MIND ASSOCIATION OCCASIONAL SERIES

This series consists of occasional volumes of original papers on predefined themes. The Mind Association nominates an editor or editors for each collection, and may co-operate with other bodies in promoting conferences or other scholarly activities in connection with the preparation of particular volumes.

Publications Officer: M. A. Stewart
Secretary: J. Butterfield

Also published in the series:

Perspectives on Thomas Hobbes
Edited by G. A. J. Rogers and A. Ryan

Reality, Representation, and Projection
Edited by J. Haldane and C. Wright

Machines and Thought
The Legacy of Alan Turing
Edited by P. J. R. Millican and A. Clark

Connectionism, Concepts, and Folk Psychology
The Legacy of Alan Turing, Volume II
Edited by A. Clark and P. J. R. Millican

Appearance versus Reality
New Essays on the Philosophy of F. H. Bradley
Edited by Guy Stock

Knowing Our Own Minds
Edited by Crispin Wright, Barry C. Smith, and Cynthia Macdonald

Transcendental Arguments
Problems and Prospects
Edited by Robert Stern

Reason and Nature

ESSAYS IN THE THEORY OF RATIONALITY

Edited by

JOSÉ LUIS BERMÚDEZ

and

ALAN MILLAR

CLARENDON PRESS • OXFORD

6

The Rational Analysis of Human Cognition[*]

NICK CHATER AND MIKE OAKSFORD

Rationality appears basic to the understanding of mind and behaviour. In practical decisions, such as whether a person is morally responsible for his or her actions, to whether a person can be hospitalized without consent, it seems crucial 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 behaviour 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 rationality assumptions go deeper still—they are embodied in the folk psychological style of explanation in which we describe each other's minds and behaviour (Fodor 1987; Stich 1983). Assumptions of rationality also appear equally essential to interpret each other's utterances and to understand texts (Davidson 1984; Quine 1960). So rationality appears basic to the explanation of human behaviour, 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 daily life.

In this informal, everyday sense, most of us, most of the time, are remarkably rational. To be sure, we 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 behaviour 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 think of them as plain common sense—implicitly assuming that common sense must be a simple thing indeed. People may not think of themselves as exhibiting high levels of

[*] Please address correspondence concerning this paper to Nick Chater, Department of Psychology, University of Warwick, Coventry CV4 7AL, UK or to Mike Oaksford, School of Psychology, Cardiff University, PO Box 901, Cardiff CF1 3YG, Wales, UK. We would like to thank José Luis Bermúdez and Alan Millar for their valuable comments on an earlier version of this paper.

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 speak, think, or act appropriately in complex, real-world situations—in short, they are labels for everyday rationality.

Indeed, so much do we tend to take the rationality of common-sense thought for granted, that only recently has it been appreciated that common-sense reasoning is immensely difficult. This realization emerged from the project of attempting to formalize everyday knowledge and reasoning in artificial intelligence, where initially high hopes that common-sense 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 common sense 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 non-monotonic reasoning, brought the project of formalizing common sense to an effective standstill (e.g. McDermott 1987). Thus the cognitive processes underlying plain 'common sense' far outperform any artificial computational system we can devise. 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 behaviour, but to mathematical theories of good reasoning. These theories represented 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; 'rational choice' theories aim to explain people's preferences and decisions, under uncertainty and in strategic interaction with other 'players' (Nash 1950; von Neumann and Morgenstern 1944). 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 behaviour in the everyday world.

How are the two sides of rationality related? 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 develop a viewpoint rooted in a style of explanation in the behavioural sciences, *rational analysis* (Anderson 1990). We suggest that rational analysis provides a good characterization of how the concept of rationality is used in explanations in psychology, economics, and animal behaviour, and usefully explicates the relationship between everyday and formal rationality.

The discussion falls into four main parts. First, we discuss formal and everyday rationality, and various possible relationships between them. Second, we describe the programme of rational analysis as a mode of explanation of mind and behaviour, which views everyday rationality as underpinned by formal rationality. Third, we consider a case study of rational analysis, concerning a celebrated laboratory reasoning task, Wason's (1966, 1968) selection task. Fourth, we defend the use of formal rationality in explaining mind and behaviour from some critical attacks (Evans and Over 1996*a*, 1997; Gigerenzer and Goldstein 1996; McDermott 1987).

RELATIONS BETWEEN FORMAL AND EVERYDAY RATIONALITY

Formal rationality concerns formal principles of good reasoning—the mathematical laws of logic, probability, decision, or game theory. These principles appear, at first sight, to be far removed from everyday rationality—from how people think and act in everyday life. Rarely in daily life do we praise or criticize each other for obeying or violating the laws of logic or probability. 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, Newstead, and Byrne 1993), probabilistic fallacies (e.g. Tversky and Kahneman 1974), and to make inconsistent decisions (Kahneman, Slovic and Tversky 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.

How can we relate the astonishing fluency and success of everyday reasoning and decision-making, exhibiting remarkable levels of everyday rationality, to our faltering and confused grasp of the principles of formal rationality? The problem is especially pressing in view of the fact that psychologists model almost all human cognition as involving inference. Thus, in deciding to cross the road, in parsing a sentence, or in catching a ball, the complex information-processing involved is standardly modelled as involving complex inferential processes concerning relevant knowledge about the movements of cars and cyclists, the lexical and grammatical structure of the language, or the trajectory of the ball and the forces generated by, and inertia tensors of, the motor system. Indeed, the view that cognition is, across the board, to be viewed as a matter of inference over representations of knowledge, is close to a fundamental assumption of cognitive science. And more specifically, the kinds of reasoning processes that are typically invoked involve precisely the formal models of reasoning (probability, decision theory, and so on) that we have discussed. Hence, almost every impressively fluent and successful aspect of human cognition is typically viewed by psychologists as involving reasoning processes—which suggests that the cognitive system must have remarkable facility at such reasoning. But this contrasts bizarrely with results in direct experiment tests of human formal reasoning—which appear to reveal that people have only the most blundering ability in formal reasoning. So we return to the question: how can some reconciliation be found between the effectiveness of everyday reasoning exhibited across cognitive processes and the ineffectiveness of performance on experimental reasoning tasks? We sketch three important possibilities, which have been influential in the literature in philosophy, psychology, and the behavioural sciences.

Everyday Rationality is Primary

This viewpoint takes everyday rationality as fundamental, and views formal theories as flawed in so far as they fail to match up with human everyday reasoning intuitions.

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 appear to reveal apparent blunders in human reasoning are taken at face value.

Where 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. A certain measure of tension may be granted between the goal of constructing a satisfyingly concise normalization of intuitions, and the goal of capturing every last intuition successfully, just as linguists allow complex centre-embedded constructions to be grammatical (e.g. 'the fish the man the dog bit ate swam'), even though most people reject them as ill-formed gibberish. But the dissonance between formal rationality and everyday reasoning appears 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; see also Ellsberg 1961, and May 1954, for a similar challenge to 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 tickets worth $0

Most people 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 people also prefer lottery C to lottery D—we don't think it is worth losing what would otherwise be a certain $1,000,000, just for the possibility of winning $5,000,000. This *combination* of responses, for all its intuitive appeal, is inconsistent with decision theory, which demands 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:

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

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

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

But the preference for lottery C over D means that:

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

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

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

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—decision theory should be revised to fit with our intuitions (see Chew 1983; Fishburn 1983; Kahneman and Tversky 1979; Loomes and Sugden 1982; Machina 1982).

