# Psychology of Reasoning

## Theoretical and Historical Perspectives

**edited by**
**Ken Manktelow & Man Cheung Chung**

Psychology Press
Taylor & Francis Group
HOVE AND NEW YORK

Wigton, R. S. (1996). Social judgement theory and medical judgment. *Thinking and Reasoning*, *2*, 175–190.
Woodworth, R. S. (1938). *Experimental psychology*. New York: Holt.

# 3 Rationality, rational analysis, and human reasoning

*Nick Chater and Mike Oaksford*

The idea of rationality is central to the explanation of human behaviour. Only on the assumption that people are at least typically rational can we attribute beliefs, motives, and desires to people – the assumption of rationality provides the "glue" that holds disparate beliefs, desires, and actions together in a coherent system. Imagine explaining a routine event, such a motorist slowing down when approaching a pedestrian crossing. The motorist, we might suggest, noticed that some people were near the crossing, believed that they were about to cross, wanted to avoid colliding with them, believed that collision might occur if the car continued at its current speed, and so on. The goal of such explanation is to provide a *rationale* for a person's behaviour, explaining how they understood and acted upon the world, from their point of view. But constructing a rationale for a piece of behaviour will only provide an explanation for it if we assume that people are sensitive to such rationales; that is, unless people exhibit rationality.

This style of explanation is, of course, ubiquitous in our everyday explanation of the thoughts and behaviour of ourselves and others – and it is embodied not merely in everyday discourse, but is also fundamental to explanation in the humanities and in literature. In attempting to *interpret* and *understand* other people's decisions and utterances, we are attempting to provide rationales for those decisions and utterances. The historian explaining the actions of a military general, the scholar interpreting a Biblical text, and the novelist conjuring up a compelling character all rely, fundamentally, on the assumption that people are, by and large, rational (Davidson, 1984; Quine, 1960).

The rationales that we provide for each other's behaviour are typically extremely subtle and elaborate, but at the same time incomplete and unsystematic. For example, in explaining why the general made a particular military decision, the historian may spell out some of the general's relevant beliefs and desires, e.g., beliefs concerning the location of enemy forces, the desire to be viewed as a hero by future generations. But the explanation will inevitably be partial. The historian will leave out beliefs such as the background assumptions that future generations will admire victory more than defeat, that to weaken enemy forces, shelling should be directed at them

rather than at the surrounding countryside, that shells travel in the direction of fire and explode on impact, that explosions are injurious to those nearby, that people fight less well when deprived of supplies, and so on. To reconstruct a rationale for the general's actions in full detail would appear to be an intractable task. This is because explaining the basis for any aspect of the general's thought appears to draw on still further beliefs and desires. Thus, the general's beliefs about the motion of shells will depend on endless beliefs about naïve physics, about the approximate weight and size of a shell (e.g., that shells are denser than balloons), about how shells are fired and so on. Understanding how shells are fired leads on to understanding the properties of the gun, the properties of gunpowder, and so on indefinitely. The historian need not, of course, bother to enunciate this apparently endless store of knowledge in order to communicate with the reader – because this indefinitely large store of knowledge can be assumed to be *common knowledge* between historian and reader. But the fact that we can and do rely on common knowledge to underpin everyday explanation of human behaviour can obscure just how partial and incomplete everyday explanations are.

In this informal, and somewhat ill-defined everyday sense, most of us, most of the time, are remarkably rational. In daily life, of course, we tend to focus on occasions when reasoning or decision making breaks down. But our failures of reasoning are only salient because they occur against the background of rational thought and behaviour that 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 tend to think of them as just plain common sense – with the implicit assumption that common sense must be a simple thing indeed. People may not think of themselves as exhibiting high levels of rationality – instead, we think of each other as "intelligent", performing "appropriate" actions, being "reasonable", or making "sensible" decisions. But these labels refer to human abilities to make the right decisions, or to say or think the right thing in complex, real-world situations – in short, they are labels for everyday rationality.

Indeed, so much do we tend to take the rationality of commonsense thought for granted, that realizing that commonsense reasoning is immensely difficult, and hence our everyday rationality is thereby immensely impressive, has been a surprising *discovery*, and a discovery made only in the latter part of the twentieth century. The discovery emerged from the project of attempting to formalize everyday knowledge and reasoning in artificial intelligence, where initially high hopes that commonsense knowledge could readily be formalized were replaced by increasing desperation at the impossible difficulty of the project. The nest of difficulties referred to under the "frame problem" (see, e.g., Pylyshyn, 1987), and the problem that each aspect of knowledge appears inextricably entangled with the rest (e.g., Fodor, 1983) so that commonsense does not seem to break down into manageable "packets" (whether schemas, scripts, or frames, Minsky, 1977; Schank & Abelson, 1977), and the deep problems of defeasible, or non-monotonic reasoning



(e.g., McDermott, 1987), brought the project of formalizing commonsense effectively to a standstill. So the discovery is now made – it is now clear that everyday, commonsense reasoning is remarkably, but mysteriously, successful in dealing with an immensely complex and changeable world and that no existing artificial computational system begins to approach the level of human performance.

Let us contrast this informal, everyday sense of rationality concerning people's ability to think and act in the real world, with a concept of rationality originating not from human behaviour, but from mathematical theories of good reasoning. These mathematical theories represent one of the most important achievements of modern thought: Logical calculi formalize aspects of deductive reasoning; axiomatic probability formalizes probabilistic reasoning; a 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 relating hypotheses to data; utility and decision theory attempt to characterize rational preferences and rational choice between actions under uncertainty; game theory and its variants (e.g., Harsanyi & Selten, 1988; von Neumann & Morgenstern, 1944) aim to provide a precise framework for determining the rational course of action in situations in which the reasoning of other agents must be taken into account. 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 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 an attractive account of the relationship between everyday and formal rationality, which has implications for both. Moreover, this view of rationality leads to a re-evaluation of the implications of data from psychological experiments which appear to undermine human rationality. A wide range of empirical results in the psychology of reasoning have been taken to cast doubt on human rationality, because people appear to persistently make elementary logical blunders. We show that, when the tasks people are given are viewed in terms of probability, rather than logic, people's responses can be seen as rational.

The discussion falls into 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 apply rational analysis to re-evaluating experimental data in the psychology of reasoning, from a

probabilistic standpoint. Finally, we consider implications, problems, and prospects for project of building a more adequate psychology of reasoning.

## RELATIONS BETWEEN FORMAL AND EVERYDAY RATIONALITY

Formal rationality concerns formal principles of good reasoning – the mathematical laws of logic, probability, or decision theory. At an intuitive level, these principles seem distant from the domain of everyday rationality – how people think and act in daily life. Rarely, in daily life, do we accuse one another of violating the laws of logic or probability theory, or praise each other for obeying them. Moreover, when people are given reasoning problems that explicitly require use of these formal principles, their performance appears to be remarkably poor, a point we touched on above. People appear to persistently fall for logical blunders (Evans, Newstead, & Byrne, 1993) and probabilistic fallacies (e.g., Tversky & Kahneman, 1974), and to make inconsistent decisions (Kahneman, Slovic, & Tversky, 1982; Tversky & 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 appear to reveal blunders in human reasoning are taken at face value.

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



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

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

| A. | B. |
|---|---|
| 10 tickets worth £1,000,000 | 1 ticket worth £5,000,000 |
| 90 tickets worth £0 | 8 tickets worth £1,000,000 |
| | 91 tickets worth £0 |

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

| C. | D. |
|---|---|
| 100 tickets worth £1,000,000 | 1 ticket worth £5,000,000 |
| | 98 tickets worth £1,000,000 |
| | 1 ticket worth £0 |

Most of us prefer lottery B to lottery A – the slight reduction in the probability of becoming a millionaire is offset by the possibility of the really large prize. But most of us also prefer lottery C to lottery D – we don't think 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:

$$10/100.U(£1{,}000{,}000) + 90/100.U(£0) < 1/00.U(£5{,}000{,}000) + 8/100.U(£1{,}000{,}000) + 91/100.U(£0) \quad (1)$$

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

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

But the preference for lottery C over D means that:

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

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

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

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 – that is, Allais argues that everyday rational intuitions take precedence over the dictates of a formal calculus.

