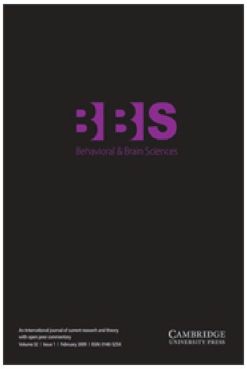
**Behavioral and Brain Sciences**
http://journals.cambridge.org/BBS



# Précis of *Bayesian Rationality: The Probabilistic Approach to Human Reasoning*

Mike Oaksford and Nick Chater

Behavioral and Brain Sciences / Volume 32 / Issue 01 / February 2009, pp 69 - 84
DOI: 10.1017/S0140525X09000284, Published online: 12 February 2009

**Link to this article:** http://journals.cambridge.org/abstract_S0140525X09000284

**How to cite this article:**
Mike Oaksford and Nick Chater (2009). Précis of *Bayesian Rationality: The Probabilistic Approach to Human Reasoning*. Behavioral and Brain Sciences,32, pp 69-84 doi:10.1017/S0140525X09000284

BEHAVIORAL AND BRAIN SCIENCES (2009) **32**, 69–120
*Printed in the United States of America*
doi:10.1017/S0140525X09000284

# Précis of *Bayesian Rationality: The Probabilistic Approach to Human Reasoning*

**Mike Oaksford**
*School of Psychology, Birkbeck College London, London, WC1E 7HX, United Kingdom*
**m.oaksford@bbk.ac.uk**
**www.bbk.ac.uk/psyc/staff/academic/moaksford**

**Nick Chater**
*Division of Psychology and Language Sciences, and ESRC Centre for Economic Learning and Social Evolution, University College London, London, WC1E 6BT, United Kingdom*
**n.chater@ucl.ac.uk**
**www.psychol.ucl.ac.uk/people/profiles/chater_nick.htm**

**Abstract:** According to Aristotle, humans are the rational animal. The borderline between rationality and irrationality is fundamental to many aspects of human life including the law, mental health, and language interpretation. But what is it to be rational? One answer, deeply embedded in the Western intellectual tradition since ancient Greece, is that rationality concerns reasoning according to the rules of logic – the formal theory that specifies the inferential connections that hold with certainty between propositions. Piaget viewed logical reasoning as defining the end-point of cognitive development; and contemporary psychology of reasoning has focussed on comparing human reasoning against logical standards.

*Bayesian Rationality* argues that rationality is defined instead by the ability to reason about *un*certainty. Although people are typically poor at numerical reasoning about probability, human thought is sensitive to subtle patterns of qualitative Bayesian, probabilistic reasoning. In Chapters 1–4 of *Bayesian Rationality* (Oaksford & Chater 2007), the case is made that cognition in general, and human everyday reasoning in particular, is best viewed as solving probabilistic, rather than logical, inference problems. In Chapters 5–7 the psychology of "deductive" reasoning is tackled head-on: It is argued that purportedly "logical" reasoning problems, revealing apparently irrational behaviour, are better understood from a probabilistic point of view. Data from conditional reasoning, Wason's selection task, and syllogistic inference are captured by recasting these problems probabilistically. The probabilistic approach makes a variety of novel predictions which have been experimentally confirmed. The book considers the implications of this work, and the wider "probabilistic turn" in cognitive science and artificial intelligence, for understanding human rationality.

**Keywords:** Bayes' theorem, conditional inference, logic, non-monotonic reasoning, probability, rational analysis, rationality, reasoning, selection task, syllogisms

*Bayesian Rationality* (Oaksford & Chater 2007, hereafter *BR*) aims to re-evaluate forty years of empirical research in the psychology of human reasoning, and cast human rationality in a new and more positive light. Rather than viewing people as flawed logicians, we focus instead on the spectacular success of human reasoning under uncertainty. From this perspective, everyday thought involves astonishingly rich and subtle probabilistic reasoning – but probabilistic reasoning which is primarily qualitative, rather than numerical. This viewpoint leads to a radical re-evaluation of the empirical data in the psychology of reasoning. Previously baffling logical "errors" in reasoning about even the simplest statements can be understood as arising naturally from patterns of qualitative probabilistic reasoning.

Why "Bayesian" rationality, rather than mere "probabilistic" rationality? The answer is that our approach draws crucially on a particular interpretation of probability, not merely on the mathematics of probability itself. Probability is often taught as capturing "objective" facts about something, for example, gambling devices such as dice or cards. It is sometimes presumed to be a fact, for example, that the probability of a fair coin producing three consecutive heads is 1/8. However, in the context of cognitive science, probability refers not to objective facts about gambling devices or anything else, but rather, it describes a reasoner's *degrees of belief*. Probability theory is then a calculus not for solving mathematical problems about objects in the world, but a calculus for rationally updating beliefs. This perspective is the *subjective*, or *Bayesian* view of probability. We thus argue that human rationality, and the coherence of human thought, is defined not by logic, but by probability.

The Bayesian perspective on human reasoning has radical implications. It suggests that the meaning of even the most elementary natural language sentences may have been fundamentally mischaracterized: many such statements may make probabilistic rather than logical

0140-525X/09 $40.00

claims. And the most elementary aspects of human reasoning may have been misunderstood – what appeared to be logically certain inferences may often instead be better understood as plausible, probabilistic reasoning. Shifting from a logical to a Bayesian perspective entirely changes our predictions concerning the *patterns* of reasoning that we should expect people to exhibit. And experimental work in the psychology of reasoning provides the data against which these predictions can be compared.

This Précis outlines the argument of *BR* chapter by chapter; the section numbering corresponds to the chapter numbering of the book, with occasional modifications to assist the flow of what is now a somewhat compressed argument. The first section of the book, Chapters 1–4, outlines the theoretical background of our shift from logical to Bayesian rationality as an account of everyday human reasoning, drawing on relevant areas of psychology, philosophy, and artificial intelligence. The second section of the book, Chapters 5–7, relates this approach to the key empirical data in the psychology of reasoning: conditional reasoning, Wason's selection task, and syllogistic reasoning. We argue that the patterns of results observed in the empirical data consistently favour a Bayesian analysis, even for purportedly paradigmatically "logical" reasoning problems. Chapter 8 reflects on the implications of this approach.

## 1. Logic and the Western conception of mind

Since the Greeks, the analysis of mind has been deeply entwined with logic. Indeed, the study of logical argument and the study of mind have often been viewed as overlapping substantially. One swift route to such a deep connection is to argue that minds are distinctively rational; and that rationality is partly, or perhaps even wholly, characterized by logic. That is, logical relations are viewed primarily as unbreakable, inferential relations between thoughts; and a coherent, intelligible agent must respect such relations. In particular, logic aims to specify inferential relations that hold *with absolute certainty*: logical inference is *truth preserving*, that is, if the premises are true, the conclusion must also be true.

MIKE OAKSFORD is Professor of Psychology and Head of the School of Psychology at Birkbeck College, London. He is the author of over one hundred scientific publications in psychology, philosophy, and cognitive science, and has written or edited six books. He currently serves on the Editorial Boards of the journals *Psychological Review*, *Memory and Cognition*, and *Thinking and Reasoning*. He has also served as Associate Editor for the *Quarterly Journal of Experimental Psychology* and on the Editorial Board of the *Journal of Experimental Psychology*: *Learning, Memory, & Cognition*. His research explores the nature of reasoning, argumentation, and the influence of emotion on cognition.

NICK CHATER is Professor of Cognitive and Decision Sciences at University College London. He is the author of over one hundred and fifty scientific publications in psychology, philosophy, linguistics, and cognitive science, and has written or edited eight books. He currently serves as Associate Editor for *Psychological Review*. He has also served as Associate Editor for *Cognitive Science* and on the Editorial Boards of *Psychological Review* and *Trends in Cognitive Sciences*. His research explores formal models of inference, choice, and language.

But which inferences *are* absolutely certain? Which can be relied upon to preserve truth reliably? We may feel confident, from our knowledge of science, that, for example, all women are mortal. We might generalize from the mortality of all other living things; or note that even the most long-lived creature will succumb in the heat-death of the universe of the far future. But such considerations, however convincing, do not give the certainty of logic – they depend on contingent facts, and such facts are not themselves certain. Aristotle answered these questions by providing the first logical system: the theory of the syllogism. Syllogisms involve two premises, such as, *All women are people; All people are mortal*. Aristotle argued that these premises imply with absolute certainty that *All women are mortal*.

Logical certainty is more than mere overwhelming confidence or conviction. A logical argument depends purely on its *structure*: thus, Aristotle noted, our logical argument put forth here is of the form *All A are B; All B are C;* therefore, *All A are C*. And this argument is valid whatever *A*, *B*, or *C* stand for; hence there is no appeal to contingent facts of any kind. Aristotle's spectacular discovery was, therefore, that patterns of reliable reasoning could be obtained merely by identifying the *structure* of that reasoning. Logic, then, aims to provide a theory that determines which argument structures are truth-preserving, and which are not. In a very real sense, in a logical inference, if you believe the premises, you already believe the conclusion – the meaning of the conclusion is, somehow, contained in the meaning of the premises. To deny the constraints of logic would thus be *incoherent*, rather than merely mistaken. Thus, logic can be viewed as providing crucial constraints on the thoughts that any rational agent can entertain (Davidson 1984; Quine 1953).

Aristotle's theory of the logical structure of the syllogism proceeded by enumeration: Aristotle identified 64 forms of the syllogism, along with a systematic, though intuitive, approach to deciding which of these syllogisms had a valid conclusion, and if so, what the nature of this conclusion is. For more than two thousand years, Aristotle's theory of the syllogism almost exhausted logical theory—and indeed, Kant considered all logical questions to have been decisively resolved by Aristotle's account, stating: "It is remarkable also, that to the present day, it has not been able to make one step in advance, so that, to all appearance, it [i.e., logic] may be considered as completed and perfect" (Kant 1787/1961, p. 501).

As we have suggested, although Aristotle's logic is defined over patterns of verbally stated arguments (converted from everyday language into the appropriate formal structure), it is nonetheless tempting to view the primary subject matter of logic as thought itself. If the mind is viewed as constituted by rational thought, and logic captures patterns of rational thought, it seems natural to view logic as a central part of psychology. Such was Boole's perspective, in going beyond Aristotle's enumeration of patterns of logical argument. Boole aimed to describe the "The Laws of Thought" (Boole

1854/1958); and, in doing so, provided, for the first time, explicit mathematical rules for logical reasoning. This allowed him to develop a *calculus* for logical reasoning, albeit limited in scope. Boole also opened up the possibility that logical reasoning could be carried out mechanistically, purely by the manipulation of logical symbols. This insight provided a partial foundation for modern computation and, by extension, cognitive science.

The view that rational thought is governed by logic, which we term the *logicist* conception of the mind (Oaksford & Chater 1991), was adopted wholeheartedly by early cognitive theorists such as Piaget (e.g., Inhelder & Piaget 1955). Piaget viewed the pinnacle of cognitive development as attaining the "formal operational" stage, at which point the mind is capable of reasoning according to a particular formal system of logic: propositional logic. He viewed the process of cognitive development as a series of stages of enrichment of the logical apparatus of the child, enabling increasingly abstract reasoning, which is less tied to the specific sensory-motor environment. Similarly, the early foundations of cognitive science and artificial intelligence involved attempting to realize logical systems practically, by building computer programs that can explicitly derive logical proofs. Tasks such as mathematical reasoning and problem solving were then viewed as exercises in logic, as in Newell and Simon's *Logic Theorist* and *General Problem Solver* (see Newell & Simon 1972; Newell et al. 1958). Moreover, Chomsky's (1957; 1965) revolutionary work in linguistics showed how the syntactic structure of language could be organized in a deductive logical system, from which all and only the grammatical sentences of the language could be generated. And in the psychology of adult reasoning, this logical conception of mind was again used as the foundation for explaining human thought.

