NICK CHATER* and MIKE OAKSFORD

THE RATIONAL ANALYSIS OF MIND AND BEHAVIOR

ABSTRACT. Rational analysis (Anderson 1990, 1991a) is an empirical program of attempting to explain why the cognitive system is adaptive, with respect to its goals and the structure of its environment. We argue that rational analysis has two important implications for philosophical debate concerning rationality. First, rational analysis provides a model for the relationship between formal principles of rationality (such as probability or decision theory) and everyday rationality, in the sense of successful thought and action in daily life. Second, applying the program of rational analysis to research on human reasoning leads to a radical reinterpretation of empirical results which are typically viewed as demonstrating human irrationality.

Rationality appears fundamental to the understanding of minds and behavior. In clinical psychology, as well in the law, it appears to be of fundamental importance to be able to draw a boundary between sanity and madness, between rationality and irrationality. In economics, and increasingly, other areas of social science, human behavior is explained as the outcome of "rational choice", concerning which products to buy, whom to marry, or how many children to have (Becker 1975, 1981; Elster 1986). But assumptions of rationality may go much deeper still – they seem to lie at the heart of the folk psychological style of explanation in which we describe each other's minds and behavior (Cherniak 1986; Fodor 1987). Assumptions of rationality also appear equally essential to interpret each other's utterances and to understand texts (Davidson 1984; Quine 1960). So rationality, in an intuitive sense, appears to be at the heart of the explanation of human behavior, whether from the perspective of social science or of everyday life. Let us call this everyday rationality: rationality concerned with people's beliefs and actions in specific circumstances.

In this informal, everyday sense, most of us, most of the time, are remarkably rational. In daily life, of course, we tend to focus on occasions when reasoning or decision making breaks down. But our failures of reasoning are only salient because they occur against the background of rational thought and behavior which is achieved with such little apparent effort that we are inclined to take it for granted. Rather than thinking of our patterns of everyday thought and action as exhibiting rationality, we

Synthese **122:** 93–131, 2000.
© 2000 *Kluwer Academic Publishers. Printed in the Netherlands.*

think of them as just plain common sense – with the implicit assumption that common sense must be a simple thing indeed. People may not think of themselves as exhibiting high levels of rationality – instead, we think of people as "intelligent", performing "appropriate" actions, being "reasonable" or making "sensible" decisions. But these labels refer to human abilities to make the right decisions, or to say or think the right thing in complex, real-world situations – in short, they are labels for everyday rationality.

Indeed, so much do we tend to take the rationality of commonsense thought for granted, that realizing that commonsense reasoning is immensely difficult, and hence our everyday rationality is thereby immensely impressive, has been a surprising *discovery*, and a discovery made only in the latter part of the twentieth century. The discovery emerged from the project of attempting to formalize everyday knowledge and reasoning in artificial intelligence, where initially high hopes that commonsense knowledge could readily be formalized were replaced by increasing desperation at the impossible difficulty of the project. The nest of difficulties referred to under the "frame problem" (see, e.g., Pylyshyn 1987), and the problem that each aspect of knowledge appears inextricably entangled with the rest (e.g., Fodor 1983) so that commonsense does not seem to break down into manageable "packets" (whether schemas, scripts, or frames, Minsky 1977; Schank and Abelson 1977), and the deep problems of defeasible, or nonmonotonic reasoning, brought the project of formalizing commonsense to an effective standstill (e.g., McDermott 1987). So the discovery is now made – it is now clear that everyday, commonsense reasoning is remarkably, but mysteriously, successful in dealing with an immensely complex and changeable world and that no artificial computational system can begin to approach the level of human performance. Hence, the sentiment with which we began: Most of us, most of the time, are remarkably rational.

But in addition to this informal, everyday sense of rationality, concerning people's ability to think and act in the real world, the concept of rationality also has another root, linked not to human behavior, but to mathematical theories of good reasoning. These theories represent one of the most important achievements of modern mathematics: Logical calculi formalize aspects of deductive reasoning; axiomatic probability formalizes probabilistic reasoning; the variety of statistical principles, from sampling theory (Fisher 1922, 1925/1970) to Neyman–Pearson statistics (Neyman 1950), to Bayesian statistics (Keynes 1921; Lindley 1971), aim to formalize the process of interpreting data in terms of hypotheses; utility and decision theory attempt to characterize rational preferences and rational choice between actions under uncertainty. According to these calculi,

rationality is defined, in the first instance, in terms of conformity with specific formal principles, rather than in terms of successful behavior in the everyday world.

The two sides of rationality raise the fundamental question of how they relate to each other: How are the general principles of formal rationality related to specific examples of rational thought and action described by everyday rationality? This question, in various guises, has been widely discussed – in this article, we shall outline a particular conception of the relation between these two notions, focussing on a particular style of explanation in the behavioral sciences, rational analysis (Anderson 1990). We will argue that rational analysis provides a good characterization of how the concept of rationality is used in explanations in psychology, economics and animal behaviour, and provides an account of the relationship between everyday and formal rationality, which has implications for both. Moreover, this view of rationality leads to a re-evaluation of the implications of data from psychological experiments which appear to undermine human rationality. We argue that, on the contrary, experimental evidence demands a change concerning which formal account defines the normative standard in experimental tasks.

This paper thus has two linked goals. The first goal is to outline what we take to be the standard role of rationality in the explanation of mind and behavior, in disciplines as diverse as experimental psychology, animal behavior and economics – we take rational analysis to be a paradigm for such an explanation. The second goal is to draw out some of the implications of the rational analysis perspective for the interpretation of experimental data which appears to show that human behavior is non-rational. We argue, instead, that human behavior is rational, if the appropriate normative standard for that behavior is adopted. Specifically, a wide range of empirical results in the psychology of reasoning have been taken to cast doubt on human rationality, because people appear to persistently make elementary logical blunders. We show that, when the tasks people are given are viewed in terms of probability, rather than logic, people's responses can be seen as rational.

The discussion falls into three main parts. First, we discuss formal and everyday rationality, and various possible relationships between them. Second, we describe the program of rational analysis as a mode of explanation of mind and behavior, which views everyday rationality as underpinned by formal rationality. Third, we apply rational analysis to re-evaluating experimental data in the psychology of reasoning.

1. RELATIONS BETWEEN FORMAL AND EVERYDAY RATIONALITY

Formal rationality concerns formal principles of good reasoning – the mathematical laws of logic, probability, or decision theory. At an intuitive level, these principles seem distant from the domain of everyday rationality – how people think and act in daily life. Rarely, in daily life, do we accuse one another of violating the laws of logic or probability theory or praise each other for obeying them. Moreover, when people are given reasoning problems that explicitly require use of these formal principles, their performance appears to be remarkably poor. People appear to persistently fall for logical blunders (Evans et al. 1993), probabilistic fallacies (e.g., Tversky and Kahneman 1974) and to make inconsistent decisions (Kahneman et al. 1982; Tversky and Kahneman 1986). Indeed, the concepts of logic, probability and the like do not appear to mesh naturally with our everyday reasoning strategies: these notions took centuries of intense intellectual effort to construct, and present a tough challenge for each generation of students.

We therefore face a stark contrast: the astonishing fluency and success of everyday reasoning and decision making, exhibiting remarkable levels of everyday rationality; and our faltering and confused grasp of the principles of formal rationality. What are we to conclude from this contrast? Let us briefly consider, in caricature, some of the most important possibilities, which have been influential in the literature in philosophy, psychology and the behavioral sciences.

1.1. *The Primacy of Everyday Rationality*

This viewpoint takes everyday rationality as fundamental, and dismisses the apparent mismatch between human reasoning and the formal principles of logic and probability theory as so much the worse for these formal theories.

This standpoint appears to gain credence from historical considerations – formal rational theories such as probability and logic emerged as attempts to systematize human rational intuitions, rooted in everyday contexts. But the resulting theories appear to go beyond, and even clash with, human rational intuitions – at least if empirical data which appears to reveal blunders in human reasoning is taken at face value.

To the extent that such clashes occur, the advocates of the primacy of everyday rationality argue that the formal theories should be rejected as inadequate systematizations of human rational intuitions, rather than condemning the intuitions under study as incoherent. It might, of course, be granted that a certain measure of tension may be allowed between

the goal of constructing a satisfyingly concise formalization of intuitions, and the goal of capturing every last intuition successfully, rather as, in linguistic theory, complex centre embedded constructions are held to be grammatical (e.g., 'the fish the man the dog bit ate swam'), even though most people would reject them as ill-formed gibberish. But the dissonance between formal rationality and everyday reasoning appears to be much more profound than this. As we have argued, fluent and effective reasoning in everyday situations runs alongside halting and flawed performance on the most elementary formal reasoning problems.

The primacy of everyday rationality is implicit in an important challenge to decision theory by the mathematician Allais (1953). Allais outlines his famous "paradox", which shows a sharp divergence between people's rational intuitions and the dictates of decision theory. One version of the paradox is as follows. Consider the following pair of lotteries, each involving 100 tickets. Which would you prefer to play?

A.	B.
10 tickets worth \$1,000,000	1 ticket worth \$5,000,000
90 tickets worth \$0	8 tickets worth \$1,000,000
	91 tickets worth \$0

Now consider which you would prefer to play of lotteries C and D:

C.	D.
100 tickets worth \$1,000,000	1 ticket worth \$5,000,000
	98 tickets worth \$1,000,000
	1 ticket worth \$0

Most of us prefer lottery B to lottery A – the slight reduction in the probability of becoming a millionaire is offset by the possibility of the really large prize. But most of us also prefer lottery C to lottery D – we don't think its worth losing what would otherwise be a certain \$1,000,000, just for the possibility of winning \$5,000,000. This combination of responses, although intuitively appealing, is inconsistent with decision theory, as we shall see. Decision theory assumes that people should choose whichever alternative has the maximum expected utility. Denote the utility associated with a sum of $\$X$ by $U(\$X)$. Then the preference for lottery B over A means that:

$$(1) \quad 10/100.U(\$1,000,000) + 90/100.U(\$0) < 1/100.U(\$5,000,000) + 8/100.U(\$1,000,000) + 91/100.U(\$0)$$

and, subtracting 90/100.U($0) from each side:

$$(2) \quad 10/100.U(\$1,000,000) < 1/100.U(\$5,000,000) + 8/100.U(\$1,000,000) + 1/100.U(\$0)$$

But the preference for lottery C over D means that:

$$(3) \quad 100.U(\$1,000,000) > 1/100.U(\$5,000,000) + 98/100.U(\$1,000,000) + 1/100.U(\$0)$$

and, subtracting 90/100.U($1,000,000) from each side:

$$(4) \quad 10.U(\$1,000,000) > 1/100.U(\$5,000,000) + 8/100.U(\$1,000,000) + 1/100.U(\$0)$$

But (2) and (4) are in contradiction.

Allais's paradox is very powerful – the appeal of the choices that decision theory rules out is considerable. Indeed, rather than condemning people's intuitions as incorrect, Allais argues that the paradox undermines the normative status of decision theory – that is, Allais argues that everyday rational intuitions take precedence over the dictates of a formal calculus.

