistic to expect one's pocket calculator to accurately compute the answer to arithmetic problems entered with roman numerals, it may be unreasonable to judge the general competence of human judgement under uncertainty on the performance of problems requiring the assessment of subjective probabilities rather than frequencies.

Kahneman and Tversky, and others (see Kahneman et al., 1982), assumed that mental heuristics were utilized to reduce the complex tasks of assessing probabilities to simpler judgmental operations. Such an assumption followed very naturally from Simon's (1957) proposal of bounded rationality—the notion that, because of their limited information processing capacities, people don't use optimal methods for reasoning but instead take short cuts, or "satisfice", in order to produce judgements and decisions accurately and efficiently. Simon's work was concerned with the simplifying mental strategies that reduced the complexity of tasks to make them manageable by the kinds of minds that people actually have. However, Gigerenzer's view of reasoning under uncertainty is based on the premise that the human mind has no need to make use of such heuristic judgements *for assessing probabilities*. The claim is that people's memory for frequency is reliable enough for them to utilize past records of events as statistics for their judgements. There is, in fact, considerable evidence that human memory for frequency information is automatically encoded and extremely accurate. The evidence has been well summarized by Hasher and Zacks (1979; 1984), who concluded that experiments "reliably and unequivocally demonstrate remarkable knowledge of the frequency of occurrence of all events so far tested." (1984, p 1373).

Hasher and Zacks (1984) were aware that their views about the veracity of the storage of frequency information might be seen as in conflict with the evidence from Tversky and Kahneman (1973) that people might use an availability heuristic for judging frequency. Tversky and

Kahneman (1973) reported a number of experiments which suggested that people judged likelihood or frequency by the ease with which instances of the event could be brought to mind. Instances of frequent events are typically easier to bring to mind than instances of less frequent events, so the availability heuristic would often be valid. However, availability and frequency are not always perfectly correlated. For example, an experiment showed that subjects were more likely to judge that the letter R was more likely to appear in the first position of a word than in the third position. In fact, the letter R occurs more often in the third position of a word than the first, and Tversky and Kahneman attributed the error to the fact that words beginning with R are easier to recall.

Hasher and Zacks (1984) comment that: ". . . the conflict between our view and that of Tversky and Kahneman is more apparent than real. First of all, in most instances frequency and availability (like frequency and probability) are highly correlated: More frequent events are, other things being equal, more recallable or "available" than less frequent events. In such situations, any biasing effects of the availability heuristic will not be seen. Use of availability will bias frequency estimates most clearly when the retrieval cue (i.e. the event to be judged) is a weak one (as in the example of words having a particular letter in the middle position). It is worth noting, in addition that in other illustrations of the availability heuristic the actual frequency differentials are small (e.g. 19 versus 20) and the countervailing effects of other variables (e.g. stimulus familiarity) are strong. Such is the case in Tversky and Kahneman's experiment in which subjects misjudged as being more frequent 19 famous names scattered in a list that also included 20 non-famous names." (pp. 1383–1384.)

Hasher and Zacks's comment that there is no real conflict in their frequency proposal and Kahneman and Tversky's heuristic model of judgement cannot be assumed to apply to those who argue that judgements of likelihood are made by reference to memorised frequencies. Either a particular judgement is based on stored frequencies or it is heuristic.[6] There are some experiments reported in the literature which attempt to unconfound the correlation between frequency and probability to determine if judgement relies on heuristics or stored frequencies. For example, Estes (1976) reported a study which led him to conclude that the basis for predictive behaviour is not a probability estimate, but rather a record in memory of the past frequencies of events. In a simulated opinion poll, on each of a series of trials, subjects observed the outcome of a mini-poll contrasting pairs of alternatives (e.g. two political candidates). The exposure sequence was so designed that in some cases one candidate had a lower probability of "winning" than another but, because the candidate appeared more often, a higher overall frequency of winning. When two such candidates were paired directly, subjects picked as the likely winner the candidate whose absolute frequency of winning was higher.