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 to the extent that it fits 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:

> If John's automobile is a Mini, John is poor, and
> if John's automobile is a Rolls, John is rich. (5)
>
> Either, if John's automobile is a Mini, John is rich, or
> if John's automobile is a Rolls, John is poor. (6)

Clearly, however, this violates intuition. Most people would agree with (5) as at least highly plausible; but would reject (6) as implausible. 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: everyday intuitions have primacy over formal theories.

But this viewpoint is not without problems. A key danger is of losing any normative force to the notion of rationality—if rationality is merely conformity to each other's predominant intuitions, then there seems no standpoint from which to assess which of our intuitions is rational. On this view, being rational is like a musician being in tune: 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 reintroduce 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. For example, the equilibrium point could conceivably leave the individual content in the idea that logical fallacies are sound principles 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 satisfied. Thagard's (1988) concerns were very general, in order 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 fulfil (i) to (iii) but contain mutually contradictory principles. The point about formalization 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 (roughly, expected number of offspring), and argue that behaviour 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 behaviour, and digestive processes that lead to successful metabolic activity. Both increase the inclusive fitness of an individual (roughly, the expected number of children of that individual); but intuitively we want to say that the first is concerned with rationality, while 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 stomachs, that are not in the business of rationality at all.

Formal Rationality is Primary

Arguments for the primacy of formal rationality take a different starting point. This viewpoint is standard with the mathematics, statistics, operations research, and the 'decision sciences' (e.g. Kleindorfer, Kunreuther and Schoemaker 1993). The idea is that everyday reasoning is fallible, and that it must be corrected by following the dictates of formal theories of rationality. In this light, for example, the Allais paradox may be viewed as revealing a flaw in human reasoning rather than exposing a problem for decision theory.

The viability of this viewpoint depends, in part, on the scope of formal theories of rationality—are they really able to handle the richness of inferences that everyday reasoning actually involves? This issue arises particularly in the context of formal logic, because the principles of logic do not give a general model of how beliefs should be revised (particularly when there is some inconsistency in the knowledge base—which is, of course, the normal situation in cognition) (e.g Harman 1986; McDermott 1987; Oaksford and Chater 1991). But it also arises more generally—for example, although inductive inference can, in many contexts, be usefully modelled in terms of probabilistic inference, there are no clear principles concerning how to set prior probabilities from which inference begins; and the choice of prior probabilities will be crucially important given any finite set of data (though see e.g. Jaynes 1989; Jeffreys 1939; Paris 1992; Rissanen 1987, 1989 for discussion). We shall touch on these issues below—but for now let us leave aside the concern that formal principles of rationality are simply too limited to engage with the principles that underlie the full complexity of everyday reasoning.

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 they may potentially override 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 behaviour 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 1926; 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 can be shown 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-averse, 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 clear-cut 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; Savage 1954; von Neumann and Morgenstern 1944). Moreover, the same general approach can be used as a justification for logic, if avoiding inconsistency is taken as axiomatic. Thus, there may have been good reasons for accepting formal theories of rationality, even if, much of the time, human intuitions and behaviour strongly violate their recommendations (see Dawes 1988, for an exposition of this viewpoint from within psychology).

[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).

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 programme (e.g. Kahneman, Slovic and Tversky 1982; Kahneman and Tversky 1979; Thaler 1987), 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's (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 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 behaviour 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 programme 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; see Chater and Oaksford 1990, for a critique).

The mental models theory of reasoning concurs that logical inference provides the computational level theory for reasoning, but instead of standard proof-theoretic rules, this view uses a 'semantic' method of proof. Such methods involve a 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 of a model.[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 counter-example exists, and that the inference is valid (this is similar to 'negation as failure' in logical programming (Clark 1978)).

[2] E.g., 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, it is not clear how this could be done. Instead, the semantics of mental models is left, rather uncomfortably, up to the theorist's intuitions.

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 behaviour 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.

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 from 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 usefulness 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 behaviour 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 behaviour so that their reasoning makes sense.

Crucially, then, the formal standards of rationality appropriate for explaining some particular cognitive processes or aspect of behaviour 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 behaviour. 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 behaviour 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 behaviour 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 behaviour 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 have 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.

[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).

THE PROGRAMME OF RATIONAL ANALYSIS

The project of providing a rational analysis for some aspect of thought or behaviour has been described by the cognitive psychologist John Anderson (e.g. Anderson 1990, 1991*a*). 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 behaviour. 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 behaviour 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 behaviour 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 behaviour, where the behaviour of people or animals is viewed as optimizing some goal, such as money, utility, inclusive fitness, food intake, or the like. But Anderson (1990, 1991*a*) was concerned to extend this approach not just to the behaviour of whole agents, but to structure and performance of particular cognitive processes of which agents are composed. Anderson's programme has led to a flurry of research in cognitive psychology (see Chater and Oaksford 1999*a*; Oaksford and Chater 1998*a*, for overviews of recent research), from areas as diverse as categorization (Anderson 1991*b*; Anderson and Matessa 1998; Lamberts and Chong 1998), memory (Anderson and Milson 1989; Anderson and Schooler 1991; Schooler 1998), reasoning (Oaksford and Chater 1994, 1995*a*, 1996, 1998*b*), searching computer menus (Young 1998), and natural language parsing (Chater, Crocker, and Pickering 1998). This research has shown that a great many empirical generalizations about cognition can be viewed as arising from the rational adaptation of the 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 behaviour, and economics.

The Importance of Goals

Everyday thought and action is focused 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 behaviour 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 behaviour 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.

Note that 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 (rationality$_2$)' (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 what 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.

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 modelling the structure of that environment. This requires using principles of formal rationality to specify 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 (1979) '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 behaviour (McCloskey 1985). In microeconomics, modelling 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 macroeconomic 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.

Computational Limitations

In rational analysis, deriving the optimal behaviour function (Anderson's Step 4) is frequently very complex. Models based on optimizing, 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 optimization 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 behaviour 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 behaviour 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 1995*b*).

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 behaviour 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 behaviour is highly flexible with respect to environmental variation, then calculation is required to determine the correct behaviour, 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 maximize the chance of winning. However, this optimization problem is known to be completely intractable because the search space is so large. But optimizing 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 maximize 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 maximizing the amount of useful information remembered, maximizing 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 maximizing expected utility (from an economic perspective) or maximizing inclusive fitness (from a biological perspective). At any level, it is possible that optimization 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 optimize or nearly optimize some more local, but relevant, quantity.

The observation that the local goals may be optimized 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 optimizing *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 optimized must be relevant to some larger goal of the agent. Thus, it seems *reasonable* that animals may attempt to optimize the amount of food they obtain, or that the categories used by the cognitive system are optimized to lead to the best predictions. This is because, for example, optimizing the amount of food obtained is likely to enhance inclusive fitness, in a way that, for example, maximizing 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 optimization. Therefore rationality requires not just optimizing something but optimizing something reasonable. As a definition of rationality, this is clearly circular. But by viewing rationality in terms of optimization, general conceptions of what are reasonable cognitive goals can be turned into specific and detailed models of cognition. Thus, the programme 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 programme. 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 behaviour 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 behaviour 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.