Another example arises in Cohen's (1981) discussion of the psychology of reasoning literature. Following similar arguments of Goodman (1954), Cohen argues that a normative or formal theory is "acceptable ... only so far as it accords, at crucial points with the evidence of untutored intuition" (Cohen 1981, 317). That is, a formal theory of reasoning is acceptable only in so far as it accords with everyday reasoning. Cohen uses the following example to demonstrate the primacy of everyday inference. According to standard propositional logic the inference from (5) to (6) is valid:

(5) If John's automobile is a Mini, John is poor, and
if John's automobile is a Rolls, John is rich.

(6) Either, if John's automobile is a Mini, John is rich, or
if John's automobile is a Rolls, John is poor.

Clearly, however, this violates intuition. Most people would agree with (5) as at least highly plausible; but would reject (6) as absurd. A fortiori, they would not accept that (5) *implies* (6) otherwise they would have to judge (6) to be at least as plausible as (5)). Consequently. Cohen argues that standard logic simply does not apply to the reasoning that is in evidence in

people's intuitions about (5) and (6). Like Allais, Cohen argues that rather than condemn people's intuitions as irrational, this mismatch reveals the inadequacy of propositional logic as a rational standard. That is, everyday intuitions have primacy over formal theories.

But this viewpoint is not without problems. For example, how can rationality be assessed? If formal rationality is viewed as basic, then the degree to which people behave rationally can be assessed by comparing performance against the canons of the relevant normative theory. But if everyday rationality is viewed as basic, assessing rationality appears to be down to intuition. There is a danger here of losing any normative force to the notion of rationality – if rationality is merely conformity to each other's predominant intuitions, then being rational is like a musician being in tune. On this view, rationality has no absolute significance; all that matters is that we reason harmoniously with our fellows. But there is a strong intuition that rationality is not like this at all – that there is some absolute sense in which some reasoning or decision making is good, and other reasoning and decision making is bad. So, by rejecting a formal theory of rationality, there is the danger that the normative aspect of rationality is left unexplained.

One way to re-introduce the normative element is to define a procedure that derives normative principles from human intuitions. Cohen appealed to the notion of reflective equilibrium (Goodman 1954; Rawls 1971) where inferential principles and actual inferential judgements are iteratively bought into a "best fit" until further judgements do not lead to any further changes of principle (narrow reflective equilibrium). Alternatively, background knowledge may also figure in the process, such that not only actual judgements but also the way they relate to other beliefs are taken into account (wide reflective equilibrium). These approaches have, however, been subject to much criticism (e.g., Stich and Nisbett 1980; Thagard 1988). For example, there is no guarantee that an individual (or indeed a set of experts) in equilibrium will have accepted a set of rational principles, by any independent standard of rationality. The equilibrium point could, for example, leave the individual content in the idea that the Gambler's fallacy is a sound principle of reasoning.

Thagard (1988) proposes that instead of reflective equilibrium, developing inferential principles involves progress towards an optimal system. This involves proposing principles based on practical judgements and background theories, and measuring these against criteria for optimality. The criteria Thagard specifies are (i) robustness: principles should be empirically adequate; (ii) accommodation: given relevant background knowledge, deviations from these principles can be explained; and (iii) efficacy: given relevant background knowledge, inferential goals are sat-

isfied. Thagard's (1988) concerns were very general: to account for the development of scientific inference. From our current focus on the relationship between everyday and formal rationality, however, Thagard's proposals seem to fall down because the criteria he specifies still seem to leave open the possibility of inconsistency, i.e., it seems possible that a system could fulfill (i) to (iii) but contain mutually contradictory principles. The point about formalisation is of course that it provides a way of ruling out this possibility and hence is why a tight relationship between formality and normativity has been assumed since Aristotle. From the perspective of this paper, accounts like reflective equilibrium and Thagard's account, which attempts to drive a wedge between formality and normativity, may not be required. We argue that many of the mismatches observed between human inferential performance and formal theories are a product of using the wrong formal theory to guide expectations about how people should behave.

An alternative normative grounding for rationality seems intuitively appealing good everyday reasoning and decision making should lead to *successful action* for example, from an evolutionary perspective, we might define success as inclusive fitness, and argue that behavior is rational to the degree that it tends to increase inclusive fitness. But now the notion of rationality appears to collapse into a more general notion of adaptiveness. There seems to be no particular difference in status between cognitive strategies which lead to successful behavior, and digestive processes that lead to successful metabolic activity. Both increase inclusive fitness; but intuitively we want to say that the first is concerned with rationality, which the second is not. More generally, defining rationality in terms of outcomes runs the risk of blurring what appears to be a crucial distinction – between minds, which may be more or less rational, and stomaches, that are not in the business of rationality at all.

1.2. *The Primacy of Formal Rationality*

Arguments for the primary of formal rationality take a different starting point. This viewpoint is standard with mathematics, statistics, operations research and the "decision sciences" (e.g., Kleindorfer et al. 1993). The idea is that everyday reasoning is fallible, and that it must be corrected by following the dictates of formal theories of rationality.

The immediate problem for advocates of the primacy of formal rationality concerns the justification of formal calculi of reasoning: Why should the principles of some calculus be viewed as principles of good reasoning, so that may even be allowed to overturn our intuitions about what is rational? Such justifications typically assume some general, and apparently

incontrovertible, cognitive goal; or seemingly undeniable axioms about how thought or behavior should proceed. They then use these apparently innocuous assumptions and aim to argue that thought or decision making must obey specific mathematical principles.

Consider, for example, the "Dutch book" argument for the rationality of the probability calculus as a theory of uncertain reasoning (de Finetti 1937; Ramsey 1931; Skyrms 1977). Suppose that we assume that people will accept a "fair" bet: that is, a bet where the expected financial gain is 0, according to their assessment of the probabilities of the various outcomes. Thus, for example, if a person believes that there is a probability of 1/3 that it will rain tomorrow, then they will be happy to accept a bet according to which they win two dollars if it does rain tomorrow, but they lose one dollar if it does not. Now, it is possible to prove that, if a person's assignment of probabilities to different possible outcomes violates the laws of probability theory in any way whatever, then it is possible to offer them a combination of different bets, such that they will happily accept each individual bet as fair, in the above sense, but where *whatever the outcome* they are certain to lose money. Such a combination of bets – where one side is certain to lose – is known as a Dutch book; and it is seems incontrovertible that accepting a bet that you are certain to lose must violate rationality. Thus, if violating the laws of probability theory leads to accepting Dutch books, which seems clearly irrational, then obeying the laws of probability theory seems to be a condition of rationality.

The Dutch book theorem might appear to have a fundamental weakness – that it requires that a person willingly accepts arbitrary fair bets. But, in reality of course, this might not be so – many people will, in such circumstances, be risk aversive, and choose not to accept such bets. But the same argument applies even if the person does not bet at all. Now the inconsistency concerns a hypothetical – the person believes that if the bet were accepted, it would be fair (so that, a win, as well as a loss, is possible). But in reality, the bet is guaranteed to result in a loss – the person's belief that the bet is fair is guaranteed to be wrong. Thus, even if we never actually bet, but simply aim to avoid endorsing statements that are guaranteed to be false, we should follow the laws of probability.

We have considered the Dutch book justification of probability theory in some detail to make it clear that justifications of formal theories of rationality can have considerable force.[1] Rather than attempting to simultaneously satisfy as well as possible a myriad of uncertain intuitions about good and bad reasoning, formal theories of reasoning can be viewed, instead, as founded on simple and intuitively clearcut principles, such as that accepting bets that you are certain to lose is irrational. Similar justifications

can be given for the rationality of the axioms of utility theory and decision theory (Cox 1961; von Neumann and Morgenstern 1944; Savage 1954). Moreover, the same general approach can be used as a justification for logic, if avoiding inconsistency is taken as axiomatic. Thus, there may been good reasons for accepting formal theories of rationality, even if much of the time, human intuitions and behavior strongly violates their recommendations.

If formal rationality is primary, what are we to make of the fact that, in explicit tests at least, people seem to be such poor probabilists and logicians? One line would be to accept that human reasoning is badly flawed. Thus, the heuristics and biases program (Kahneman and Tversky 1973; Kahneman Slovic and Tversky 1982), which charted systematic errors in human probabilistic reasoning and decision making under uncertainty, can be viewed as exemplifying this position (see Gigerenzer and Goldstein 1996), as can Evans' (1982, 1989) heuristic approach to reasoning. Another line follows the spirit of Chomsky's (1965) distinction between linguistic competence and performance – the idea is that the people's reasoning competence accords with formal principles, but in practice, performance limitations (e.g., limitations of time or memory) lead to persistently imperfect performance when people are given a reasoning task.

Reliance on a competence/performance distinction, whether implicitly or explicitly, has been very influential in the psychology of reasoning: for example, mental logic (Braine 1978; Rips 1994) and mental models (Johnson-Laird 1983; Johnson-Laird and Byrne 1991) theories of human reasoning assume that classical logic provides the appropriate competence theory for deductive reasoning; and flaws in actual reasoning behavior are explained in terms of "performance" factors.

Mental logic assumes that human reasoning algorithms correspond to proof-theoretic operations (specifically, in the framework of natural deduction, e.g., Rips 1994). This viewpoint is also embodied in the vast program of research in artificial intelligence, especially in the 1970s and 1980s, which attempted to axiomatize aspects of human knowledge, and view reasoning as a logical inference (e.g., McCarthy 1980; McDermott 1982; McDermott and Doyle 1980; Reiter 1980, 1985). Moreover, in the philosophy of cognitive science, it has been controversially suggested that this viewpoint is basic to the computational approach to mind: the fundamental claim of cognitive science, according to this viewpoint, is that "cognition is proof theory" (Fodor and Pylyshyn 1988, 29–30; see also Chater and Oaksford 1990).

Mental model views concur that logical inference provides the computational level theory for reasoning, but provides an alternative method of proof. Instead of standard proof theoretic rules, this view uses a "semantic" method of proof. Such methods involve search for models (in the logical sense) – a semantic proof that A does not imply B might involve finding a model in which A and B both hold. Mental models theory uses a similar idea, although the notion of model in play is rather different from the logical notion.[2] How can this approach show that A does imply B? The mental models account assumes that the cognitive system attempts to construct a model in which A is true and B is false; if this attempt fails, then it is assumed that no counterexample exists, and that the inference is valid (this is similar to "negation as failure" in logical programming (Clark 1978)).

Mental logic and mental models assume that formal principles of rationality–specifically classical logic – (at least partly) define the standards of good reasoning. They explain the non-logical nature of people's actual reasoning behavior in terms of performance factors, such as memory and processing limitations.

Nonetheless, despite its popularity, the view that formal rationality has priority in defining what good reasoning is, and that actual reasoning is systematically flawed with respect to this formal standard, suffers a fundamental difficulty. If formal rationality is the key to everyday rationality, and if people are manifestly poor at following the principles of formal rationality (whatever their "competence" with respect to these rules), even in simplified reasoning tasks, then the spectacular success of everyday reasoning in the face of an immensely complex world seems entirely baffling.