In summarizing the probability learning literature, Estes commented: ". . . the results suggest that the term "probability learning is in a sense a misnomer. I have found nothing to encourage the tendency to think of probability learning as a basic or unitary process or as a direct manifestation of

[6]However, it may be that there are different mechanisms which underlie different judgements. Heuristics are very successful at accounting for the responses produced by some tasks while the notion that people can generate probabilities from stored frequencies is helpful for explaining performance on other tasks. Tiegen (1994) argues that subjective judgements of probability can be arrived at by a number of different processes. These processes may be at variance with one another or some standard. For example, if Tom is competing for a job with four other equally qualified candidates, his chances will be estimated as about 20%. However, when asked to select the most appropriate verbal phrase describing his chances, most will prefer high probability expressions (e.g. "he has a good chance") to low probability expressions (e.g. "it is doubtful"). However few people consider 20% as a "good chance". The suggestion is not so much that people are inconsistent, but that different concepts of subjective probability are invoked under the different circumstances. It may well be that people have a range of strategies for coping with uncertainty, and these are selected depending on the precise task demands (cf. Payne, 1982). Fox (1994) has emphasized that there may well be a range of alternative theories that could be developed for probability and proposes a non-numerical calculus of *argumentation* for reasoning about uncertainty.

a capacity for perceiving the statistical structure of sequences of events. The subjects clearly are extremely efficient at acquiring information concerning relative frequencies of events." (p.51).

Other research is less supportive of the notion that human judgements under uncertainty rely on stored frequencies and not heuristics. A number of studies have shown that mere repetition of the presentation of large sets of statements causes the subjective degree of belief in their validity to increase compared to non-repeated control statements (e.g. Hasher et al., 1977; Gigerenzer, 1984). This phenomenon has been attributed to the automatic and accurate encoding of frequency information (e.g. Gigerenzer et al., 1991). However, two studies suggest a heuristic account for the effect. Bacon (1979) demonstrated that higher levels of rated validity occurred for statements that subjects judged to be repeated—whether they had been or not. Arkes et al. (1989) found that the effect did not generalize to all repeated statements. Specifically, they found that statements concerning topics with which their subjects were unfamiliar did not increase in perceived validity with repetition. Arkes et al. concluded that the repetition-validity effect was attributable to a familiarity heuristic like availability. Note however that it is possible to argue that Bacon's and Arkes' et al. results are due to effects on memory for frequency (cf. Estes, 1976), and that therefore an availability heuristic explanation may not be required.

With regard to the third general heuristic discussed by Tversky and Kahneman—the anchor and adjust heuristic—we know of no evidence or argument that specifically militates against its existence. There is of course a large body of research that has found evidence for its operation. A recent study of judgmental forecasting by Bolger and Harvey (1993) finds that the general anchor and adjust heuristic, operating in different forms depending on context, was very useful for explaining their subjects' responses. Their subjects were required to make forecasts of future data points given previous ones in the same series, and appeared to alter their forecasting strategy depending on the presence and absence of trends and serial dependency. Bolger and Harvey questioned whether such an account might conflict with Gigerenzer's more "ecological" account of statistical reasoning. They suggest that, although we might not have evolved to perform the type of judgmental extrapolations required of their subjects, their tasks are now ecologically valid ones. Among business people, judgmental extrapolation is the most popular method of forecasting.

The frequentist approach adopted by Gigerenzer argues strongly against the operation of mental heuristics and biases but there is, as yet, for those choosing to adopt a heuristic approach to judgement, plainly still scope for invoking general heuristics to account for judgmental behaviour in situations where a frequentist representation is inappropriate, i.e. the assessment of subjective probabilities for *unique* one-off events. However, the experimental evidence of fallacies of subjective probability, usually attributed to the operation of the representativeness heuristic, is obviously compromised by the disappearance of these effects when subjects are able to contemplate uncertainty from a frequentist perspective.