Having outlined the general rational analysis approach, and argued that the approach is prevalent in the social and biological sciences, we now consider how the programme of rational analysis provides a very different perspective on human reasoning than has been traditionally obtained from laboratory studies. Specifically, apparently non-deductive reasoning performance in laboratory reasoning tasks can be shown to make coherent sense if it is recognized that people may not be treating the reasoning tasks as deductive at all. A probabilistic rational analysis of these tasks provides a simple and powerful framework for explaining a wide variety of empirical data on human reasoning.

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 non-rational, 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 behaviour, economics, or psychology, assumes that the agent is well adapted to its normal environment. However, almost all psychological data are 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, Newstead, and Byrne 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 propose first that everyday rationality is founded on uncertain rather than certain reasoning. This suggests that probablity 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 probabilitistic 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.

Our discussion will focus on Wason's selection task (Wason 1966, 1968), 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), although the approach we describe has been applied in other areas of reasoning, including other areas in the psychology of 'deductive' reasoning: reasoning with conditionals and syllogisms (e.g. Anderson 1995; Chater and Oaksford 1999*c*; Oaksford and Chater 1998*b*).

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 bear 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, Newstead, and Byrne 1993), although a number of authors have argued for a non-deductive conception of the task (Fischhoff and Beyth-Marom 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: to only 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 be 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 selections 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 different 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. *p*s are invariably associated with *q*s, and (ii) that the occurrence of *p*s and *q*s 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 are 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:

$$\text{Information before receiving } D\text{:} \quad I(H) = -\sum_{i=1}^{n} P(H_i)\log_2 P(H_i)$$

$$\text{Information after receiving } D\text{:} \quad I(H|D) = -\sum_{i=1}^{n} P(H_i|D)\log_2 P(H_i|D)$$

$$\text{Information gain:} \quad I_g = I(H) - I(H|D)$$

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 'rarity assumption' is an appropriate default because in a typical everyday rule such as *if it's a raven then it's 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:

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

where E represents the expectation operator. This corresponds to the observed frequency of card selections in Wason's task: $p > q > not\text{-}q > not\text{-}p$ and thus explains the predominance of *p* and *q* card selections as a rational inductive strategy.

This result might seem paradoxical: it might seem that the Bayesian analysis suggests that finding falsifying instances of the rule (which may occur by turning the *not-q* card to reveal a *p*) is not important. And this would seem to be bizarre, because from any reasonable point of view, falsifying instances should be especially significant (because they decisively answer the question of whether or not the rule is correct); and any method of testing a rule should put an emphasis on finding such instances if they exist. Fortunately, there is no puzzle here. The Bayesian analysis does rate falsifying instances as highly informative—indeed, as maximally informative, because uncertainty concerning whether the rule is true drops to 0 as soon as a falsifier is discovered. But the *expected* amount of information obtained by turning the *not-q* card is, nonetheless, low, because, according to the rarity assumption, mentioned above, the probability of finding a falsifying instance on the back of a *not-q* card is low.

To get an intuitive feel for how this works, consider the following scenario. Suppose that the hypothesis under test is 'if a saucepan falls from the kitchen shelf (*p*) it makes a clanging noise (*q*).' This rule, like the vast majority of everyday rules, conforms to the rarity assumption—saucepans fall quite rarely (most of the time no saucepan is falling); and clangs are heard quite rarely (most of the time no clang is audible). The four cards in the selection task can be seen as analogous to the following four scenarios. Suppose I am in the kitchen, and see the saucepan beginning to fall (*p* card); should I bother to take off my headphones and listen for a clang (i.e. should I turn the *p* card?)? Intuitively, it seems that I should, because, whether there is a clang or not, I will learn something useful concerning the rule (if there is no clang, the rule is falsified; if there is a clang, then my estimate of the probability that the rule is true increases). Suppose, on the other hand, I am next door and I hear a clang (*q* card); should I bother to come into the kitchen to see whether the saucepan has fallen (should I turn the *q* card?)? Intuitively, this is also worth doing—if the saucepan has not fallen then I have learned nothing (something else must have caused the clang); but if the saucepan has fallen, then this strongly confirms the rule. This is the intuitive explanation for why the *q* card is worth turning, even though there is no possibility that turning this card can falsify the rule.

Now consider the analogue of the turning of the *not-q* card: I am next door and I hear *no* clang. This time should I bother to come into the kitchen to see whether the saucepan has fallen (should I turn the *not-q* card?)? Intuitively, to bother to do so seems crazy—I'll be in and out to the kitchen all day if I adopt this strategy! And I will probably learn nothing whatever, as the saucepan will remain unmoved on the shelf. Of course, in the very unlikely event that I find that the saucepan has fallen (*p*), then I can falsify the rule—because if the rule were true I should have heard a clang (*q*) and I did not. But in everyday reasoning contexts, where the rarity assumption holds, the expected information gain for the analogue of turning the *not-q* card is typically very low—because the probability of obtaining falsification is so low. Crucially, intuitively (and in Oaksford and Chater's 1994 formal analysis) the expected informational value of turning the *q* card is greater than turning the *not-q* card, even though turning the *q* card cannot lead to falsification—I will be more inclined to bother to check whether the saucepan has fallen if I hear a clang than if I do not. To complete the example, the *not-p* card corresponds to the case in which I see that the saucepan is sitting safely on the shelf; should I bother to take off my headphones and listen for a clang. Clearly not, because the rule only makes a claim about what happens *if* the saucepan falls.

Oaksford and Chater (1994) also show how their model generalizes 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*b*; Laming 1996; Klauer, in press; and for responses and developments see Oaksford and Chater 1996, 1998*b*, 1998*c*; Chater and Oaksford, 1999*c*). 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 (Oaksford and Chater 1995*a*; Sperber, Cara, and Girotto 1995).

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.

Having seen how rational analysis can be applied in a specific case, and how the approach may have radical implications for standard interpretations of laboratory data on human reasoning, we now defend the rational analysis approach against theorists who argue that formal rationality has no useful role in explaining everyday rationality.

COULD FORMAL AND EVERYDAY RATIONALITY BE UNRELATED?

The first part of this paper considered various possible relations between formal and everyday rationality. The second part developed a particular conception of this relationship, framed in terms of Anderson's methodology of rational analysis, and the third provided an illustration of the approach. This section considers recent viewpoints which suggest that the whole enterprise may have been misconceived from the beginning—because there is no useful relationship between formal and everyday rationality. We shall argue that formal rationality does indeed form an indispensable part of the explanation of everyday rationality, and that the nature of this explanation is best understood in terms of rational analysis.

The view that formal and everyday rationality can be disconnected has been advanced by a number of theorists. In artificial intelligence, McDermott (1987) argues that the attempt to build knowledge representation systems based on logical principles persistently fails to capture human everyday reasoning, and (with some sense of despair!) recommends a 'procedural' approach—the researcher simply aims to specify algorithms that seem to work, without attempting to ground these in formal logic or probability. In robotics, there has been much interest in so-called behaviour-based robotics (Brooks 1991; McFarland and Bösser 1993), where perceptual and motor functions are linked directly together, using essentially heuristic methods, rather than attempting to use general principles of perceptual analysis and motor control (as exemplified in e.g. Marr 1982).