1.3. *Everyday and Formal Rationality Are Completely Separate*

Recently, a number of theorists have suggested what is effectively a hybrid of the two approaches outlined above. They argue that formal rationality and everyday rationality are entirely separate enterprises. For example, Evans and Over (1997) distinguish between two notions of rationality:

> Rationality$_1$: Thinking, speaking, reasoning, making a decision, or acting in a way that is generally reliable and efficient for achieving one's goals.
>
> Rationality$_2$: Thinking, speaking, reasoning, making a decision, or acting when one has a reason for what one does sanctioned by a normative theory. (Evans and Over 1997, 2)

They argue that "people are largely rational in the sense of achieving their goals (rationality$_1$) but have only a limited ability to reason or act for good reasons sanctioned by a normative theory (rationality$_2$)" (Evans and Over 1997, 1). If this is right, then achieving one's goals can be achieved without following a formal normative theory – i.e., without there being a justification for the actions, decisions or thoughts which lead to success: rationality$_1$ does not require rationality$_2$. That is, Evans and Over are committed to the view that thoughts, actions or decisions which cannot be normatively justified can, nonetheless, consistently lead to practical success.

But this hybrid view does not tackle the fundamental problem we outlined for the first view sketched above. It does not answer the question: *why* do the cognitive processes underlying everyday rationality consistently work? If everyday rationality is somehow based on formal rationality, then this question can be answered, at least in general terms. The principles of formal rationality are provably principles of good inference and decision making; and the cognitive system is rational in everyday contexts to the degree that it approximates the dictates of these principles. But if everyday and formal rationality are assumed to be unrelated, then this explanation is not available. Unless some alternative explanation of the basis of everyday rationality can be provided, the success of the cognitive system is again left entirely unexplained.

1.4. *Everyday Rationality is Based on Formal Rationality: An Empirical Approach*

We seem to be at an impasse. The success of everyday rationality in guiding our thoughts and actions must somehow be explained; and it seems that there are no obvious alternative explanations, aside form arguing that everyday rationality is somehow based on formal reasoning principles, for which good justifications can be given. But the experimental evidence appears to show that people do not follow the principles of formal rationality.

There is, however, a way out of this impasse. Essentially, the idea is to reject the idea that rationality is a monolithic notion that can be defined a priori, and compared with human performance. Instead, we treat the problem of explaining everyday rationality as an empirical problem of explaining why people's cognitive processes are successful in achieving their goals, given the constraints imposed by their environment. Formal rational theories are used in the development of these empirical explanations for the success of cognitive processes – but which formal principles are appropriate, and how they should be applied, is not decided a priori;

but in the light of the empirical success of the explanation of the adaptive success of the cognitive process under consideration.

According to this viewpoint, the apparent mismatch between normative theories and reasoning behavior suggests that the wrong normative theories may have been chosen; or the normative theories may have been misapplied. Instead, the empirical approach to the grounding of rationality aims to "do the best" for human everyday reasoning strategies – by searching for a rational characterization of how people actually reason. There is an analogy here with rationality assumptions in language interpretation (Davidson 1984; Quine 1960). We aim to interpret people's language so that it makes sense; similarly, the empirical approach to rationality aims to interpret people's reasoning behavior so that their reasoning makes sense.

Crucially, then, the formal standards of rationality appropriate for explaining some particular cognitive processes or aspect of behavior are not prior to, but are rather developed as part of; the explanation of empirical data. Of course, this is not to say that, in some sense, formal rationality may be prior to, and separate from, empirical data. The development of formal principles of logic, probability theory, decision theory and the like may proceed independently of attempting to explain people's reasoning behavior. But which element of this portfolio of rational principles should be used to define a normative standard for particular cognitive processes or tasks, and how the relevant principles should be applied, is constrained by the empirical human reasoning data to be explained.

It might seem that this approach is flawed from the outset. Surely, any behavior can be viewed as rational from some point of view. That is, by cooking up a suitably bizarre set of assumptions about the problem that a person thinks they are solving, surely their rationality can always be respected; and this suggests the complete vacuity of the approach. But this objection ignores the fact that the goal of empirical rational explanation is to provide an empirical account of data on human reasoning. Hence, such explanations must not be merely possible, but also simple, consistent with other knowledge, independently plausible, and so on. In short, such explanations are to be judged in the light of the normal canons of scientific reasoning (Howson and Urbach 1989).[3] Thus, rational explanations of cognition and behavior can be treated as on a par with other scientific explanations of empirical phenomena.

This empirical view of the explanation of rationality is attractive, to the extent that it builds in an explanation of the success of everyday rationality. It does this by attempting to recruit formal rational principles to explain why cognitive processes are successful. But how can this empirical approach to rational explanation be conducted in practice? And can plausible

rational explanations of human behavior be found? The next two sections of the paper answer these questions. First, we outline a methodology for the rational explanation of empirical data – rational analysis. We also illustrate a range of ways in which this approach is used, in psychology, and the social and biological sciences. We then use rational analysis to re-evaluate the psychological data which has appeared to show human reasoning performance to be hopelessly flawed, and argue that, when appropriate rational theories are applied, reasoning performance may, on the contrary, be rational.

2. THE PROGRAM OF RATIONAL ANALYSIS

The project of providing a rational analysis for some aspect of thought or behavior has been described by the cognitive psychologist John Anderson (e.g., Anderson 1990, 1991a). This methodology provides a framework for explaining the link between principles of formal rationality and the practical success of everyday rationality not just in psychology, but throughout the study of behavior. This approach involves six steps:

1. Specify precisely the goals of the cognitive system.
2. Develop a formal model of the environment to which the system is adapted.
3. Make minimal assumptions about computational limitations.
4. Derive the optimal behavior function given 1-3 above. (This requires formal analysis using rational norms, such as probability theory and decision theory.)
5. Examine the empirical evidence to see whether the predictions of the behavior function are confirmed.
6. Repeat, iteratively refining the theory.

According to this viewpoint, formal rational principles relate to explaining everyday rationality, because they specify the optimal way in which the goals of the cognitive system can be attained in a particular environment, subject to "minimal" computational limitations. The assumption is that the cognitive system exhibits everyday rationality – i.e., successful thought and action in the everyday world – to the extent that it approximates the optimal solution specified by rational analysis.

The framework of rational analysis aptly fits the methodology in many areas of economics and animal behavior, where the behavior of people or animals is viewed as optimizing some goal, such as money, utility, inclusive fitness, food intake, or the like. But Anderson (1990, 1991a) was concerned to extend this approach not just to the behavior of whole agents,

but to structure and performance of particular cognitive processes of which agents are composed. Anderson's program has led to a flurry of research in cognitive psychology (see Oaksford and Chater 1998a, for an overview of recent research), from areas as diverse as categorization (Anderson 1991b; Anderson and Matessa 1998; Lamberts and Chong 1998), memory (Anderson and Milson 1989; Anderson and Schooler 1991; Schooler 1998), searching computer menus (Young 1998) and natural language parsing (Chater et al. 1998). This research has shown that a great many empirical generalizations about cognition can be viewed as arising from the rational adaptation of cognitive system to the problems and constraints that it faces. We shall argue below that the cognitive processes involved in reasoning can also be explained in this way.

The three inputs to the calculations using formal rational principles, goals, environment, and computational constraints, each raise important issues regarding the connection between formal rational principles and everyday rationality. We discuss these in turn, and in doing so, illustrate rational analysis in action in psychology, animal behavior and economics.

2.1. *The Importance of Goals*

Everyday thought and action is focussed on achieving goals relevant to the agent. Formal principles of rationality can help specify how these goals are achieved, but not, of course, what those goals are. The simplest cases are economic in spirit. For example, consider a consumer, wondering which washing machine to buy. Goals are coded in terms of the subjective "utilities" associated with objects or events for this particular consumer. Each washing machine is associated with some utility (high utilities for the effective, attractive, or low energy washing machines, for example); and money is also associated with utility. Simple decision theory will specify which choice of machine maximizes subjective utility. Thus goals enter very directly; people with different goals (here, different utilities) will be assigned different "rational" choices. Suppose instead that the consumer is wondering whether to take out a service agreement on the washing machine. Now the negative utility associated with the cost of the agreement must be balanced with the positive utility of saving possible repair costs. But what are the possible repairs; how likely, and how expensive, is each type? Decision theory again recommends a choice, given utilities associated with each outcome, and subjective probabilities concerning the likelihood of each outcome.

But not all goals may have the form of subjective utilities. In evolutionary contexts, the goal of inclusive fitness might be more appropriate (Dawkins 1977); in the context of foraging behavior in animals, amount

of food intake or nutrition gained might be the right goal (Stephens and Krebs 1986). Moreover, in some cognitive contexts, the goal of thought or action may be disinterested curiosity, rather than the attempt to achieve some particular outcome. Thus, from exploratory behavior in children and animals to the pursuit of basic science, a vast range of human activity appears to be concerned with finding out information, rather than achieving particular goals. Of course, having this information may ultimately prove important for achieving goals; and this virtue may at some level explain the origin of the disinterested search for knowledge (just as the prospect of unexpected applications may partially explain the willingness of the state to fund fundamental research). Nonetheless, disinterested inquiry is conducted without any particular goal in mind. In such contexts, gaining, storing or retrieving information, rather than maximizing utility, may be the appropriate specification of cognitive goals. If this is the goal, then information theory and probability theory may be the appropriate formal normative tools, rather than decision theory.

This aspect of rational analysis is at variance with Evans and Over's distinction between two forms of rationality, mentioned above. They argue that "people are largely rational in the sense of achieving their goals (rationality$_1$) but have only a limited ability to reason or act for good reasons sanctioned by a normative theory (rationalty$_2$)" (Evans and Over 1997, 1). But the approach of rational analysis attempts to explain *why* people exhibit the everyday rationality involved in achieving their goals by assuming that their actions approximate with would be sanctioned by a formal normative theory. Thus, formal rationality helps *explain* everyday rationality, rather than being completely separate from it.

To sum up, everyday rationality is concerned with goals (even if the goal is just to "find things out"); knowing which formal theory of rationality to apply, and applying formal theories to explaining specific aspects of everyday cognition, requires an account of the nature of these goals.

2.2. *The Role of the Environment*

Everyday rationality is concerned with achieving particular goals, in a particular environment. Everyday rationality requires thought and action to be adapted (whether through genes or through learning) to the constraints of this environment. The success of everyday rationality is, crucially, success relative to a specific environment – to understand that success requires modeling the structure of that environment. This requires using principles of formal rationality to specific the optimal way in which the agent's goals can be achieved in that environment (Anderson's Step 4) and showing that the cognitive system approximates this optimal solution.

In psychology, this strategy is familiar from perception, where a key part of understanding the computational problem solved by the visual system involves describing the structure of the visual environment (Marr 1982). Only then can optimal models for visual processing of that environment be defined. Indeed, Marr (1982) explicitly allies this level of explanation with Gibson's "ecological" approach to perception, where the primary focus is on environmental structure.

Similarly, in zoology, environmental idealizations of resource depletion and replenishment of food stocks, patch distribution and time of day are crucial to determining optimal foraging strategies (Gallistel 1990; McFarland and Houston 1981; Stephens and Krebs 1986).

Equally, in economics, idealizations of the "environment" are crucial to determining rational economic behavior (McCloskey 1985). in microeconomics, modeling the environment (e.g., game-theoretically) involves capturing the relation between each actor and the environment of other actors. In macroeconomics, explanations using rational expectations theory (Muth 1961) begin from a formal model of the environment, as a set of equations governing macro-economic variables.