We noted that Kahneman and Tversky suggested that people are not conservative Bayesians but judge and reason with probabilities using mental heuristics. However, Gigerenzer argues that people do not use heuristics when experimental problems are re-cast into a relative frequency paradigm and, indeed, are not equipped to reason about uncertainty using single-event probabilities at all—but, nonetheless, they can reason successfully about uncertainty with frequencies. These frequency estimates can, under appropriate task conditions, be translated into a valid probability metric. We should note though one emergent point of consensus in this dispute about human inference under uncertainty. Gigerenzer's conclusion from experiments that test subjects with frequentist versions of Kahneman and Tversky's problem is, in one sense at least, strikingly similar to that of Kahneman and Tversky's. All parties would appear to agree that human reasoning under uncertainty is ". . . not Bayesian at all."

5 Likelihoods of unique events

The evidence for the use of heuristics in judgements of likelihood is based on tasks where subjects were required to estimate the likelihood of single events. How do they do this? According to the

frequentist statistician the task is nonsense. Probability only applies to the relative frequency of events and it makes no sense to consider the probability attached to the truth of a single statement which is either true or not. Plainly, though, people do feel different degrees of confidence in the truth of individual propositions and, on the face of it, don't object to providing descriptions of their confidence in terms of probability. From the standpoint of the Bayesian statistician, there is no reason to discourage this practice. The difficulty according to Gigerenzer is that human information processing simply isn't suited to the task. Nonetheless, it follows that difficulties may arise if subjects are asked to assess veridical probabilities for single events, or assume that they can do so.

We could ask ourselves why it is that subjects produce responses that are so well predicted by the heuristic approach. For example, in one of their studies, Kahneman and Tversky (1982b) asked one group of subjects to judge the representativeness of personality descriptions with respect to a whole series of different professions. A separate group of subjects rated the likelihood that each of the described individuals really was a member of each of the listed professions. The correlation between the two was 0.96; plainly, the judgements of likelihood were quite indistinguishable from those of representativeness. It would seem that there is a danger that when asked to assess subjective probabilities for single events subjects will report a measure of representativeness or availability.

Kahneman and Tversky (1979) explained that one way to avoid the biases of subjective probability implied by the heuristic account was to take an external rather than an internal view, by contemplating the target event in relation to a reference class of similar events and considering the distribution of likelihoods for the whole class of events. The strategy looks very much like a way of attempting to invoke a frequentist set for judging likelihood. (Indeed, Tversky & Kahneman (1983) themselves found evidence that the conjunction fallacy was largely eliminated when subjects were presented with the problem expressed with frequencies rather than percentages.) This analysis is amplified and extended by Kahneman and Lovallo (1993), who argue that people have a strong tendency to see problems as unique when they would be more advantageously viewed as instances of a broader class. They claim that the natural tendency in thinking about a particular problem, such as the likelihood of success of a business venture, is to take the "inside" rather than the "outside" view. People will pay particular attention to the distinguishing features of a particular case and reject analogies to other instances of the same general type as crudely superficial and unappealing. Consequently, they will fall prey to fallacies of planning—anchoring their estimates on present values or extrapolations of current trends.

Once a forecaster takes the inside view they will not seek out relevant statistical knowledge, will be less likely to formulate a realistic estimate and will be overconfident about their forecasts. It would seem then that proponents of the heuristic view are persisting with a pessimistic view about human judgement of likelihood. However, in advising that anyone attempting to assess probabilities should take an "outside view", it also seems that there is very little in practical terms that separates the advocates of the heuristic and the ecological frequentist approaches in terms of their attitudes to the quality of subjective probabilities. For example, Kahneman and Lovallo review evidence which suggests that, because they take an inside view, people can be unrealistically optimistic or, if failure is easier to imagine, pessimistic). They cite a study by Cooper et al. (1988), which showed that entrepreneurs interviewed about their chances of business success produced assessments that were unrelated to objective predictors such as college education, prior supervisory experience and initial capital. Moreover, more than 80% of them described their chances as 70% or better, whilst the survival rate for new businesses is as low as 33%. Such findings might be taken as evidence for poor judgement under uncertainty—or alternatively, as evidence that people are better off not attempting to assess probabilities for single events.

