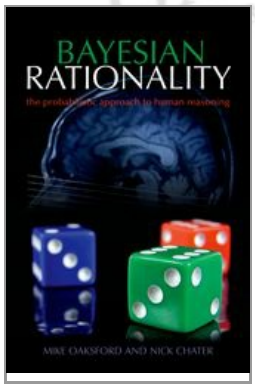

Bayesian Rationality: The probabilistic approach to human reasoning
Mike Oaksford and Nick Chater

Print publication date: 2007
Print ISBN-13: 9780198524496
Published to Oxford Scholarship Online: April 2010
DOI: 10.1093/acprof:oso/9780198524496.001.0001

Rationality and rational analysis

Mike Oaksford
Nick Chater

DOI:10.1093/acprof:oso/9780198524496.003.0002

[–] Abstract and Keywords

One of the central goals of this book is to show how empirical data on human reasoning can be reconciled with the notion that people are rational. This raises two questions: first, the general theoretical question of how the concept of rationality relates to human behaviour; and, second, the methodological question of how to develop 'rational' explanations of behaviour. The answer to the first question will provide a starting point for tackling the second; for which John Anderson's methodology of rational analysis is advocated. This chapter is divided into two parts. First, it discusses formal and everyday rationality, and various possible relationships between them. Second, it outlines how the programme of rational analysis, which is the framework of the research in this book, leads to a new conception of how formal and everyday rationality are related.

Keywords: rationality, human reasoning, John Anderson, rational analysis

One of the central goals of this book is to show how empirical data on human reasoning can be reconciled with the notion that people are rational. This raises two questions: first, the general theoretical question of how the concept of rationality relates to human behaviour; and, second, the methodological question of how to develop 'rational' explanations of behaviour. The answer to the first question will provide a starting point for tackling the second—for which we advocate John Anderson's (1990) methodology of rational analysis.

To get the discussion started, we first consider what it *means* to be rational. Immediately reflection suggests that the task is a difficult one, because the notion of rationality appears in a wide variety of contexts, and is used in a wide variety of apparently rather loosely related senses. For example, in clinical psychology, as well as in the law, rationality enters immediately, to the extent that we attempt to draw a boundary between sanity and madness, or to determine when people are not to be held responsible for their actions. In economics and, increasingly, other areas of social science, human behaviour is explained as the outcome of 'rational choice' (Becker 1976, 1991; Elster 1986). This approach to human behaviour, which we alluded to in the previous chapter, involves assuming that rationality is spelled out in terms of beliefs and desires (with associated probabilities and utilities). The idea is that people's behaviour can be explained as rationally justified, in relation to these postulated degrees of beliefs and levels of desire. But rationality assumptions go much deeper still—such assumptions seem to lie at the heart of the folk psychological style of explanation in which we describe each other's minds and behaviour (Fodor 1987; Sellars 1956). Assumptions of rationality also appear equally essential to interpret each others utterances and to understand texts (Davidson 1984a; Quine 1960). So rationality, in an intuitive sense, appears to be at the heart of 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 specific circumstances.

In this informal, everyday sense, most of us, most of the time, are remarkably rational. In daily life, of course, we tend to focus on occasions when reasoning or decision-making breaks down. But our failures of reasoning are **(p.20)** only salient because they occur against the background of rational thought and behaviour. The rationality of this thought and behaviour is achieved with such little apparent effort that we are inclined to take it for granted—to view it as emerging from plain common sense, where such common sense must be a simple thing indeed. People may not think of themselves as exhibiting high levels of rationality—instead, we think of people as 'intelligent', performing 'appropriate' actions, being 'reasonable' or making 'sensible' decisions. But these labels refer to human abilities to make the right decisions or to say or think the right thing in complex, real-world situations—in short they are labels for everyday rationality.