This aspect of rational analysis contrasts with the view that the concerns of formal rationality are inherently disconnected from environmental constraints. For example, Gigerenzer and Goldstein (1996) propose that "the minds of living systems should be understood relative to the environment in which they evolved *rather than* to the tenets of classical [i.e., formal] rationality . . . " (p. 651) (emphasis added). Instead, rational analysis aims to explain *why* agents succeed in their environment by understanding the structure of that environment, and using formal principles of rationality to understand what thought or action will succeed in that environment.

2.3. *Computational Limitations*

In rational analysis, deriving the optimal behavior function (Anderson's Step 4) is frequently very complex. Models based on optimising, whether in psychology, animal behaviour or economics, need not, and typically do not, assume that agents are able to find the perfectly optimal solutions to the problems that they face. Quite often, perfect optimisation is impossible even in principle, because the calculations involved in finding a perfect optimum are frequently computationally intractable (Simon 1955, 1956), and, moreover, much crucial information is typically not available. Indeed, formal rational theories in which the optimization calculations are made, including probability theory, decision theory and logic are typically computationally intractable for complex problems (Cherniak 1986; Garey and Johnson 1979; Good 1971; Paris 1992; Reiner 1995). Intractability results

imply that no computer algorithm could perform the relevant calculations given the severe time and memory limitations of a "fast and frugal" cognitive system. The agent must still act, even in the absence of the ability to derive the optimal solution (Gigerenzer and Goldstein 1996; Simon 1956). Thus it might appear that there is an immediate contradiction between the limitations of the cognitive system and the intractability of rational explanations.

There is no contradiction, however, because the optimal behavior function is an explanatory tool, not part of an agent's cognitive equipment. Using an analogy from Marr (1982), the theory of aerodynamics is a crucial component of explaining why birds can fly. But clearly birds know nothing about aerodynamics, and the computational intractability of aerodynamic calculations does not in any way prevent birds from flying. Similarly, people do not need to calculate their optimal behavior functions in order to behave adaptively. They simply have to use successful algorithms; they do not have to be able to make the calculations that would show that these algorithms are successful. Indeed, it may be that many of the algorithms that the cognitive system uses may be very crude "fast and frugal" heuristics (Gigerenzer and Goldstein 1996) which generally approximate the optimal solution in the environments that an agent normally encounters. In this context, the optimal solutions will provide a great deal of insight into why the agent behaves as it does. However, an account of the algorithms that the agent uses will be required to provide a full explanation of their behaviour (e.g., Anderson 1993; Oaksford and Chater 1995a).

This viewpoint is standard in rational explanations across a broad range of disciplines. Economists do not assume that people make complex game-theoretic or macroeconomic calculations (Harsanyi and Selten 1988); zoologists do not assume that animals calculate how to forage optimally (e.g., McFarland and Houston 1981); and, in psychology, rational analyses of, for example, memory, do not assume that the cognitive system calculates the optimal forgetting function with respect to the costs of retrieval and storage (Anderson and Schooler 1991). Such behavior may be built in by evolution or be acquired via a long process of learning – but it need not require on-line computation of the optimal solution.

In some contexts, however, some on-line computations may be required. Specifically, if behavior is highly flexible with respect to environmental variation, then calculation is required to determine the correct behavior, and this calculation may be intractable. Thus the two leading theories of perceptual organization assume that the cognitive system seeks to optimize on-line either the *simplicity* (e.g., Leeuwenberg and Boselie, 1988) or *likelihood* (Helmholtz 1910/1962; see Pomerantz and Kubovy

1987) of the organization of the stimulus array. These calculations are recognized to be computationally intractable (see Chater 1996). This fact does not invalidate these theories, but it does entail that they can only be approximated in terms of cognitive algorithms. Within the literature on perceptual organization, there is considerable debate concerning the nature of such approximations, and which perceptual phenomena can be explained in terms of optimization, and which result from the particular approximations that the perceptual system adopts (Helm and Leeuwenberg 1996).

It is important to note also that, even where a general cognitive goal is intractable, a more specific cognitive goal relevant to achieving the general goal may be tractable. For example, the general goal of moving a piece in chess is to maximise the chance of winning. However, this optimisation problem is known to be completely intractable because the search space is so large. But optimising local goals, such as controlling the middle of the board, weakening the opponent's king, and so on, may be tractable. Indeed, most examples of optimality-based explanations, whether in psychology, animal behaviour or economics, are defined over a local goal, which is assumed to be relevant to some more global aims of the agent. For example, evolutionary theory suggests that animal behaviour should be adapted so as to increase an animal's inclusive fitness, but specific explanations of animals' foraging behaviour assume more local goals. Thus, an animal may be assumed to forage so as to maximise food intake, on the assumption that this local goal is generally relevant to the global goal of maximising inclusive fitness. Similarly, the explanations concerning cognitive processes discussed in rational analysis in cognitive psychology concern local cognitive goals such as maximising the amount of useful information remembered, maximising predictive accuracy, or acting so as to gain as much information as possible. All of these local goals are assumed to be relevant to more general goals, such as maximising expected utility (from an economic perspective) or maximising inclusive fitness (from a biological perspective). At any level, it is possible that optimisation is intractable; but it is also possible that by focusing on more limited goals, evolution or learning may have provided the cognitive system with mechanisms that can optimise or nearly optimise some more local, but relevant, quantity.

The observation that the local goals may optimised as surrogates for the larger aims of the cognitive system raises another important question about providing rational models of cognition. The fact that a model involves optimising something does not mean that the model is a rational model. Optimality is not the same as rationality. It is crucial that the local goal

that is optimised must be relevant to some larger goal of the agent. Thus, it seems reasonable that animals may attempt to optimise the amount of food they obtain, or that the categories used by the cognitive system are optimised to lead to the best predictions. This is because, for example, optimising the amount of food obtained is likely to enhance inclusive fitness, in a way that, for example, maximising the amount of energy consumed in the search process would not. Determining whether some behaviour is rational or not therefore depends on more than just being able to provide an account in terms of optimisation. Therefore rationality requires not just optimising something but optimising something reasonable. As a definition of rationality, this is clearly circular. But by viewing rationality in terms of optimisation, general conceptions of what are reasonable cognitive goals can be turned into specific and detailed models of cognition. Thus, the program of rational analysis, while not answering the ultimate question of what rationality is, nonetheless provides the basis for a concrete and potentially fruitful line of empirical research.

This flexibility of what may be viewed as rational, in building a rational model, may appear to raise a fundamental problem for the entire rational analysis program. It seems that the notion of rationality may be so flexible that whatever people do, it is possible that it may seem rational under some description. So for example, to pick up an example we have already mentioned, it may be that our stomachs are well adapted to digesting the food in our environmental niche. Indeed they may even prove to be optimally efficient in this respect. However, we would not therefore describe the human stomach as rational, because stomachs presumably cannot usefully be viewed as information processing devices, which approximate, to any degree, the dictates of normative theories of formal rationality. Stomachs may be well or poorly adapted to their function (digestion), but they have no beliefs, desires or knowledge, and make no decisions or inferences. Thus, their behavior cannot be given a rational analysis and hence they cannot be related to the optimal performance provided by theories of formal rationality. Hence the question of the stomach's rationality does not arise.

In this section, we have seen that rational analysis provides a mode of explaining behavior which clarifies the relationship between the stuff of everyday rationality, reasoning with particular goals, in a specific environment, with specific computational constraints, and apparently abstract principles of formal rationality in probability theory, decision theory or logic. Formal rational principles spell out the optimal solution for the information processing problem that the agent faces. The assumption is that a well-adapted agent will approximate this solution to some degree.

3. RE-EVALUATING EMPIRICAL DATA ON HUMAN REASONING

We began by discussing the controversy concerning the relationship between formal theories of rationality and the everyday notion of the rationality that underlies effective thought and action in the world. We have seen how everyday rationality can be underpinned by principles of formal rationality in rational analysis. We now consider how rational analysis can be applied to explaining data on human reasoning gained from laboratory tasks. The rational analysis approach allows us to see laboratory performance, which has typically been viewed as systematically nonrational, as having a rational basis. This diffuses a crucial tension at the heart of the psychology and philosophy of rationality – between the manifest success of cognition in dealing with the complexities of the everyday world, and the apparently stumbling and flawed performance on laboratory reasoning tasks.

Everyday rationality is a matter of being adapted to the structure and goals in the real world. Thus, rational explanation, whether in animal behavior, economics or psychology, assumes that the agent is well-adapted to its normal environment. However, almost all psychological data is gained in a very unnatural setting, where a person performs an artificial task in the laboratory. Any laboratory task will recruit some set of cognitive mechanisms that determine the participant's behaviour. But it is not obvious what problem these mechanisms are adapted to solving. This adaptive problem is not likely to be directly related to the problem given to the participant by the experimenter, precisely because adaptation is to the natural world, not to laboratory tasks. In particular, this means that participants may fail with respect to the task that the experimenter thinks they have set. But this may be because this task is unnatural with respect to the participant's normal environment. Consequently people may assimilate the task that they are given to a more natural task, recruiting adaptively appropriate mechanisms which solve this, more natural, task successfully.

In the area of research known as the "psychology of deductive reasoning" (e.g., Evans et al. 1993; Johnson-Laird and Byrne 1991; Rips 1994), people are given problems that the experimenters conceive of as requiring logical inference. But they consistently respond in a non-logical way. Thus, human rationality appears to be called into question (Stein 1996; Stich 1985, 1990).

But the perspective of rational analysis suggests an alternative view. We argue first that everyday rationality is founded on uncertain rather than certain reasoning. This suggests that probability provides a better starting point for a rational analysis of human reasoning than logic. Second, we

argue that a probabilistic rational analysis of classic "deductive" reasoning tasks provides an excellent empirical fit with observed performance. The upshot is that much of the experimental research in the "psychology of deductive reasoning" does not engage people in deductive reasoning at all – but rather engages strategies suitable for probabilistic reasoning. Thus, the field of research appears to be crucially misnamed! But more importantly, probabilistic rational analysis helps resolve the tension between apparently poor laboratory reasoning performance, and the conspicuous success of everyday rationality. Laboratory performance is rational after all, once the appropriate rational standard is adopted.

We now illustrate this approach by sketching, in varying degrees of detail, the probabilistic rational analysis of three key "deductive" reasoning tasks: Wason's selection task, conditional inference, and syllogistic reasoning. We then briefly reconsider empirical evidence on human probabilistic reasoning, and how it relates to the probabilistic reasoning framework that we have developed.

3.1. *Wason's Selection Task*

Wason's selection task (Wason 1966, 1968) is perhaps the most intensively studied task in the psychology of reasoning, and perhaps the "deductive" reasoning task that has raised the greatest concerns about human rationality (e.g., Cohen 1981; Stein 1996; Stich 1985, 1990; Sutherland 1992).