If we compare studies of the calibration of probability assessments concerning individual unique events (e.g. Wright & Ayton, 1992) with those where assessments have been made for repetitive predictions of weather events, e.g. rain (see Murphy & Brown, 1985), we can observe that relatively poor calibration has been observed in the former, whereas relatively good calibration has been observed in the latter. Bolger and Wright (1994) argue that this differential forecasting

performance is due, in part, to the existence of rapid and meaningful feedback to the weather forecasters in terms of both the relative frequency of probability predictions *and* the predicted events occurrence. Such prediction feedback frequency information may well be ideal for the achievement of frequentistic based accuracy. An empirical study by Benson and Önkal (1992) found that simple outcome feedback had very little impact on the performance of forecasters probabilistic judgements; however, performance feedback, i.e. information about the accuracy of the forecasters judgements *and* the outcomes that occur, did improve forecasting. While, doubtless, people register outcome feedback about the idiosyncratic "one off" events in their lives it is harder to believe that they will ordinarily be in receipt of performance feedback for their informal forecasts of these unrelated events.

We conclude our perusal of the issues surrounding the evidence of human judgements of probability with a final point concerning a possible special case for subjective probabilities. On occasion there will be single events for which no obvious reference class exists, and then one will plainly be unable to assess likelihood according to an outside view, or by taking the frequentist approach. Consider, for example, the possibility, in 1991, that Saddam Hussein would attack Kuwait. How could President Bush's administration have gone about assessing a subjective probability for this unique proposition? As van der Heijden (1994) discusses, such an assessment task may place unrealistic demands on the forecaster. He argues that in planning for such plausible, high consequence, unique events, application of scenario planning techniques aid the creation of a robust strategy that works well under a *range* of *plausible* futures. As a methodology for dealing with uncertainty, scenario planning accepts and downplays the decision maker's poor ability to make realistic probability assessments for single events. Ecologically, the acceptability of scenario planning techniques to senior mangers, and the relative disdain with which decision analysis is viewed, may reflect an intuitive appreciation of the poor quality of probabilistic judgements of the occurrence of unique events. By contrast, in the psychological laboratory, subjects will, helpfully, produce probabilities for unique future events as required by the experimenter. Since most of our knowledge about probabilistic judgement has been derived from laboratory studies, the documentation of the heuristics and biases implicated in the assessment of probability may be valid *but* unrelated to the way in which decision makers choose to deal with uncertainty given a free choice.

How *should* a person go about assessing numerical subjective probabilities for such unique events? By definition, it is difficult to see how any reference class of similar events could be selected for such events. However, one might perfectly be able to account for the (no doubt varying) subjective probabilities offered by a sample of people by referring to various judgmental heuristics. But, note that, without any reference class, we have no means of evaluating the validity of any judgements that might be offered. A single probability that is unconstrained by reference to any parent distribution admits no standard for evaluation. Consequently, the probability of unique events remains something of a mystery.

6 Conclusions

Our examination of the evidence for cognitive biases in human judgements under uncertainty discusses a controversy. Although there has been a substantial amassing of evidence for the view that humans are inept at dealing with uncertainty using probability we also find evidence for a counter-argument. It seems that disparities with basic requirements of probability theory can be observed when people are asked to make judgements of probability as a measure of propensity or strength of belief. Explanations of these errors in terms of postulated mental heuristics provide an account of these errors. The research into heuristics and biases provides a convincing methodology, a very vivid explanatory framework and a strong suggestion that judgement is not as good as it might be.

The counter-argument proposes that people may be very much better at reasoning under uncertainty than all this research suggests when they are presented with tasks in a manner that permits them to conceive of probability as a relative frequency. We reviewed Gigerenzer's claim

that putative biases in human judgement "disappear" when the usual experimental tasks are altered to refer to frequencies rather than single-event probabilities. If we further assume that probability laws only apply to the relative frequency conception of probability then it can be claimed that no statement about confidence in a single event can violate any laws of probability (cf. Gigerenzer, 1994). Such a view leads to a serious undermining of the case for incompetence in human uncertainty judgements; if probability theory is no standard for comparison for these judgements then who can say when error has occurred?