Indeed, so much do we tend to take the rationality of common-sense thought for granted, that the immense subtlety and sophistication of common-sense reasoning has only been discovered in the latter part of the twentieth century. This discovery 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. Indeed, the project of formalizing common-sense has been brought to an effective standstill (e.g. McDermott 1987) in the face of a nest of difficulties, sometimes grouped under the heading of the 'frame problem' (see Pylyshyn 1987). Two related difficulties are worth highlighting. First, that it does not seem possible to break down the knowledge underlying common-sense into manageable chunks (whether schemas, scripts, or frames; Minsky 1977; Schank and Abelson 1977), which can be understood separately. Instead, each aspect of common-sense knowledge appears inextricably entangled with the rest (e.g. Fodor 1983), so that it seems difficult or even impossible to represent common-sense knowledge in an incremental fashion. Alongside this difficulty stand deep problems in understanding how the plethora of partial constraints embodied in any incomplete and inconsistent knowledge-base can be used to make reliable inferences. This problem is particularly difficult because common-sense inferences are typically defeasible—that is, the addition of new information can overturn a current conclusion. This appears to imply that the entire knowledge-base must somehow be accessed, if a conclusion is to be asserted, because otherwise that conclusion may be overturned by some other piece of knowledge. We have discussed these complex issues in detail elsewhere (Oaksford and Chater 1991, 1995b, 1998a). But in the present context what is important is the upshot of the discussion: that everyday, common-sense reasoning is remarkably, but mysteriously, successful in dealing with an immensely complex and changeable world and that no artificial computational system can begin to approach the level of human performance. This is **(p.21)** the root of the theoretical starting point of this book: that most people, most of the time, are remarkably rational; and hence that the principal challenge in the psychology of reasoning is to understand how such impressive levels of rationality are achieved.

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, such as logic and probability. 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.

The two sides of rationality raise the fundamental question of how they relate to each other: how are the general principles of formal rationality related to specific examples of rational thought and action described by everyday rationality? This question, in various guises, has been widely discussed—in this chapter, we shall outline a particular conception of the relation between these two notions, focusing on a particular style of explanation in the behavioural sciences, *rational analysis* (Anderson 1990). We will argue that rational analysis provides a good characterization of how the concept of rationality is used in explanations in psychology, economics, and animal behaviour, and provides an account of the relationship between everyday and formal rationality, which has implications for both. Moreover, this view of rationality leads to a re-evaluation of the

implications of data from psychological experiments that appear to undermine human rationality. As we shall argue in detail in Chapters 5, 6, and 7, the experimental evidence demands a change concerning which formal account defines the normative standard in experimental tasks.

The discussion in this chapter falls into two parts. First, we discuss formal and everyday rationality, and various possible relationships between them. Second, we outline how the programme of rational analysis, which is the framework of the research in this book, leads to a new conception of how formal and everyday rationality are related.

Relations between formal and everyday rationality

Formal rationality concerns formal principles of good reasoning—the mathematical laws of logic, probability, or decision theory. At an intuitive level, these principles seem distant from the domain of everyday rationality—how people think and act in daily life. Rarely, in daily life, do we accuse one another of violating the laws of logic or probability theory, or praise each other for obeying them. Moreover, when people are given reasoning problems that explicitly require use of these formal principles, their performance appears to be **(p.22)** remarkably poor as we have mentioned. People appear to persistently fall for logical blunders (Evans *et al*. 1993), probabilistic fallacies (e.g. Tversky and Kahneman 1974) and to make inconsistent decisions (Kahneman *et al*. 1982; Tversky and Kahneman 1986). Indeed, the concepts of logic, probability, and the like do not appear to mesh naturally with our everyday reasoning strategies: these notions took centuries of intense intellectual effort to construct, and present a tough challenge for each generation of students.

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

The primacy of everyday rationality

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

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

To the extent that such clashes occur, the advocates of the primacy of everyday rationality argue that the formal theories should be rejected as inadequate systematizations of human rational intuitions, rather than condemning the intuitions under

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

The primacy of everyday rationality is implicit in an important challenge to decision theory by the mathematician Allais (1953). Allais outlines his famous 'paradox', which shows a sharp divergence between people's rational intuitions **(p.23)** and the dictates of decision theory. One version of the paradox is as follows. Consider the following pair of lotteries, each involving 100 tickets. Which would you prefer to play?

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

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

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

Most of us prefer lottery B to lottery A—the slight reduction in the probability of becoming a millionaire is offset by the possibility of the really large prize. But most of us also prefer lottery C to lottery D—we do not 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, although intuitively appealing, is inconsistent with decision theory, as we shall see. Decision theory assumes that people should choose whichever alternative has the maximum expected utility. Denote the utility associated with a sum of $X by U($X). Then the preference for lottery B over A means that:

$$10/100.U(\$1{,}000{,}000) + 90/100.U(\$0) < 1/100.U(\$5{,}000{,}000) + 8/100.U(\$1{,}000{,}000) + 91/100.U(\$0) \quad (2.1)$$

$$10/100.U(\$1{,}000{,}000) < 1/100.U(\$5{,}000{,}000) + 8/100.U(\$1{,}000{,}000) + 1/100.U(\$0) \quad (2.2)$$

But the preference for lottery C over D means that:

$$100.U(\$1{,}000{,}000) > 1/100.U(\$5{,}000{,}000) + 98/100.U(\$1{,}000{,}000) + 1/100.U(\$0) \quad (2.3)$$

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

$$10/100.U(\$1{,}000{,}000) > 1/100.U(\$5{,}000{,}000) + 8/100.U(\$1{,}000{,}000) + 1/100.U(\$0) \quad (2.4)$$

But (2.2) and (2.4) are in contradiction.