Another example arises in Cohen's (1981) discussion of the psychology of reasoning literature. Following similar arguments of Goodman (1954), Cohen argues that a normative or formal theory is "acceptable . . . only so far as it accords, at crucial points with the evidence of untutored intuition," (Cohen, 1981, 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 (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 absurd. *A fortiori*, they would not accept that (5) *implies* (6) – otherwise they would have to judge (6) to be at least as plausible as (5). Consequently, Cohen argues that standard logic simply does not apply to the reasoning that is in evidence in people's intuitions about (5) and (6). Like Allais, Cohen argues that rather than condemn people's intuitions as irrational, this mismatch reveals the inadequacy of propositional logic as a rational standard. That is, everyday intuitions have primacy over formal theories.

But this viewpoint is not without problems. For example, how can rationality be assessed? If formal rationality is viewed as basic, then the degree to



which people behave rationally can be evaluated 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 with each other's predominant intuitions, then being rational is like a musician being in tune. On this view, rationality has no absolute significance; all that matters is that we reason harmoniously with our fellows. But there is a strong intuition that rationality is not like this at all – that there is some absolute sense in which some reasoning or decision making is good, and other reasoning and decision making is bad. So, by rejecting a formal theory of rationality, there is the danger that the normative aspect of rationality is left unexplained.

One way to re-introduce the normative element is to define a procedure that derives normative principles from human intuitions. Cohen appealed to the notion of reflective equilibrium (Goodman, 1954; Rawls, 1971) where inferential principles and actual inferential judgements are iteratively bought into a "best fit" until further judgements do not lead to any further changes of principle (narrow reflective equilibrium). Alternatively, background knowledge may also figure in the process, such that not only actual judgements but also the way they relate to other beliefs are taken into account (wide reflective equilibrium). These approaches have, however, been subject to much criticism (e.g., Stich & 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 leave the individual content in the idea that the Gambler's Fallacy is a sound principle of reasoning.

Thagard (1988) proposes that instead of reflective equilibrium, developing inferential principles involves progress towards an optimal system. This involves proposing principles based on practical judgements and background theories, and measuring these against criteria for optimality. The criteria Thagard specifies are (i) robustness: principles should be empirically adequate; (ii) accommodation: given relevant background knowledge, deviations from these principles can be explained; and (iii) efficacy: given relevant background knowledge, inferential goals are 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 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 chapter, accounts like reflective equilibrium and Thagard's account, which attempts to drive a wedge between formality and normativity, may not be required. We argue that many of the

mismatches observed between human inferential performance and formal theories are a product of using the wrong formal theory to guide expectations about how people should behave.

An alternative normative grounding for rationality seems intuitively appealing: good everyday reasoning and decision making should lead to *successful action*; for example, from an evolutionary perspective, we might define success as inclusive fitness, and argue that 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, which the second is not. More generally, defining rationality in terms of outcomes runs the risk of blurring what appears to be a crucial distinction – between minds, which may be more or less rational, and stomachs, which 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 within mathematics, statistics, operations research, and the "decision sciences" (e.g., Kleindorfer, Kunreuther, & 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.

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 even be allowed to overturn our intuitions about what is rational? Such justifications typically assume some general, and apparently incontrovertible, cognitive goal; or seemingly undeniable axioms about how thought or 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 1/3 that it will rain tomorrow, then they will be happy to accept a bet according to which they win two dollars if it does rain tomorrow, but they lose one dollar if it does not. Now, it is possible to prove that, if a person's assignment of probabilities to different possible outcomes violates the laws of probability theory in any way whatever, then the following curious state of affairs holds. It is possible



to offer the person a combination of different bets, such that they will happily accept each individual bet as fair, in the above sense. But, despite being happy that each of the bets is fair, it turns out that *whatever the outcome* the person will lose money. Such a combination of bets – where one side is certain to lose – is known as a Dutch book; and it is seems incontrovertible that accepting a bet that you are certain to lose must violate rationality. Thus, if violating the laws of probability theory leads to accepting Dutch books, which seems clearly irrational, then obeying the laws of probability theory seems to be a condition of rationality.

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

We have considered the Dutch book justification of probability theory in some detail to make it clear that justifications of formal theories of rationality can have considerable force. 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; von Neumann & Morgenstern, 1944; Savage, 1954). Moreover, the same general approach can be used as a justification for logic, if avoiding inconsistency is taken as axiomatic. Thus, there may be good reasons for accepting formal theories of rationality, even if, much of the time, human intuitions and behaviour strongly violate 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 & 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 & 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 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 & 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 & 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 & Pylyshyn, 1988, pp. 29–30; see also Chater & 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 searching 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. 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).

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 nonlogical 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 (1996a, 1997) distinguish between two notions of rationality (1997, p. 2):

> 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 & Over, 1997, p. 1). If this is right, then 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 led to success: rationality$_1$ does not require rationality$_2$. That is, Evans and Over are committed to the view that thoughts, actions, or decisions that cannot be normatively justified can, nonetheless, consistently lead to practical success.

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

### 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 notion 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 – however, which formal principles are appropriate, and how they should be applied, is not decided *a priori* but in the light of the empirical success of the explanation of the adaptive success of the cognitive process under consideration.

According to this viewpoint, the apparent mismatch between normative theories and reasoning 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 people think 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 & Urbach, 1989). 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 chapter aim to 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 that 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.

## RATIONAL ANALYSIS

As with all good ideas, rational analysis has a long history. The roots of rational analysis derive from the earliest attempts to build theories of rational thought or choice. For example, probability theory was originally developed as a theory of how sensible people reason about uncertainty (Gigerenzer, Swijtnik, Porter, Daston, Beatty & Krüger, 1989). Thus, the early literature on probability theory treated the subject both as a description of human psychology and as a set of norms for how people ought to reason when dealing with uncertainty. Similarly, the earliest formalisations of logic (Boole, 1951/1854) viewed the principles as describing the laws governing thought, as well providing a calculus for good reasoning. This early work in probability theory and logic is a precursor of rational analysis, because it aims both to describe how the mind works, and to explain why the mind is rational.

The twentieth century, however, saw a move away from this "psychologism" (Frege, 1879; Hilbert, 1925) and now mathematicians, philosophers, and psychologists sharply distinguish between normative theories, such as a probability theory and logic, which are about how people *should* reason, and descriptive theories of the psychological mechanisms by which people actually *do* reason. Moreover, a major finding in psychology has been that the rules by which people *should* and *do* reason are not merely conceptually distinct; but they appear to be empirically very different (Kahneman & Tversky, 1973; Kahneman et al., 1982; Wason, 1966; Wason & Johnson-Laird, 1972). Whereas very early research on probability theory and logic took their project as codifying how people think, the psychology of reasoning has suggested that probability theory and logic are profoundly at

variance with how people think. If this viewpoint is correct, then the whole idea of rational models of cognition is misguided: cognition simply is not rational.

Rational analysis suggests a return to the earlier view of the relationship between descriptive and normative theory, i.e., that a single theory can, and should, do both jobs. A rational model of cognition can therefore explain both how the mind works and why it is successful. But why is rational analysis not just a return to the conceptual confusion of the past? It represents a psychological proposal for explaining cognition that recognizes the conceptual distinction between normative and descriptive theories, but explicitly suggests that in explaining cognitive performance a single account that has both functions is required. Moreover, contemporary rational analyses are explicit scientific hypotheses framed in terms of the computer metaphor, which can be tested against experimental data. Consequently a rational model of cognition is an empirical hypothesis about the nature of the human cognitive system and not merely an a priori assumption.