Some have proposed that there is a whole family of possible ways of dealing with uncertainty (cf. Hacking, 1975; Fox, 1994). The idea that there are viable alternative methods for dealing with uncertainty other than one particular version of probability may be crucial for understanding human reasoning. Tiegen (1994) proposes that judgements of probability can be, and are, arrived at via a number of different processes. Researchers in the heuristics and biases tradition have sometimes generated shock and astonishment that people seem so bad at reasoning with probability despite the fact that we all live in a rather uncertain world. Tiegen argues that the variety of terms and concepts that we use to communicate uncertainty suggests that this is a misleading conclusion; perhaps, he provocatively suggests, we are even quite sophisticated at dealing with uncertainty in all respects except the quantitative. Given the evidence for good frequentist judgements, even this optimistic sounding argument might be claimed to be an understatement.

For those contemplating the design and scope of decision support systems, it seems crucial to take note of these arguments. For example, we have seen how people may be assisted in their forecasting by taking an "outside view" of their particular situation. Such a proposal will have implications for the refinement of techniques for soliciting expert opinions regarding uncertainty. Moreover, one might feel more confident that the judgements obtained will not be "plagued" with biases. More significant than this though, we should feel more justified in exploring and developing alternative methods for representing uncertainty in expert knowledge. If human judgements and knowledge about uncertainty are more meaningful and more exploitable when elicited and modelled in the appropriate fashion, then we should try and discover the appropriate methods and representations as soon as possible. Scenario planning is one approach to managing uncertainty without reference to probability but there are a number of others particularly in the field of expert system design (see Krause & Clark, 1993). We hope our review has gone some way to suggest that human judgements of uncertainty are worth considering as a valuable resource rather than as objects to be regarded with suspicion or disdain.

Acknowledgements

Preparation of this paper was supported by a grant from the EPSRC Supported under the DTI Intelligent Systems Integration Programme.

References

Arkes, HR, Hackett, C and Boehm, L, 1989. "The generality of the relation between familiarity and validity" *Journal of Behavioral Decision Making* **2** 81–94.

Ayton, P, 1992. "On the competence and incompetence of experts" In: Wright, G and Bolger, F (eds.), *Expertise and Decision Support*, Plenum.

Ayton, P and Wright, G, 1987. "Assessing and improving judgmental probability forecasts" *OMEGA: International Journal of Management Science* **15** 191–196.

Ayton, P, Hunt, A and Wright, G, 1989. "Psychological conceptions of randomness" *Journal of Behavioral Decision Making* **2** 221–238.

Ayton, P, Hunt, A and Wright, G, 1991. "Randomness and reality" *Journal of Behavioral Decision Making* **4** 222–226.

Bacon, FT, 1979. "Credibility of repeated statements: memory for trivia" *Journal of Experimental Psychology: Human Learning and Memory* **5** 241–252.

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

Beach, LR and Braun, G, 1994. "Laboratory studies of subjective probability: A status report" In: G Wright, and P Ayton (eds.), *Subjective Probability*, Wiley.

Benson, PG and Önkal, D, 1992. "The effects of feedback and training on the performance of probability forecasters" *International Journal of Forecasting* **8** 559–573.

Bolger, F and Harvey, N, 1993. "Context-sensitive heuristics in statistical reasoning" *Quarterly Journal of Experimental Psychology* **46A** 779–811.

Bolger, F and Wright, G, 1994. "Assessing the quality of expert judgement: Issues and analysis" *Decision Support Systems* **11** 1–24.

Cohen, LJ, 1981. "Can human irrationality be experimentally demonstrated?" *The Behavioral and Brain Sciences* **4** 317–370.