In the selection task, people must assess whether some evidence is relevant to the truth or falsity of a conditional rule of the form *if p then q*, where by convention "*p*" stands for the antecedent clause of the conditional and "*q*" for the consequent clause. In the standard abstract version of the task, the rule concerns cards, which have a number on one side and a letter on the other. The rule is *if there is a vowel on one side* (*p*), *then there is an even number on the other side* (*q*). Four cards are placed before the subject, so that just one side is visible; the visible faces show an "A" (*p* card), a "K" (*not-p* card), a "2" (*q* card) and a "7" (*not-q* card). Subjects then select those cards they must turn over to determine whether the rule is true or false. Typical results were: *p* and *q* cards (46%); *p* card only (33%), *p*, *q* and *not-q* cards (7%), *p* and *not-q* cards (4%) (Johnson-Laird and Wason 1970).

The task subjects confront is analogous to a central problem of experimental science: the problem of which experiment to perform. The scientist has a hypothesis (or a set of hypotheses) which they must assess (for the subject, the hypothesis is the conditional rule); and must choose which experiment (card) will be likely to provide data (i.e., what is on the reverse of the card) which bears on the truth of the hypothesis.

In the light of the epistemological arguments we have already considered, it may seem unlikely that this kind of scientific reasoning will be deductive in character. Nonetheless, the psychology of reasoning has viewed the selection task as paradigmatically deductive (e.g., Evans 1982; Evans et al. 1993), although a number of authors have argued for a non-deductive conception of the task (Fischhoff and Beyth-Marorn 1983; Kirby 1994; Klayman and Ha 1987; Rips 1990).

The assumption that the selection task is deductive in character arises from the fact that psychologists of reasoning have tacitly accepted Popper's hypothetico-deductive philosophy of science. Popper (1959/1935) assumes that evidence can falsify but not confirm scientific theories. Falsification occurs when predictions that follow deductively from the theory do not accord with observation. This leads to a recommendation for the choice of experiments: only to conduct experiments that have the potential to falsify the hypothesis under test.

Applying the falsificationist account to the selection task, the recommendation is that subjects should only turn cards that are potentially logically incompatible with the conditional rule. When viewed in these terms, the selection task has a deductive component, in that the subject must deduce which cards would be logically incompatible with the conditional rule. According to the rendition of the conditional as material implication (which is standard in the propositional and predicate calculi, see Haack 1978), the only observation that is incompatible with the conditional rule *if p then q* is a card with *p* on one side and *not-q* on the other. Hence the subject should select only cards that could potentially produce such an instance. That is, they should turn the *p* card, since it might have a *not-q* on the back; and the *not-q* card, since it might have a *p* on the back.

This pattern of selection is rarely observed in the experimental results outlined above. Subjects typically select cards that could *confirm* the rule, i.e., the *p* and *q* cards. However, according to falsification the choice of the *q* card is irrational, and is an example of so-called "confirmation bias" (Evans and Lynch 1973; Wason and Johnson-Laird 1972). The rejection of confirmation as a rational strategy follows directly from the falsificationist perspective.

We have argued that the usual standard of "correctness" in the selection task follows from Popper's hypothetico-deductive view of science. Rejecting the falsificationist picture would eliminate the role of logic, and hence deduction, in the selection task. The hypothetico-deductive view faces considerable difficulties as a theory of scientific reasoning (Kuhn 1962; Lakatos 1970; Putnam 1974). This suggests that psychologists should explore alternative views of scientific inference that may provide differ-

ent normative accounts of experiment choice, and hence might lead to a different "correct" answer in the selection task. Perhaps the dictates of an alternative theory might more closely model human performance, and hence be consistent with the possibility of human rationality.

Oaksford and Chater (1994) adopted this approach, adapting the Bayesian approach to philosophy of science (Earman 1992; Horwich 1982; Howson and Urbach 1989), rather than the hypothetico-deductive view, to provide a rational analysis of the selection task. They view the selection task in probabilistic terms, as a problem of Bayesian optimal data selection (Good 1966; Lindley 1956; MacKay 1992). Suppose that you are interested in the hypothesis that eating tripe makes people feel sick. Should known tripe-eaters or tripe-avoiders be asked whether they feel sick? Should people known to be, or not to be, sick be asked whether they have eaten tripe? This case is analogous to the selection task. Logically, you can write the hypothesis as a conditional sentence, if you eat tripe (p) then you feel sick (q). The groups of people that you may investigate then correspond to the various visible card options, p, *not-p*, q and *not-q*. In practice, who is available will influence decisions about which people you question. The selection task abstracts away from this factor by presenting one example of each potential source of data. In terms of our everyday example, it is like coming across four people, one known tripe eater, one known not to have eaten tripe, one known to feel sick, and one known not to feel sick. The task is to decide whom to question about how they feel or what they have eaten.

Oaksford and Chater (1994, 1996) suggest that hypothesis testers should choose experiments (select cards) to provide the greatest "expected information gain" in deciding between two hypotheses: (i) that the task rule, if p then q, is true, i.e., ps are invariably associated with qs, and (ii) that the occurrence of ps and qs are independent. For each hypothesis, Oaksford and Chater (1994) define a probability model that derives from the prior probability of each hypothesis (which for most purposes they assume to be equally likely, i.e., both = 0.5), and the probabilities of p and of q in the task rule. They define information gain as the difference between the uncertainty *before* receiving some data and the uncertainty after receiving that data where they measure uncertainty using Shannon–Wiener information. Thus Oaksford and Chater define the information gain of data D as:

Information before receiving D:

$$I(H_i) = -\sum_{i=1}^{n} P(H_i) \log_2 P(H_i)$$

Information after receiving D:

$$I(H_i|D) = -\sum_{i=1}^{n} P(H_i|D) \log_2 P(H_i|D)$$

Information gain:

$$I_g = I(H_i) - I(H_i|D)$$

They calculate the $P(H_i|D)$ terms using Bayes' theorem. Thus information gain is the difference between the information contained in the *prior* probability of a hypothesis (H_i) and the information contained in the *posterior* probability of that hypothesis given some data D.

When choosing which experiment to conduct (that is, which card to turn), the subject does not know what that data will be (that is, what will be on the back of the card). So they cannot calculate actual information gain. However, subjects can compute *expected* information gain. Expected information gain is calculated with respect to all possible outcomes, e.g., for the p card, the possible outcomes with regard to what will be found on the back of the card are q and *not-q*; and the calculation also averages over both hypotheses (that the rule is true, or that p and q are independent).

Oaksford and Chater (1994) calculated the expected information gain of each card assuming that the properties described in p and q are rare. This is an appropriate default because in typical everyday rule such as *if its a raven then its black*, only a small minority of things satisfy the antecedent (most things are not ravens) or the consequent (most things are not black). (Klayman and Ha 1987, make a similar assumption in accounting for related data on Wason's (1960), 2-4-6 task.) With this 'rarity' assumption, the order in expected information gain is:

$$E(I_g(p)) > E(I_g)) > E(I_g(\text{not-}q)) > E(I_g(\text{not-}p))$$

This corresponds to the observed frequency of card selections in Wason's task: $p > q >$ *not-q* $>$ *not-p* and thus explains the predominance of p and q card selections as a rational inductive strategy. Oaksford and Chater (1994) also show how their model generalises to all the main patterns of results in the selection task (for discussions of this account see Almor and Sloman 1996; Evans and Over 1996; Laming 1996; Klauer, in press, and for responses and developments see Oaksford and Chater 1996, 1998b, 1998c; Chater and Oaksford, in press, a). Specifically, it accounts for the non-independence of card selections (Pollard 1985), the negations paradigm (e.g., Evans and Lynch 1973), the therapy experiments (e.g.,

Wason 1969), the reduced array selection task (Johnson-Laird and Wason 1970), work on so-called fictional outcomes (Kirby 1994) and deontic versions of the selection task (e.g., Cheng and Holyoak 1985) including perspective and rule-type manipulations (e.g., Cosmides 1989; Gigerenzer and Hug 1992), the manipulation of probabilities and utilities in deontic tasks (Kirby 1994), and effects of relevance (Sperber et al. 1995; Oaksford and Chater 1995b).

We noted above that the philosophy of science that underlies the "deductive" conception of the selection task can be questioned. The current consensus is that scientific theories do not deductively imply predictions, and hence that the general problem of choosing which experiment to perform (or analogously, which card to turn in the selection task) cannot be reconstructed deductively. Further, Oaksford and Chater's (1994) probabilistic account provides a better model of human performance on the selection task. According to this model, people do not use deduction when solving the selection task, rather they use a probabilistic inferential strategy.

3.2. *Conditional Inference*

The selection task is perhaps the most celebrated "deductive" reasoning task. However, the conditional inference task is perhaps the task that seems most unequivocally to engage deductive reasoning processes. For example, Rips (1994) uses this task in introducing his mental logic theory of reasoning. If human reasoning is not deductive even in this task, then it seems unlikely that other areas of human reasoning will be well explained in deductive terms. For this reason, the conditional reasoning task is a particularly crucial test-case for theories of reasoning that employ deductive logic as a computational level theory.

In the standard conditional inference task, participants see a conditional rule, *if p then q*, an additional premise (*p*, *q*, *not-p* or *not-q*) and are asked whether a given conclusion (again, *p*, *q*, *not-p* or *not-q*) follows. Consider the simplest form of the task, where the premises are *p* and *if p then q*, and participants decide whether *q* follows. This appears to be an example of the paradigmatic deductive inference of *modus ponens.* Rips's (1994) central example of deductive inference has this form:

(1) If Calvin deposits 50 cents, he'll get a coke.
Calvin deposits 50 cents
Therefore, Calvin will get a coke

Interpreting this natural language argument involves applying a standard logical analysis, which presupposes that it should be viewed in deduct-

ive terms. However, this inference seems to be a typical example of a probabilistic or uncertain inference, and not an instance of the deductive reasoning at all, despite Rips. Calvin won't get the coke if the machine is broken, if the cokes have run out, if the power is turned off, and so on. That is, additional premises can overturn the conclusion, which deductive inference does not allow. Thus, although the task is *intended* as a test of deductive reasoning, the subject may be more likely to *interpret* the reasoning materials so that it involves uncertain, probabilistic reasoning.

The question for the psychology of reasoning, then, is which account of how people interpret and reason with the materials in the task provides the best fit with reasoning performance. It turns out that the experimental data support the claim that people treat such inferences as defeasible rather than deductive. Work on conditional inference indicates that subjects interpret conditional sentences as default rules (Holyoak and Spellman 1993) even in laboratory tasks (Oaksford et al. 1990). Byrne (1989) and Cummins et al. (1991) have shown that background information derived from stored world knowledge can affect inferential performance (see also, Markovits 1984, 1985). Specifically they showed that additional antecedents influence the inferences conditional statements allow. For example:

(2) If you turn the key the car starts.

(3) *Additional Antecedent:* You are out of petrol.

(2) could be used to predict that the car will start if you turn the key. This is an inference by *modus ponens.* However, including information about an additional antecedent (3) defeats this inference (Byrne 1989). Moreover, confidence reduces in this inference for rules that possess many alternative antecedents even when this information is only implicit (Cummins et al. 1991). Additional antecedents also affect inferences by *modus tollens.* If the car does not start, you can infer that you didn't turn the key, unless you are out of petrol. Explicitly providing information about alternative antecedents undermines the use of *modus tollens* (Byrne 1989) and reduces confidence in rules that possess many alternative antecedents even when this information is only implicit (Cummins et al. 1991). This result was very striking, and unexpected, within the context of the psychology of reasoning. However, in the light of the uncertainty of everyday reasoning, it is just what we would expect. Human inferences about coke machines, as about the rest of the external world, are defeasible

In sum, the experimental data seem to show that people treat conditionals in laboratory reasoning tasks as default rules. So it seems that

even the everyday inferences that some reasoning researchers regard as paradigmatic examples of deduction, like (1), are not examples of deductive inference at all. If defeasibility infects even such paradigmatic cases of deductive reasoning, then it threatens to leave the advocate of deductive reasoning with no everyday reasoning at all to explain.