The computational metaphor is important because it suggests that rational analyses should be described in terms of a scheme for computational explanation. The most well-known scheme for computational explanation was provided by Marr (1982). At Marr's highest level of explanation, the *computational* level the function that is being computed in the performance of some task is outlined. This level corresponds to a rational analysis of the cognitive task. The emphasis on computational explanation makes two points explicit. First, that in providing a computational explanation of the task that a particular device performs there is an issue about whether the computational-level theory is correct. Second, there is a range of possible computational-level theories that may apply to a given task performance, and which one is correct must be discovered and cannot be assumed a priori.

Let us consider an example. Suppose you find an unknown device and wonder what its function might be. Perhaps, observing its behaviour, you hypothesize that it may be performing arithmetical calculations. To make this conjecture is to propose a particular rational model of its performance. That is, this is a theory about what the device *should* do. In this case, the device should provide answers to arithmetical problems that conform to the laws of arithmetic, i.e., arithmetic (or some portion of it) provides the hypothesized rational model. On this assumption, you might give the device certain inputs, which you interpret as framing arithmetical problems. It may turn out, of course, that the outputs that you receive do not appear to be interpretable as solutions to these arithmetical problems. This may indicate that your rational model is inappropriate. You may therefore search for an alternative rational model – perhaps the device is not doing arithmetic, but is solving differential equations. Similarly, in rational analysis, theorists cannot derive appropriate computational-level theories by reflecting on normative considerations alone,



but only by attempting to use those theories to describe human performance. For example, it is not controversial that arithmetic is a good normative account of how numbers should be manipulated – the question is: does this device do arithmetic?

This leads to the second difference between the modern programme of rational analysis and early developments of logic and probability: that the goal is not merely to capture people's intuitions, but rather to model detailed experimental data on cognitive function. Rational models aim to capture experimental data on the rate at which information is forgotten; on the way people generalize from old to new instances; on performance on hypothesis-testing tasks; on search problems; and so on. Rational analysis as a programme in cognitive science is primarily aimed at capturing these kinds of empirical phenomena, while explaining how the cognitive system is successful. Nonetheless, rational analysis shares with early views the assumption that accounts of the mind must be both normatively justified and descriptively adequate.

So far, we have considered rationality in the abstract – as consisting of reasoning according to sound principles. But the goals of an agent attempting to survive and prosper in its ecological niche are more concrete – it must decide how to act in order to achieve its goals. So a crucial issue is how normative principles can be combined with analysis of the structure of the environment in order to provide rational explanations of successful cognitive performance. Recent research indicates that many aspects of cognition can be viewed as optimized (to some approximation) to the structure of the environment. For example, the rate of forgetting an item in memory seems to be optimized to the likelihood of encountering that item in the world (Anderson & Milson, 1989; Anderson & Schooler, 1991; Schooler, 1998); categorization may be viewed as optimizing the ability to predict the properties of a category member (Anderson, 1991b, 1998); searching computer menus (Young, 1998), parsing (Chater, Crocker, & Pickering, 1998), and selecting evidence in reasoning (Oaksford & Chater, 1994, 1996, 1998a; Over & Jessop, 1998) may all be viewed as optimizing the amount of information gained. This style of explanation is similar to optimality-based explanations that have been influential in other disciplines. In the study of animal behaviour (Stephens & Krebs, 1986), foraging, diet selection, mate selection and so on, have all been viewed as problems, which animals solve more or less optimally. In economics, people and firms are viewed as more or less optimally making decisions in order to maximize utility or profit.

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 optimized 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. The agent must still act, even in the absence of the ability to derive the optimal solution (Chater & Oaksford, 1996; Gigerenzer & Goldstein, 1996; Oaksford & Chater, 1991; Simon, 1956). Thus, there may be a tension between the theoretical goal of the rational analysis and the practical need for the agent to be able to decide how to act in real time, given the partial information available. This leads directly into the area of what Simon (1955, 1956) calls *bounded rationality*. We believe that rational analysis can be reconciled with the boundedness of cognitive systems in a number of ways.

First, the cognitive system may, in general, approximate, perhaps very coarsely, the optimal solution. Thus, the algorithms that the cognitive system uses may be fast and frugal heuristics (Gigerenzer & Goldstein, 1996) which generally approximate the optimal 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 the agent's behaviour – including those aspects that depart from the predictions from a rational analysis (Anderson, 1990, 1994).

Second, 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, but 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 explanation, 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 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 to maximize food intake, on the assumption that this local goal is generally relevant to the global goal of maximizing inclusive fitness. Similarly, explanations concerning cognitive processes may concern local cognitive goals such as maximizing the amount of useful information remembered, maximizing predictive accuracy, or acting 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 importance that the local goals be relevant to 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, 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. Stomachs may be well or poorly adapted to their function (digestion), but they have no beliefs, desires, or knowledge, and hence the question of their rationality does not arise.

Optimality approaches in biology, economics, and psychology assume 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 a very artificial task in the laboratory. Any laboratory task will recruit some set of cognitive mechanisms that determine the participants' behaviour. But it is not obvious what problem these mechanisms are adapted to solving. Clearly, 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 participants may assimilate the task that they are given to a more natural task, recruiting adaptively appropriate mechanisms which solve this, more natural, task successfully.

This issue is most pressing in reasoning tasks where human performance has been condemned as irrational. For example, hypothesis-testing tasks, where people do not adopt the supposedly "logical" strategy of falsification, have been taken to demonstrate the irrationality of human reasoning (Stich, 1985, 1990; Sutherland, 1992). However, recently a number of theorists have suggested that human reasoning should be judged against probabilistic standards, as opposed the norms of logic (e.g., Evans & Over, 1997; Fischhoff & Beyth-Marom, 1983; Kirby, 1994; Oaksford & Chater, 1994, 1998c; Over & Jessop, 1998). One powerful argument for this position is that the complex and uncertain character of the everyday world implies that real-world everyday reasoning is inevitably uncertain (Chater & Oaksford, 1996; Oaksford & Chater, 1998c), and hence better modelled by probability, the calculus of uncertain reasoning, than by logic, the calculus of certain reasoning. From this point of view, people's behaviour in laboratory reasoning tasks is (to an approximation at least) rational, even though it violates the standards set by the experimenter – but this can only be appreciated once the standard of correct performance is reconceptualized in probabilistic terms.

## RE-EVALUATING HUMAN REASONING: A PROBABILISTIC APPROACH

This section focuses on our recent attempts to develop a probabilistic analysis of laboratory reasoning tasks (Chater & Oaksford, 1999a, 1999b, 1999c, 2000; Oaksford & Chater, 1994, 1995a, 1995b, 1996, 1998a, 1998b, 1998c; Oaksford, Chater, & Grainger, 1999; Oaksford, Chater, Grainger & Larkin, 1997; Oaksford, Chater, & Larkin, 2000). But to appreciate what is distinctive about the probabilistic approach, we must first begin by considering logic-based theories in the psychology of reasoning, which have been dominant since the inception of the field. Logic-based theories of reasoning fall into two types.

According to the *mental models* view (Johnson-Laird, 1983; Johnson-Laird & Byrne, 1991), people construct one or more concrete models of the situation that is described by the premises with which they are presented, and derive conclusions from "reading off" conclusions that follow in one or more of these models. There are procedures for building, checking, and reading from models that should allow the reasoner, if all goes well, to conform with the dictates of deductive logic. According to the *mental logic* view, people reason by directly performing calculations in a particular logical system – typically assumed to be some kind of natural deduction system (Braine, 1978; Rips, 1983, 1994).

According to these viewpoints, people are rational in principle but err in practice – that is, we have sound procedures for deductive reasoning but the algorithms that we use can fail to produce the right answers because of



cognitive limitations such as working memory capacity. Such an approach seems hard to reconcile with two facts. First, these faulty algorithms can lead to error rates as high as 96% (in Wason's selection task) compared to the standard provided by formal logic. Second, our everyday rationality in guiding our thoughts and actions seems in general to be highly successful. How is this success to be understood if the reasoning system people use is prone to so much error?