Simultaneous with the construction of the logicist program in cognition, there were some discordant and puzzling observations. Specifically, researchers such as Wason, who attempted to verify the Piagetian view of the adult mind as a perfect logic engine, found that people appeared surprisingly and systematically illogical in some experiments. Given the dissonance between these results and the emerging logicist paradigm in cognitive science, these results were largely set aside by mainstream cognitive theorists, perhaps to be returned to once the logicist approach had reached a more developed state. But the general form that an account of apparent irrationality might take was that all illogical performance resulted from misunderstandings and from the faulty way in which the mind might sometimes apply logical rules. For example, Henle stated: "I have never found errors which could unambiguously be attributed to faulty reasoning" (Henle 1978, p. xviii). But the central notion that thought is based on logic was to be retained.

This fundamental commitment to logic as a foundation for thought is embodied in contemporary reasoning theory in two of the main theoretical accounts of human reasoning. The *mental logic* view (Braine 1978; Rips 1983; 1994) assumes that human reasoning involves logical calculation over symbolic representations, using systems of proof which are very similar to those developed by Hilbert in mathematics, and used in computer programs for theorem-proving in artificial intelligence and computer science. By contrast, the *mental models* view (Johnson-Laird 1983; Johnson-Laird & Byrne 1991) takes its starting point as the denial of the assumption that reasoning involves formal operations over logical formulae, and instead assumes that people reason over concrete representations of situations or "models" in which the formulae are true. This provides a different method of proof (see Oaksford & Chater 1991; 1998a, for discussion), but one that can achieve logical performance by an indirect route.

Although mental logic and mental models both give logic a central role in human reasoning, they explain apparent irrationalities in different ways. For example, mental logics may explain errors in terms of the accessibility of different rules, whereas mental models explain errors in terms of limitations in how mental models are constructed and checked, and how many models must be considered.

These logicist reactions to data appearing to show human irrationality seem entirely reasonable. Every new theory in science could be immediately refuted if the mere existence of data apparently inconsistent with the theory were assumed to falsify it decisively (Kuhn 1962; Lakatos 1970). The crucial question is: Can a more plausible explanation of these puzzling aspects of human reasoning be provided? We argue that the Bayesian approach provides precisely such an alternative.

## 2. Rationality and rational analysis

*BR* aims to promote a Bayesian, rather than a logical, perspective on human reasoning. But to make sense of any debate between the logical and Bayesian standpoints, we need first to clarify how we interpret the relationship between a normative mathematical theory of reasoning (whether logic or probability), and empirical findings about human reasoning. In particular, how do we deal with any systematic clashes between the theory's dictates concerning how people *ought* to reason, and empirical observations of how they actually *do* reason?

Various viewpoints have been explored. One option is to take observed human intuitions as basic, and hence as the arbiter of what counts as a good formal theory of reasoning (e.g., Cohen 1981). Another is to take the mathematical theory as basic, and view it as providing a standpoint from which to evaluate the quality of observed reasoning performance (e.g., Rips 1994). Still a further possibility is that clashes between the formal theory and actual reasoning may arise because human thought itself is divided between two systems of reasoning (e.g., Evans & Over 1996a).

Here, we take a different line: We view normative theory as a component of the project of providing a "rational analysis" which aims to capture empirical data concerning thought and behavior. Rational analysis (e.g., Anderson 1990; 1991a; Oaksford & Chater 1998b) has six steps:

1. Specify precisely the goals of the cognitive system.
2. Develop a formal model of the environment to which the system is adapted.
3. Make minimal assumptions about computational limitations.
4. Derive the optimal behaviour function given steps 1–3.

(This requires formal analysis using rational norms, such as probability theory, logic, or decision theory.)

5. Examine the empirical evidence to see whether the predictions of the behaviour function are confirmed.

6. Repeat, iteratively refining the theory.

So the idea of rational analysis is to understand the problem that the cognitive system faces, and the environmental and processing constraints under which it operates. Behavioral predictions are derived from the assumption that the cognitive system is solving this problem, optimally (or, more plausibly, approximately), under these constraints. The core objective of rational analysis, then, is to understand the structure of the problem *from the point of view of the cognitive system*, that is, to understand what problem the brain is attempting to solve.

In the psychology of reasoning, this point is particularly crucial. We shall see that even when the experimenter intends to confront a participant with a logical reasoning puzzle, the participant may interpret the problem in probabilistic terms. If so, the patterns of reasoning observed may be well described in a Bayesian framework, but will appear to be capriciously errorful from a logical point of view. In Chapters 5–7 of *BR*, and summarized further on here, we argue that the core data in the psychology of reasoning, which has focussed on putatively "logical" reasoning tasks, can be dramatically clarified by adopting a Bayesian rational analysis.

It might appear that Step 2, concerning the environment, could not be relevant to rational analysis of the reasoning, as opposed to, say, perception. Mathematical theories of reasoning are supposed to apply across topics, and hence should surely be independent of environmental structure. We shall see further on that the reverse is the case. Very general features of the environment, such as the fact that almost all natural language categories occur with a low probability and that arbitrarily chosen probabilistic constraints are often independent or nearly independent, turn out to have substantial implications for reasoning. Indeed, the project of providing a rational analysis of human reasoning gains its empirical purchase precisely by explaining how a "topic neutral" mathematical theory applies to a specific goal, given a particular set of environmental and computational constraints.

Two caveats are worth entering concerning Bayesian rational analysis. The first is that rational analysis is not intended to be a theory of psychological *processes*. That is, it does not specify the representations or algorithms that are used to carry out this solution. Indeed, as Anderson (1990; 1991a) points out, these representations and algorithms might take many different forms – but certain general aspects of their behavior will follow irrespective of such specifics; they will arise purely because the cognitive system is well-adapted to solving this particular problem. Hence, the correct analysis of the rational structure of the cognitive problem at hand can have considerable explanatory power.

The second caveat is that the aim of understanding the structure of human reasoning, whether from a logical or a Bayesian perspective, should be carefully distinguished from the goal of measuring people's performance on logical or probabilistic problems (Evans et al. 1993; Kahneman et al. 1982). Indeed, both logic and probability provide a fresh and bracing challenge to each generation of students; performance on logical and probability problems results from explicit instruction and study, rather than emerging from capacities that are immanent within the human mind. But this observation need not impact our evaluation of logic or probability as explanations for patterns of everyday thought. Even if the mind is a probabilistic or logical calculating engine, it may not be possible to engage that engine with verbally, symbolically, or numerically stated probabilistic or logical puzzles, which it is presumably not adapted to handle. This point is no deeper than the observation that, although the early visual processes in the retina may compute elaborate convolutions and decorrelations of the image, this does not mean that people can thereby readily apply this machinery to solve mathematics problems concerning convolution or decorrelation. Thus, empirical evidence from the psychology of reasoning is not used, in *BR*, to evaluate people's logical or probabilistic reasoning competence. Rather, this evidence is used to explore the patterns of reasoning that people find natural; and to relate such patterns to how people reason outside the experimental laboratory.

From the standpoint of rational analysis, the question of whether logic or probability is the appropriate framework for understanding reasoning is an empirical question: Which rational analysis of human reasoning best captures the data? In Chapters 5–7 of *BR*, we argue, case-by-case, that a Bayesian rational analysis provides a better account of core reasoning data than its logicist rivals. First, though, we consider why. In Chapter 3, we argue that real-world, informal, everyday, reasoning is almost never deductive, that is, such reasoning is almost always logically *in*valid. In Chapter 4, we consider what has driven the broader "probabilistic turn" in cognitive science and related fields, of which the Bayesian analysis of human reasoning is a part.

## 3. Reasoning in the real world: How much deduction is there?

Logic provides a calculus for certain reasoning – for finding conclusions which follow, of necessity, from the premises given. But in everyday life, people are routinely forced to work with scraps of knowledge, each of which may be only partially believed. Everyday reasoning seems to be more a matter of tentative conjecture, rather than of water-tight argument.

Notice, in particular, that a successful logical argument cannot be overturned by any additional information that might be added to the premises. Thus, if we know that *All people are mortal*, and *All women are people*, then we can infer, with complete certainty, that *All women are mortal*. Of course, on learning new information we may come to doubt the premises – but we cannot come to doubt that the conclusion follows from the premises. This property of classical logic is known as *monotonicity*, meaning that adding premises can never overturn existing conclusions.

In reasoning about the everyday world, by contrast, *non*-monotonicity is the norm: almost any conclusion can be overturned, if additional information is acquired. Thus, consider the everyday inference from *It's raining* and *I am about to go outside* to *I will get wet*. This inference is uncertain – indefinitely many additional premises (*the rain is about to stop*; *I will take an umbrella*; *there is a covered walkway*) can overturn the conclusion, even if the premises are correct. The nonmonotonicity of everyday inference is problematic for the application of logical methods to modelling thought. Nonmonotonic inferences are not logically valid and hence fall outside the scope of standard logical methods.

The nonmonotonicity of everyday reasoning often strikes in subtle and unexpected ways. Most notorious is the "frame problem" (McCarthy & Hayes 1969), which arose in early applications of logical methods in artificial intelligence. Suppose an agent, with knowledge base *K*, makes an action *A* (e.g., it turns a cup upside down). Which other information in *K* needs to be updated to take account of this action? Intuitively, almost all other knowledge should be unchanged (e.g., that the street is empty, or that the burglar alarm is off). But, from a logical point of view, the "interia" of such everyday knowledge does not follow, because it is logically possible that *A* may have all manner of consequences. For example, given the additional information that the cup is valuable and placed in an alarmed glass case, then turning it over *may* trigger the burglar alarm and may fill the street with curious bystanders. The difficulties generated by the frame problem have had a paralyzing effect on logical approaches to planning, action, and knowledge representation in artificial intelligence.

Analogous problems arise more generally (Fodor 1983; Pylyshyn 1987). Given a database with knowledge *K*, adding a new fact *F* (not necessarily concerning an action) can typically overthrow many of the previous consequences of *K*, in highly idiosyncratic ways. It proves to be impossible to delimit the inferential consequences of a new fact in advance. Learning a new fact about football can, for example, readily modify my beliefs about philosophy. For example, suppose one has been told footballing facts and philosophical facts by the same person, of uncertain trustworthiness. Then learning that a footballing fact is incorrect may cause one to doubt a putative philosophical fact. Thus, nonmonotonicty may apply to arbitrarily remote pieces of knowledge. And note, of course, that an inference that can be overturned by additional premises cannot be logically valid – because standard logic is monotonic by definition.

Inferences which are nonmonotonic, and hence cannot be captured by conventional logic, are described in different literatures using a variety of terms: *non-demonstrative* inference, *informal* argument, and *common-sense* reasoning. For the purposes of our arguments, these terms are interchangeable. But their import, across psychology, artificial intelligence, and philosophy, is the same: nonmonotonic arguments are outside the scope of deductive logic.

This conclusion has alarming implications for the hypothesis that thought is primarily based on logical inference. This is because the scope of monotonic inference is vanishingly small – indeed, it scarcely applies anywhere outside mathematics. As we shall see in Chapters 5–7, this point applies even to verbally stated inferences that are typically viewed as instances of deduction. For example, consider the argument from *if you put 50p in the coke machine, you will get a coke* and *I've put 50p in the coke machine*, to *I'll get a coke*. This argument appears to be an instance of a canonical monotonic logical inference: *modus ponens*.