Conditional inferences, like the everyday examples with which we introduced this paper, involve two premises, one conditional, If *A* then *B*, and one categorical, either, *A*, not-*A*, *B*, or not-*B*. For example, given If *A* then *B*, and not-*A*, people are asked to say whether, not-*B* follows. Endorsing this argument is to endorse the logical fallacy of denying the antecedent (DA). Interesting biases arise when negations are used in the conditional premise, e.g., If not-*A*, then not-*B*, and not-*A*, therefore not-*B* is an instance of the valid inference form modus ponens (MP). Evans (1977, 1993) observed a bias towards accepting conclusions containing a negation, like the MP inference above (using a different rule an affirmative conclusion follows by MP, e.g., If not-*A*, then *B*, not-*A*, therefore *B*).

This effect, which Evans et al. (1993) calls *negative conclusion bias*, may have a straightforward explanation on the assumption that people endorse arguments to the extent that the conditional probability of the conclusion given the categorical premise is high. This will depend on the probabilities of *A* and of *B* and on the conditional probability relating the two. So if we look at DA, the conditional probability that reeds to be high is $P(\text{not-}B \mid \text{not-}A)$. The probability of a negated category is higher than an affirmative category (Oaksford and Chater 1994; Oaksford and Stenning 1992), e.g., the probability that you are not drinking whiskey as you read this paper is higher than the probability that you are. To illustrate very simply how negative conclusion bias could arise, let us assume that you believe the rule is false. On the account of Wason's selection task outlined above, this means that you believe that *A* and *B* are independent. Consequently $P(\text{not-}B \mid \text{not-}A) = P(\text{not-}B)$, i.e., you should endorse the DA inference if the probability of the conclusion is high. And because negated conclusions have a higher probability than affirmative conclusions, the former should be endorsed more often. In sum, the probabilistic rational analysis that we developed for the selection task appears to carry over relatively directly to conditional reasoning.

3.3. *Syllogistic Reasoning*

Syllogisms may appear paradigms of deductive logic. Aristotle's theory of syllogisms constituted the only account of valid argumentation for more than 2000 years. Indeed, until the 19th century, the theory of syllogisms was widely viewed as exhausting the study of argument. For example,

Kant (1961, 501) argued that since Aristotle "it is remarkable . . . that to the present day [logic] has not been able to make one step in advance, so that, to all appearance, it may be considered as completed and perfect". Nonetheless, we shall argue that people do not treat even syllogistic reasoning as a deductive task.

Syllogistic reasoning involves two quantified statements of the form, All X are Y, No X are Y, Some X are Y, or Some X are not Y. Some combinations of premises yield logically valid conclusions, e.g., All X are Y, All Y are Z yields the logically valid conclusion, All X are Z; others do not, e.g., No Y are X, Some Y are not Z, has no valid conclusion. If people were reasoning logically then they should be able to draw all and only the valid conclusions indicating that nothing necessarily follows from the invalid syllogisms. However, people have graded difficulty with drawing the valid syllogisms. Moreover, they make systematic errors on the invalid syllogisms, offering conclusions where none follow.

Chater and Oaksford (in press, b; see also Manktelow, in press, for an exposition) adopt a probabilistic approach to syllogisms. Thus, the quantified statement All X are Y is interpreted as a conditional probability, $P(Y \mid X) = 1$; the statement Some X are not Y is interpreted as the joint probability, $P(X, \text{not-}Y) > 0$, and so on. They then develop a notion of informational strength (probabilistically defined) of premises to guide conclusion construction. *All . . .* statements are the most informative (roughly, the most unlikely to the true of arbitrarily chosen predicates); *Some . . . not . . .* statements are the least informative.

It turns out that the most informative conclusion that can follow from a syllogism is given by the least informational premise. Moreover, for most valid syllogisms the least informational premise also provides the form of the conclusion. Thus selecting the form of the least information premise as the form of the conclusion will usually procluce a valid conclusion if there is one. If this strategy is overgeneralised it can also explain the systematic errors made on the invalid syllogisms. Consequently Chater and Oaksford (in press, b) show that a very simple strategy can be explain syllogistic reasoning performance. Moreover, this probabilistic account has the advantage that not only can it explain the data from the 64 syllogisms that use the standard logical quantifiers (see above), it also extends naturally to the 144 syllogisms that result from combining these with the *generalised* quantifiers, Most and Few which have no logical interpretation.

These analyses raise the apparently paradoxical possibility that explaining all of the key experimental paradigms for studying human deductive reasoning requires viewing people's performance as approximating to probabilistic rather than deductive inference. In short, people reason prob-

abilistically even when faced with what the experimenter intends to be a deductive reasoning task. Reasoning strategies are adapted to deal with uncertainty in everyday life – and therefore these strategies are likely to be carried over by people into laboratory settings. Thus, paying closer attention to everyday reasoning may provide the key to giving a detailed analysis of laboratory performance.

3.4. *Probabilistic Reasoning*

The approach we have outlined might be characterised as arguing that although people are poor at logical reasoning they are nonetheless good at probabilistic reasoning. However, this viewpoint seems to be at odds with established results that appear to show that people are also very poor probabilistic reasoners (e.g., Tversky and Kahneman 1974; Kahneman et al. Tversky, 1982). For example, people seem to be insensitive to base rates, i.e., in applying Bayes's theorem people often provide estimates of posterior probabilities that seem to reflect only the likelihoods and not the priors. People also seem to be overconfident in their probability judgements, i.e., they do not seem to be well calibrated to the actual frequencies of events in the world. Moreover, people also seem prone to the conjunction fallacy. That is, they violate the probabilistic law that the joint probability of any two events can not he greater than either individual event, i.e., $P(A) \geq P(A, B)$.

There are two reasons why tension between a probabilistic rational analysis of people's reasoning strategies does not conflict with understanding reasoning in terms of proabilistic rational analysis, however.

First, recall that the probabilistic rational analysis is a way of assessing what strategies will be adaptively successful. There is no assumption that the cognitive system actually makes probabilistic calculations, simply that the strategies that it adopts are adaptively successful.

Second, according to recent analyses, many of the apparent errors and biases observed in probabilistic reasoning are a consequence of presenting the probabilistic information in an unnatural format (Gigerenzer and Hoffrage 1995). Most often in experiments of this type people are given the probabilistic information in terms of explicit probability statements or percentages, e.g., 0.05 or 5%. However, Gigerenzer et al. (1995) argue that this is unnatural given the normal sampling situation where we build up frequency information as a result of multiple encounters with objects and events. What you discover by such a process is, for example, that something like 95 out of the 100 ravens you have examined are black. Mathematically this information can be expressed as 95% of ravens are black, or the probability of a bird being black given it is a raven is 0.95.

However, this loses information about sample size and moreover, it seems unnecessary to make this conversion of the information format. Gigerenzer and Hoffrage suggest that if people naturally represent frequencies then presenting probabilistic information in this form should facilitate reasoning. We illustrate research showing that Gigerenzer and Hoffrage appear to be correct in the three areas where biases have been observed and which we introduced above.

Experiments revealing base rate neglect usually present the information as follows, using the mammogram problem:

> A thirty year old woman discovers a lump in her breast and goes to her doctor. The doctor knows that only 5% of women of the patient's age and health have breast cancer (C). A mammogram (breast X ray) is taken. It indicates cancer 80% of the time in women who have breast cancer but falsely indicates breast cancer in healthy patients 20% of the time. The mammogram (M) comes out positive. What is the probability that the patient has cancer?

Most participants in an experiment such as this give estimates that the woman has cancer given a positive mammogram of around 0.80, which appears to ignore the prior that most women of her age, i.e., 95% do not have breast cancer. However, a simple change in the instructions reverses this finding:

> A thirty year old woman discovers a lump in her breast and goes to her doctor. The doctor knows that only 5 out of every 100 women of the patient's age and health have breast cancer (C). A mammogram (breast X ray) is taken. For 80 out of every 100 women who have breast cancer it gives a positive result but falsely indicates breast cancer in 20 out of every 100 healthy patients. The mammogram (M) comes out positive. What is the probability that the patient has cancer?

Gigerenzer and Hoffrage argue that the frequency information also allows a simpler version of Bayes theorem to be used hence reducing cognitive load.

In discussing overconfidence Gigerenzer points out that like is not being compared with like. People are typically asked a series of general knowledge questions and are asked to rate their confidence in each answer. To determine overconfidence, their average confidence rating is compared with the frequency of correct answers. That is, people are asked repeatedly about their beliefs in single events, and then their average performance

on this task is compared with their relative frequency of correct answers. Gigerenzer observes that these can be independent judgements. To test whether overconfidence arises when like is compared with like, at the end of the task Gigerenzer also asked people to estimate their relative frequency of correct answers. Comparing their estimates with their actual frequency of correct answers revealed no evidence of overconfidence. That is, when like is compared with like, people seem well calibrated in judging their own likelihood of success.

The conjunction fallacy seems also to emerge because of the unnatural presentation of probabilities. People are typically given information such as:

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

They are then asked to estimate the probability that (i) Linda is a bankteller, and (ii) Linda is a feminist bankteller. People typically estimate (ii) as more likely than (i), violating the conjunction rule. However, if people are asked this question using a frequency format such as: There are 100 people who fit the description above; How many of them are: (i) Bank tellers, (ii) Bank tellers and active in the feminist movement, then they do not estimate (i) as less likely than (ii), conforming to the conjunction rule.

In summary, it would appear that people are not as bad at probabilistic reasoning as the evidence from the heuristics and biases programme had led us to believe. Moreover, as we noted above, the theoretical accounts of reasoning we have discussed do not require that people possess quantitatively accurate probabilistic reasoning abilities. Thus, any apparent tension between the probabilistic approach to the rational analysis of reasoning that we advocate and experimental data on human probabilistic reasoning is illusory.

4. CONCLUSION

This paper has aimed to establish two theses. The first is that the empirical program of rational analysis provides an account of the relationship between everyday rationality and formal rationality. The connection is that formal rational principles are used to derive the optimal solution to achieving the cognitive system's goals given environmental and computational constraints. Rational analysis is an empirical, rather than an a priori,

enterprise. This is because the choice of rational principles and the goals and constraints to which they are applied constitute empirical hypotheses, which are intended to account for psychological data.

The second thesis that we have aimed to establish is that human laboratory reasoning does not demonstrate human irrationality. People carry over probabilistic reasoning strategies which are appropriate to dealing with the uncertainty of the everyday world into the laboratory. These strategies are rationally justified, once an appropriate probabilistic standard of rationality has been adopted. Thus, the apparent tension between psychological data on human reasoning and the conspicuous success of everyday rationality is illusory. Even in laboratory tasks, people may not be logical; but they are rational.