As we discussed above, we attempt to resolve this problem by arguing that people's everyday reasoning can be understood from the perspective of probability theory and that people make errors in so-called deductive tasks because they generalize their everyday strategies to these laboratory tasks. The psychology of deductive reasoning involves giving people problems that the experimenters conceive of as requiring logical inference. But people consistently respond in a non-logical way, thus calling human rationality into question (Stein, 1996; Stich, 1985, 1990). In our view, everyday rationality is founded on uncertain rather than certain reasoning (Oaksford & Chater, 1991, 1998c) and so probability provides a better starting point for an account of human reasoning than logic. It also resolves the problem of explaining the success of everyday reasoning: it is successful to the extent that it approximates a probabilistic theory of the task. Second, we suggest that a probabilistic analysis of classic "deductive" reasoning tasks provides an excellent empirical fit with observed performance. The upshot is that much of the experimental research in the "psychology of deductive reasoning" does not engage people in deductive reasoning at all but rather engages strategies suitable for probabilistic reasoning. According to this viewpoint, the field of research appears to be crucially misnamed!

We illustrate our probabilistic approach in the three main tasks that have been the focus of research into human reasoning: conditional inference, Wason's selection task, and syllogistic inference.

### Conditional inference

Conditional inference is perhaps the simplest inference form investigated in the psychology of reasoning. It involves presenting participants with a conditional premise, *if p then q*, and then one of four categorical premises, *p*, *not-p*, *q*, or *not-q*. Logically, given the categorical premise *p* participants should draw the conclusion *q* and given the categorical premise *not-q* they should draw the conclusion *not-p*. These are the logically valid inferences of modus ponens ("MP") and modus tollens ("MT") respectively. Moreover, given the categorical premise *not-p* participants should *not* draw the conclusion *not-q* and given the categorical premise *q* they should *not* draw the conclusion *p*. These are the logical fallacies of denying the antecedent ("DA") and affirming the consequent ("AC") respectively. So, logically, participants should endorse MP and MT in equal proportion and they should refuse to

endorse DA or AC. However, they endorse MP significantly more than MT and they endorse DA and AC at levels significantly above zero.

Following a range of other researchers (Anderson, 1995; Chan & Chua, 1994, George, 1997; Liu, Lo, & Wu, 1996; Stevenson & Over, 1995), Oaksford, Chater, and Larkin (2000) proposed a model of conditional reasoning based on conditional probability. The greater the conditional probability of an inference the more it should be endorsed. On their account the meaning of a conditional statement can be defined using a 2 by 2 contingency table as in Table 3.1 (see Oaksford & Chater, 1998c).

*Table 3.1* Contingency table for a conditional rule

| | *q* | *not-q* |
|---|---|---|
| *p* | $a\,(1-\varepsilon)$ | $a\varepsilon$ |
| *not-p* | $b-a\,(1-\varepsilon)$ | $(1-b)-a\varepsilon$ |

This table represents a conditional rule, if $p$ then $q$, where there is a dependency between the $p$ and $q$ that may admit exceptions ($\varepsilon$) and where $a$ is the probability of the antecedent, $P(p)$, $b$ is the probability of the consequent, $P(q)$, and $\varepsilon$ is the probability of exceptions, i.e., the probability that $q$ does not occur even though $p$ has, $P(not\text{-}q|p)$. It is straightforward to then derive conditional probabilities for each inference. For example, the conditional probability associated with MP, i.e., $P(q|p) = 1 - \varepsilon$, only depends on the probability of exceptions. If there are few exceptions the probability of drawing the MP inference will be high. However, the conditional probability associated with MT, i.e.,

$$P(not\text{-}p|not\text{-}q) = \frac{1-b-a\varepsilon}{1-b}$$

depends on the probability of the antecedent, $P(p)$, and the probability of the consequent, $P(q)$, as well the probability of exceptions. As long as there are exceptions ($\varepsilon > 0$) and the probability of the antecedent is greater than the probability of the consequent not occurring ($P(p) > 1 - P(q)$), then the probability of MT is less than MP ($P(not\text{-}p|not\text{-}q) < P(q|p)$). For example, if $P(p) = .5$, $P(q) = .8$ and $\varepsilon = .1$, then $P(q|p) = .9$ and $P(not\text{-}p|not\text{-}q) = .75$. This behaviour of the model accounts for the preference for MP over MT in the empirical data. In the model conditional probabilities associated with DA and AC also depend on these parameters, which means that they can be non-zero. Consequently the model also predicts that the fallacies should be endorsed to some degree.



Oaksford et al. (2000) argue that this simple model can also account for other effects in conditional inference. For example, using Evans' Negations Paradigm in the conditional inference task leads to a bias towards negated conclusions. Oaksford and Stenning (1992; see also Oaksford & Chater, 1994) proposed that negations define higher-probability categories than their affirmative counterparts, e.g., the probability that an animal is not a frog is much higher than the probability that it is. Oaksford et al. (2000) show that according to their model the conditional probability of an inference increases with the probability of the conclusion. Consequently the observed bias towards negated conclusions may actually be a rational preference for high-probability conclusions. If this is right then, when given rules containing high- and low-probability categories, people should show a preference to draw conclusions that have a high probability analogous to negative conclusion bias, a prediction later confirmed experimentally (Oaksford et al., 2000).

## Wason's selection task

The probabilistic approach was originally applied to Wason's selection task, which we introduced above (Oaksford & Chater, 1994, 1995b, 1996, 1998a, 1998c; Oaksford et al., 1999; Oaksford et al., 1997). According to Oaksford and Chater's (1994) optimal data selection model people select evidence (i.e., turn cards) to determine whether $q$ depends on $p$, as in Table 3.1, or whether $p$ and $q$ are statistically independent (i.e., the cell values would simply be the products of the marginal probabilities, rather than as in Table 3.1). What participants are looking for in the selection task is evidence that gives the greatest probability of discriminating between these two possibilities. Initially participants are assumed to be maximally uncertain about which possibility is true, i.e., a prior probability of .5 is assigned to both the possibility of a dependency (the dependence hypothesis, $H_D$) and to the possibility of independence (the independence hypothesis, $H_I$). The participants' goal is to select evidence (turn cards) that would be expected to produce the greatest reduction in this uncertainty. This involves calculating the posterior probabilities of the hypotheses, $H_D$ or $H_I$, being true given some evidence. These probabilities are calculated using Bayes' theorem, which requires information about prior probabilities ($P(H_D) = P(H_I) = .5$) and the likelihoods of evidence given a hypothesis, e.g., the probability of finding an A when turning the 2 card assuming $H_D$ ($P(A|2, H_D)$). These likelihoods can be calculated directly from the contingency tables for each hypothesis: for $H_D$, Table 3.1, and for $H_I$, the independence model. With these values it is possible to calculate the reduction in uncertainty that can be expected by turning any of the four cards in the selection task. Oaksford and Chater (1994) observed that assuming that the marginal probabilities $P(p)$ and $P(q)$ were small (their "rarity assumption"), the $p$ and the $q$ cards would be expected to provide the greatest reduction in uncertainty about which hypothesis was true. Consequently, the

selection of cards that has been argued to demonstrate human irrationality may actually reflect a highly rational data selection strategy. Indeed this strategy may be optimal in an environment where most properties are rare, e.g., most things are not black, not ravens, and not apples (but see Klauer, 1999, and Chater & Oaksford, 1999b, for a reply).