Yet in the context of commonsense reasoning, this argument does not appear to be monotonic at all. There are innumerable possible additional factors that may block this inference (power failure, the machine is empty, the coin or the can become stuck, etc.). Thus, you can put the money in, and no can of coke may emerge. Attempting to maintain a logical analysis of this argument, these cases could be interpreted as indicating that, from a logical point of view, the conditional rule is simply false – precisely because it succumbs to counterexamples (Politzer & Braine 1991). But this is an excessively rigorous standpoint, from which almost all everyday conditionals will be discarded as false. But how could a plethora of false conditional statements provide a useful basis for thought and action. From a logical point of view, after all, we can only make inferences from *true* premises; a logical argument tells us nothing, if one or more of its premises are false.

In sum, there appears to be a fundamental mismatch between the nonmonotonic, uncertain character of everyday reasoning, and the monotonicity of logic; and this mismatch diagnoses the fundamental problem with logic-based theories of reasoning and logicist cognitive science more broadly. In *BR*, we draw a parallel with a similar situation in the philosophy of science, where there has been a gradual retreat from early positive claims that theoretical claims somehow logically follow from observable premises, to Popper's (1935/1959) limitation of logical deduction, to the process of drawing predictions from theories, to the abandonment of even this position, in the light of the nonmonotonicity of predictive inference (there are always additional forces, or factors, that can undo any prediction; Putnam 1974). Indeed, modern philosophy of science has taken a resolutely Bayesian turn (e.g., Bovens & Hartmann 2003; Earman 1992; Horwich 1982; Howson & Urbach 1993). *BR* also considers attempts to deal with the apparent mismatch by attempting to deal with uncertainty by developing nonmonotonic logics (e.g., Reiter 1980), a project that rapidly became mired in difficulties (see, e.g., Oaksford & Chater 1991). Perhaps it is time to shift our attention to a calculus that deals directly with uncertainty: probability theory.

## 4. The probabilistic turn

We have seen how uncertainty, or nonmonotonicity, is a ubiquitous feature of everyday reasoning. Our beliefs, whether arising from perception, commonsense thought, or scientific analysis, are tentative and provisional. Our expectation that the car will start, that the test tube will turn blue, or that one arrow is longer than another, are continually being confounded by faulty batteries, impure chemicals, or visual illusions.

Interestingly, Aristotle, the founder of logic, was keenly aware of the limits of the logical enterprise. After all, he was interested not only in mathematical and philosophical reasoning, but also with the scientific description and analysis of the everyday world, and with practical affairs and human action. An often quoted passage from the *Nicomachean Ethics* (1094b, Aristotle 1980, p. 3) notes that "it is the mark of an educated man to look for precision in each class of things just so far as the nature of the subject admits: it is evidently equally foolish to accept probable reasoning from a mathematician and to demand from a rhetorician demonstrative reasoning."

Indeed, one key motivation for developing a theory of probability was closely connected with Aristotle's rhetorician. The goal in rhetoric, in its traditional sense, is to provide reasoned arguments for why people should hold certain opinions concerning matters about which certainty is impossible. Thus, in deciding court cases by jury, a different piece of evidence (e.g., eye-witness testimony, forensic evidence, evidence of previous good character) must somehow be combined to yield a degree of belief concerning the likely guilt of the defendant. Here, probability is interpreted subjectively, in terms of a person's strength of opinion, rather than concerning an assumption about the external world. Indeed, the very word "probability" initially referred to the degree to which a statement was supported by the evidence at hand (Gigerenzer et al. 1989). Jakob Bernoulli explicitly endorsed this interpretation when he entitled his definitive book *Ars Conjectandi*, or the *Art of Conjecture* (Bernoulli 1713). This *subjectivist*, or *Bayesian*, conception of probability ran through the eighteenth and into the nineteenth centuries (Daston 1988), frequently without clear distinctions being drawn between probability theory as a model of actual thought (or more usually, the thought of "rational", rather than common, people [Hacking 1975; 1990]) or as a set of normative canons prescribing how uncertain reasoning should be conducted. As with logic, early probability theory itself was viewed as a model of mind.

Over the latter part of the twentieth century, the Bayesian perspective has been increasingly influential across the cognitive sciences and related disciplines. Chapter 4 of *BR* surveys some of these developments. For example, if everyday inference is inherently probabilistic, this raises the possibility that natural language statements should be interpreted as making probabilistic, rather than logical, claims. So, for example, Adams (e.g., 1975; 1998) directly imports probability into logical theory, arguing that the conditional *If A then B* should, roughly, be interpreted as saying that *B* is probable, if *A* is true. Later we shall see how this, and other probabilistic analyses of familiar "logical" structures (e.g., concerning the quantifiers *All*, *Some*, etc.), cast new light on the empirical reasoning data.

It is, we suggest, significant that three key domains in which uncertain inference is ubiquitous, philosophy of science, artificial intelligence, and cognitive psychology, have all embraced the Bayesian approach. *BR* reviews some of the key developments: the application of Bayes' theorem to hypothesis confirmation (e.g., Earman 1992); the development of graphical models for knowledge representation and causal reasoning (Pearl 1988; 2000); and the application of Bayesian methods in rational models of cognitive phenomena (Chater & Oaksford 2008b; Oaksford & Chater 1998b) in areas as diverse as categorization (Anderson 1991b; Anderson & Matessa 1998), memory (Anderson & Milson 1989; Anderson & Schooler 1991), conditioning (Courville et al. 2006; Kakade & Dayan 2002), causal learning (Griffiths & Tenenbaum 2005; Novick & Cheng 2004), natural language processing (Chater et al. 1998; Chater & Manning 2006), and vision (Knill & Richards 1996; Yuille & Kersten 2006).

There has, in short, been a "probabilistic turn" across a broad range of domains – a move away from the attempt to apply logical methods to uncertain reasoning, and towards dealing with uncertainty by the application of probability theory. In Chapters 5–7, we illustrate how the switch from logical to Bayesian rationality leads to a radical re-evaluation of the psychology of human reasoning – so radical, in fact, that even apparently paradigmatic "logical" reasoning tasks turn out to be better understood from a probabilistic point of view.

## 5. Does the exception prove the rule? How people reason with conditionals

In Chapters 5–7 of *BR*, we describe Bayesian probabilistic models for the three core areas of human reasoning research: conditional inference (Ch. 5), data selection (Ch. 6), and quantified syllogistic inference (Ch. 7). The key idea behind all these models is to use conditional probability, $P(q|p)$, to account for the meaning of conditional statements, *if p then q* (e.g., *if you turn the key then the car starts*). The aim is to show that what appear to be "errors and biases" from a logicist standpoint are often entirely rational from a Bayesian point of view. In this Précis, for each area of reasoning, we introduce the task, the standard findings, and existing logicist accounts. We then introduce a Bayesian rational analysis for each problem, show how it accounts for the core data, and provide a snapshot of some of the further data that we discuss in *BR*. Finally, for each area of reasoning, we summarise and describe one or two outstanding problems confronting the Bayesian approach.

Chapter 5 of *BR* begins with conditional inference, that is, inferences directly involving the conditional *if p then q*. In the conditional, $p$ is called the "antecedent" and $q$ is called the "consequent." Four inference patterns have been extensively studied experimentally (see Fig. 1). Each inference consists of the *conditional* premise and one of four possible *categorical* premises, which relate either to the antecedent or consequent of the conditional, or their negations ($p$, $\neg p$, $q$, $\neg q$ where "$\neg$" = not). For example, the inference *Modus Ponens* (MP) combines the conditional premise *if p then q* with the categorical premise $p$; and yields the conclusion $q$.

According to standard logic, two of these inferences are logically valid (MP and *Modus Tollens* [MT], see Fig. 1), and two are fallacies (Denying the Antecedent [DA] and

$$\text{(MP)}\quad \frac{p \Rightarrow q,\, p}{\therefore q} \qquad \text{(MT)}\quad \frac{p \Rightarrow q,\, -q}{\therefore \neg p}$$

$$\text{(DA)}\quad \frac{p \Rightarrow q,\, -p}{\therefore \neg q} \qquad \text{(AC)}\quad \frac{p \Rightarrow q,\, q}{\therefore p}$$

Figure 1. The four inference patterns investigated in the psychology of conditional inference: *Modus Ponens* (MP) and *Modus Tollens* (MT) are logically valid. *Denying the Antecedent* (DA) and *Affirming the Consequent* (AC) are logically fallacious. These inference schemata read that if the premises above the line are true then so must be the conclusion below the line. "$p \Rightarrow q$" signifies the "material conditional" of standard logic, which is true unless $p$ is true and $q$ is false.

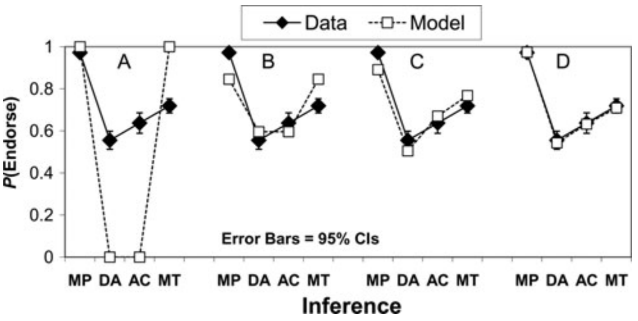


Figure 2. The fits to the experimental data (Schroyens & Schaeken 2003) of standard logic (Panel A), standard logic plus the biconditional interpretation and error (Panel B), the original probabilistic model (Panel C), and the probabilistic model adjusted for rigidity violations (Panel D).

Affirming the Consequent [AC], see Fig. 1). Figure 2 (Panel A) shows the results of a meta-analysis of experiments where people are asked whether they endorse each of these four inferences (Schroyens & Schaeken 2003). Panel A also shows the predictions of the standard logical model, revealing a large divergence.

From a logicist standpoint, this divergence may be reduced by assuming that some people interpret the conditional as a *biconditional*, that is, that *if p then q* also means that *if q then p*. This move from conditional to biconditional is, of course, logically invalid. For example, *if a bird is a swan, then it is white* clearly does not entail that *if a bird is white, then it is a swan*. Nonetheless, the biconditional interpretation may be pragmatically reasonable, in some cases. For example, promises such as *if you mow the lawn, I will pay you £5* do seem to allow this pragmatic inference; it seems reasonable to assume that I will *only* pay you £5 if you mow the lawn (or, at least, that I will not pay you £5 if you refuse). By assuming that people make this pragmatic inference for the stimuli used in experimental tasks and by making some allowance for random error, the best fit that standard logic can provide is shown in Figure 2 (Panel B) (see Oaksford & Chater 2003a).

To further close the gap with the data in Figure 2, logicist theories of conditional inference typically assume not only that people adopt the pragmatic inference to the biconditional interpretation, but also that they fail to represent logic completely in their cognitive system. For example, mental logic (e.g., Rips 1994) is typically assumed to involve an MP inference rule, but no MT rule. This means that MT inferences must be drawn in a more complex way, often leading to error. Similarly, according to mental models theory, people do not initially represent the full meaning of the conditional (Johnson-Laird & Byrne 2002). To draw an MT inference, they must "flesh out" their representations to fully capture the meaning of the conditional. In both cases, logically unwarranted pragmatic inferences and assumptions about cognitive limitations are invoked to explain the data.