Christensen-Szalanski, JJJ and Bushyhead, JB, 1981. "Physicians use of probabilistic information in a real clinical setting" *Journal of Experimental Psychology, Human Perception and Performance* **7** 928–935.

Cooper, A, Woo, C and Dunkelberger, W, 1988. "'Entrepreneurs' perceived chances for success" *Journal of Business Venturing* **3** 97–108.

Cosmides, L and Tooby, J, in press. "Are humans good intuitive statisticians after all? Rethinking some conclusions from the literature on judgment under uncertainty" *Cognition*.

Dalkey, N, 1972. "An elementary cross impact model" *Technological Forecasting and Social Changes* **3** 341–351.

Dawes, RM, 1980. "Confidence in intellectual judgments" In: ED Lantermann and H Feger (eds.), *Similarity and Choice*, Hans Huber.

Dawes, RM, 1988. *Rational Choice in an Uncertain World*. Harcourt.

Dostoyevsky, F, 1966. *The Gambler*, Penguin.

DuCharme, WM and Peterson, CR, 1968. "Intuitive inference about normally distributed populations" *Journal of Experimental Psychology* **78** 269–275.

Duda, RO, Hart, PE and Nilsson, NJ, 1976. "Subjective Bayesian methods for rule-based inference systems" *Proc. Nat. Computer. Cong (AFIPS)* **45** 1075–1082.

Eddy, DM, 1982. "Probabilistic reasoning in clinical medicine: Problems and opportunities" In: D Kahneman, P Slovic and A Tversky (eds.), *Judgement under Uncertainty: Heuristics and biases*, Cambridge University Press.

Edwards, W, 1954. "The theory of decision making" *Psychological Bulletin* **51** 380–417.

Edwards, W, 1968. "Conservatism in human information processing" In: B Kleinmuntz (ed.), *Formal Representation of Human Judgment*, Wiley.

Estes, W, 1976. "The cognitive side of probability learning" *Psychological Review* **83** 37–64.

Ferrell, WR and McGooey, PJ, 1980. "A model of calibration for subjective probabilities" *Organizational Behavior and Human Performance* **26** 32–53.

Finke, RA, 1984. "Strategies for being random" *Bulletin of the Psychonomic Society* **22** 40–41.

Fischhoff, B and Beyth-Marom, R, 1983. "Hypothesis evaluation from a Bayesian perspective" *Psychological Review* **90** 239–260.

Fischhoff, B, Slovic, P and Lichtenstein, R, 1983. "Fault trees: Sensitivity of estimated failure probabilities to problem representation" *Journal of Experimental Psychology: Human Perception and Performance* **4** 330–344.

Fox, J, 1994. "On the necessity of probability: Reasons to believe and grounds for doubt" In: G Wright and P Ayton (eds.), *Subjective Probability*, Wiley.

Gigerenzer, G, 1984. "External validity of laboratory experiments: The frequency-validity relationship" *American Journal of Psychology* **97** 185–195.

Gigerenzer, G, 1994. "Why the distinction between single event probabilities and frequencies is important for Psychology and vice-versa" In: G Wright and P Ayton (eds.), *Subjective Probability*, Wiley.

Gigerenzer, G, Hell, W and Blank, H, 1988. "Presentation and content: The use of base rates as a continuous variable" *Journal of Experimental Psychology: Human Perception and Performance* **14** 513–525.

Gigerenzer, G, Hoffrage, U and Kleinbolting, H, 1991. "Probabilistic mental models: A Brunswikian theory of confidence" *Psychology Review* **98** 506–528.

Goodwin, P and Wright, G, 1990. *Decision Analysis for Business*, Wiley.

Hacking, 1975. *The Emergence of Probability*. Cambridge University Press.

Hasher, L, Goldstein, D and Toppino, T, 1977. "Frequency and the conference of referential validity" *Journal of Verbal Learning and Verbal Behavior* **21** 127–141.

Hasher, L and Zacks, RT, 1979. "Automatic and effortful processes in memory" *Journal of Experimental Psychology: General* **108** 356–388.