Oaksford and Chater (1994) argued that this model can account for most of the evidence on the selection task, and Oaksford and Chater (1996) defended the model against a variety of objections. For example, Evans and Over (1996b) criticized the notion of information used in the optimal data selection model and proposed their own probabilistic model. This model made some predictions that diverged from Oaksford and Chater's model and these have been experimentally tested by Oaksford et al. (1999). Although the results seem to support the optimal data selection model, there is still much room for further experimental work in this area. Manktelow and Over have been exploring probabilistic effects in deontic selection tasks (Manktelow, Sutherland, & Over, 1995). Moreover, Green and Over have also been exploring the probabilistic approach to the standard selection task (Green, Over, & Pyne, 1997; see also Oaksford, 1998; Green & Over, 1998). They have also extended this approach to what they refer to as "causal selection tasks" (Green & Over, 1997, 2000; Over & Green, 2001). This is important because their work develops the link between research on causal estimation (e.g., Anderson & Sheu, 1995; Cheng, 1997) and research on the selection task suggested by Oaksford and Chater (1994).

## Syllogistic reasoning

Chater and Oaksford (1999c) have further extended the probabilistic approach to the more complex inferences involved in syllogistic reasoning, which we discussed in looking at mental models. In their probability heuristics model (PHM), they extend their probabilistic interpretation of conditionals to quantified claims, such as, All, Some, None, and Some . . . not. In Table 3.1, if there are no exceptions, then the probability of the consequent given the antecedent, ($P(q|p)$), is 1. The conditional and the universal quantifier "All" have the same underlying logical form: $\forall x(P(x) \Rightarrow Q(x)$. Consequently Chater and Oaksford interpreted universal claims such as All *P*s are *Q*s as asserting that the probability of the predicate term ($Q$) given the subject term ($P$) is 1, i.e., $P(Q|P) = 1$. Probabilistic meanings for the other quantifiers are then easily defined: None, $P(Q|P) = 0$; Some, $P(Q|P) > 0$; Some . . . not, $P(Q|P) < 1$. Given these probabilistic interpretations it is possible to show which conclusions follow probabilistically for all 64 syllogisms (i.e., which syllogisms are "*p*-valid"). Moreover, given these interpretations and again making the rarity assumption (see above on the selection task), the quantifiers can be ordered in terms of how informative they are: All > Some > None > Some . . . not. It turns out that a simple set of heuristics defined over



the informativeness of the premises can successfully predict the *p*-valid conclusion, if there is one. The most important of these heuristics is the *min*-heuristic, which states that the conclusion will have the form of the least informative premise. So for example, a *p*-valid syllogism such as, *All B are A*, *Some B are not C*, yields the conclusion *Some A are not C*. Note that the conclusion has the same form as the least informative premise. This simple heuristic captures the form of the conclusion for most *p*-valid syllogisms. Moreover, if overgeneralized to the invalid syllogisms, the conclusions it suggests match the empirical data very well. Other heuristics determine the confidence that people have in their conclusions and the order of terms in the conclusion.

Perhaps the most important feature of PHM is that it can generalise to syllogisms containing quantifiers such as Most and Few that have no logical interpretation. In terms of Table 3.1 the suggestion is that these terms are used instead of All when there are some (Most) or many (Few) exceptions. So the meaning of Most is: $1 - \Delta < P(Q|P) < 1$, and the meaning of Few is: $0 < P(Q|P) < \Delta$, where $\Delta$ is small. These interpretations lead to the following order of informativeness: All > Most > Few > Some > None > Some . . . not. Consequently, PHM uniquely makes predictions for the 144 syllogisms that are produced when Most and Few are combined with the standard logical quantifiers. Chater and Oaksford (1999c) (i) show that their heuristics pick out the *p*-valid conclusions for these new syllogisms, and (ii) they report experiments confirming the predictions of PHM when Most and Few are used in syllogistic arguments.

There has already been some work on syllogistic reasoning consistent with PHM. Newstead, Handley, and Buck (1999) found that the conclusions participants drew in their experiments were mainly as predicted by the *min*-heuristic, although they found little evidence of the search for counterexamples predicted by mental models theory for multiple model syllogisms. Evans, Handley, Harper, and Johnson-Laird (1999) also found evidence consistent with PHM, indeed they found that an important novel distinction they discovered between strong and weak possible conclusions could be captured as well by the *min*-heuristic as by mental models theory. A conclusion is necessarily true if it is true in all models of the premises, a conclusion is possibly true if it is true in at least one model of the premises, and a conclusion is impossible if it is not true in any model of the premises. Evans et al. (1999) found that some possible conclusions were endorsed by as many participants as necessary conclusions and that some were endorsed by as few participants as impossible conclusions. According to mental models theory this happens because strong possible conclusions are those that are true in the initial model constructed but not in subsequent models, and weak possible conclusions are those that are only true in non-initial models. Possible strong conclusions all conform to the *min*-heuristic, i.e., they either match the *min*-premise or are less informative than the *min*-premise. Possible weak conclusions all violate the *min*-heuristic (bar one), i.e., they have conclusions that are

more informative than the *min*-premise. In sum, PHM would appear to be gaining some empirical support.

## WHERE NEXT FOR THE PSYCHOLOGY OF REASONING?

Despite the intensive research effort over the last 40 years, human reasoning remains largely mysterious. While there is increased understanding of some aspects of laboratory performance, deep puzzles over the nature of everyday human reasoning processes remain. We suggest that three key issues may usefully frame the agenda for future research: (1) establishing the relation between reasoning and other cognitive processes; (2) developing formal theories that capture the full richness of everyday reasoning; (3) explaining how such theories can be implemented in real-time in the brain.

### Reasoning and cognition

From an abstract perspective, almost every aspect of cognition can be viewed as involving inference. Perception involves inferring the structure of the environment from perceptual input; motor control involves inferring appropriate motor commands from proprioceptive and perceptual input, together with demands of the motor task to be performed; learning from experience, in any domain, involves inferring general principles from specific examples; understanding a text or utterance typically requires inferences relating the linguistic input to an almost unlimited amount of general background knowledge. Is there a separate cognitive system for *reasoning*, or are the processes studied by reasoning researchers simply continuous with the whole of cognition? A key sub-question concerns the modularity of the cognitive system. If the cognitive system is non-modular, then reasoning would seem, of necessity, to be difficult to differentiate from other aspects of cognition. If the cognitive system is highly modular, then different principles may apply in different cognitive domains. Nonetheless, it might still turn out that, even if modules are informationally sealed off from each other (e.g., Fodor, 1983), the inferential principles that they use might be the same; the same underlying principles and mechanisms might simply be reused in different domains. Even if the mind is modular, it seems unlikely that there could be a module for *reasoning* in anything like the sense studied in psychology. This is because everyday reasoning (in contrast to some artificial laboratory tasks) requires engaging arbitrary world knowledge. Consequently, understanding reasoning would appear to be part of the broader project of understanding central cognitive processes and the knowledge they embody in full generality.

This is an alarming prospect for reasoning researchers because current formal research is unable to provide adequate tools for capturing even limited amounts of general knowledge, let alone reasoning with it effectively and in real-time, as we shall discuss below. Reasoning researchers often attempt to



seal off their theoretical accounts from the deep waters of general knowledge, by assuming that these problems are solved by other processes – e.g., processes constraining how mental models are "fleshed out" (Johnson-Laird & Byrne, 1991) or when particular premises can be used in inference (Politzer & Braine, 1991), what information is relevant (Evans, 1989; Sperber, Cara, & Girotto, 1995) or how certain probabilities are determined (Oaksford & Chater, 1994). Whether or not this strategy is methodologically appropriate in the short term, substantial progress in understanding everyday reasoning will require theories that address, rather than avoid, these crucial issues, i.e., theories that explicate, rather than presupposing, our judgements concerning what is plausible, probable, or relevant. Moreover, as we have seen, recent empirical work seems to strongly suggest that progress in understanding human reasoning even in the laboratory requires the issue of general knowledge to be tackled.