In contrast, the Bayesian approach only invokes probability theory. There are four key ideas behind the probabilistic account of conditional inference. First, the probability of a conditional is the conditional probability, that is, $P(\textit{if p then q}) = P(q|p)$. In the normative literature, this identification is simply called "The Equation" (Adams 1998; Bennett 2003; Edgington 1995). In the psychological literature, the Equation has been confirmed experimentally by Evans et al. (2003) and by Oberauer and Wilhelm (2003). Second, as discussed earlier, probabilities are interpreted "subjectively," that is, as degrees of belief. It is this interpretation of probability that allows us to provide a probabilistic theory of inference as belief updating. Third, conditional probabilities are determined by a psychological process called the "Ramsey Test" (Bennett 2003; Ramsey 1931/1990b). For example, suppose you want to evaluate your conditional degree of belief that *if it is sunny in Wimbledon, then John plays tennis*. By the Ramsey test, you make the hypothetical supposition that *it is sunny in Wimbledon* and revise your other beliefs so that they fit with this supposition. You then "read off" your hypothetical degree of belief that *John plays tennis* from these revised beliefs.

The final idea concerns standard conditional inference: how we reason when the categorical premise is not merely *supposed*, but is actually believed or known to be true. This process is known as *conditionalization*. Consider an MP inference, for example, *If it is sunny in Wimbledon, then John plays tennis*, and *It is sunny in Wimbledon*, therefore, *John plays tennis*. Conditionalization applies when we know (instead of merely supposing) that *it is sunny in Wimbledon*; or when a high degree of belief can be assigned to this event (e.g., because we know that it is sunny in nearby Bloomsbury). By conditionalization, our new degree of belief that *John plays tennis* should be equal to our prior degree of belief that *if it is sunny in Wimbledon, then John plays tennis* (here "prior" means before learning that *it is sunny in Wimbledon*). More formally, by the Equation, we know that $P_0$(*if it is sunny in Wimbledon, then John plays tennis*) equals $P_0$(*John plays tennis*|*it is sunny in Wimbledon*), where "$P_0(x)$" = prior probability of $x$. When we learn *it is sunny in Wimbledon*, then $P_1$(*it is sunny in Wimbledon*) = 1, where "$P_1(x)$" = posterior probability of $x$. Conditionalizing on this knowledge tells us that our new degree of belief in *John plays tenniş* $P_1$(*John plays tennis*), should be equal to $P_0$(*John plays tennis*|*it is sunny in Wimbledon*). That is, $P_1(q) = P_0(q|p)$, where $p$ = *it is sunny in Wimbledon*, and $q$ = *John plays tennis*.[1] So from a probabilistic perspective, MP provides a way of updating our degrees of belief in the consequent, $q$, on learning that the antecedent, $p$, is true.

So, quantitatively, if you believe that $P_0$(*John plays tennis*|*it is sunny in Wimbledon*) = 0.9, then given you discover that *it is sunny in Wimbledon* ($P_1$(*it is sunny in Wimbledon*) = 1) your new degree belief that *John plays tennis* should be 0.9, that is, $P_1$(*John plays tennis*) = 0.9. This contrasts with the logical approach in which believing the conditional premise entails with *certainty* that the conclusion is true, so that $P_0$(*John plays tennis*|*it is sunny in Wimbledon*) = 1. This is surely too strong a claim.

The extension to the other conditional inferences is not direct, however. Take an example of AC, *if it is sunny in Wimbledon, John plays tennis* and *John plays tennis*, therefore, *it is sunny in Wimbledon*. In this case, one knows or strongly believes that *John play tennis* (perhaps we were told by a very reliable source), so $P_1(q) = 1$. But to use Bayesian conditionalization to infer one's new degree of belief that *it is sunny in Wimbledon*, $P_1(p)$, one needs to know one's conditional degree of belief that *it is sunny in Wimbledon* given *John plays tennis*, that is, $P_0(p|q)$. However, the conditional premise of AC, like that of MP, is about $P_0(q|p)$ *not* about $P_0(p|q)$ (Sober 2002). The solution proposed by Oaksford et al. (2000; see also Wagner 2004) is that that people also know the prior marginal probabilities (at least approximately). That is, they know something about the probability of a sunny day in Wimbledon, $P_0(p)$, and the probability that John plays tennis, $P_0(q)$, *before* learning that it is in fact a sunny day in Wimbledon. With this additional information, $P_0(p|q)$ can be calculated from the converse conditional probability, $P_0(q|p)$, using Bayes' Theorem.[2] The same approach also works for DA and MT where the relevant probabilities are $P_0(\neg q|\neg p)$ and $P_0(\neg p|\neg q)$, respectively. The fact that the conditional premises of AC, DA, and MT do not determine the appropriate conditional probability marks an important asymmetry with MP. For these inferences, further knowledge is required to infer the relevant conditional degrees of belief.

The rest of Chapter 5 in *BR* shows how the errors and biases observed in conditional inference are a consequence of this rational probabilistic model. The first set of "biases" relates directly to the data in Figure 2. These are what, in *BR*, we call "the inferential asymmetries." That is, MP is drawn more than MT and AC is drawn more than DA (MT is also drawn more than AC). Figure 2, Panel C shows how well a probabilisitic account can explain these asymmetries. Here we have calculated the values of $P_0(q|p)$, $P_0(p)$, and $P_0(q)$ that best fit the data, that is, they minimize the sum of squared error between the data and the models predictions ("model" in Fig. 2). As Panel C shows, a probabilistic account can capture the asymmetries without pragmatic inference or appeal to process limitations. Panel C also shows, however, that this probabilistic model (Oaksford et al. 2000) does not capture the magnitudes of the inferential asymmetries (Evans & Over 2004; Schroyens & Schaeken 2003). It underestimates the MP – MT asymmetry and overestimates the DA – AC asymmetry.

In *BR*, we argue that this is because learning that the categorical premise is true can have two inferential roles. The first inferential role is in conditionalization, as we have described. The second inferential role is based on the pragmatic inference that *being told that the categorical premise is true* often suggests that there is a counterexample to the conditional premise. For example, consider the MT inference on the rule: *If I turn the key, the car starts*. If you were told that *the car did not start*, it seems unlikely that you would immediately infer that the *key was not turned*. Telling someone that *the car did not start* seems to presuppose that an attempt has been made to start it, presumably by turning the key. Consequently, the categorical premise here seems to suggest a counterexample to the conditional itself, that is, a case where the key was turned but the car did not start. Hence, one's degree of belief in the conditional should be reduced on *being told* that the car did not start. Notice, here, the contrast between being told that the car did not start (and drawing appropriate pragmatic inferences), and merely *observing* a car that has not started (e.g., a car parked in the driveway). In this latter situation, it is entirely natural to use the conditional rule to infer that the key has not been turned.

Where the second, pragmatic, inferential role of the categorical premise is operative, this violates what is called the *rigidity condition* on conditionalization, $P_0(q|p) = P_1(q|p)$ (Jeffrey 1983). That is, learning the categorical premise alters one's degree of belief in the conditional premise. In *BR*, we argue that taking account of such rigidity violations helps capture the probability of the conditional; and that, for MT this modified probability is then used in conditionalization. Furthermore, we argue that DA and AC also suggest violations of the rigidity condition, concerning the case where the car starts without turning the key. These violations lead to reductions in our degree of belief that the cars starts, given that the key is turned ($P_0(q|p)$). Using this lower estimate to calculate the relevant probabilities for DA, AC, and MT can rationally explain the relative magnitudes of the MP – MT and DA – AC asymmetries (see Fig. 2, Panel D).

We now turn to one of the other biases of conditional inference that we explain in Chapter 5 of *BR*: *negative conclusion* bias. This bias arises when negations are used in conditional statements, for example, *If a bird is a swan, then it is not red*. In Evans' (1972) *Negations Paradigm*, four such rules are used: *If p then q*; *if p then not-q*; *if not-p then q*; *if not-p then not-q*. The most robust finding is that people endorse DA, AC, and MT more when the conclusion contains a negation (see Fig. 3). So, for example, DA on *if p then q* (see Panel A in Fig. 3) yields a negated conclusion, *not-q*. Whereas, DA on *if p then not-q* (see Panel B in Fig. 3) yields an affirmative conclusion, *q* (because *not-not-q* = *q*). In Figure 3, it is clear that the frequency with which DA is endorsed for *if p then q* is much higher than for *if p then not-q*.

To explain negative conclusion bias, we appeal to the idea that most categories apply only to a minority of objects (Oaksford & Stenning 1992). Hence, the probability of an object being, say, red is lower than the probability of it not being red, that is, $P_0(Red) < P_0(\neg Red)$. Consequently, the marginal probabilities ($P_0(p)$ and $P_0(q)$) will take on higher values when $p$ or $q$ are negated. Higher values of the prior probabilities of the conclusion imply higher values of the relevant conditional probabilities for DA, AC, and MT, that is, to higher values of the posterior probability of the conclusion. So, for example, for our rule *if a bird is a swan, then it is white*, the prior probability of the conclusion of the DA inference ($P_0(\neg White)$) is high. This means that the conditional probability ($P_0(\neg White|\neg$

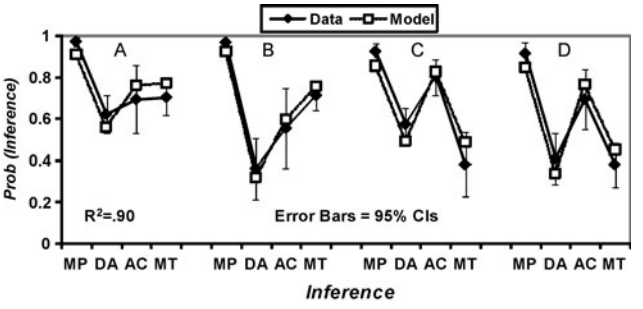


Figure 3. The results of Oaksford et al.'s (2000) meta-analysis of the negations paradigm conditional inference task for *if p then q* (Panel A), *if p then* $\neg q$ (Panel B), *if* $\neg p$ *then q* (Panel C), and *if* $\neg p$ *then* $\neg q$ (Panel D), showing the fit of the original probabilistic model.

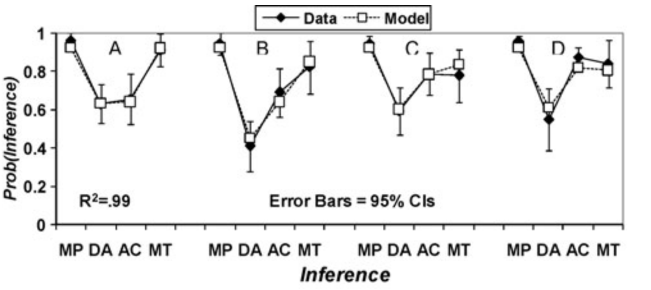


Figure 4. The results of Oaksford et al.'s (2000) Experiment 1 for the *low P*($p$), *low P*($q$) rule (Panel A), the *low P*($p$), *high P*($q$) rule (Panel B), the *high P*($p$), *low P*($q$) rule (Panel C), and the *high P*($p$), *high P*($q$) rule (Panel D), showing the fit of the original probabilistic model.

*Swan*)) is also high and, consequently, so is the probability of the conclusion ($P_1(\neg White)$). Therefore, an apparently irrational negative conclusion bias can be seen as a rational "high probability conclusion" effect. Oaksford et al. (2000) tested this explanation by manipulating directly $P_0(p)$ and $P_0(q)$ rather than using negations and showed results closely analogous to negative conclusion bias (see Fig. 4).