As we noted above, in psychology, Evans and Over (1996*a*, 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.

A similar view is advocated by Gigerenzer and Goldstein (1996) who claim to provide an 'existence proof' for algorithms which *work* in the real world, but have no apparent justification in terms of formal theories of reasoning (such an algorithm is therefore intended to be a candidate for explaining part of rationality$_1$, in Evans and Over's terms, even though they are held to be unrelated to rationality$_2$).

The domain they consider is one of cognitive estimation: deciding which is the larger of two cities, based on a list of features of each city. Their 'non-rational' algorithm, Take-the-Best, works in two steps. First, it uses a 'recognition principle' and—if one of the cities is not known, it is assumed to be the smaller. Second, the algorithm sequentially considers features of the cities, one by one, in decreasing order of 'diagnosticity' for size (the diagnosticity ordering is a prior calculation). So, for example, the feature 'is a national capital' may be most diagnostic of size—if one city has this property it is declared to be the larger. Hence this will be the first feature to be considered. If the cities 'tie' on this property (e.g. neither is a national capital), then another feature is examined (e.g. has the city been the site of an exposition), and so on, until the tie is broken. This algorithm is designed to be 'fast and frugal'—i.e. to consume little time or memory resources; but it has no obvious rational basis. Nonetheless, in a competition with other algorithms, including multiple regression from statistics, Gigerenzer and Goldstein show that Take-the-Best performs as well as these apparently more rationally justified algorithms (and indeed, at levels that appear comparable with human performance).

As well as arguing that Take-the-Best is an existence proof that algorithms can succeed in real environments, without any basis in formal rational theories, Gigerenzer and Goldstein (1996) argue that, more generally, human reasoning works by fast and frugal algorithms which work in the real world, but have no justification in terms of probability, statistics, or other normative principles.

But this viewpoint does not tackle the fundamental problem we outlined for advocates of the primacy of everyday rationality 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.

There is, though, an interesting lesson to be learned from the success of 'fast and frugal' algorithms such as Take-the-Best, which do a good job in the real world without being directly based on formal rational principles. This is that explanation in terms of cognitive algorithms can run ahead of rational explanation—i.e. we can specify algorithms that *do* work, without knowing *why* they work. We shall see shortly that the projects of developing rational and algorithm explanations of cognition quite frequently run at different speeds—each approach may run ahead of the other. But this does not undermine the importance of ultimately being able to provide both styles of explanation. In particular, it does not undermine the utility of accounts based on formal rationality, in explaining the real-world everyday rationality of the cognitive system.

Consider, first, cases where rational explanations of behaviour have proceeded without considering how they might be approximated by cognitive algorithms. The vast bulk of 'rational choice' explanation, whether in social behaviour (Crawford, Smith, and Krebs 1987; Messick 1991), economics (e.g. Muth 1961; von Neumann and Morgenstern 1944) or animal behaviour (Maynard-Smith and Price 1973) has this character. The programme of rational analysis, outlined above, has the same character—indeed, one of Anderson's (1990) motivations for developing the rational analysis approach was precisely that it abstracts away from specifying underlying cognitive algorithms, which can often be underdetermined by empirical data (e.g. Anderson 1978; Pylyshyn 1984). In all these explanations, formal rational principles specify what *should* occur, given a specific goal and environment, but the particular cognitive algorithms which underlie behaviour in these contexts may be entirely unknown.

Gigerenzer and Goldstein (along with others who advocate separating formal rational explanation from the explanation of everyday, real-world thought and behaviour) focus on the opposite case, where algorithmic explanation has run ahead of rational explanation. This occurs in much of cognitive psychology, which has focused on describing cognitive algorithms and the representations over which they operate. Equally, the study of animal cognition has resulted in accounts such as the Rescorla–Wagner associative learning algorithm for classical conditioning (Rescorla and Wagner 1972). Indeed, explanation in terms of algorithms, whether specified in terms of sequential operations, 'box and arrow' diagrams, or neural networks, is arguably the dominant mode of explanation in many areas of psychology.

Similarly, in the technical study of machine learning, neural networks, and much practical (rather than theoretical) mathematical statistics, algorithms have been constructed which address complex and poorly understood real-world problems, with at least some success. But the rational theory of why these algorithms are successful lags behind these developments. To choose an example of current psychological interest, it has recently been shown that a neural network can learn to map from orthography to phonology, dealing successfully both with exception words and non-words (Bullinaria 1994; Plaut, McClelland, Seidenberg, and Patterson 1996; Seidenberg and McClelland 1989).[4] But there is no known rational theory of the nature of the orthography–phonology mapping, or how it should be learned. A different kind of example of psychological interest concerns the vast range of practical statistical tests which are widely used, although the assumptions under which they apply are not known (Gigerenzer and Murray 1987). Thus, Take-the-Best seems unnecessary as an 'existence proof' that we can design successful algorithms without knowing why they work, because there are already many examples of such algorithms in the psychological, computational, and statistical literatures.

However, even where algorithmic theories have predominated, it remains an important goal to provide rational explanations of *why* they succeed. For example, in psychology, the adaptiveness of the Rescorla–Wagner learning algorithm (Rescorla and Wagner 1972) has been explained by showing that it asymptotically approximates the optimal solution in a normative probabilistic account of causal reasoning (Cheng 1997; Shanks 1995*a*, 1995*b*). Rescorla–Wagner learning therefore approximates a rational standard, using limited computational resources. Equally, classification by similarity to stored exemplars, for which there is considerable empirical evidence (Medin and Schaffer 1978), can be shown to be adaptive because it approximates Anderson's (1991*b*) Bayesian classification model (Nosofsky 1991). A further example provided by McKenzie (1994) who has shown that so-called 'linear combination heuristics', which are good descriptions of human causal reasoning performance, also provide good approximations to a normative Bayesian solution (see also Anderson 1990; Cheng 1997). For many years, the fact that people appear to use such heuristics has been cited as evidence for the irrationality of human causal reasoning. Recent analyses suggest that this was premature: these heuristics provide a 'fast and frugal' approximation to rational norms.

Even relatively ill-defined heuristics for probabilistic reasoning like 'availability' (Kahneman, Slovic, and Tversky 1982) may have a rational basis. The concept of availability has been developed to explain a range of systematic

[4] We take no stand here on whether such models are compatible with detailed psychological and neuropsychological data. See e.g. Coltheart, Curtis, Atkins, and Haller (1993) for discussion.

biases in people's probability and frequency judgements. In a famous study, Tversky and Kahneman (1974) asked people how many seven letter words have the form:

_ _ _ _n_

and how many have the form

_ _ _ _ing

People typically estimate that there are more words of the second form than the first. But this cannot be correct, because all the words that are examples of the second form are necessarily examples of the first! Tversky and Kahneman's explanation is that the second form provides a better cue to memory—words ending 'ing' are more 'available'. The assumption is that people estimate frequencies and probabilities by using availability—the more available an item is, the more frequent or probable it is assumed to be. But this heuristic seems to have a sound rational basis: to the extent that memory retrieval reflects an unbiased sample of the environment, availability will conform to a rational probabilistic analysis. Biased sampling (e.g. because items are stored or retrieved differentially) may lead to errors, but generally, this heuristic will be successful. Indeed, the power of the 'cognitive illusion' in Tversky and Kahneman's study arises precisely because sampling is so biased in this case.