Hasher, L and Zacks, RT, 1984. "Automatic processing of fundamental information. The case of frequency of occurrence" *American Psychologist* **39** 1372–1388.

Henrion, M, 1990. "Towards Efficient Probabilistic Diagnosis in Multiply Connected Belief Network" In: RM Olivier and JQ Smith (eds.), *Influence Diagrams. Belief Nets and Decision Analysis*, Wiley.

Jacob, VS, Gaulteney, LD and Salvendy, G, 1986. "Strategies and biases in human decision making and their implications for expert systems" *Behaviour and Information Technology* **5** 119–140.

Juslin, P, 1994. "The overconfidence phenomenon as a consequence of informal experimenter-guided selection of almanac items" *Organizational Behavior and Human Decision Processes* **57** 226–246.

Kahneman, D, Slovic, P and Tversky, A (eds.), 1982, *Judgement under Uncertainty: Heuristics and Biases*, Cambridge University Press.

Kahneman, D and Tversky, A, 1972. "Subjective probability: A judgement of representativeness" *Cognitive Psychology* **3** 430–454.

Kahneman, D and Tversky, A, 1973. "On the psychology of prediction" *Psychological Review* **80** 237–251.

Kahneman, D and Tversky, A, 1979. "Intuitive prediction: Biases and corrective procedures" *Management Science* **12** 313–327.

Kahneman, D and Tversky, A, 1982a. "On the study of statistical intuitions" In: D Kahneman, P Slovic and A Tversky (eds.), *Judgement under Uncertainty: Heuristics and Biases*, Cambridge University Press.

Kahneman, D and Tversky, A, 1982b. "Judgements of and by representativeness" In: D Kahneman, P Slovic and A Tversky (eds.), *Judgement under Uncertainty: Heuristics and Biases*, Cambridge University Press.

Kahneman, D and Lovallo, D, 1993. "Timid choices and bold forecasts. A cognitive perspective on risk taking" *Management Science* **39** 17–31.

Kareev, Y, 1992. "Not that bad after all: Generation of random sequences" *Journal of Experimental Psychology: Human Perception and Performance* **18** 1189–1194.

Keren, GB, 1987. "Facing uncertainty in the game of bridge: A calibration study" *Organizational Behavior and Human Decision Processes* **39** 98–114.

Keren, G, 1991. "Calibration and probability judgments: Conceptual and methodological issues" *Acta Psychologica* **77** 217–273.

Krause, PJ and Clark, DA, 1993. *Representing Uncertain Knowledge: An Artificial Intelligence Approach*, Intellect.

Lenat, DL, 1982. "AM: Discovery in mathematics as heuristic search" In: R Davis and D Lenat (eds.), *Knowledge Based Systems in Artificial Intelligence*, McGraw Hill.

Lichtenstein, S, Fischhoff, B and Phillips, LD, 1982. "Calibration of probabilities: The state of the art to 1980" In: D Kahneman, P Slovic and A Tversky (eds.), *Judgement under Uncertainty: Heuristics and Biases*, Cambridge University Press.

McClelland, AGR and Bolger, F, 1994. "The calibration of subjective probabilities: Theories and Models 1980–1994" In: G Wright and P Ayton (eds.), *Subjective Probability*, Wiley.

Meehl, PE, 1954. *Clinical versus Statistical Prediction: A theoretical analysis and a review of the evidence*, University of Minnesota Press.

Meehl, PE, 1986. "Causes and effects of my disturbing little book" *Journal of Personality Assessment* **50** 370–375.

Milburn, MA, 1983. "Sources of bias in the prediction of future events" *Organizational Behavior and Human Performance* **21** 17–26.

Murphy, AH and Brown, BG, 1985. "A comparative evaluation of objective and subjective weather forecasters in the U.S." In: G Wright (ed.), *Behavioral Decision Making*, Plenum.

Murphy, AH and Winkler, RL, 1984. "Probability forecasting in meteorology" *Journal of the American Statistical Association* **79** 489–500.