To conclude this section on conditional inference, we briefly review one of the most cited problems for a probabilistic account. Like any rational analysis, this account avoids theorising about the specific mental representations or algorithms involved in conditional reasoning. This may seem unsatisfactory. We suggest, by contrast, that it is premature to attempt an algorithmic analysis. The core of our approach interprets conditionals in terms of conditional probability, that is, using the Equation; and our current best understanding of conditional probability is given by the Ramsey test (Bennett 2003). But there is currently no possibility of building a full algorithmic model to carry through the Ramsey test, because this involves solving the notorious frame problem, discussed in Chapter 3. That is, it involves knowing how to update one's knowledge-base, in the light of a new piece of information – and this problem has defied 40 years of artificial intelligence research.

Nonetheless, an illustrative small-scale implementation of the Ramsey test is provided by the operation of a constraint satisfaction neural network (Oaksford 2004a). In such a model, performing a Ramsey test means clamping on or off the nodes or neurons corresponding to the categorical premise of a conditional inference. Network connectivity determines relevance relations and the weight matrix encodes prior knowledge. Under appropriate constraints, such a network can be interpreted as computing true posterior probabilities (McClelland 1998). A challenge for the future is to see whether such small-scale implementations can capture the full range of empirically observed effects in conditional inference.

## 6. Being economical with the evidence: Collecting data and testing hypotheses

Chapter 6 of *BR* presents a probabilistic model of Wason's selection task. In this task, people see four double-sided cards, with a number on one side and a letter on the other. They are asked which cards they should turn over, in order to test the hypothesis that *if there is an A* ($p$) *on one side of a card, then there is a 2* ($q$) *on the other*. The upturned faces of the four cards show an A ($p$), a K ($\neg p$), a 2 ($q$), and a 7 ($\neg q$) (see Fig. 5). The typical pattern of results is shown in Figure 6 (Panel A, Data).

As Popper (1935/1959) argued, logically one can never be *certain* that a scientific hypothesis is true in the light of observed evidence, as the very next piece of evidence one discovers could be a counterexample. So, just because all the swans you have observed up until now have been white, is no guarantee that the next one will not be black. Instead, Popper argues that the only logically sanctioned strategy for hypothesis testing is to seek *falsifying* cases. In testing a conditional rule *if* $p$ *then* $q$, this means seeking out $p$, $\neg q$ cases. This means that, in the standard selection task, one should select the A ($p$) and the 7 ($\neg q$) cards, because these are the only cards that could potentially falsify the hypothesis. Figure 6 (Panel A, Model) shows the logical prediction, and, as for conditional inference, the divergence from the data is large. Indeed, rather than seek *falsifying* evidence, participants seem to select the cases that *confirm* the conditional ($p$ and $q$). This is called "confirmation bias."

The range of theories of the selection task parallels the range of accounts of the conditional inference task described earlier. Mental logic theories (e.g., Rips 1994) assume that people attempt to perform conditional inferences, using the upturned face as the categorical premise to infer what is on the hidden face. Again, a biconditional interpretation is invoked: that *if A then 2* may pragmatically imply *if 2 then A*. If people perform an MP inference on both conditionals, this will yield a confirmatory response pattern. To infer that the 7 card should be turned, involves considering the hidden face. If people consider the possibility that the hidden face is not an A, then the complex inference pattern required for MT can be applied. A problem for mental logic is that, on this explanation, selection task performance

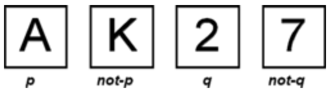


Figure 5. The four cards in Wason's Selection Task.

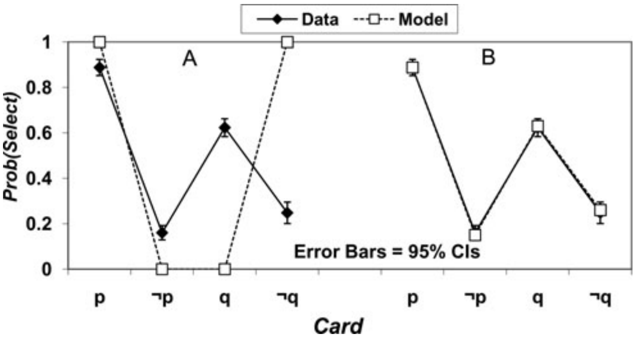


Figure 6. The fits to the experimental data on the Wason Selection Task of standard logic (Panel A) and of the optimal data selection model (Oaksford & Chater 1994) (Panel B).

should look like conditional inference task performance where selecting the 2 ($q$) card corresponds to AC and selecting the 7 ($\neg q$) card corresponds to MT. However, in conditional inference, MT is endorsed more than AC, but in the selection task this is reversed, that is, $q$ (AC) is selected more than $\neg q$ (MT).[3] For mental models, similar predictions are made if people initially represent the conditional as a biconditional and do not "flesh out" this representation.

The optimal data selection (ODS) model of this task (Oaksford & Chater 1994; 1996; 2003b) is a rational analysis derived from the normative literature on optimal experimental design in Bayesian statistics (Lindley 1956). The idea again relies on interpreting a conditional in terms of conditional probability. For example, the hypothesis, *if swan* ($p$) *then white* ($p$), is interpreted as making the claim that the probability of a bird being white given that it is a swan, $P(q|p)$, is high, certainly higher than the base rate of being a white bird, $P(q)$. This hypothesis is called the *dependence* hypothesis ($\mathrm{H_D}$). Bayesian hypothesis testing is comparative rather than exclusively concentrating on falsification. Specifically, in the ODS model, it is assumed that people compare $\mathrm{H_D}$ with an *independence* hypothesis ($\mathrm{H_I}$) in which the probability of a bird being white, given it is a swan, is the same as the base rate of a bird being white, that is, $P(q|p) = P(q)$. We assume that, initially, people are maximally uncertain about which hypothesis is true ($P(\mathrm{H_D}) = P(\mathrm{H_I}) = 0.5$) and that their goal in selecting cards is to reduce this uncertainty as much as possible while turning the fewest cards.

Take, for example, the card showing *swan* ($p$). This card could show *white* on the other side ($p$, $q$) or another color ($p$, $\neg q$). The probabilities of each outcome will be quite different according to the two hypotheses. For example, suppose that the probability of a bird being white, given that it is a swan is 0.9 ($P(q|p,\ \mathrm{H_D}) = 0.9$) in the dependence hypothesis; the marginal probability that a bird is swan is 0.2 ($P(p) = 0.2$); and the marginal probability that a bird is white is 0.3 ($P(q) = 0.3$). Then, according to the dependence hypothesis, the probability of finding *white* ($q$) on the other side of the card is 0.9, whereas according to the independence hypothesis it is 0.3 (as the antecedent and consequent are, in this model, independent, we need merely consult the relevant marginal probability). And, according to the dependence hypothesis, the probability of finding a color other than white ($\neg q$) on the other side of the card is 0.1, whereas according to the independence hypothesis it is 0.7. With this information it is now possible to calculate one's new degree of uncertainty about the dependence hypothesis after turning the *swan* card to find *white* on the other side ($P(\mathrm{H_D}|p,\ q)$). According to Bayes' theorem (see Note 2), this probability is 0.75. Hence, one's new degree of belief in the dependence model should be 0.75 and one's degree of belief in the independence model should be 0.25. Hence, the degree of uncertainty about which hypothesis is true has been reduced. More specifically, the ODS model is based on *information gain*, where information is measured in *bits* as in standard communication theory. Here, the initial uncertainty is 1 bit (because $P(\mathrm{H_D}) = P(\mathrm{H_I}) = 0.5$, equivalent to the uncertainty of a single fair coin flip) and in this example this is reduced to 0.81 bits (because now $P(\mathrm{H_D}) = 0.75$ and $P(\mathrm{H_I}) = 0.25$). This is an *information gain* of 0.19 bits.

In Wason's task, though, participants do not actually turn the cards, and hence they cannot know how much information they will gain by turning a card before doing so. Consequently, they must base their decision on *expected* information gain, taking both possible outcomes ($p$, $q$ and $p$, $\neg q$) into account. The ODS model assumes that people select each card in direct proportion to its expected information gain.

The ODS model also makes a key assumption about the task environment – that is, Step 2, in rational analysis: The properties that occur in the antecedents and consequents of hypotheses are almost always rare and so have a low base rate of occurrence. For example, most birds are not swans and most birds are not white. This assumption has received extensive independent verification (McKenzie et al. 2001; McKenzie & Mikkelsen 2000; 2007).

The ODS model predicts that the two cards that lead to the greatest expected information gain are the $p$ and the $q$ cards. Figure 6 (Panel B) shows the fit of the model to the standard data (Oaksford & Chater 2003b). The value of $P(q|p,\ \mathrm{H_D})$ was set to 0.9 and the best fitting values of $P(p)$ and $P(q)$ were 0.22 and 0.27 respectively, that is, very close to the values used in the earlier example. The ODS model suggests that performance on the selection task displays rational hypothesis testing behavior, rather than irrational confirmation bias. Taking rarity to an extreme provides a simple intuition here. Suppose we consider the (rather implausible) conditional: *If a person is bitten by a vampire bat* ($p$), *they will develop pointed teeth* ($q$). Clearly, we should check people who we know to have been bitten, to see if their teeth are pointed (i.e., turn the $p$ card); and, uncontroversially, we can learn little from people we know have not been bitten (i.e., do not turn the $\neg p$ card). If we see someone with pointed teeth, it is surely worth finding out whether they have been bitten – if they have, this raises our belief in the conditional, according to a Bayesian analysis (this is equivalent to turning the $q$ card). But it seems scarcely productive to investigate someone without pointed teeth (i.e., do not turn the $\neg q$ card) to see if they have been bitten. To be sure, it is *possible* that such a person might have been bitten, which would disconfirm our hypothesis, and lead to maximum information gain; but this has an almost infinitesimal probability. Almost certainly, we shall find that they have not been bitten, and learn nothing. Hence, with rarity, the expected informativeness of the $q$ card is higher than that of the $\neg q$ card, diverging sharply from the falsificationist perspective, but agreeing with the empirical data.

It has been suggested, however, that behaviour on this task might be governed by what appears to be a wholly non-rational strategy: *matching bias*. This bias arises in the same context as negative conclusion bias that we discussed earlier, that is, in Evans' (1972) negations paradigm. Take, for example, the rule *if there is an A on one side*, *then there is not a 2 on the other side* (*if p then* $\neg q$). The cards in this task are described using their logical status, so for this rule, 2 is the *false consequent* (FC) card and 7 is the *true consequent* (TC) card. For this negated consequent rule, participants tend to select the A card (TA: *true antecedent*) and the 2 card (FC). That is, participants now seem to make the falsifying response. However, as Evans and Lynch (1973) pointed out, participants may simply ignore the negations entirely and *match* the values named in the conditional, that is, A and 2. Prima facie, this is completely irrational. However, the "contrast set" account of negation shows that because of the rarity assumption – that most categories apply to a minority of items – *negated* categories are high probability categories (discussed earlier). Having a high probability antecedent or consequent alters the expected information gains associated with the cards. If the probability of the consequent is high, then the ODS model predicts that people *should* make the falsifying TA and FC responses, because these are associated with the highest information gain. Consequently, matching bias is a rational hypothesis testing strategy after all.