More generally, the programme of rational analysis has shown why a wide range of empirically derived algorithmic processes are successful, by showing that they approximate normative Bayesian standards, given certain assumptions about environmental structure. This approach to explaining why cognitive algorithms succeed has been adopted by a wide range of researchers in the cognitive sciences (Oaksford and Chater 1998*b*). In each case, success is explained because the algorithm approximates, however crudely, some rational norm for optimal behaviour in that environment. Moreover, in line with the mutual constraint between the levels mentioned above, rational level explanations have been used to develop new algorithmic accounts (e.g. Anderson 1993; Chater and Oaksford, 1999*c*).

Similarly, in other domains where the algorithmic theory has run ahead, there has been enormous effort to develop complementary rational theories. The goal of the research programmes of computational learning theory (Valiant 1984) and statistical learning theory (Vapnik 1995) is to provide a rational foundation for practical learning algorithms. Moreover, there has been great interest in interpreting neural networks as probabilistic inference devices, to give insight into the rational basis for their success (e.g. Chater 1995; MacKay 1992; McClelland 1998; Neal 1993). Furthermore, statistical theory has been developed as a rational basis for practical statistical algorithms (e.g. Bernado and Smith 1995). In each case, algorithms have been assumed to approximate rational standards to some degree. Typically, algorithms will be shown to be rational, given a certain goal (e.g. minimizing prediction error), on the assumption that the environment has a certain structure (e.g. that samples are independent, that variance is constant, that different causal factors interact linearly, and so on). Moreover, as in psychology, rational theories in these areas have not merely shown why, and in what environments, existing algorithms will succeed, but also served to develop new algorithms. In sum, across domains where algorithmic theories have run ahead of rational accounts, there has been vigorous and important research on developing complementary rational explanations. This indicates the desirability of both levels of explanation in providing complete accounts of cognitive phenomena.

Thus it seems that the real-world success of algorithms such as Take-the-Best, apparently disconnected from a formal rational theory, does not imply that formal rational explanation is unnecessary. Algorithmic and rational levels of explanation are complementary: without an algorithmic account, we do not know *how* cognition works; without a rational account, we do not know *why* cognition works.

Is There an Alternative Style of 'Why' Explanation?

A possible counter-attack by those advocating the view that formal rationality has no role in explaining everyday thought and behaviour is to argue that there is an alternative, 'ecological' or 'adaptive' explanation of *why* cognition works, which makes no reference to formal rational principles. This is one interpretation of Gigerenzer and Goldstein's statement that 'the minds of living systems should be understood relative to the environment in which they evolved *rather than* to the tenets of classical rationality' (p. 651) (emphasis added). This suggests a notion of 'adaptive rationality', i.e. success in relation to an environment, as an alternative to classical rationality. But to see that this notion does not provide an *alternative* explanation, consider the question: Why does a cognitive algorithm succeed in a particular environment? To reply that this is because it is adaptively rational is clearly circular; because for an algorithm to be adaptively rational means *by definition* that it succeeds in the environment. In contrast, the rational level explains behavioural success by showing how that behaviour approximates optimal performance given appropriate assumptions about the agent's goals and environment.

It is, of course, conceivable that there may be some other alternative way of explaining *why* cognitive algorithms succeed, which Gigerenzer and Goldstein might advert to as an alternative to rational explanation. An obvious suggestion is to appeal to evolution or learning. Perhaps natural selection has ensured that our cognitive algorithms succeed; or perhaps our learning mechanisms have simply favoured algorithms that work. But explanations in terms of evolution or learning do not explain *why* specific cognitive algorithms are adaptive. Instead, they explain why we possess adaptive rather than non-adaptive algorithms—essentially because adaptive algorithms, by definition, perform better in the natural environment, and processes of natural selection or learning will tend to favour algorithms which are successful. But this still leaves open the question of *why* some algorithms *are* successful in the environment whereas some are not. Answering this question requires analysing the structure of the environment, the goals of the agent, and studying how these goals can be achieved given that environment. In short, it involves rational level explanation. To choose an example from a domain in which evolutionary explanation is widely accepted, an account of optimal foraging in behavioural ecology may explain *why* particular foraging strategies are successful and others are not. Zoologists assume evolution explains why animals possess good foraging strategies, but do not take evolutionary explanation to provide an alternative to the rational level explanation given by optimal foraging theory.

CONCLUSIONS

This paper has considered the relation between everyday and formal rationality, and has developed a particular view of the relation between the two, based on Anderson's programme of rational analysis. We have illustrated this approach with a rational analysis of performance on Wason's selection task, and defended the approach against the view that formal rational explanation is unnecessary in explaining cognition. We have argued that formal rational explanation is indispensable in explaining *why* human cognitive mechanisms are able to succeed in the real world—i.e. why they are able to exhibit everyday rationality.

The relation that we have identified between rationality and algorithmic accounts, which is apparent in examples from rational analysis in psychology, and from work in zoology and economics, has broad application. It promises to reconcile rational and mechanistic constraints in a range of contexts where the debate focuses on the different level of emphasis placed on these constraints. Both rational and mechanistic factors are important, because the system under study is presumed only to approximate, perhaps quite accurately or perhaps very coarsely, a rational solution. Within this framework, the debate between rationality-based versus mechanistic explanation becomes a matter of emphasis and degree, rather than a fundamental divide. We suggest that in any debate of this kind, there should be a methodological imperative to explore rationality-based explanations—only by doing so can the scope of this level of explanation be assessed; and we caution that rationality-based explanation cannot be abandoned wholesale, without losing the ability to explain *why* the cognitive system is adaptive or successful.

The tension between the limited scope of current formal theories of reasoning and the astonishing richness and flexibility of human reasoning should not, however, be underestimated. There are presently no adequate formal theories of simple default inference in everyday reasoning, let alone formal theories of induction, analogical reasoning, or reasoning by comparison with past cases—and it is not clear that formal explanation will be possible at all in all of these cases (e.g. Goodman 1954). Explaining thought and behaviour both in terms of formal rational principles, and at the level of cognitive algorithms, will be one of the principal intellectual challenges of the third millennium.

REFERENCES

Allais, M. (1953), 'Le comportement de l'homme rationnel devant le risque: Critique des postulats et axiomes de l'école américaine', *Econometrica*, 21, 503–46.

Almor, A., and Sloman, S. (1996), 'Is deontic reasoning special?', *Psychological Review*, 103, 374–80.

Anderson, J. R. (1978), 'Arguments concerning representations for mental imagery', *Psychological Review*, 85, 249–77.

——(1990), *The Adaptive Character of Thought* (Hillsdale, NJ: Lawrence Erlbaum Associates).

——(1991*a*), 'Is human cognition adaptive?' *Behavioral and Brain Sciences*, 14, 471–517.

——(1991*b*), 'The adaptive nature of human categorisation', *Psychological Review*, 98, 409–29.

——(1993), *Rules of the Mind* (Hillsdale, NJ: Lawrence Erlbaum).

——(1995), *Cognitive Psychology and Its Implications*, 4th edn. (San Francisco: W. H. Freeman).

——and Matessa, M. (1998), 'The rational analysis of categorisation and the ACT-R architecture', in M. Oaksford and N. Chater (eds), *Rational Models of Cognition* (Oxford: Oxford University Press): 197–217.