### Formal theories of everyday reasoning

Explaining the cognitive processes involved in everyday reasoning requires developing a formal theory that can capture everyday inferences. Unfortunately, however, this is far from straightforward, because everyday inferences are *global*: whether a conclusion follows typically depends not just on a few circumscribed "premises" but on arbitrarily large amounts of general world knowledge (see, e.g., Fodor, 1983; Oaksford & Chater, 1991, 1998c). From a statement such as *While John was away, Peter changed all the locks in the house*, we can provisionally infer, e.g., that Peter did not want John to be able to enter the house; that John possesses a key; that Peter and John have had a disagreement, and so on. But such inferences draw on background information, such as that the old key will not open the new lock, that locks secure doors, that houses can usually only be entered through doors, and a host of more information about the function of houses, and the nature of human relationships, and the law concerning breaking and entering. Moreover, deploying each piece of information requires an inference that is just as complex as the original one. Thus, even to infer that John's key will not open the new lock requires background information concerning the way in which locks and keys are paired together, the convention that when locks are replaced, they will not fit the old key, that John's key will not itself be changed when the locks are changed, that the match between lock and key is stable over time, and so on. This is what we call the "fractal" character of commonsense reasoning (Oaksford & Chater, 1998c) – just as, in geometry, each part of a fractal is as complex as the whole, each part of an everyday inference is as complex as the whole piece of reasoning.

How can such inferences be captured formally? Deductive logic is inappropriate, because everyday arguments are not deductively valid, but can be overturned when more information is learned. The essential problem is that these methods fail to capture the global character of everyday inference

successfully (Oaksford & Chater, 1991, 1992, 1993, 1998c). In artificial intelligence, this has led to a switch to using probability theory, the calculus of *un*certain reasoning, to capture patterns of everyday inference (e.g., Pearl, 1988). This is an important advance, but only a beginning. Probabilistic inference can only be used effectively if it is possible to separate knowledge into discrete chunks – with a relatively sparse network of probabilistic dependencies between the chunks. Unfortunately, this just does not seem to be possible for everyday knowledge. The large variety of labels for the current impasse – the "frame" problem (McCarthy & Hayes, 1969; Pylyshyn, 1987), the "world knowledge" problem or the problem of knowledge representation (Ginsberg, 1987), the problem of non-monotonic reasoning (Paris, 1994), the criterion of completeness (Oaksford & Chater, 1991, 1998c) – is testimony to its fundamental importance and profound difficulty. The problem of providing a formal calculus of everyday inference presents a huge intellectual challenge, not just in psychology, but in the study of logic, probability theory, artificial intelligence, and philosophy.

## Everyday reasoning and real-time neural computation

Suppose that a calculus which captured everyday knowledge and inference could be developed. If this calculus underlies thought, then it must be implemented (probably to an approximation) in real-time in the human brain. Current calculi for reasoning, including standard and non-standard logics, probability theory, decision theory, and game theory, are computationally intractable (Garey & Johnson, 1979; Paris, 1994). That is, as the amount of information that they have to deal with increases, the amount of computational resources (in memory and time) required to derive conclusions explodes very rapidly (or, in some cases, inferences are not computable at all, even given limitless time and memory). Typically, attempts to extend standard calculi to mimic everyday reasoning more effectively make problems of tractability *worse* (e.g., this is true of "non-monotonic logics" developed in artificial intelligence). Somehow, a formal calculus of everyday reasoning must be developed that, instead, eases problems of tractability.

This piles difficulty upon difficulty for the problem of explaining human reasoning computationally. Nonetheless, there are interesting directions to explore. For example, modern "graphical" approaches to probabilistic inference in artificial intelligence and statistics (e.g., Pearl, 1988) are very directly related to connectionist computation; and more generally, connectionist networks can be viewed as probabilistic inference machines (Chater, 1995; MacKay, 1992; McClelland, 1998). To the extent that the parallel, distributed style of computation in connectionist networks can be related to the parallel, distributed computation in the brain, this suggests that the brain may be understood, in some sense, as directly implementing rational calculations. Nonetheless, there is currently little conception either of how such probabilistic models can capture the "global" quality of everyday reasoning, or of how these probabilistic calculations can be carried out in real-time to support fluent and rapid inference, drawing on large amounts of general knowledge, in a brain consisting of notoriously slow and noisy neural components (Feldman & Ballard, 1982).



## Where do we stand?

This chapter has focused on the relationship between mathematical theories of good reasoning and the everyday rational explanations of thought and behaviour. We have argued for a particular relationship between the informal and formal rationality – that patterns of informal reasoning can be explained as approximating the dictates of formal, rational theories. Rational analysis of a particular pattern of inference can serve both a descriptive and a normative role. It can describe the broad patterns of human reasoning performance; but at the same time explain why these patterns of reasoning are adaptively successful in the real world. We have also given a range of concrete examples of how this approach can be applied, showing that many aspects of human laboratory reasoning that appear to be unsystematic and irrational when viewed from the perspective of deductive logic, appear systematic and rational when re-conceptualized in terms of probability theory. But we have cautioned that the project of building a more adequate and general psychology of reasoning faces, nonetheless, enormous difficulties – most fundamentally because the performance of human everyday reasoning radically exceeds the performance of any current formal theories of reasoning. Thus, we believe that the project of understanding human reasoning requires the construction of richer normative theories of good reasoning. Hence, the apparently narrow project of the psychology of reasoning is, in fact, a joint project for disciplines that have fundamentally normative concerns (philosophy, probability theory, decision theory, artificial intelligence) in concert with the experimental, descriptive, study of human thought that has been the traditional territory of the psychologist.

## REFERENCES

Allais, M. (1953). Le comportement de l'homme rationnel devant le risque: Critique des postulats et axiomes de l'école américaine. *Econometrica*, *21*, 503–546.

Anderson, J. R. (1990). *The adaptive character of thought*. Hillsdale, NJ: Lawrence Erlbaum Associate Inc.

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

Anderson, J. R. (1991b). The adaptive nature of human categorization. *Psychological Review*, *98*, 409–429.

Anderson, J. R. (1994). *Rules of the mind*, Hillsdale, NJ: Lawrence Erlbaum Associates Inc.

Anderson, J. R. (1995). *Cognitive psychology and it implications*. New York: W. H. Freeman & Company.

Anderson, J. R., & Matessa, M. (1998). The rational analysis of categorization and the ACT-R architecture. In M. Oaksford & N. Chater (Eds.), *Rational models of cognition* (pp. 197–217). Oxford, UK: Oxford University Press.

Anderson, J. R., & Milson, R. (1989). Human memory: An adaptive perspective. *Psychological Review*, *96*, 703–719.

Anderson, J. R., & Schooler, L. J. (1991). Reflections of the environment in memory. *Psychological Science*, *2*, 396–408.

Anderson, J. R., & Sheu, C-F. (1995). Causal inferences as perceptual judgements. *Memory and Cognition*, *23*, 510–524.

Boole, G. (1951). *An investigation into the laws of thought*. New York: Dover. [Originally published in 1854.]

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

Chan, D., & Chua, F. (1994). Suppression of valid inferences: Syntactic views, mental models, and relative salience. *Cognition*, *53*, 217–238.

Chater, N. (1995). Neural networks: The new statistical models of mind. In J. P. Levy, D. Bairaktaris, J. A. Bullinaria, & P. Cairns (Eds.), *Connectionist models of memory and language* (pp. 207–227). London: UCL Press.

Chater, N., Crocker, M., & Pickering, M. (1998). The rational analysis of inquiry: The case of parsing. In M. Oaksford, & N. Chater (Eds.), *Rational models of cognition* (pp. 441–468). Oxford: Oxford University Press.

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

Chater, N., & Oaksford, M. (1996). The falsity of folk theories: Implications for psychology and philosophy. In W. O'Donohue & R. Kitchener (Eds.), *The philosophy of psychology* (pp. 244–256), London: Sage Publications.

Chater, N., & Oaksford, M. (1999a). Ten years of the rational analysis of cognition. *Trends in Cognitive Sciences*, *3*, 57–65.

Chater, N., & Oaksford, M. (1999b). Information gain vs. decision-theoretic approaches to data selection: Response to Klauer. *Psychological Review*, *106*, 223–227.