NOTES

* Please address correspondence concerning this article to Nick Chater, Department of Psychology, University of Warwick, Coventry, CV4 7AL, UK or to Mike Oaksford, School of Psychology, Cardiff University, P.O. Box 901, Cardiff CF1 3YG, Wales, UK.

1 There are also a range of other justifications of the laws of probability theories as a calculus of uncertain inference, based on preferences (Savage 1954), scoring rules (Lindley 1982) and derivation from minimal axioms (Cox 1961; Good 1950; Lucas 1970). Although each argument can be challenged individually, the fact that so many different lines of argument converge on the very same laws of probability has been taken as powerful evidence for the view that degrees of belief can be interpreted as probabilities (see, e.g., Howson and Urbach, 1989; and Earman 1992, for discussion).

2 For example, mental models correspond to mental representations of states of affairs, rather than states of affairs themselves; and these mental representations have a specific syntax, and presumably a specific semantics. The precise semantic properties of mental models representation has not been given, and indeed, and it is not clear how this could be done. Instead, the semantics of mental models is left, rather uncomfortably, in the hands of the theorists' intuitions.

3 Note also that for all reasonably rich scientific theories, any empirical data can be accommodated, by suitable changes in auxiliary assumptions (Quine 1953). Thus rational explanations are no different in this regard, from, e.g., explanations in terms of the principles of Newtonian mechanics (Putnam 1974).

REFERENCES

Allais, M.: 1953, 'Le Comportement de l'Homme Rationnel Devant le Risque: Critique des Postulats et Axiomes de l'École Américaine', *Econometrica* **21**, 503–546.

Almor, A. and S Sloman: 1996, 'Is Deontic Reasoning Special?', *Psychological Review* **103**, 374–380.

Anderson, J. R. and R. Milson: 1989, 'Human Memory: An Adaptive Perspective', *Psychological Review* **96**, 703–719.

Anderson, I. R.: 1990, *The Adaptive Character of Thought*, LEA, Hillsdale, NJ.
Anderson, I. R.: 1991a. 'Is Human Cognition Adaptive?', *Behavioral and Brain Sciences* **14**, 471–517.
Anderson, J. R.: 1991b., 'The Adaptive Nature of Human Categorization', *Psychological Review* **98**, 409–429.
Anderson, J. R.: 1993, *Rules of the Mind*, Erlbaum, Hillsdale, NJ.
Anderson, J. R. and M. Matessa: 1998, 'The Rational Analysis of Categorization and the ACT-R Architecture', in M. Oaksford and N. Chater (eds), *Rational Models of Cognition*, Oxford University Press, Oxford, pp. 197–217.
Anderson, J. R. and L. J. Schooler: 1991, 'Reflections of the Environment in Memory', *Psychological Science* **2**, 396–408.
Becker, G.: 1975, *Human Capital* (2nd Edition), Columbia University Press, New York.
Becker, G.: 1981, *A Treatise on the Family*, Harvard University Press, Cambridge, MA.
Braine, M. D. S.: 1978, 'On the Relation between the Natural Logic of Reasoning and Standard Logic', *Psychological Review* **85**, 1–21.
Byrne, R. M. J.: 1989, 'Suppressing Valid Inferences with Conditionals', *Cognition* **31**, 1–21.
Chater, N.: 1996, 'Reconciling Simplicity and Likelihood Principles in Perceptual Organization', *Psychological Review* **103**, 566–581.
Chater, N., M. Crocker, and M. Pickering: 1998, 'The Rational Analysis of Inquiry: The Case of Parsing', in M. Oaksfield and N. Chater (eds), *Rational Models of Cognition*, Oxford University Press, Oxford, pp. 441–468.
Chater, N. and M. Oaksford: 1990, 'Autonomy, Implementation and Cognitive Architecture: A Reply to Fodor and Pylyshyn', *Cognition* **34**, 93–107.
Chater, N. and M. Oaksford: 1999, a, 'Information Gain vs. Decision-Theoretic Approaches to Data Selection: Response to Klauer', *Psychological Review* **106**, 223–227.
Chater, N. and M. Oaksford: in press, b, 'The Probability Heuristics Model of Syllogistic Reasoning', *Cognitive Psychology* **38**, 191–258.
Cheng, P. W. and K. J. Holyoak: 1985, 'Pragmatic Reasoning Schemas', *Cognitive Psychology* **17**, 391–416.
Cherniak, C.: 1986, *Minimal Rationality*, MIT Press, Cambridge, MA.
Chomsky, N.: 1965, *Aspects of the Theory of Syntax*, MIT Press, Cambridge, MA.
Clark, K. L.: 1978, 'Negation as Failure', in *Logic and Databases*, Plenum Press, New York, pp. 293–322.
Cohen, L. J.: 1981, 'Can Human Irrationality be Experimentally Demonstrated?', *Behavioral and Brain Sciences* **4**, 317–370.
Cosmides, L.: 1989, 'The Logic of Social Exchange: Has Natural Selection Shaped How Humans Reason? Studies with the Wason Selection Task', *Cognition* **31**, 187–276.
Cox, R. T.: 1961, *The Algebra of Probable Inference*, The Johns Hopkins University Press, Baltimore, MD.
Cummins, D. D., T. Lubart, O. Alksnis, and R. Rist: 1991, 'Conditional Reasoning and Causation', *Memory and Cognition* **19**, 274–282.
Davidson, D.: 1984, *Inquiries into Truth and Interpretation*, Clarendon Press, Oxford.
Dawkins, R.: 1977, *The Selfish Gene*, Oxford University Press, Oxford.
Earman, J.: 1992, *Bayes or Bust? A Critical Examination of Bayesian Confirmation Theory*, MIT Press, Cambridge, MA.
Elster, J. (ed.): 1986, *Rational Choice*, Basil Blackwell, Oxford.
Evans, J. St. B. T.: 1977, 'Linguistic Factors in Reasoning', *Quarterly Journal of Experimental Psychology* **29**, 297–306.

Evans, J. St. B. T.: 1982, *The Psychology of Deductive Reasoning*, Routledge and Kegan Paul, London.
Evans, J. St. B. T.: 1989, *Bias in Human Reasoning: Causes and Consequences*, Erlbaum, Hillsdale, NJ.
Evans, J. St. B. T.: 1993, 'The Mental Model Theory of Conditional Reasoning: Critical Appraisal and Revision', *Cognition* **48**, 1–20.
Evans, J. St. B. T. and J. S. Lynch: 1973, 'Matching Bias in the Selection Task', *British Journal of Psychology* **64**, 391–397.
Evans, J. St. B. T., S. E. Newstead, and R. M. J. Byrne: 1993, *Human Reasoning*, Lawrence Erlbaum Associates, Hillsdale, N.J.
Evans, J. St. B. T. and D. Over: 1996, Rationality in the Selection Task: Epistemic Utility vs. Uncertainty Reduction', *Psychological Review* **103**, 356–363.
Evans, J. St. B. T. and D. Over: 1997, 'Rationality in Reasoning: The Problem of Deductive Competence', *Cahiers de Psychologie Cognitive* **16**, 1–35.
Finetti, B. de: 1937, 'La Prevision: Ses Lois Logiques, ses Sources Subjectives', *Annales de l'Institute Henri Poincaré* **7**, 1–68. [Translated as 'Foresight: Its Logical Laws, its Subjective Sources', in H. E. Kyburg and H. E. Smokler (eds), 1964, *Studies in Subjective Probability*, Wiley, Chichester, UK.]
Fisher, R. A.: 1922, 'On the Mathematical Foundations of Theoretical Statistics', *Philosophical Transactions of the Royal Society of London, Series A*, **222**, 309–368.
Fisher, R. A.: 1925/1970, *Statistical Methods for Research Workers*, 14th Edition, Oliver and Boyd, Edinburgh.
Fischhoff, B. and R. Beyth-Marom: 1983, 'Hypothesis Evaluation from a Bayesian Perspective', *Psychological Review* **90**, 239–260.
Fodor, J. A.: 1983, *Modularity of Mind*, MIT Press, Cambridge, MA.
Fodor, J. A.: 1987, *Psychosemantics*, MIT Press, Cambridge, MA.
Fodor, J. A. and Z. W. Pylyshyn: 1988, 'Connectionism and Cognitive Architecture: A Critical Analysis', *Cognition* **28**, 3–71.
Gallistel, C. R.: 1990, *The Organization of Learning*, MIT Press, Cambridge, MA.
Garey, M. R. and D. S. Johnson: 1979, *Computers and Intractability: A Guide to the Theory of NP-Completeness*, W. H. Freeman, San Francisco.
Gigerenzer, G. and D. Goldstein: 1996, 'Reasoning the Fast and Frugal Way: Models of Bounded Rationality', *Psychological Review* **103**, 650–669.
Gigerenzer, G. and U. Hoffrage: 1995, 'How to Improve Bayesian Reasoning without Instruction: Frequency Formats', *Psychological Review* **102**, 684–704.
Gigerenzer, G. and K. Hug: 1992, 'Domain-Specific Reasoning: Social Contracts, Cheating, and Perspective Change', *Cognition* **43**, 127–71.
Good, I. J.: 1950, *Probability and the Weighting of Evidence*, Griffin, London.
Good, I. J.: 1966, 'A Derivation of the Probabilistic Explication of Information', *Journal of the Royal Statistical Society, Series B* **28**, 578–581.
Good, I. J.: 1971, 'Twenty Seven Principles of Rationality', in V. P. Godambe and D. A. Sprott (eds), *Foundations of Statistical Inference*, Holt, Rhinehart and Wilson, Toronto.
Goodman, N.: 1954, *Fact, Fiction and Forecast*, Harvard University Press, Cambridge, MA.
Haack, S.: 1978, *Philosophy of Logics*, Cambridge University Press, Cambridge.
Harsanyi, J. C. and R. Selten: 1988, *A General Theory of Equilibrium Selection in Games*, MIT Press, Cambridge, MA.
Helm, P. A. van der and E. L.J. Leeuwenberg: 1996, 'Goodness of Visual Regularities: A Non-transformational Approach', *Psychological Review* **103**, 429–456.