(p.24) Allais's paradox is very powerful—the appeal of the choices that decision theory rules out is considerable. Indeed, rather than condemning people's intuitions as incorrect, Allais argues that the paradox undermines the normative status of decision theory—that is, Allais argues that everyday rational intuitions take precedence over the dictates of a formal calculus. Moreover, accounting for the psychological basis of Allais' paradox has become a central objective of descriptive theories of choice, in psychology and economics (e.g. Birnbaum 2004).

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, p. 317). That is, a formal theory of reasoning is acceptable only in so far as it accords with everyday reasoning. Cohen uses the following example to demonstrate the primacy of everyday inference. According to standard propositional logic, the inference from (2.5) to (2.6) is valid:

> If John's automobile is a Mini, John is poor, and (2.5)
> if Johns's automobile is a Rolls, John is rich.
>
> Either, if John's automobile is a Mini, John is rich, or (2.6)
> if Johns's automobile is a Rolls, John is poor.

Clearly, however, this violates intuition. Most people would agree with (2.5) as at least highly plausible; but would reject (2.6) as absurd. *A fortiori*, they would not accept that (2.5) *implies* (2.6) (otherwise they would have to judge (2.6) to be at least as plausible as (2.5)). Consequently, Cohen argues that standard logic simply does not apply to the reasoning that is in evidence in people's intuitions about (2.5) and (2.6). Like Allais, Cohen argues that rather than condemn people's intuitions as irrational, this mismatch reveals the inadequacy of propositional logic as a rational standard. That is, everyday intuitions have primacy over formal theories.

This viewpoint is not without problems. For example, how can rationality be assessed? If formal rationality is viewed as basic, then the degree to which people behave rationally can be assessed by comparing performance against the canons of the relevant normative theory. But if everyday rationality is viewed as basic, assessing rationality appears to be

down to intuition. There is a danger here of losing any normative force to the notion of rationality—if rationality is merely conformity to each others predominant intuitions, then being rational is like a musician being in tune. On this view, rationality has no absolute significance; all that matters is that we reason harmoniously with our fellows. But there is a strong intuition that rationality is not like this at **(p.25)** all—that there is some absolute sense in which some reasoning or decision-making is good, and other reasoning and decision-making is bad. So, by rejecting a formal theory of rationality, there is the danger that the normative aspect of rationality is left unexplained.

One way to re-introduce the normative element is to define a procedure that derives normative principles from human intuitions. Cohen appealed to the notion of reflective equilibrium (Goodman 1954; Rawls 1971), where inferential principles and actual inferential judgements are iteratively bought into a 'best fit' until further judgements do not lead to any further changes of principle (narrow reflective equilibrium). Alternatively, background knowledge may also figure in the process, such that not only actual judgements but also how 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 equilbrium point could, for example, leave the individual content in the idea that the Gambler's fallacy that an event is more likely if it has not occurred recently is a sound principle of reasoning.

Thagard (1988) proposes that instead of reflective equilibrium, developing inferential principles involves progress towards an *optimal* system. This involves proposing principles based on practical judgements and background theories, and measuring these against criteria for optimality. The criteria Thagard specifies are:

(1) *robustness:* principles should be empirically adequate;
(2) *accommodation:* given relevant background knowledge, deviations from these principles can be explained; and
(3) *efficacy:* given relevant background knowledge, inferential goals are satisfied.

Thagard's (1988) concerns were very general: to account for the development of scientific inference. From our current focus on the relationship between everyday and formal rationality, however, Thagard's proposals seem to fall down because the criteria he specifies still seem to leave open the possibility of inconsistency, i.e. it seems possible that a system could fulfill (1) to (3) 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 Euclid and Aristotle. From our perspective, 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 **(p.26)** mismatches observed between human inferential performance and formal theories are a product of using the wrong formal theory to guide expectations about how people should behave.

An alternative normative grounding for rationality seems intuitively appealing: good everyday reasoning and decision-making should lead to *successful action*. For example, from an evolutionary perspective, we might define success as inclusive fitness, and argue that 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 that lead to successful behaviour, and digestive processes that lead to successful metabolic activity. Both increase inclusive fitness; but intuitively we want to say that the first is concerned with rationality, 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.

