Rational Models of Cognition

Edited by

Mike Oaksford
School of Psychology,
University of Wales, Cardiff

and

Nick Chater
Department of Psychology,
University of Warwick

Oxford · New York · Tokyo
OXFORD UNIVERSITY PRESS
1998

1 *An introduction to rational models of cognition*

Mike Oaksford and Nick Chater

In 1990, John Anderson summarized a new methodology for understanding cognition: rational analysis. This book brings together leading researchers from a range of areas in cognitive science, whose work addresses issues concerned with rational analysis. It provides a state-of-the-art overview of the variety and fecundity of research using this approach, and addresses fundamental theoretical issues in four key areas of cognitive science: memory, categorization, reasoning and search. We believe that research using rational analysis is an important and exciting development, the implications of which are beginning to bear fruit in a variety of areas. Moreover, rational analysis forges important connections with other branches of the biological and social sciences, and suggests new directions for the methodology and philosophy of cognitive science.

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

The twentieth century has, however, seen a move away from this 'psychologism' (Frege, 1879; Hilbert, 1925), and now mathematicians, philosophers and psychologists sharply distinguish between normative theories, such as probability theory and logic, which are about how people *should* reason, and descriptive theories of the psychological mechanisms by which people actually reason. Moreover, a major finding in psychology has been that the rules by which people *should* and *do* reason are not merely conceptually distinct, but that they appear to be empirically very different (Wason, 1966; Wason and Johnson-Laird, 1972; Kahneman and Tversky,

1973; Kahneman *et al.*, 1982). Whereas very early research on probability theory and logic took their project as codifying how people think, the psychology of reasoning has suggested that probability theory and logic are profoundly at variance with how people think. If this viewpoint is correct, then the whole idea of rational models of cognition is misguided: cognition simply is not rational.

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

The computational metaphor is important because it suggests that rational analyses should be described in terms of a scheme for computational explanation. The most well-known scheme for computational explanation was provided by David Marr (1982). At Marr's highest *computational* level the function that is being computed in the performance of some task is outlined. This level corresponds to a rational analysis of the cognitive task. The emphasis on computational explanation makes two points explicit. First, that in providing a computational explanation of the task that a particular device performs there is an issue about whether the computational level theory is correct. Second, there is a range of possible computational level theories that may apply to a given task performance; which one is correct must be discovered and cannot be assumed a priori. Let us consider an example. Suppose you find an unknown device and wonder what its function might be. Perhaps, observing its behaviour, you hypothesize that it may be performing arithmetical calculations. To make this conjecture is to propose a particular rational model of its performance. That is, this is a theory about what the device *should* do. In this case, the device should provide answers to arithmetical problems that conform to the laws of arithmetic, i.e. arithmetic (or some portion of it) provides the hypothesized rational model. On this assumption, you might give the device certain inputs, which you interpret as framing arithmetical problems. It may turn out, of course, that the outputs that you receive do not appear to be interpretable as solutions to these, or perhaps any other, arithmetical problems. This may indicate that your rational model is inappropriate, particularly if you cannot interpret most of the outputs as correct answers. You may therefore search for an alternative rational model—perhaps the device is not doing arithmetic, but is solving differential equations. Similarly, in rational analysis, theorists cannot derive appropriate computational level theories by reflecting on normative considerations alone, but only by attempting to use those theories to describe human performance. For example, it is not controversial that arithmetic is a good normative account of how numbers should be manipulated—the question is: Does this device do arithmetic?

This leads to the second difference between the modern programme of rational analysis and early developments of logic and probability: that the goal is not merely to capture people's intuitions, but rather to model detailed experimental data on cognitive function. The rational models in this book aim to capture experimental data: on the rate at which information is forgotten; on the way people generalize from old to new instances; on performance on hypothesis testing tasks; and on search problems. Rational analysis as a programme in cognitive science is primarily aimed at capturing these kinds of empirical phenomena, while explaining how the cognitive system is successful. None the less, rational analysis shares with early views the assumption that accounts of the mind must be both normatively justified and descriptively adequate.