Chater, N., & Oaksford, M. (1999c). The probability heuristics model of syllogistic reasoning. *Cognitive Psychology*, *38*, 191–258.

Chater, N., & Oaksford, M. (2000). The rational analysis of mind and behaviour. *Synthese*, 93–131.

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

Chomsky, N. (1965). *Aspects of the theory of syntax*. Cambridge, MA: MIT Press.

Clark, K. L. (1978). Negation as failure. In *Logic and databases* (pp. 293–322). New York: Plenum Press.

Cohen, L. J. (1981). Can human irrationality be experimentally demonstrated? *Behavioral and Brain Sciences*, *4*, 317–370.

Cox, R. T. (1961). *The algebra of probable inference*. Baltimore: The Johns Hopkins University Press.

Davidson, D. (1984). *Inquiries into truth and interpretation*. Oxford: Clarendon Press.

de Finetti, B. (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 & H. E. Smokler (Eds.) (1964). *Studies in subjective probability*. Chichester, UK: Wiley.]

Evans, J. St. B. T. (1982). *The psychology of deductive reasoning*. London: Routledge & Kegan Paul.

Evans, J. St. B. T. (1989). *Bias in human reasoning: Causes and consequences*. Hove, UK: Lawrence Erlbaum Associates Ltd.

Evans, J. St. B. T., Handley, S. J., Harper, C. N. J., & Johnson-Laird, P. N. (1999). Reasoning about necessity and possibility: A test of the mental model theory of deduction. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *25*, 1495–1513.

Evans, J. St. B. T., Newstead, S. E., & Byrne, R. M. J. (1993). *Human reasoning*. Hove, UK: Lawrence Erlbaum Associates Ltd.

Evans, J. St. B. T., & Over, D. (1996a). *Rationality and reasoning*. Hove, UK: Psychology Press.

Evans, J. St. B. T., & Over, D. (1996b). Rationality in the selection task: Epistemic utility vs. uncertainty reduction. *Psychological Review*, *103*, 356–363.

Evans, J. St. B. T., & Over, D. (1997). Rationality in reasoning: The problem of deductive competence. *Cahiers de Psychologie Cognitive*, *16*, 1–35.

Feldman, J., & Ballard, D. (1982). Connectionist models and their properties. *Cognitive Science*, *6*, 205–254.

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

Fisher, R. A. (1922). On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society of London, Series A*, *222*, 309–368.

Fisher, R. A. (1925/1970). *Statistical methods for research workers* (14th Ed.). Edinburgh: Oliver & Boyd.

Fodor, J. A. (1983). *Modularity of mind*. Cambridge MA: MIT Press.

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

Frege, G. (1879). *Begriffschrift*. Halle, Germany: Nebert.

Garey, M. R., & Johnson, D. S. (1979). *Computers and intractability: A guide to the theory of NP-completeness*. San Fransisco: W. H. Freeman.

George, C. (1997). Reasoning from uncertain premises. *Thinking and Reasoning*, *3*, 161–190.

Gigerenzer, G., & Goldstein, D. (1996). Reasoning the fast and frugal way: Models of bounded rationality. *Psychological Review*, *103*, 650–669.

Gigerenzer, G., Swijtnik, Z., Porter, T., Daston, L., Beatty, J., & Krüger, L. (1989). *The empire of chance*. Cambridge: Cambridge University Press.

Ginsberg, M. L. (Ed.). (1987). *Readings in nonmonotonic reasoning*. Los Altos, CA: Morgan Kaufman.

Goodman, N. (1954). *Fact, fiction and forecast*. Cambridge, MA: Harvard University Press.

Green, D. W., & Over, D. E. (1997). Causal inference, contingency tables and the selection task. *Current Psychology of Cognition*, *16*, 459–487.

Green, D. W., & Over, D. E. (1998). Reaching a decision: A reply to Oaksford. *Thinking and Reasoning*, *4*, 231–248.

Green, D. W., & Over, D. E. (2000). Decision theoretic effects in the selection task. *Current Psychology of Cognition*, *19*, 51–68.

Green, D. W., Over, D. E., & Pyne, R. A. (1997). Probability and choice in the selection task. *Thinking and Reasoning*, *3*, 209–235.
Harsanyi, J., & Selten, R. (1988). *A general theory of equilibrium selection in games*. Cambridge, MA: MIT Press.
Hilbert, D. (1925). Über das unendliche. *Mathematische Annalen*, *95*, 161–190.
Howson, C., & Urbach, P. (1989). *Scientific reasoning: The Bayesian approach*. La Salle, IL: Open Court.
Johnson-Laird, P. N. (1983). *Mental models*. Cambridge: Cambridge University Press.
Johnson-Laird, P. N., & Byrne, R. M. J. (1991). *Deduction*. Hillsdale, NJ: Lawrence Erlbaum Associates Inc.
Kahneman, D., Slovic, P., & Tversky, A. (Eds.). (1982). *Judgement under uncertainty: Heuristics and biases*. Cambridge: Cambridge University Press.
Kahneman, D., & Tversky, A. (1973). On the psychology of prediction. *Psychological Review*, *80*, 237–251.
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). On the normative justification for information gain in Wason's selection task. *Psychological Review*, *106*, 215–222.
Kleindorfer, P. R., Kunreuther, H. C., & Schoemaker, P. J. H. (1993). *Decision sciences: An integrated perspective*. Cambridge: Cambridge University Press.
Lindley, D. V. (1971). *Bayesian statistics: A review*. Philadelphia, PA: Society for Industrial & Applied Mathematics.
Liu, I., Lo, K., & Wu, J. (1996). A probabilistic interpretation of "If-then." *Quarterly Journal of Experimental Psychology*, *49*A, 828–844.
MacKay, D. J. C. (1992). Information-based objective functions for active data selection. *Neural Computation*, *4*, 590–604.
Manktelow, K. I., Sutherland, E. J., & Over, D. E. (1995). Probabilistic factors in deontic reasoning. *Thinking and Reasoning*, *1*, 201–220.
Marr, D. (1982). *Vision*. San Francisco: W. H. Freeman.
McCarthy, J. M. (1980). Circumscription: A form of nonmonotonic reasoning. *Artificial Intelligence*, *13*, 27–39.
McCarthy, J. M., & Hayes, P. (1969). Some philosophical problems from the standpoint of Artificial Intelligence. In B. Meltzer, & D. Michie (Eds.), *Machine intelligence, Volume 4* (pp. 463–502). Edinburgh: Edinburgh University Press.
McClelland, J. L. (1998). Connectionist models and Bayesian inference. In M. Oaksford & N. Chater, (Eds.), *Rational models of cognition* (pp. 21–53). Oxford: Oxford University Press.
McDermott, D. (1982). Non-monotonic logic II: Nonmonotonic modal theories. *Journal of the Association for Computing Machinery*, *29*, 33–57.
McDermott, D. (1987). A critique of pure reason. *Computational Intelligence*, *3*, 151–160.
McDermott, D., & Doyle, J. (1980). Non-monotonic logic I. *Artifical Intelligence*, *13*, 41–72.
Minsky, M. (1977). Frame system theory. In P. N. Johnson-Laird, & P. C. Wason (Eds.), *Thinking: Readings in cognitive science* (pp. 355–376). Cambridge: Cambridge University Press.
Newstead, S. E., Handley, S. J., & Buck, E. (1999). Falsifying mental models: Testing the predictions of theories of syllogistic reasoning. *Memory & Cognition*, *27*, 344–354.