The primacy of formal rationality

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

The immediate problem for advocates of the primacy of formal rationality concerns the *justification* of formal calculi of reasoning: why should the principles of some calculus be viewed as principles of good reasoning, so that they may be allowed to overturn our intuitions about what is rational? Such justifications typically assume some general, and apparently incontrovertible, cognitive goal; or seemingly undeniable axioms about how thought or 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 1931; Skyrms 1977). Suppose that we assume that people will accept a 'fair' bet: that is, a bet where the expected financial gain is 0, according to their assessment of the probabilities of the various outcomes. Thus, for example, if a person believes that there is a probability of one in three 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.

(p.27) Now, it is possible to prove that, if a person's assignment of probabilities to different possible outcomes violates the laws of probability theory in any way whatever, then it is possible to offer them a combination of different bets, such that they will happily accept each individual bet as fair, in the above sense, but where *whatever the outcome* they are certain to lose money. Such a combination of bets—where one side is certain to lose—is known as a Dutch book; and it seems incontrovertible that accepting a bet that you are certain to lose must violate rationality. Thus, if violating the laws of probability theory leads to accepting Dutch books, which seems clearly irrational, then obeying the

laws of probability theory seems to be a condition of rationality.

The Dutch book theorem might appear to have a fundamental weakness—that it requires that a person willingly accepts arbitrary fair bets. But, in reality of course, this might not be so—many people will, in such circumstances, be risk aversive and choose not to accept such bets. But the same argument applies even if the person does not bet at all. Now the inconsistency concerns a hypothetical—the person believes that *if* the bet were accepted, it would be fair (and 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 what may be a myriad of possibly conflicting 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 **(p.28)** if, much of the time, human intuitions and behaviour strongly violates their recommendations.

If formal rationality is primary, what are we to make of the fact that, in explicit tests at least, people seem to be such poor probabilists and logicians? One line would be to accept that human reasoning is badly flawed. Thus, the heuristics and biases programme (Kahneman and Tversky 1973; Kahneman *et al*. 1982), which charted systematic errors in human probabilistic reasoning and decision-making under uncertainty, can be viewed as exemplifying this position (see Gigerenzer and Goldstein 1996), as can Evans' (1982, 1989) heuristic approach to reasoning. Another line follows the spirit of Chomsky's (1965) distinction between linguistic competence and performance—the idea is that the people's reasoning competence accords with formal principles, but in practice, performance limitations (e.g. limitations of time, memory, and language comprehension) 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. In Chapter 1, we noted that two of the leading theoretical frameworks for modelling human reasoning, mental logic (Braine 1978; Rips 1994) and mental models (Johnson-Laird 1983; Johnson-Laird and Byrne 1991) rely on a distinction between reasoning competence and performance. Both these frameworks assume that classical logic provides the appropriate competence theory for deductive reasoning—they differ only over how the dictates of this competence theory

are implemented in mental processes. Thus, according to both theories, logically errors in people's 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 views 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 also Chater and Oaksford 1990).

Mental models concurs that logical inference provides the computational-level theory for reasoning, but provides an alternative method of proof. Instead of standard proof theoretic rules, this view uses a 'semantic' method of proof. Such methods involve search for models (in the logical sense)—a **(p.29)** 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 logic notion. 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 logic programming; Clark 1978).

Mental logic and mental models assume that formal principles of rationality—specifically classical logic—(at least partly) define the standards of good reasoning. They explain the non-logical nature of people's actual reasoning 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 and formal rationality are completely separate

Recently, a number of theorists have suggested what is effectively a hybrid of the two approaches outlined above. They argue that formal rationality and everyday rationality are entirely separate enterprises. For example, Evans and Over (1997, p. 2) 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.

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, p. 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 that lead to success: rationality$_1$ does not require rationality$_2$. That is, Evans and Over seem committed to the view that thoughts, actions, or decisions that cannot be normatively justified can, nonetheless, consistently lead to practical success.

(p.30) But this hybrid view does not tackle the fundamental problem we outlined for the first view sketched above. It does not answer the question: *why* do the cognitive processes underlying everyday rationality consistently work? If everyday rationality is somehow based on formal rationality, then this question can be answered, at least in general terms. The principles of formal rationality are provably principles of good inference and decision-making; and the cognitive system is rational in everyday contexts to the degree that it approximates the dictates of these principles. But if everyday and formal rationality are pressumed 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.