Probabilistic effects were first experimentally demonstrated using the *reduced array* version of Wason's selection task (Oaksford et al. 1997), in which participants can successively select up to 15 $q$ and 15 $\neg q$ cards (there are no upturned $p$ and $\neg p$ cards that can be chosen). As predicted by the ODS model, where the probability of $q$ is high (i.e., where rarity is violated), participants select more $\neg q$ cards and fewer $q$ cards. Other experiments have also revealed similar probabilistic effects (Green & Over 1997; 2000; Kirby 1994; Oaksford et al. 1999; Over & Jessop 1998).

There have also been some failures to produce probabilistic effects, however (e.g., Oberauer et al. 1999; 2004). We have argued that these arise because of weak probability manipulations or other procedural problems (Oaksford & Chater 2003b; Oaksford & Moussakowski 2004; Oaksford & Wakefield 2003). We therefore introduced a *natural sampling* (Gigerenzer & Hoffrage 1995) procedure in which participants sample the frequencies of the card categories while performing a selection task (Oaksford & Wakefield 2003). Using this procedure, we found probabilistic effects using the *same* materials as Oberauer et al. (1999), where these effects were not evident.

In further work on matching bias, Yama (2001) devised a crucial experiment to contrast the matching bias and the information gain accounts. He used rules that introduced a high and a low probability category, relating to the blood types Rhesus Negative (Rh–) and Positive (Rh+). People were told that one of these categories, Rh−, was rare. Therefore, according to the ODS model and the rule *if p then* $\neg$*Rh+* should lead participants to select the rare Rh− card. In contrast, according to matching bias they should select the Rh+ card. Yama's (2001) data were largely consistent with the information gain model. Moreover, this finding was strongly confirmed by using the natural sampling procedure with these materials (Oaksford & Moussakowski 2004).

Alternative probabilistic accounts of the selection task have also been proposed (Evans & Over 1996a; 1996b; Klauer 1999; Nickerson 1996; Over & Evans 1994; Over & Jessop 1998). Recently, Nelson (2005) directly tested the measures of information underpinning these models, including Bayesian diagnosticity (Evans & Over 1996b; McKenzie & Mikkelsen 2007; Over & Evans 1994), information gain (Hattori 2002; Oaksford & Chater 1994; 1996; 2003b), Kullback-Liebler distance (Klauer 1999; Oaksford & Chater 1996), probability gain (error minimization) (Baron 1981; 1985), and impact (absolute change) (Nickerson 1996). Using a related data selection task, he looked at a range of cases in which these norms predicted different orderings of informativeness, for various data types. Nelson found the strongest correlations between his data and information gain (.78). Correlations with diagnosticity (−.22) and log diagnosticity (−.41) were actually *negative*. These results mirrored Oaksford et al.'s (1999) results in the Wason selection task. Nelson's work provides strong convergent evidence for information gain as the index that most successfully captures people's intuitions about the relative importance of evidence.

There has been much discussion in the literature of the fact that selection task results change dramatically for conditionals that express rules of conduct, rather than putative facts, about the world (Cheng & Holyoak 1985; Manktelow & Over 1991). In such tasks, people typically do select the $p$ and $\neg q$ cards – the apparently "logical" response. One line of explanation is that reasoning is domain-specific, rather than applying across-the-board; a further claim, much discussed in evolutionary psychology, is that such tasks may tap basic mechanisms of social reasoning, such as "cheater-detection" (Cosmides 1989), which enables "correct" performance.

A Bayesian rational analysis points, we suggest, in a different direction – that such *deontic* selection tasks (i.e., concerning norms, not facts) require a different rational analysis. In the deontic selection task, participants are given conditionals describing rules concerning how people *should* behave, for example, *if you enter the country, you must have an inoculation against cholera*. The rule is not a hypothesis under test, but a regulation that should be obeyed (Manktelow & Over 1987). Notice, crucially, that it makes no sense to confirm or disconfirm a rule concerning how people *should* behave: People entering the country *should* be inoculated, whether or not they actually *are*. The natural interpretation of a deontic task is for the participant to check whether the rule is being disobeyed – that is, to look for $p$, $\neg q$ cases (people who enter the country, but are not inoculated); and indeed, in experiments, very high selections of the $p$ and $\neg q$ cards are observed. This is not because people have suddenly become Popperian falsifiers. This is because the task is no longer about attempting to gain information about whether the conditional is true or false. The conditional now concerns how people *should* behave, and hence can neither be confirmed nor disconfirmed by any observations of actual behavior.

We adopted a decision theoretic approach to these tasks (Oaksford & Chater 1994; Perham & Oaksford 2005). Violators are people who enter the country ($p$) without a vaccination ($\neg q$). Thus, we assume that participants whose role it is to detect violators attach a high utility to detecting these cases, that is, U($p$, $\neg q$) is high. However, every other case represents a cost, as it means wasted effort. We argue that people calculate the expected utility associated with each card. So, for example, take the case where someone does not have an inoculation ($\neg q$). She could be either entering the country ($p$, $\neg q$) or not entering the country ($\neg p$, $\neg q$). Just as in calculating expected information gain, both possible outcomess have to be taken into account in calculating *expected* utility (EU($x$)):

$$EU(\neg q) = P(p|\neg q)U(p, \neg q) + P(\neg p|\neg q)U(\neg p, \neg q)$$

We argue that people select cards in the deontic selection task to maximise expected utility. As only the utility of detecting a violator – someone trying to enter without an inoculation – is positive, this means that only the $p$ and the $\neg q$ cards will have positive utilities (because only these cards could show a violator). This model can account for a broad range of effects on deontic reasoning, including the effects of switched rules (Cosmides 1989), perspective changes (Manktelow & Over 1991), utility manipulations (Kirby 1994), and probability manipulations (Manktelow et al. 1995).

Recently, we have also applied this model to rules that contain emotional content, for example, *if you clean up blood, then you must wear gloves* (Perham & Oaksford 2005). With the goal of detecting cheaters (Cosmides 1989), you will look at people who are not cleaning up blood but who are wearing gloves ($\neg p$, $q$). With the goal of detecting people who may come to harm, you will want to check people who are cleaning up blood but who are not wearing gloves ($p$, $\neg q$). Perham and Oaksford (2005) set up contexts in which cheater detection should dominate, but in which the goal of detecting people who may come to harm may still be in play. That is, $\mathrm{U}(\neg p,\ q) > \mathrm{U}(p,\ \neg q) > 0$. The threatening word "blood" can appear for either the $p$, $q$ case or the $p$, $\neg q$ case. In calculating *generalized expected utility* (Zeelenberg et al. 2000), a *regret* term (*Re*) is subtracted from the expected utility of an act of detection, if the resulting state of the world is anticipated to be threatening. For example, by checking someone who is not wearing gloves ($\neg q$), to see if they are at risk of harm, one must anticipate encountering blood ($p$). Because "blood" is a threatening word, the utility for the participant of turning a $\neg q$ card is reduced; that is, the utility of encountering a $p$, $\neg q$ card is now $\mathrm{U}(p, \neg q) - Re$, for regret term *Re*. Consequently, selections of the "not wearing gloves" card ($\neg q$) should be lower for our *blood* rule than for a rule that does not contain a threatening antecedent, such as, *if you clean up packaging, then you must wear gloves*.

In two experiments, Perham and Oaksford (2005) observed just this effect. When participants' primary goal was to detect cheaters, their levels of $\neg p$ and $q$ card selection were the same for the threat (blood rule) as for the no-threat rule. However, their levels of $p$ and $\neg q$ card selection were significantly lower for the threatening than for the non-threatening rules. This finding is important because it runs counter to alternative theories, in particular the evolutionary approach (Cosmides 1989; Cosmides & Tooby 2000), which makes the opposite prediction, that $p$ and $\neg q$ card selections should, if anything, increase for threat rules.

Of the models considered in *BR*, the optimal data selection and expected utility models have been in the literature the longest, and have been subject to most comment. In the rest of Chapter 6, we respond in detail to these comments, pointing out that many can be incorporated into the evolving framework, and that some concerns miss their mark.

## 7. An uncertain quantity: How people reason with syllogisms

Chapter 7 of *BR* presents a probabilistic model of quantified syllogistic reasoning. This type of reasoning relates two quantified premises. Logic defines four types of quantified premise: *All*, *Some*, *Some…not*, and *None*. An example of a logically valid syllogistic argument is:

*Some Londoners* (*P*) *are soldiers* (*Q*)
*All soldiers* (*Q*) *are well fed* (*R*)
*Therefore* *Some Londoners* (*P*) *are well fed* (*R*)

In this example, *P* and *R* are the *end terms* and *Q* is the *middle term*, which is common to both premises. In the premises, these terms can only appear in four possible configurations, which are called *figures*. When one of these terms appears before the copula verb ("are") it is called the *subject* term (in the example, *P* and *Q*) and when one appears after this verb it is called the *predicate* term (*Q* and *R*). As the premises can appear in either order, there are 16 combinations, and as each can be in one of four figures, there are 64 different syllogisms.

There are 22 logically valid syllogisms. If people are reasoning logically, they should endorse these syllogisms and reject the rest. However, observed behavior is graded, across both valid and invalid syllogisms; and some invalid syllogisms are endorsed more than some valid syllogisms. Table 1 shows the graded behaviour over the 22 logically valid syllogisms. There are natural breaks dividing the valid syllogisms into three main groups. Those above the single line are endorsed most, those below the double line are endorsed least, and those in between are endorsed at an intermediate level.

Table 1. *Meta-analysis of the logically valid syllogisms showing the form of the conclusion, the number of mental models (MMs) needed to reach that conclusion, and the percentage of times the valid conclusion was drawn, in each of the five experiments analyzed by Chater and Oaksford (1999b)*

| Syllogism | Conclusion | MMs | Mean |
|---|---|---|---|
| *All(Q,P), All(R,Q)* | *All* | 1 | 89.87 |
| *All(P,Q), All(Q,R)* | *All* | 1 | 75.32 |
| *All(Q,P), Some(R,Q)* | *Some* | 1 | 86.71 |
| *Some(Q,P), All(Q,R)* | *Some* | 1 | 87.97 |
| *All(Q,P), Some(Q,R)* | *Some* | 1 | 88.61 |
| *Some(P,Q), All(Q,R)* | *Some* | 1 | 86.71 |
| *No(Q,P), All(R,Q)* | *No* | 1 | 92.41 |
| *All(P,Q), No(R,Q)* | *No* | 1 | 84.81 |
| *No(P,Q), All(R,Q)* | *No* | 1 | 88.61 |
| *All(P,Q), No(Q,R)* | *No* | 1 | 91.14 |
| *All(P,Q), Some…not(R,Q)* | *Some…not* | 2 | 67.09 |
| *Some…not(P,Q), All(R,Q)* | *Some…not* | 2 | 56.33 |
| *All(Q,P), Some…not(Q,R)* | *Some…not* | 2 | 66.46 |
| *Some…not(Q,P), All(Q,R)* | *Some…not* | 2 | 68.99 |
| *Some(Q,P), No(R,Q)* | *Some…not* | 3 | 16.46 |
| *No(Q,P), Some(R,Q)* | *Some…not* | 3 | 66.46 |
| *Some(P,Q), No(R,Q)* | *Some…not* | 3 | 30.38 |
| *No(P,Q), Some(R,Q)* | *Some…not* | 3 | 51.90 |
| *Some(Q,P), No(Q,R)* | *Some…not* | 3 | 32.91 |
| *No(Q,P), Some(Q,R)* | *Some…not* | 3 | 48.10 |
| *Some(P,Q), No(Q,R)* | *Some…not* | 3 | 44.30 |
| *No(P,Q), Some(Q,R)* | *Some…not* | 3 | 26.56 |

*Note.* The means in the final column are weighted by sample size.