So far, we have considered rationality in the abstract—as consisting of reasoning according to sound principles. But the goals of an agent attempting to survive and prosper in its ecological niche are more concrete—it must decide how to act in order to achieves its goals. The chapters in this book explain how normative principles can be combined with analysis of the structure of the environment in order to provide rational explanations of successful cognitive performance. Indeed, what many of these chapters show is that many aspects of cognition can be viewed as optimized (to some approximation) to the structure of the environment. For example, the rate of forgetting an item in memory seems to be optimized to the likelihood of encountering that item in the world (Schooler, Chapter 7), categorization may be viewed as optimizing the ability to predict the properties of a category member (J. Anderson and Matessa, Chapter 10), searching computer menus (Young, Chapter 21), parsing (Chater *et al.*, Chapter 20), and selecting evidence in reasoning (Oaksford and Chater, Chapter 17; Over and Jessop, Chapter 18) may all be viewed as optimizing the amount of information gained. This style of explanation is similar to optimality based explanations which have been influential in other disciplines. In the study of animal behaviour (Kacelnik, Chapter 3), foraging, diet selection, mate selection and so on, have all been viewed as problems which animals solve more or less optimally. In economics, people and firms are viewed as more or less optimally making decisions in order to maximize utility or profit.

Models based on optimizing, whether in psychology, animal behaviour or economics, need not, and typically do not, assume that agents are able to find the perfectly optimized solutions to the problems that they face. Quite often, perfect optimization is impossible even in principle, because the calculations involved in finding a perfect optimum are frequently computationally intractable (Simon, 1955, 1956), and, moreover, much crucial information is typically not available. The agent must still act, even in the absence of the ability to derive the optimal solution (Simon, 1956; Chater and Oaksford, 1996; Gigerenzer and Goldstein, 1996). Thus, there may be a tension between the theoretical goal of the rational analysis and the practical need for the agent to be able to decide how to act in real time, given the partial information available (R. Anderson, Chapter 8; McGonigle and Chalmers, Chapter 8; Shiffrin and Steyvers, Chapter 4). This leads directly into the area of what Simon (1955, 1956) calls *bounded rationality*. We believe that rational analysis can be reconciled with the boundedness of cognitive systems in a number of ways.

First, the cognitive system may, in general, approximate, perhaps very coarsely, the optimal solution. Thus, the algorithms that the cognitive system uses may be fast and frugal heuristics (Gigerenzer and Goldstein, 1996) which generally approximate the optimal in the environments that an agent normally encounters. In this context, the optimal solutions will provide a great deal of insight into why the agent behaves as it does. However, an account of the algorithms that the agent uses will be required to provide a full explanation of its behaviour. Issues concerning algorithmic explanations in conjunction with rational analysis are discussed in many chapters in this book (e.g. R. Anderson, Chapter 8; Dennis and Humphreys, Chapter 6; López, *et al.*, Chapter 15; McGonigle and Chalmers, Chapter 15; Shiffrin and Steyvers, Chapter 4) and have been extensively discussed by J. Anderson (1990, 1994).

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

The importance that the local goals be relevant to the larger aims of the cognitive system, raises another important question about providing rational models of cognition. The fact that a model involves optimizing *something* does not mean that the model is a *rational* model. Optimality is not the same as rationality. It is crucial that the local goal that is optimized must be relevant to some larger goal of the agent. Thus, it seems *reasonable* that animals may attempt to optimize the amount of food they obtain, or that the categories used by the cognitive system are optimized to lead to the best predictions. This is because, for example, optimizing the amount of food obtained is likely to enhance inclusive fitness, in a way that, for example, maximizing the amount of energy consumed in the search process would not. Determining whether some behaviour is rational or not therefore depends on more than just being

able to provide an account in terms of optimization. Therefore rationality requires not just optimizing something but optimizing something reasonable. As a definition of rationality, this is clearly circular. But by viewing rationality in terms of optimization, general conceptions of what are reasonable cognitive goals can be turned into specific and detailed models of cognition. Thus, the programme of rational analysis, while not answering the ultimate question of what rationality is, none the less provides the basis for a concrete and potentially fruitful line of empirical research.