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, this 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 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 that 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; this **(p.31)** is Davidson's (1984a) *principle of charity*. 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 explanation here 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 1993). 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 rational explanation 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 this chapter 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. In Chapters 5–7, we will use rational analysis to re-evaluate the psychological data, which has appeared to show human reasoning performance to be hopelessly flawed, and argue that, when appropriate rational theories are applied, reasoning performance may, on the contrary, be rational.

(p.32) 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, 1991a). This methodology provides a framework for explaining the link between principles of formal rationality and the practical success of everyday rationality not just in

psychology, but throughout the study of 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, 1991a) 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 (for an overview of recent research see: Oaksford and Chater 1998a), from areas as diverse as categorization (Anderson 1991b; Anderson and Matessa 1998; Lamberts and Chong 1998), memory (Anderson and Milson 1989; Anderson and Schooler 1991; Schooler and Anderson 1997), searching computer menus (Young 1998), and natural language parsing (Chater *et al*. 1998). This research has shown that a great many empirical generalizations about cognition can be viewed as arising from the rational adaptation of the cognitive system to the problems and constraints that it faces. We shall **(p.33)** 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 are 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). **(p.34)** 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 the appropriate formal normative tools, rather than decision theory.

This aspect of rational analysis is at variance with Evans and Over's distinction between two forms of rationality, mentioned above. They argue that 'people are largely rational in the sense of achieving their goals (rationality$_1$) but have only a limited ability to reason or act for good reasons sanctioned by a normative theory (rationality$_2$)' (Evans and Over 1997, p. 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*. Moreover, 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 '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).

(p.35) Equally, in economics, idealizations of the 'environment' are crucial to determining rational economic behaviour (McCloskey 1985). In micro-economics, modelling the environment (e.g. game-theoretically) involves capturing the relation between each actor and the environment of other actors. In macro-economics, explanations using rational expectations theory (Muth 1961) begin from a formal model of the environment, as a set of equations governing macro-economic variables.

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

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 **(p.36)** 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 also be required to provide a full explanation of their behaviour (e.g. Anderson 1993; Oaksford and Chater 1995b).

This viewpoint is standard in rational explanations across a broad range of disciplines. Economists do not assume that people make complex game-theoretic or macro-economic 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* (von Helmholtz 1910; 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 (Van der 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 **(p.37)** 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 narrower 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 maximizing 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. Therefore, determining whether some behaviour is rational or not depends on more than just being able to provide an account in terms of optimization. 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 **(p.38)** 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. 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 that 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.

Later, we shall see how the rational analysis approach can lead to specific accounts of the three key areas of the psychology of deductive reasoning: conditional reasoning (Chapter 5), Wason's selection task (Chapter 6), and syllogistic reasoning (Chapter 7). In the present chapter, we have set out an empirical programme for investigating the rationality of any particular human behaviour. But to carry out this programme in practice requires choosing a particular framework for developing rational analyses of reasoning.

One key clue that probability theory, rather than logic, will provide more appropriate rational analyses, is the nature of everyday reasoning. Perhaps our theories of laboratory tasks should be inspired by our theories of human everyday reasoning, on the assumption that the cognitive system is adapted to reasoning in the everyday world rather than to reasoning in the laboratory.

This suggests that it would be useful to consider the degree to which real-world reasoning can be modelled using logic. If it can, then we would be confident that **(p.39)** people's underlying logical competence must be quite impressive and that their apparently dismal laboratory performance must somehow be explicable in terms of the impoverished or unrealistic nature of the tasks, just as the visual system is subject to a wide range of perceptual illusions using impoverished stimuli (Cohen 1981). If, on the other hand, everyday reasoning does not involve logical reasoning to any substantial degree, then the possibility must be admitted that people are genuinely poor at logical reasoning—because this is not the kind of reasoning to which the cognitive system is adapted. It also raises the further possibility that people might tackle supposedly 'logical'

reasoning tasks by co-opting non-logical reasoning strategies that they use in everyday life. This would lead to the paradoxical conclusion that 'logical' reasoning tasks may not be treated as logical tasks by experimental participants at all.

In the next chapter we address the question of the nature of everyday reasoning head-on. The traditional assumption, in both philosophy and psychology, has been that logic is at the core of everyday reasoning; but we shall suggest research in a range of disciplines suggests that, on the contrary, human everyday reasoning is fundamentally uncertain. **(p.40)**

Notes:

(1) Dutch-book arguments remain, however, controversial–e.g. there are, however, a range of alternative justifications, based on theories of 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 (for discussion see: Earman 1992; Howson and Urbach 1993).