Alternative theories of syllogistic reasoning invoke similar processes to explain these data as for conditional inference and the selection task. However, here both mental logic and mental models have to introduce new machinery to deal with quantifiers. For mental logic (Rips 1994), this requires new logical rules for *All* and *Some*, and a guessing mechanism to account for the systematic pattern of responses for the invalid syllogisms. For mental models, dealing with quantifiers requires reinterpreting the lines of a mental model as objects described by their properties ($P$, $Q$, and $R$) rather than as conjunctions of propositions. For the different syllogisms different numbers of mental models are consistent with the truth of the premises. Only conclusions that are true in all of these possible models are logically valid. As Table 1 shows, for the most endorsed valid syllogisms, there is only one model consistent with the truth of the premises, and so the conclusion can be immediately read off. For the remaining valid syllogisms, more than one model needs to be constructed. If people only construct an initial model, then errors will occur. As Table 1 shows, mental models theory provides a good qualitative fit for the valid syllogisms, that is, the distinction between 1, 2, and 3 model syllogisms maps on to the key qualitative divisions in the data.

The probabilistic approach to syllogisms was developed at both the computational and the algorithmic levels in the *Probabilistic Heuristics Model* (PHM, Chater & Oaksford 1999b). One of the primary motivations for this model was the hypothesis that, from a probabilistic point of view, reasoning about *all* and *some* might be continuous with reasoning about more transparently probabilistic quantifiers, such as *most* and *few*. By contrast, from a logical standpoint, such *generalised quantifiers* require a different, and far more complex, treatment (Barwise & Cooper 1981), far beyond the resources of existing logic-based accounts in psychology. Perhaps for this reason, although generalised quantifiers were discussed in early mental models theory (Johnson-Laird 1983), no empirical work on these quantifiers was carried out in the psychology of reasoning.

In deriving PHM, the central first step is to assign probabilistic meanings to the central terms of quantified reasoning using conditional probability. Take the universally quantified statement, *All P are Q* (we use capitals to denote predicates; these should be applied to variables $x$, which are bound by the quantifier, e.g., $P(x)$, but we usually leave this implicit). Intuitively, the claim that *All soldiers are well fed* can naturally be cast in probabilistic terms: as asserting that the probability that a person is well fed given that they are a soldier is 1. More generally, the probabilistic interpretation of *All* is straightforward: because its underlying logical form can be viewed as a conditional, that is, $All(x)(if\ P(x)\ then\ Q(x))$. Thus, the meaning is given as $P(Q|P) = 1$, as specifying the conditional probability of the predicate term ($Q$), given the subject term ($P$).

Similar constraints can be imposed on this conditional probability to capture the meanings of the other logical quantifiers. So, *Some P are Q* means that $P(Q|P) > 0$; *Some P are not Q* means that $P(Q|P) < 1$; and *No P are Q* means that $P(Q|P) = 0$. Thus, for example, "*Some Londoners are soldiers*" is presumed to mean that the probability that a person is a soldier given that he or she is a Londoner is greater than zero, and similarly for the other quantifiers. Such an account generalises smoothly to the generalised quantifiers *most* and *few*. *Most P are Q* means that $1 - \Delta < P(Q|P) < 1$ and *Few P are Q* means that $0 < P(Q|P) < \Delta$, where $\Delta$ is small. So, for example, *Most soldiers are well fed* may be viewed as stating that the probability that a person is well fed, given that they are a soldier, is greater than, say, 0.8, but less than 1.

At the level of rational analysis, these interpretations are used to build very simple graphical models (e.g., Pearl 1988) of quantified premises, to see if they impose constraints on the conclusion probability. For example, take the syllogism:

|  |  |  |
|---|---|---|
|  | *Some P are Q* |  |
|  | *All Q are R* | $P \rightarrow Q \rightarrow R$ |
| *Therefore* | *Some P are R* |  |

The syllogistic premises on the left define the dependencies on the right because of their *figure*, that is, the arrangement of the middle term ($Q$) and the end terms ($P$ and $R$) in the premises. There are four different arrangements or *figures*. The different figures lead to different dependencies, with different graphical structures. Note that these dependency models all imply that the end terms ($P$ and $R$) are *conditionally independent*, given the middle term, because there is no arrow linking $P$ and $R$, except via the middle term $Q$. Assuming conditional independence as a default is a further assumption about the environment (Step 2 in rational analysis). This is an assumption not made in, for example, Adams' (1998) probability logic.

These dependency models can be parameterised. Two of the parameters will always be the conditional probabilities associated with the premises. One can then deduce whether the constraints on these probabilities, implied by the earlier interpretations, impose constraints on the possible conclusion probabilities, that is, $P(R|P)$ or $P(P|R)$. In this example, the constraints that $P(Q|P) > 0$, and $P(R|Q) = 1$, and the conditional independence assumption, *entail* that $P(R|P) > 0$. Consequently, the inference to the conclusion *Some P are R* is probabilistically valid ($p$-valid). If each of the two possible conclusion probabilities, $P(R|P)$ or $P(P|R)$, can fall anywhere in the [0, 1] interval given the constraints on the premises, then no $p$-valid conclusion follows. It is then a matter of routine probability to determine which inferences are $p$-valid, of the 144 two premise syllogisms that arise from combining *most* and *few* and the four logical quantifiers (Chater & Oaksford 1999b).

In the PHM, however, this rational analysis is also supplemented by an algorithmic account. We assume that people approximate the dictates of this rational analysis by using simple heuristics. Before introducing these heuristics, though, we introduce two key notions: the notions of the *informativeness* of a quantified claim, and the notion of *probabilistic entailment* between quantified statements.

According to communication theory, a claim is informative in proportion to how surprising it is: informativeness varies inversely with probability. But what is the probability of an arbitrary quantified claim? To make sense of this idea, we begin by making a rarity assumption, as in our models of the conditional reasoning and the selection task, that is, the subject and predicate terms apply to only small subsets of objects. On this assumption, if we selected subject term *P*, and predicate term, *Q*, at random, then it is very likely that they will not cross-classify any object (this is especially true, given the hierarchical character of classification; Rosch 1975). Consequently, $P(Q|P) = 0$ and so *No P are Q* is very likely to be true (e.g., *No toupees are tables*). Indeed, for any two randomly chosen subject and predicate terms it is probable that *No P are Q*. Such a statement is therefore quite uninformative. *Some P are not Q* is even more likely to be true, and hence still less informative, because the probability interval it covers includes that for *No P are Q*. The quantified claim least likely to be true is *All P are Q*, which is therefore the most informative. Overall, the quantifiers have the following order in informativeness: I(*All*) > I(*Most*) > I(*Few*) > I(*Some*) > I(*None*) > I(*Some-not*) (see Oaksford et al. 2002, for further analysis and discussion).

Informativeness applies to individual quantified propositions. The second background idea, probabilistic entailment, concerns inferential relations *between* quantified propositions. Specifically, the use of one quantifier frequently provides evidence that another quantifier could also have been used. Thus, the claims that *All swans are white* is strong evidence that *Some swans are white* – because $P(white|swan) = 1$ is included in the interval $P(white|swan) > 0$ (according to standard logic, this does not follow logically, as there may be no swans). Thus, we say that *All* probabilistically entails (or *p*-entails) *Some*. Similarly, *Some* and *Some...not* are mutually *p*-entailing because the probability intervals $P(Q|P) > 0$ and $P(Q|P) < 1$ overlap almost completely.

With this background in place, we can now state the probabilistic heuristics model (PHM) for syllogistic reasoning. There are two types of heuristic: *generate* heuristics, which produce candidate conclusions, and *test* heuristics, which evaluate the plausibility of the candidate conclusions. The PHM account also admits the possibility that putative conclusions may also be tested by more analytic test procedures such as mental logics or mental models. The generate heuristics are:

(G1) *Min*-heuristic: The conclusion quantifier is the same as that of the least informative premise (*min*-premise)

(G2) *P-entailments*: The next most preferred conclusion quantifier will be the *p*-entailment of the *min*-conclusion

(G3) *Attachment*-heuristic: If just one possible subject noun phrase (e.g., *Some R*) matches the subject noun phrase of just one premise, then the conclusion has that subject noun phrase.

The two test heuristics are:

(T1) *Max*-heuristic: Be confident in the conclusion generated by G1 – G3 in proportion to the informativeness of the most informative premise (*max*-premise)

(T2) *Some_not*-heuristic: Avoid producing or accepting *Some_not* conclusions, because they are so uninformative.

We show how the heuristics combine in the following example:

| | | |
|---|---|---|
| | *All P are Q* | (*max*-premise) |
| | *Some R are not Q* | (*min*-premise) |
| *Therefore* | *Some_not* | (by *min*-heuristic) |
| | *Some R are not P* | (by *attachment*-heuristic) |

and a further conclusion can be drawn:

| | | |
|---|---|---|
| | *Some R are P* | [by *p-entailment*] |

In *BR*, we compare the results of these heuristics with probabilistic validity, and show that where there is a *p*-valid conclusion, the heuristics generally identify it. For example, the idea behind the *min*-heuristic is to identify the most informative conclusion that validly follows from the premises. Out of the 69 *p*-valid syllogisms, the *min*-heuristic identifies that conclusion for 54; for 14 syllogisms the *p*-valid conclusion is less informative than the *min*-conclusion. There is only one violation; that is, where the *p*-valid conclusion is more informative than the *min*-conclusion.

In turning to the experimental results, in *BR* we first show how all the major distinctions between standard syllogisms captured by other theories are also captured by PHM. So, returning to Table 1, all the syllogisms above the double line have the most informative *max*-premise, *All* (see heuristic T1). Moreover, all the syllogisms below the single line have uninformative conclusions, *Some-not* (see heuristic T2), and those below the double line violate the *min*-heuristic (heuristic G1) and require a *p*-entailment (heuristic G2), that is, *Some...not* ↔ *Some*. Consequently, this simple set of probabilistic heuristics makes the same distinctions among the valid syllogisms as the mental models account.

In this Précis, we concentrate on novel predictions that allow us to put clear water between PHM and other theories. As we discussed earlier, the most important feature of PHM is the extension to *generalised quantifiers*, like *most* and *few*. No other theory of reasoning has been applied to syllogistic reasoning with generalised quantifiers. Table 2 shows the *p*-valid syllogisms involving generalised quantifiers showing the conclusion type and the percentage of participants selecting that conclusion type in Chater and Oaksford's (1999b) Experiments 1 and 2. The single lines divide syllogisms with different *max*-premises, showing a clear ordering in levels of endorsements dependent on heuristic T1. All those above the double line conform to the *min*-heuristic (heuristic G1), whereas those below it do not and require a *p*-entailment (heuristic G2). As Chater and Oaksford (1999b) pointed out, one difference with experiments using standard logical quantifiers was that the *Some...not* conclusion was not judged to be as uninformative, that is, heuristic T2 was not as frequently in evidence. However, in general, in experiments using generalised quantifiers in syllogistic arguments the heuristics of PHM predict the findings just as well as for the logical quantifiers (Chater & Oaksford 1999b).