Neyman, J. (1950). *Probability and statistics*. New York: Holt.
Oaksford, M. (1998). Task demands and revising probabilities in the selection task. *Thinking and Reasoning*, *4*, 179–186.
Oaksford, M., & Chater, N. (1991). Against logicist cognitive science. *Mind & Language*, *6*, 1–38.
Oaksford, M., & Chater, N. (1992). Bounded rationality in taking risks and drawing inferences. *Theory and Psychology*, *2*, 225–230.
Oaksford, M., & Chater, N. (1993). Reasoning theories and bounded rationality. In K. I. Manktelow, & D. E. Over (Eds.), *Rationality* (pp. 31–60). London: Routledge.
Oaksford, M., & Chater, N. (1994). A rational analysis of the selection task as optimal data selection. *Psychological Review*, *101*, 608–631.
Oaksford, M., & Chater, N. (1995a). Theories of reasoning and the computational explanation of everyday inference. *Thinking and Reasoning*, *1*, 121–152.
Oaksford, M., & Chater, N. (1995b). Information gain explains relevance which explains the selection task. *Cognition*, *57*, 97–108.
Oaksford, M., & Chater, N. (1996). Rational explanation of the selection task. *Psychological Review*, *103*, 381–391.
Oaksford, M., & Chater, N. (1998a). A revised rational analysis of the selection task: Exceptions and sequential sampling. In M. Oaksford & N. Chater (Eds.), *Rational models of cognition* (pp. 372–398). Oxford: Oxford University Press.
Oaksford, M., & Chater, N. (Eds.). (1998b). *Rational models of cognition*. Oxford: Oxford University Press.
Oaksford, M., & Chater, N. (1998c). *Rationality in an uncertain world*. Hove, UK: Psychology Press.
Oaksford, M., Chater, N., & Grainger, B. (1999). Probabilistic effects in data selection. *Thinking and Reasoning*, *5*, 193–243.
Oaksford, M., Chater, N., Grainger, B., & Larkin, J. (1997). Optimal data selection in the reduced array selection task (RAST). *Journal of Experimental Psychology: Learning, Memory and Cognition*, *23*, 441–458.
Oaksford, M., Chater, N., & Larkin, J. (2000). Probabilities and polarity biases in conditional inference. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *26*, 883–899.
Oaksford, M., & Stenning, K. (1992). Reasoning with conditionals containing negated constituents. *Journal of Experimental Psychology: Learning, Memory & Cognition*, *18*, 835–854.
Over, D. E., & Green, D. W. (2001). Contingency, causation, and adaptive inference. *Psychological Review*, *108*, 682–684.
Over, D. E., & Jessop, A. (1998). Rational analysis of causal conditionals and the selection task. In M. Oaksford & N. Chater (Eds.), *Rational models of cognition* (pp. 399–414). Oxford: Oxford University Press.
Paris, J. (1994). *The uncertain reasoner's companion*. Cambridge: Cambridge University Press.
Pearl, J. (1988). *Probabilistic reasoning in intelligent systems: Networks of plausible inference*. San Mateo, CA: Morgan Kaufman.
Politzer, G., & Braine, M. D. S. (1991). Responses to inconsistent premises cannot count as suppression of valid inferences. *Cognition*, *38*, 103–108.
Pylyshyn, Z. W. (Ed.). (1987). *The robot's dilemma: The frame problem in artificial intelligence*. Norwood, NJ: Ablex.

Quine, W. V. O. (1960). *Word and object*. Cambridge, MA: MIT Press.
Ramsey, F. P. (1931). *The foundations of mathematics and other logical essays*. London: Routledge & Kegan Paul.
Rawls, J. (1971). *A theory of justice*. Cambridge, MA: Harvard University Press.
Reiter, R. (1980). A logic for default reasoning. *Artificial Intelligence*, *13*, 81–132.
Reiter, R. (1985). One reasoning by default. In R. Brachman, & H. Levesque (Eds.), *Readings in knowledge representation* (pp. 401–410). Los Altos, CA: Morgan Kaufman. [First published in 1978.]
Rips, L. J. (1983). Cognitive processes in propositional reasoning. *Psychological Review*, *90*, 38–71.
Rips, L. J. (1994). *The psychology of proof*. Cambridge, MA: MIT Press.
Savage, L. J. (1954). *The foundations of statistics*. New York: Wiley.
Schank, R. C., & Abelson, R. P. (1977). *Scripts, plans, goals, and understanding*. Hillsdales, NJ: Lawrence Erlbaum Associates Inc.
Schooler, L. J. (1998). Sorting out core memory processes. In M. Oaksford & N. Chater (Eds.), *Rational models of cognition* (pp. 128–155). Oxford: Oxford University Press.
Simon, H. A. (1955). A behavioral model of rational choice. *Quarterly Journal of Economics*, *69*, 99–118.
Simon, H. A. (1956). Rational choice and the structure of the environment. *Psychological Review*, *63*, 1298–1138.
Skyrms, B. (1977). *Choice and chance*. Belmont: Wadsworth.
Sperber, D., Cara, F., & 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.
Stevenson, R. J., & Over, D. E. (1995). Deduction from uncertain premises. *Quarterly Journal of Experimental Psychology*, *48*A, 613–643.
Stich, S. (1985). Could man be an irrational animal? *Synthese*, *64*, 115–135.
Stich, S. (1990). *The fragmentation of reason*. Cambridge, MA: MIT Press.
Stich, S., & Nisbett, R. (1980). Justification and the psychology of human reasoning. *Philosophy of Science*, *47*, 188–202.
Sutherland, N. S. (1992). *Irrationality: The enemy within*. London: Constable.
Thagard, P. (1988). *Computational philosophy of science*. Cambridge, MA: MIT Press.
Tversky, A., & Kahneman, D. (1974). Judgement under uncertainty: Heuristics and biases. *Science*, *125*, 1124–1131.
Tversky, A., & Kahneman, D. (1986). Rational choice and the framing of decisions. *Journal of Business*, *59*, 251–278.
von Neumann, J., & Morgenstern, O. (1944). *Theory of games and economic behaviour*. Princeton, NJ: Princeton University Press.
Wason, P. C. (1966). Reasoning. In B. Foss (Ed.), *New horizons in psychology*. Harmondsworth, UK: Penguin.
Wason, P. C., & Johnson-Laird, P. N. (1972). *The psychology of reasoning: Structure and content*. Cambridge, MA: Harvard University Press.
Wilson, E. O. (1975). *Sociobiology: The new synthesis*. Cambridge, MA: Belknap Press.
Young, R. (1998). Rational analysis of exploratory choice. In M. Oaksford & N. Chater (Eds.), *Rational models of cognition* (pp. 469–500). Oxford: Oxford University Press.

# 4 The psychology of conditionals

*David Over*

The use of conditionals is central to human reasoning, and any psychological theory of reasoning worthy of the name must have an adequate account of conditionals in natural language. Yet even taking the first steps towards a theory of the ordinary indicative conditional immediately entangles the psychologist in formidable logical and philosophical, as well as psychological, problems. It is a good test of any psychological theory of reasoning to go straight to its account of ordinary indicative conditionals. The theory is in serious trouble if it does not have an adequate account of this conditional, and it is very easy to fail this test, as I will try to show in what follows. Further challenges are presented by counterfactual and deontic conditionals.

Present psychological theories of all these conditionals suffer from what has been called logicism (Evans, 2002; Oaksford & Chater, 1998). One way to characterize logicism is that it is the attempt to account for a significant aspect of human reasoning by using logic alone. Logicism restricts itself to the study of logical inference from assumptions, i.e., premises that are supposed to be taken, in effect, as certain. But inference in the real world is usually from premises that people rightly think of as uncertain to some degree. Effective reasoning from premises, whether scientific or everyday, essentially depends on judgements of probability, and sometimes of utility, even when it partly consists of logically valid inferences. Conditional premises are prominent in both scientific and everyday inference, and consequently psychological theories of reasoning should include an acceptable account of the subjective probability of conditionals. More generally, the psychological study of conditional reasoning (as well as of other types of reasoning) should be fully integrated with research on probability, utility, and decision making (Evans & Over, 1996; Evans & Over, 2004). Psychological theories of the ordinary conditional will be severely limited until this fact is fully appreciated.