Payne, JW, 1982. "Contingent Decision Behavior" *Psychological Bulletin* **92** 382–402.

Pearl, J, 1988. *Probabilistic Reasoning in Intelligent Systems: Networks of plausible inference*, Morgan Kaufmann.

Phillips, LD, 1987. "On the adequacy of judgmental probability forecasts" In: G Wright and P Ayton (eds.), *Judgmental Forecasting*, Wiley.

Phillips, LD and Edwards, W, 1966. "Conservatism in simple probability inference tasks" *Journal of Experimental Psychology* **72** 346–357.

Pitz, GF, 1974. "Subjective probability distributions for imperfectly known quantities" In: LW Gregg (ed.), *Knowledge and Cognition*, Erlbaum.

Nisbett, R and Ross, L, 1980. *Human Inference: Strategies and Shortcomings*, Prentice Hall.

Peterson, CR, Schneider, RJ and Miller, AJ, 1965. "Sample size and the revision of subjective probability" *Journal of Experimental Psychology* **69** 522–527.

Pitz, GF, Downing, L and Rheinold, H, 1967. "Sequential effects in the revision of subjective probabilities" *Canadian Journal of Psychology* **21** 381–393.

Poulton, EC, 1989. *Bias in Quantifying Judgments*, Erlbaum.

Raiffa, H, 1968. *Decision Analysis*, Addison-Wesley.

Rapoport, A and Budescu, D, 1992. "Generation of random series in two-person strictly competitive games" *Journal of Experimental Psychology: General* **121** 352–363.

Savage, LJ, 1954. *The Foundation of Statistics*, Wiley.

Schwartz, S and Griffin, T, 1986. *Medical Thinking: The Psychology of Medical Judgment and Decision Making*, Springer-Verlag.

Shortliffe, EH and Buchanan, BG, 1976. "A model of inexact reasoning in medicine" *Mathematical Biosciences* **23** 351–379.

Simon, H, 1957. *Models of Man: Social and Rational*, Wiley.

Slatter, PE, 1987. "Cognitive emulation in expert system design" *Knowledge Engineering Review* **2** 27–42.

Tiegen, KH, 1994. "Variants of subjective probabilities: Concepts, norms and biases" In: G Wright and P Ayton (eds.), *Subjective Probability*, Wiley.

Tversky, A and Kahneman, D, 1971. "Belief in the law of small numbers" *Psychological Bulletin* **76** 105–110.

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

van der Heijden, C, 1994. "Probabilistic planning and scenario planning" In: G Wright and P Ayton (eds.), *Subjective Probability*, Wiley.

von Winterfeldt, D and Edwards, W, 1986. *Decision Analysis and Behavioral Research*, Cambridge University Press.

Wagenaar, WA and Keren, GB, 1986. "Does the expert know? The reliability of predictions and confidence ratings of experts" In: E Hollnagel, G Mancini and DD Woods (eds.), *Intelligent Decision Support In Process Environments*, Springer-Verlag.

Winefield, AH, 1966. "Negative recency and event dependence" *Quarterly Journal of Experimental Psychology* **18** 47–54.

Winkler, RL and Murphy, AM, 1973. "Experiments in the laboratory and the real world" *Organizational Behavior and Human Performance* **20** 252–270.

Wright, G and Ayton, P, 1992. "Judgmental probability forecasting in the immediate and medium term" *Organizational Behavior and Human Decision Processes* **51** 344–363.

Yates, JF, 1982. "External correspondence: decompositions of the mean probability score" *Organizational Behavior and Human Performance* **30** 132–156.

Yates, JF, McDaniel, LS and Brown, ES, 1991. "Probabilistic forecasts of stock prices and earnings: The hazards of nascent expertise" *Organizational Behavior and Human Decision Processes* **40** 60–79.

Yousseff, ZI and Peterson, CR, 1973. "Intuitive cascaded inferences" *Organizational Behavior and Human Performance* **10** 349–358.