Helmholtz, H. von: 1910/1962, *Treatise on Physiological Optics*, Vol. 3, J. P. Southall (ed.) and translation, Dover, New York.
Holyoak, K. J. and B. A. Spellman: 1993, 'Thinking,' *Annual Review of Psychology* **44**, 265–315.
Horwich, P.: 1982, *Probability and Evidence*, Cambridge University Press, Cambridge.
Howson, C. and P. Urbach: 1989, *Scientific Reasoning: The Bayesian Approach*, Open Court, La Salle.
Johnson-Laird, P. N.: 1983, *Mental Models*, Cambridge University Press, Cambridge.
Johnson-Laird, P. N. and R. M. J. Byrne: 1991, *Deduction*, Erlbaum, Hillsdale, N.J.
Johnson-Laird, P. N. and P. C. Wason: 1970, 'Insight into a Logical Relation', *Quarterly Journal of Experimental Psychology* **22**, 49–61.
Kahneman, D., P. Slovic, and A. Tversky (eds): 1982, *Judgement under Uncertainty: Heuristics and Biases*, Cambridge University Press, Cambridge.
Kant, I.: 1961, *Critique of Pure Reason*, translated by F. M. Muller, Doubleday, Dolphin Books, New York.
Keynes, J. M.: 1921, *A Treatise on Probability*, Macmillan, London.
Kirby, K. N.: 1994, 'Probabilities and Utilities of Fictional Outcomes in Wason's Four Card Selection Task', *Cognition* **51**, 1–28.
Klauer, K. C.: 1999, 'On the Normative Justification for Information Gain in Wason's Selection Task', *Psychological Review* **106**, 215–222.
Klayman, J. and Y. Ha: 1987, 'Confirmation, Disconfirmation and Information in Hypothesis Testing', *Psychological Review* **94**, 211–228.
Kleindorfer, P. R., H. C. Kunreuther, and P. J. H. Schoemaker: 1993, *Decision Sciences: An Integrated Perspective*, Cambridge University Press, Cambridge.
Kuhn, T.: 1962, *The Structure of Scientific Revolutions* University of Chicago Press, Chicago.
Lakatos, I.: 1970, 'Falsification and the Methodology of Scientific Research Programmes', in I. Lakatos and A. Musgrave (eds), *Criticism and the Growth of Knowledge*, Cambridge University Press, Cambridge, pp. 91–196.
Lamberts, K. and S. Chong: 1998, 'Dynamics of Dimension Weight Distribution and Flexibility in Categorization' in M. Oaksford and N. Chater (eds), *Rational Models of Cognition*, Oxford University Press, Oxford, pp. 275–92.
Laming, D.: 1996, 'On the Analysis of Irrational Data Selection: A Critique of Oaksford and Chater (1994)', *Psychological Review* **103**, 364–73.
Leeuwenberg, E. and F. Boselie: 1988, 'Against the Likelihood Principle in Visual Form Perception', *Psychological Review* **95**, 485–91.
Lindley, D. V.: 1956, 'On a Measure of the Information Provided by an Experiment', *Annals of Mathematical Statistics* **21**, 986–1005.
Lindley, D. V.: 1971, *Bayesian Statistics: A Review*, Society for Industrial and Applied Mathematics, Philadelphia, PA.
Lindley, D. V.: 1982, 'Scoring Rules and the Inevitability of Probability', *International Statistical Review* **50**, 11-26.
Lucas, J. R.: 1970, *The Concept of Probability*, Oxford University Press, Oxford.
MacKay, D. J. C.: 1992, 'Information-based Objective Functions for Active Data Selection', *Neural Computation* **4**, 590–604.
Manktelow, K.: in press, *Reasoning and Thinking*, Psychology Press, Hove, Sussex, UK.
Markovits, H.: 1984, 'Awareness of the "Possible" as a Mediator of Formal Thinking in Conditional Reasoning Problems', *British Journal of Psychology* **75**, 367–376.

Markovits, H.: 1985, 'Incorrect Conditional Reasoning Among Adults: Competence or Performance', *British Journal of Psychology* **76**, 241–247.
Marr, D.: 1982, *Vision*, W. H. Freeman, San Francisco.
McCarthy, J. M.: 1980, 'Circumscription: A Form of Nonmonotonic Reasoning', *Artificial Intelligence* **13**, 27–39.
McCloskey, D. N.: 1985, *The Rhetoric of Economics*, University of Wisconsin Press, Madison.
McDermott, D.: 1982, 'Non-monotonic Logic II: Nonmonotonic Model Theories', *Journal of the Association for Computing Machinery* **29**, 33–57.
McDermott, D.: 1987, 'A Critique of Pure Reason', *Computational Intelligence* **3**, 151–160.
McDermott, D. and J. Doyle: 1980, 'Non-monotonic Logic I', *Artificial Intelligence* **13**, 41–72.
McFarland, D. and A. Houston: 1981, *Quantitative Ethology: The State Space Approach*, Pitman, London.
Minsky, M.: 1977, 'Frame System Theory', in P. N. Johnson-Laird and P. C. Wason (eds), *Thinking: Reading in Cognitive Science*, Cambridge University Press, Cambridge, pp. 355–376.
Muth, J. F.: 1961, 'Rational Expectations and the Theory of Price Movements', *Econometrica* **29**, 315–335.
Neumann, J. von and O. Morgenstern: 1944, *Theory of Games and Economic Behavior*, Princeton University Press, Princeton, NJ.
Neyman, J.: 1950, *Probability and Statistics*, Holt, New York.
Oaksford, M. and N. Chater: 1994, 'A Rational Analysis of the Selection Task as Optimal Data Selection', *Psychological Review* **101**, 608–631.
Oaksford, M. and N. Chater: 1995a, 'Theories of Reasoning and the Computational Explanation of Everyday Inference', *Thinking and Reasoning* **1**, 121–152.
Oaksford, M. and N. Chater: 1995b, 'Information Gain Explains Relevance which Explains the Selection Task', *Cognition* **57**, 97–108.
Oaksford, M. and N. Chater: 1996, 'Rational Explanation of the Selection Task', *Psychological Review* **103**, 381–391.
Oaksford, M. and N. Chater (eds): 1998a, *Rational Models of Cognition*, Oxford University Press, Oxford.
Oaksford, M. and N. Chater: 1998b, *Rationality in an Uncertain World*, Psychology Press Hove, England.
Oaksford, M. and N. Chater: 1998c, 'A Revised Rational Analysis of the Selection Task: Exceptions and Sequential Sampling', in M. Oaksford and N. Chater (eds), *Rational Models of Cognition*, Oxford University Press, Oxford, pp. 372–98.
Oaksford, M., N. Chater, and K. Stenning: 1990, 'Connectionism, Classical Cognitive Science and Experimental Psychology', *AI & Society* **4**, 73–90.
Oaksford, M. and K. Stenning: 1992, 'Reasoning with Conditionals Containing Negated Constituents', *Journal of Experimental Psychology: Learning Memory and Cognition* **18**, 835–854.
Paris, J.: 1992, *The Uncertain Reasoner's Companion*, Cambridge University Press, Cambridge.
Pollard, P.: 1985, 'Nonindependence of Selections on the Wason Selection Task', *Bulletin of the Psychonomic Society* **23**, 317–320.
Pomerantz, J. R. and M. Kubovy: 1987, 'Theoretical Approaches to Perceptual Organization', in K. R. Boff, L. Kaufman and J. P. Thomas (eds), *Handbook of Perception*

and Human Performance, Volume II: Cognitive Processes and Performance, Wiley, New York, pp. 36.1–36.
Popper, K. R.: 1959/1935, *The Logic of Scientific Discovery*, Hutchinson, London.
Putnam, H.: 1974, 'The "Corroboration" of Theories', in P. A. Schilpp (ed.), *The Philosophy of Karl Popper*, Vol. 1, Open Court Publishing, La Salle, pp. 221–40.
Pylyshyn, Z. W. (ed.): 1987, *The Robot's Dilemma: The Frame Problem in Artificial Intelligence*, Ablex, Norwood, NJ.
Quine, W. V. O.: 1953, 'Two Dogmas of Empiricism', in *From a Logical Point of View*, Harvard University Press, Cambridge, MA, pp. 20–46.
Quine, W. V. O.: 1960, *Word and Object*, MIT Press, Cambridge, MA.
Ramsey, F. P.: 1931, *The Foundations of Mathematics and Other Logical Essays*, Routledge and Kegan Paul, London.
Rawls, J.: 1971, *A Theory of Justice*, Harvard University Press, Cambridge, MA.
Reiner, R.: 1995, 'Arguments Against the Possibility of Perfect Rationality', *Minds and Machines* **5**, 373–89.
Reiter, R.: 1980, 'A Logic for Default Reasoning', *Artificial Intelligence* **13**, 81–132.
Reiter, R.: 1985/1978, 'On Reasoning by Default', in R. Brachan and H. Levesque (eds), *Readings in Knowledge Representation*, Morgan Kaufman, Los Altos, CA, pp. 401–10.
Rips, L. J.: 1990, 'Reasoning', *Annual Review of Psychology* **41**, 321–353.
Rips, L. J.: 1994, *The Psychology of Proof*, MIT Press, Cambridge, MA.
Savage, L. J.: 1954, *The Foundations of Statistics*, Wiley, New York.
Schank, R. C. and R. P. Abelson: 1977, *Scripts, Plans, Goals, and Understanding*, Erlbaum, Hillsdales, N.J.
Schooler, L. J.: 1998, 'Sorting out Core Memory Processes', in M. Oaksford and N. Chater (eds), *Rational Models of Cognition*, Oxford University Press, Oxford, pp. 128–55.
Simon, H. A.: 1955, 'A Behavioral Model of Rational Choice', *Quarterly Journal of Economics* **69**, 99–118.
Simon, H. A.: 1956, 'Rational Choice and the Structure of the Environment', *Psychological Review* **63**, 129–138.
Skyrms, B: 1977, *Choice and Chance*, Wadsworth, Belmont.
Sperber, D., F. Cara, and V. Girotto: 1995, 'Relevance Theory Explains the Selection Task', *Cognition* **57**, 31–95.
Stein, E.: 1996, *Without Good Reason*, Oxford University Press, Oxford.
Stephens, D. W. and J. R. Krebs: 1986, *Foraging Theory*, Princeton University Press, Princeton, NJ.
Stich, S.: 1985, 'Could Man be an Irrational Animal?', *Synthese* **64**, 115–135.
Stich, S.: 1990, *The Fragmentation of Reason*, MIT Press, Cambridge, MA.
Stich, S. and R. Nisbett: 1980, 'Justification and the Psychology of Human Reasoning', *Philosophy of Science* **47**, 188–202.
Sutherland, S.: 1992, *Irrationality: The Enemy Within*, Constable, London.
Thagard, P.: 1988, *Computational Philosophy of Science*, MIT Press, Cambridge, MA.
Tversky, A. and D. Kahneman, D.: 1974, 'Judgement under Uncertainty: Heuristics and Biases', *Science* **125**, 1124–1131.
Tversky, A. and D. Kalineman: 1986, 'Rational Choice and the Framing of Decisions', *Journal of Business* **59**, 251–278.
Wason, P. C.: 1960, 'On the Failure to Eliminate Hypotheses in a Conceptual Task', *Quarterly Journal of Experimental Psychology* **12**, 129–40.
Wason, P. C.: 1966, 'Reasoning', in B. Foss (ed.), *New Horizons in Psychology*, Penguin, Harmondsworth, Middlesex.

Wason, P. C.: 1968, 'Reasoning about a Rule', *Quarterly Journal of Experimental Psychology* **20**, 273–81.
Wason, P. C.: 1969, 'Regression in Reasoning', *British Journal of Psychology* **60**, 471–80.
Wason, P. C. and P. N. Johnson-Laird: 1972, *The Psychology of Reasoning: Structure and Content*, Harvard University Press, Cambridge, MA.
Young, R.: 1998, 'Rational Analysis of Exploratory Choice', in M. Oaksford and N. Chater (eds), *Rational Models of Cognition*, Oxford University Press, Oxford, pp. 469–500.

Nick Chater
Department of Psychology,
University of Warwick
U.K.
E-mail: nick.chater@warwick.ac.uk

Mike Oaksford
School of Psychology
Cardiff University
U.K.
E-mail: oaksford@cardiff.ac.uk