——and Milson, R. (1989), 'Human memory: An adaptive perspective', *Psychological Review*, 96, 703–19.

——and Schooler, L. J. (1991), 'Reflections of the environment in memory', *Psychological Science*, 1, 396–408.

Becker, G. (1975), *Human Capital*, 2nd edn. (New York: Columbia University Press).

——(1981), *A Treatise on the Family* (Cambridge, Mass.: Harvard University Press).

Bernado, J. M., and Smith, A. F. M. (1995), *Bayesian Theory* (Chichester, Sussex: Wiley).

Braine, M. D. S. (1978), 'On the relation between the natural logic of reasoning and standard logic', *Psychological Review*, 85, 1–21.

Brooks, R. A. (1991), 'How to build complete creatures rather than isolated cognitive siumulators', in K. Van Lehn (ed.), *Architectures for Intelligence* (Hillsdale, NJ: Lawrence Erlbaum Associates): 225–39.

Bullinaria, J. A. (1994), 'Internal representations of a connectionist model of reading aloud', *Proceedings of the Sixteenth Annual Conference of the Cognitive Science Society* (Hillsdale, NJ: Erlbaum): 84–9.

Chater, N. (1995), 'Neural networks: The new statistical models of mind', in J. P. Levy, D. Bairaktaris, J. A. Bullinaria, and P. Cairns (eds), *Connectionist Models of Memory and Language* (London: UCL Press): 207–28.

——(1996), 'Reconciling simplicity and likelihood principles in perceptual organization', *Psychological Review*, 103, 566–81.

——Crocker, M., and Pickering, M. (1998), 'The rational analysis of inquiry: The case of parsing', in M. Oaksford and N. Chater (eds), *Rational Models of Cognition* (Oxford: Oxford University Press): 441–68.

——and Oaksford, M. (1990), 'Autonomy, implementation and cognitive architecture: A reply to Fodor and Pylyshyn', *Cognition*, 34, 93–107.

————(1999*a*), 'Ten years of the rational analysis of cognition', *Trends in Cognitive Sciences*, 3, 57–65.

————(1999*b*), 'The probability heuristics model of syllogistic reasoning', *Cognitive Psychology*, 38, 191–258.

————(1999*c*), 'Information gain vs. decision-theoretic approaches to data selection', *Psychological Review*, 106, 223–7.

Cheng, P. W. (1997), 'From covariation to causation: A causal power theory', *Psychological Review*, 104, 367–405.

——and Holyoak, K. J. (1985), 'Pragmatic reasoning schemas', *Cognitive Psychology*, 17, 391–416.

Cherniak, C. (1986), *Minimal Rationality* (Cambridge, Mass.: MIT Press).

Chew, S. H. (1983), 'A generalization of the quasilinear mean with applications to the measurement of income inequality and decision theory resolving the Allais paradox', *Econometrica*, 51, 1065–92.

Chomsky, N. (1965), *Aspects of the Theory of Syntax* (Cambridge, Mass.: MIT Press).

Clark, K. L. (1978), 'Negation as failure', in *Logic and Databases* (New York: Plenum Press): 293–322.

Cohen, L J. (1981), 'Can human irrationality be experimentally demonstrated?' *Behavioral and Brain Sciences*, 4, 317–70.

Coltheart, M., Curtis, B., Atkins, P., and Haller, M. (1993), 'Models of reading aloud: Dual-route and parallel-distributed-processing approaches', *Psychological Review*, 100, 589–608.

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* (Baltimore: Johns Hopkins University Press).

Crawford, C., Smith, M., and Krebs, D. (1987), *Sociobiology and Psychology* (Hillsdale, NJ: Erlbaum).

Davidson, D. (1984), *Inquiries into Truth and Interpretation* (Oxford: Clarendon Press).

Dawes, R. M. (1988), *Rational Choice in an Uncertain World* (San Diego, Calif.: Harcourt, Brace, Jovanovich).

Dawkins, R. (1977), *The Selfish Gene* (Oxford: Oxford University Press).

Earman, J. (1992), *Bayes or Bust? A Critical Examination of Bayesian Confirmation Theory* (Cambridge, Mass.: MIT Press).

Ellsberg, D. (1961), 'Risk, ambiguity and the Savage axioms', *Quarterly Journal of Economics*, 75, 643–69.

Elster, J. (ed.) (1986), *Rational Choice* (Oxford: Blackwell).

Evans, J. St. B. T. (1982), *The Psychology of Deductive Reasoning* (London: Routledge & Kegan Paul).

——(1989), *Bias in Human Reasoning: Causes and Consequences* (Hillsdale, NJ: Erlbaum).

——and Lynch, J. S. (1973), 'Matching bias in the selection task', *British Journal of Psychology*, 64, 391–7.

——Newstead, S. E., and Byrne, R. M. J. (1993), *Human Reasoning* (Hillsdale, NJ: Erlbaum).

——and Over, D. E. (1996*a*), *Rationality and Reasoning* (Hove, Sussex: Psychology Press).

————(1996*b*), 'Rationality in the selection task: Epistemic utility vs. uncertainty reduction', *Psychological Review*, 103, 356–63.

————(1997), 'Rationality in reasoning: The problem of deductive competence', *Cahiers de Psychologie Cognitive*, 16, 1–35.

Finetti, B. de (1937), 'La Prévision: Ses lois logiques, ses sources subjectives' (Foresight: Its logical laws, its subjective sources), *Annales de l'Institute Henri Poincaré*, 7, 1–68; translated in H. E. Kyburg and H. E. Smokler (1964) (eds), *Studies in Subjective Probability* (Chichester: Wiley).

Fischhoff, B., and Beyth-Marom, R. (1983), 'Hypothesis evaluation from a Bayesian perspective', *Psychological Review*, 90, 239–60.

Fishburn, P. C. (1983), 'Transitive measurable utility', *Journal of Economic Theory*, 31, 293–317.

Fisher, R. A. (1922), 'On the mathematical foundations of theoretical statistics', *Philosophical Transactions of the Royal Society of London, Series A*, 222: 309–68.

——(1925/1970), *Statistical Methods for Research Workers*, 14th edn. (Edinburgh: Oliver & Boyd).

Fodor, J. A. (1983), *Modularity of Mind* (Cambridge, Mass.: MIT Press).

——(1987), *Psychosemantics* (Cambridge, Mass.: MIT Press).

——and Pylyshyn, Z. W. (1988), 'Connectionism and cognitive architecture: A critical analysis', *Cognition*, 28, 3–71.

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

Garey, M. R., and Johnson, D. S. (1979), *Computers and Intractability: A Guide to the Theory of NP-Completeness* (San Francisco: W. H. Freeman).

Gibson, J. J. (1979), *The Ecological Approach to Visual Perception* (Boston: Houghton Mifflin).

Gigerenzer, G., and Goldstein, D. (1996), 'Reasoning the fast and frugal way: Models of bounded rationality', *Psychological Review*, 103, 650–69.

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

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

Good, I. J. (1950), *Probability and the Weighting of Evidence* (London: Griffin).

——(1966), 'A derivation of the probabilistic explication of information', *Journal of the Royal Statistical Society, Series B*, 28, 578–81.