Many further results have emerged that confirm PHM. The *min*-heuristic captures an important novel distinction between strong and weak possible conclusions introduced by Evans et al. (1999). They distinguished conclusions that are necessarily true, possibly true, or impossible. For example, taking the syllogism discussed earlier (with premises, *Some P are Q, All Q are R*), the conclusion *Some P are R* follows *necessarily*, *No P are R* is *impossible*, and *Some P are not R* is *possible*. Some possible conclusions are endorsed by as many participants as the necessary conclusions (Evans et al. 1999). Moreover, some of the possible conclusions were endorsed by as few participants as the impossible conclusions. Evans et al. (1999) observed that possible conclusions that are commonly endorsed all conform to the *min*-heuristic, whereas those which are rarely endorsed violate the

Table 2. *The* p-*valid syllogisms less the syllogisms that are also logically valid (shown in Table 1), showing the form of the conclusion and the proportion of participants picking the* p-*valid conclusion in Chater and Oaksford's (1999b) Experiments 1 and 2*

| Syllogism | Conclusion | Mean |
|---|---|---|
| *All(Q,P), Most(R,Q)* | *Most* | 85 |
| *Most(Q,P), All(R,Q)* | *Most* | 65 |
| *All(P,Q), Most(Q,R)* | *Most* | 70 |
| *Most(P,Q), All(Q,R)* | *Most* | 55 |
| *Few(P,Q), All(R,Q)* | *Few* | 80 |
| *All(P,Q), Few(R,Q)* | *Few* | 85 |
| *Few(P,Q), All(R,Q)* | *Few* | 85 |
| *All(P,Q), Few(Q,R)* | *Few* | 75 |
| *Most(Q,P), Most(R,Q)* | *Most* | 65 |
| *Most(P,Q), Most(Q,R)* | *Most* | 50 |
| *Few(Q,R), Most(R,Q)* | *Few* | 60 |
| *Most(Q,R), Few(R,Q)* | *Few* | 75 |
| *Most(P,Q), Few(Q,R)* | *Few* | 70 |
| *Most(Q,P), Some…not(R,Q)* | *Some…not* | 80 |
| *Some…not(Q,P), Most(R,Q)* | *Some…not* | 60 |
| *Some…not(Q,P), Most(Q,R)* | *Some…not* | 75 |
| *Most(Q,P), Some…not(Q,R)* | *Some…not* | 65 |
| *Most(P,Q), Some…not(Q,R)* | *Some…not* | 75 |
| *Some…not(P,Q), Most(Q,R)* | *Some…not* | 75 |
| *Few(Q,P), Some…not(R,Q)* | *Some…not* | 60 |
| *Some…not(Q,P), Few(R,Q)* | *Some…not* | 40 |
| *Some…not(Q,P), Few(Q,R)* | *Some…not* | 30 |
| *Few(Q,P), Some…not(Q,R)* | *Some…not* | 60 |
| *Few(P,Q), Some…not(Q,R)* | *Some…not* | 60 |
| *Some…not(P,Q), Few(Q,R)* | *Some…not* | 40 |
| *All(P,Q), Most(R,Q)* | *Some…not* | 35 |
| *Most(P,Q), All(R,Q)* | *Some…not* | 35 |
| *Few(Q,P), Few(R,Q)* | *Some…not* | 35 |
| *Few(P,Q), Few(Q,R)* | *Some…not* | 30 |
| *Few(P,Q), Most(Q,R)* | *Some…not* | 30 |

*Note.* This table excludes the eight MI, IM, FI, and IF syllogisms, which have two $p$-valid conclusions only one of which was available in Chater and Oaksford's (1999b) Experiment 2.

*min*-heuristic (with one exception). Hence, PHM captures this important new set of data.

Some experiments designed to test the claim that syllogism difficulty is determined by the number of alternative mental models can also be interpreted as confirming PHM (Newstead et al. 1999). Participants wrote down or drew diagrams consistent with the alternative conclusions they entertained, during syllogistic reasoning. No relationship was found between the number of models a syllogism requires (according to mental models theory) for its solution and the number of conclusions or diagrams participants produced. This suggests that sophisticated analytical procedures, such as those described in mental models, play, at most, a limited role in the outcome of syllogistic reasoning. By contrast, participants' responses agreed with those predicted by the *min*- and *attachment*-heuristics. Furthermore, no differences in task difficulty dependent on syllogistic figure were observed, a finding consistent with PHM, but not mental models.

Recent work relating memory span measures to syllogistic reasoning has also confirmed PHM (Copeland & Radvansky 2004). PHM makes similar predictions to mental models theory because the number of heuristics that need to be applied mirrors the one, two, and three model syllogism distinction (see Table 1). For one model syllogisms, just the *min*-heuristic and *attachment* is required (two heuristics). For two model syllogisms, the *some__not*-heuristic is also required (three heuristics). In addition, for three model syllogisms a $p$-entailment is required (four heuristics). The more mental operations that need to be performed, the more complex the inference will be, and the more working memory it will require. Copeland and Radvansky (2004) found significant correlations between working memory span and strategy use, for both mental models and PHM. While not discriminating between theories, this work confirmed the independent predictions of each theory for the complexity of syllogistic reasoning and its relation to working memory span.

As with Chapters 5 and 6, Chapter 7 of *BR* closes by addressing the critiques of PHM that have arisen since the theory first appeared. One criticism is that PHM does not generalise to *cardinal* quantifiers (Geurts 2003) such as *Exactly three P are Q*, which have no probabilistic interpretation. Yet, such quantifiers can, nonetheless, naturally mesh with the generalized quantifiers, to yield interesting inferences. For example, suppose you are told that exactly three birds in the aviary are black. If there are twenty birds in the aviary, then *few* of the birds are black; if there are four, then *most* of the birds are black; and, in either case, further inferences from these generalized quantifiers can be drawn, as appropriate.

## 8. Conclusion

As we have seen, Chapters 5 to 7 of *BR* provide the empirical support for taking a probabilistic approach to human reasoning and rationality. The final chapter provides further arguments for pursuing this research strategy in the form of a dialogue between an adherent of the probabilistic approach and a sceptic. In this Précis, we concentrate on two key issues that emerge from that debate.

The first topic we consider is whether the brain is a probabilistic inference machine. *BR* focuses primarily, as we have seen, on providing rational analyses of human reasoning – and we have noted that rational analysis does not make direct claims about underlying computational operations. But, to what extent can the mind or brain be viewed as a probabilistic (or for that matter, a logical) calculating engine? Although not the primary focus in this book, this is nonetheless a fundamental question for the behavioural and brain sciences. We suspect that, in general, the probabilistic problems faced by the cognitive system are too complex to be solved by direct probabilistic calculation. Instead, we suspect that the cognitive system has developed relatively computationally "cheap" methods for reaching solutions that are "good enough" probabilistic solutions to be acceptable. In particular, where we propose a specific processing theory (in our account of syllogistic reasoning) this account consists of simple, but surprisingly effective, heuristics – heuristics that, however, lead to errors, which we argue are revealed in the empirical data. Moreover, in related work on the topic of rational choice and decision making, which we do not consider here, we and others have proposed models that solve probabilistic/decision making problems, but do so using relatively cheap, and hence approximate, methods (Gigerenzer & Goldstein 1996; Gigerenzer et al. 1999; Stewart et al. 2006).

To the degree that algorithmic models can be formulated, is rational analysis simply redundant? We argue that it is not. Rational analysis is essential because it explains *why* the particular algorithms used by the cognitive system are appropriate. That is, without a characterization of what problem the cognitive system solves, we cannot ask, let alone answer, the questions of why the algorithm has its particular form, or how effectively it works. Moreover, it may be that a good deal of empirical data about human reasoning (and indeed, human cognition more generally) can be understood as arising from the structure of the problem itself – that is, the nature of the problem drives any reasonable algorithmic solution to have particular properties, which may be evident in the data. This idea is a core motivation for the rational analysis approach (Anderson 1990; 1991a); and we have seen that a broad spectrum of data on human reasoning can be understood purely at the rational level – that is, without formulating an algorithmic theory of any kind.

The second topic we consider is the importance of qualitative patterns of probabilistic reasoning, rather than precise numerical calculations. Suppose, for concreteness, we consider a person reasoning about a game of dice. If the dice are unbiased, then it is easy, of course, for the theorist to formulate a probabilistic model specifying that each throw is independent, and that each face has a probability of 1/6. But this model is both too strong and too weak. It is too strong because it generates all manner of subtle mathematical predictions, concerning, say, the relative probabilities of rolling at least one six out of six dice rolls versus rolling at least two sixes out of twelve dice rolls, predictions that are not available to everyday intuition. And it is too weak because it ignores many factors of crucial importance in everyday reasoning. For example, watching a dice being thrown, we have not only a model of the probability that each face will be uppermost, but a rough model of where it will land, how likely it is to fall off the table, how loud that impact is likely to be, how another player is likely to react to a particular outcome, given their temperament, the gamble they have placed, and so on.

This observation implies that, if the cognitive system is indeed building probabilistic models of the world, then it is building models of considerable complexity – models that can take into account any aspect of knowledge, from naïve physics to folk psychology. This implies that the probabilistic turn does not resolve the difficulty of representing knowledge – rather it provides a framework into which this knowledge must be integrated. The advantage of the probabilistic viewpoint, though, is that it provides a powerful framework for dealing with an uncertain world; and, indeed, for assessing competing explanations of observed phenomena (rival interpretations of perceptual input; competing grammars; alternative interpretations of sentences, stories, or court-cases). Moreover, probabilistic models of complex domains do not need to be fully specified, at a numerical level – most critical is that the *functional* relationships between pieces of information are represented. What tends to cause what? What is evidence for what? The direction and existence of functional dependencies between pieces of information may be mentally represented, even though precise numerical probabilities may be unknown. Thus, probability theory can provide a framework for *qualitative* reasoning, without using numerical values (e.g., Pearl 2000). We tentatively suggest that much of the power, and limitations, of human reasoning about the everyday world flows from this qualitative style of reasoning. From this point of view, it is perhaps not surprising that people are not good at explicit reasoning with probabilities – indeed, they fall into probabilistic fallacies just as readily as they fall into logical contradictions (e.g., Kahneman et al. 1982).

The probabilistic mind is not, of course, a machine for solving verbally or mathematically specified problems of probability theory. Instead, we suggest, the mind is a qualitative probabilistic reasoner, in the sense that the rational analysis of human reasoning requires understanding how the mind deals qualitatively with uncertainty. As we have stressed, this does not imply that the mind is a probabilistic calculating machine (although it may be); still less does it imply that the mind can process probabilistic problems posed in a verbal or mathematical format. Nonetheless, the concepts of probability are, we suggest, as crucial to understanding the human mind as the concepts of aerodynamics are in understanding the operation of a bird's wing.

ACKNOWLEDGMENTS

The book *Bayesian Rationality* owes huge debts to more people and institutions than we can summarize here. We would especially like to express our gratitude to the University of Edinburgh and the Centre for Cognitive Science (formerly the School of Epistemics, and now part of the Division of Informatics) for starting us on the adventure of understanding human reasoning from a formal and empirical standpoint, which has engaged so much of our lives for the last twenty years.

## NOTES

[1] The case where the categorical premise is uncertain can be accommodated using a generalization of this idea, Jeffrey conditionalization (Jeffrey 1983). The new degree of belief that *John plays tennis* ($q$), on learning that *it is sunny in Bloomsbury* (which confers only a high probability that *it is sunny in Wimbledon* [$p$]), is:

$$P_1(q) = P_0(q|p)P_1(p) + P_0(q|\neg p)P_1(\neg p)$$

[2] Bayes' theorem is an elementary identity of probability theory that allows a conditional probability to be calculated from its converse conditional probability and the priors: $P(p|q) = (P(q|p)P(p))/P(q)$.

[3] However, this may be because of the different way that negations are used in each task (see Evans & Handley 1999; Oaksford 2004b).