——(1971), 'Twenty seven principles of rationality', in V. P. Godambe and D. A. Sprott (eds), *Foundations of Statistical Inference* (Toronto: Holt, Rhinehart & Wilson).

Goodman, N. (1954), *Fact, Fiction and Forecast* (Cambridge, Mass.: Harvard University Press).

Haack, S. (1978), *Philosophy of Logics* (Cambridge: Cambridge University Press).

Harman, G. (1986), *Change in View* (Cambridge, Mass.: MIT Press).

Harsanyi, John C., and Selten, Reinhard (1988), *A General Theory of Equilibrium Selection in Games* (Cambridge, Mass.: MIT Press).

Helm, P. A. van der, and Leeuwenberg, E. L. J. (1996), 'Goodness of visual regularities: A non-transformational approach', *Psychological Review*, 103, 429–56.

Helmholtz, H. von (1910/1962), *Treatise on Physiological Optics*, iii (J. P. Southall (ed). and translation) (New York: Dover).

Horwich, P. (1982), *Probability and Evidence* (Cambridge: Cambridge University Press).

Howson, C., and Urbach, P. (1989), *Scientific Reasoning: The Bayesian Approach* (La Salle: Open Court).

Jeffreys, H. (1939), *Theory of Probability* (Oxford: Oxford University Press).

Jaynes, E. T. (1989), *Papers on Probability, Statistics, and Statistical Physics*, 2nd edn. (Amsterdam: North-Holland).

Johnson-Laird, P. N. (1983), *Mental Models* (Cambridge: Cambridge University Press).

——and Byrne, R. M. J. (1991), *Deduction* (Hillsdale, NJ: Erlbaum).

——and Wason, P. C. (1970), 'Insight into a logical relation', *Quarterly Journal of Experimental Psychology*, 22, 49–61.

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

——and Tversky, A. (1979), 'Prospect theory: An analysis of decision under risk', *Econometrica*, 47, 263–91.

Keynes, J. M. (1921), *A Treatise on Probability* (London: Macmillan).

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), 'The normative justification for information gain in Wason's selection task', *Psychological Review*, 106, 215–22.

Klayman, J., and Ha, Y. (1987), 'Confirmation, disconfirmation and information in hypothesis testing', *Psychological Review*, 94, 211–28.

Kleindorfer, P. R., Kunreuther, H. C., and Schoemaker, P. J. H. (1993), *Decision Sciences: An Integrated Perspective* (Cambridge: Cambridge University Press).

Kuhn, T. (1962), *The Structure of Scientific Revolutions* (Chicago: University of Chicago Press).

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: Cambridge University Press): 91–196.

Lamberts, K., and Chong, S. (1998), 'Dynamics of dimension weight distribution and flexibility in categorization', in M. Oaksford and N. Chater (eds), *Rational Models of Cognition* (Oxford: Oxford University Press): 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 Boselie, F. (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.

——(1971), *Bayesian Statistics: A Review* (Philadelphia: Society for Industrial and Applied Mathematics).

——(1982), 'Scoring rules and the inevitability of probability', *International Statistical Review*, 50, 1–26.

Loomes, G., and Sugden, R. (1982), 'Regret theory: An alternative theory of rational choice under uncertainty', *Economic Journal*, 92, 805–24.

Lopes, L. L. (1991), 'The rhetoric of irrationality', *Theory & Psychology*, 1, 65–82.

——(1992), 'Three misleading assumptions in the customary rhetoric of the bias literature', *Theory & Psychology*, 2, 231–6.

Lucas, J. R. (1970), *The Concept of Probability* (Oxford: Oxford University Press).

McCarthy, J. M. (1980), 'Circumscription: A form of nonmonotonic reasoning', *Artificial Intelligence*, 13, 27–39.

McClelland, J. L. (1998), 'Connectionist models of Bayesian inference', in M. Oaksford and N. Chater (eds), *Rational Models of Cognition* (Oxford: Oxford University Press): 21–53.

McCloskey, D. N. (1985), *The Rhetoric of Economics* (Madison: University of Wisconsin Press).

McDermott, D. (1982), 'Non-monotonic logic II: Nonmonotonic modal theories', *Journal of the Association for Computing Machinery*, 29, 33–57.

——(1987), 'A critique of pure reason', *Computational Intelligence*, 3, 151–60.

——and Doyle, J. (1980), 'Non-monotonic logic I', *Artifical Intelligence*, 13, 41–72.

McFarland, D. J., and Bösser, T (1993), *Intelligent Behaviour in Animals and Robots (Complex Adaptive Systems)* (Cambridge, Mass.: MIT Press).

McFarland, D., and Houston, A. (1981), *Quantitative Ethology: The State Space Approach* (London: Pitman).

Machina, M. J. (1982). '"Expected utility" analysis without the independence axiom', *Econometrica*, 39, 277–323.

MacKay, D. J. C. (1992), 'A Practical Bayesian Framework for Backpropagation Networks', *Neural Computation*, 4, 448–62.

McKenzie, C. R. M. (1994), 'The accuracy of intuitive judgement strategies: Covariation assessment and Bayesian inference', *Cognitive Psychology*, 26, 209–39.

Marr, D. (1982), *Vision* (San Francisco: W. H. Freeman).

May, K. O. (1954), 'Intransitivity, utility, and the aggregation of preference patterns', *Econometrica*, 22, 1–13.

Maynard-Smith, J., and Price, G. R. (1973), 'The logic of animal conflict', *Nature*, 246, 15–18.

Medin, D. L., and Schaffer, M. M. (1978), 'Context theory of classification learning', *Psychological Review*, 85, 201–38.

Messick, D. M. (1991), 'On the evolution of group-based altruism', in R. Selten (ed.), *Game Equilibrium Models I: Evolution and Game Dynamics* (Berlin: Springer-Verlag): 304–28.

Minsky, M. (1977), 'Frame system theory', in P. N. Johnson-Laird and P. C. Wason (eds), *Thinking: Readings in Cognitive Science* (Cambridge: Cambridge University Press): 355–76.

Muth, J. F. (1961), 'Rational expectations and the theory of price movements', *Econometrica*, 29, 315–35.

Nash, J. (1950), 'The bargaining problem', *Econometrica*, 28, 155–62.

Neal, R. (1993), 'Bayesian learning via stochastic dynamics', in S. J. Hanson, J. D. Cowan, and C. L. Giles (eds), *Advances in Neural Information Processing Systems 5* (San Mateo, Calif.: Morgan Kaufman): 475–82.

Neumann, J. von, and Morgenstern, O. (1944), *Theory of Games and Economic Behavior* (Princeton: Princeton University Press).

Neyman, J. (1950), *Probability and Statistics* (New York: Holt).

Nosofsky, R. M. (1991), 'Relation between the rational model and the context model of categorization', *Psychological Science*, 2, 416–21.

Oaksford, M., and Chater, N. (1991), 'Against logicist cognitive science', *Mind & Language*, 6, 1–38.

———(1992), 'Bounded rationality in taking risks and drawing inferences', *Theory & Psychology*, 2, 225–30.

———(1994), 'A rational analysis of the selection task as optimal data selection', *Psychological Review*, 101, 608–31.

———(1995*a*), 'Information gain explains relevance which explains the selection task', *Cognition*, 57, 97–108.

———(1995*b*), 'Theories of reasoning and the computational explanation of everyday inference', *Thinking and Reasoning*, 1, 121–52.

———(1996), 'Rational explanation of the selection task', *Psychological Review*, 103, 381–91.

———(1998*a*) (eds), *Rational Models of Cognition* (Oxford: Oxford University Press).

———(1998*b*), *Rationality in an Uncertain World* (Hove: Psychology Press).

———(1998*c*), 'A revised rational analysis of the selection task: Exceptions and sequential sampling', in M. Oaksford and N. Chater (eds), *Rational Models of Cognition* (Oxford: Oxford University Press): 372–98.

Paris, J. (1992), *The Uncertain Reasoner's Companion* (Cambridge: Cambridge University Press).

Plaut, D. C., McClelland, J. L., Seidenberg, M. S., and Patterson, K. E. (1996), 'Understanding normal and impaired word reading: Computational principles in quasi-regular domains', *Psychological Review*, 103, 56–115.

Pollard, P. (1985), 'Nonindependence of selections on the Wason selection task', *Bulletin of the Psychonomic Society*, 23, 317–20.

Pomerantz, J. R., and Kubovy, M. (1987), 'Theoretical approaches to perceptual organization', in K. R. Boff, L. Kaufman, and J. P. Thomas (eds), *Handbook of Perception and Human Performance, ii: Cognitive Processes and Performance* (New York: Wiley): 36.1–36.46.

Popper, K. R. (1959), *The Logic of Scientific Discovery* (London: Hutchinson), originally published in 1935.

Putnam, H. (1974), 'The "corroboration" of theories', in P. A. Schilpp (ed.), *The Philosophy of Karl Popper*, i (La Salle, Ill.: Open Court Publishing): 221–40.

Pylyshyn, Z. W. (1984), *Computation and Cognition* (Cambridge, Mass.: MIT Press).

——(ed.), (1987) *The Robot's Dilemma: The Frame Problem in Artificial Intelligence* (Norwood, NJ: Ablex).

Quine, W. V. O. (1953), 'Two dogmas of empiricism', in *From a Logical Point of View* (Cambridge, Mass.: Harvard University Press): 20–46.

——(1960), *Word and Object* (Cambridge, Mass.: MIT Press).

Ramsey, F. P. (1926) 'Truth and Probability', in Ramsey, *The Foundation of Mathematics and Other Logical Essays*, ed. R. B. Braithewaite (London: Kegan Paul).

Rawls, J. (1971), *A Theory of Justice* (Cambridge, Mass.: Harvard University Press).

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.

——(1985), 'One reasoning by default', in R. Brachman and H. Levesque (eds), *Readings in Knowledge Representation* (Los Altos, Calif.: Morgan Kaufman): 401–10; first published in 1978.

Rescorla, R. A., and Wagner, A. R. (1972), 'A Theory of Pavlovian Conditioning: Variations in the Effectiveness of Reinforcement and Nonreinforcement', in A. H. Black and W. F. Prokasy (eds), *Classical Conditioning II: Current Research and Theory* (New York: Appleton-Century-Crofts), 64–94.

Rips, L. J. (1990), 'Reasoning', *Annual Review of Psychology*, 41, 321–53.

——(1994), *The Psychology of Proof* (Cambridge, Mass.: MIT Press).

Rissanen, J. (1987), 'Stochastic complexity', *Journal of the Royal Statistical Society, Series B*, 49, 223–39.

——(1989), *Stochastic Complexity and Statistical Inquiry* (Singapore: World Scientific).

Savage, L. J. (1954), *The Foundations of Statistics* (New York: Wiley).

Schank, R. C., and Abelson, R. P. (1977), *Scripts, Plans, Goals, and Understanding* (Hillsdales, NJ: Erlbaum).

Schooler, L. J. (1998), 'Sorting out core memory processes', in M. Oaksford and N. Chater (eds), *Rational Models of Cognition* (Oxford: Oxford University Press): 128–55.

Seidenberg, M. S., and McClelland, J. L. (1989), 'A distributed, developmental model of word recognition and naming', *Psychological Review*, 96, 523–68.

Shanks, D. R. (1995*a*), 'Is Human Learning Rational?', *Quarterly Journal of Experimental Psychology*, 48A, 257–79.

——(1995*b*), *The Psychology of Associative Learning* (Cambridge: Cambridge University Press).

Simon, H. A. (1955), 'A behavioral model of rational choice', *Quarterly Journal of Economics* 69, 99–118.

——(1956), 'Rational choice and the structure of the environment', *Psychological Review*, 63, 129–38.

——(1991), 'Cognitive architectures and rational analysis: Comment', in K. van Lehn (ed.), *Architectures for Intelligence* (Hillsdale, NJ: Lawrence Erlbaum Associates): 25–40.

Skyrms, B. (1977), *Choice and Chance* (Belmont: Wadsworth).

Sperber, D., Cara, F., and Girotto, V. (1995), 'Relevance theory explains the selection task', *Cognition*, 57, 31–95.

Stein, E. (1996), *Without Good Reason* (Oxford: Oxford University Press).

Stephens, D. W., Krebs, J. R. (1986), *Foraging Theory* (Princeton, NJ: Princeton University Press).

Stich, S. (1983), *From Folk Psychology to Cognitive Science* (Cambridge, Mass.: MIT Press).

——(1985), 'Could man be an irrational animal?', *Synthese*, 64, 115–35.

——(1990), *The Fragmentation of Reason* (Cambridge, Mass.: MIT Press).

——and Nisbett, R. (1980), 'Justification and the psychology of human reasoning', *Philosophy of Science*, 47, 188–202.

Sutherland, S. (1992), *Irrationality: The Enemy Within* (London: Constable).

Thagard, P. (1988), *Computational Philosophy of Science* (Cambridge, Mass.: MIT Press).

Thaler, R. (1987), 'The psychology of choice and the assumptions of economics', in A. Roth (ed.). *Laboratory Experimentation in Economics: Six Points of View* (Cambridge: Cambridge University Press): 99–130.

Tversky, A., and Kahneman, D. (1974), 'Judgement under uncertainty: Heuristics and biases', *Science*, 125, 1124–31.

————(1986), 'Rational choice and the framing of decisions', *Journal of Business*, 59, 251–78.

Valiant, L. G. (1984), 'A theory of the learnable', *Communications of the Association for Computing Machinery*, 27, 1134–42.

Vapnik, V. N. (1995), *The Nature of Statistical Learning Theory* (New York: Springer-Verlag).

Wason, P. C. (1960), 'On the failure to eliminate hypotheses in a conceptual task', *Quarterly Journal of Experimental Psychology*, 12, 129–40.

——(1966), 'Reasoning', in B. Foss (ed.), *New Horizons in Psychology* (Harmondsworth, Mddx.: Penguin).

——(1968), 'Reasoning about a rule', *Quarterly Journal of Experimental Psychology*, 20, 273–81.

——(1969), 'Regression in reasoning', *British Journal of Psychology*, 60, 471–80.

——and Johnson-Laird, P. N. (1972), *The Psychology of Reasoning: Structure and Content* (Cambridge, Mass.: Harvard University Press).

Young, R. (1998), 'Rational analysis of exploratory choice', in M. Oaksford and N. Chater (eds), *Rational Models of Cognition* (Oxford: Oxford University Press): 469–500.