This flexibility of what may be viewed as rational, in building a rational model, may appear to raise a fundamental problem for the entire rational analysis programme. It seems that the notion of rationality may be so flexible that whatever people do, it is possible that it may seem rational under some description. So for example, it may be that our stomachs are well adapted to digesting the food in our environmental niche, indeed they may even prove to be optimally efficient in this respect. However, we would not therefore describe the human stomach as rational, because stomachs presumably cannot usefully be viewed as information processing devices. Stomachs may be well or poorly adapted to their function (digestion), but they have no beliefs, desires or knowledge, and hence the question of their rationality does not arise.

Optimality approaches in biology, economics and psychology, assume that the agent is well-adapted to its normal environment. However, almost all psychological data are gained in a very unnatural setting, where a person performs a very artificial task in the laboratory. Any laboratory task will recruit some set of cognitive mechanisms that determine the participants' behaviour. But it is not obvious what problem these mechanisms are adapted to solving. Clearly, this adaptive problem is not likely to be directly related to the problem given to the participant by the experimenter, precisely because adaptation is to the natural world, not to laboratory tasks. In particular, this means that participants may fail with respect to the task that the experimenter thinks they have set. But this may be because this task is unnatural with respect to the participants' normal environment. Consequently, participants may assimilate the task that they are given to a more natural task, recruiting adaptively appropriate mechanisms which solve this, more natural, task successfully. This issue is most pressing in reasoning tasks where human performance has been condemned as irrational. For example, hypothesis testing tasks, where people do not adopt the supposedly 'logical' strategy of falsification have been taken to demonstrate the irrationality of human reasoning (Stich, 1985, 1990; Sutherland, 1992). However, recently a number of theorists have suggested that these tasks are more likely to engage cognitive mechanisms which are adapted to different real world problems. In particular, several researchers have suggested that people are maximizing the amount of information that they can gain in selecting evidence, as opposed to following logical rules (Oaksford and Chater, Chapter 17; Over and Jessop, Chapter 18). On these accounts, people's behaviour is (to an approximation at least) rational, even though it violates the standards set by the experimenter.

In summary, rational analysis as a general approach to understanding cognition has come a long way since the nineteenth century. Through the computer metaphor

and the detailed experimental investigations of cognitive phenomena that have dominated much of psychology during the second half of the twentieth century we are now in a position to apply this methodology in a much more rigorous way. The chapters in this book, although not shying away from the problems, reveal the fecundity and promise of this approach. We now outline the organization and contents of the contributions to this book.

The book is organized into five parts: general issues; memory; categorization; reasoning; and search. The first part on general issues contains two chapters that discuss issues about rational analysis which cut across subject boundaries. In Chapter 2, *Connectionist models and Bayesian inference*, McClelland traces the fundamental connection between rational Bayesian analysis and connectionist networks. Although this connection has been discussed in the technical connectionist literature, it is not widely known in cognitive science (although see Chater, 1995). This connection holds at two levels. First, connectionist networks can be viewed as performing probabilistic calculations, in a parallel, distributed fashion. Thus, connectionist psychological models, such as the models of word recognition developed by McClelland and colleagues (see Chapter 2), can be interpreted as integrating information according to Bayesian principles. Crucially, interpreting networks in this way reveals their underlying assumptions (e.g. various kinds of independence between different sources of information), leading to a deeper understanding of how such models work and what predictions they make. The second level at which a connection between networks and Bayesian analysis holds concerns *learning*. The process of training connectionist networks from examples can be understood in terms of Bayesian updating in the light of new data. In both cases, the probabilistic 'rational' interpretation is not merely *post hoc*—it has driven new developments in connectionist research. More fundamentally, the Bayesian interpretation of both network behaviour and learning may provide a crucial bridge between the kinds of probabilistic calculations postulated in rational analysis, and the parallel, distributed neural substrate on which cognitive processes must run.

In Chapter 3, *Normative and descriptive models of decision making: time discounting and risk sensitivity*, Kacelnik considers how the adaptiveness of cognition relates to evolutionary considerations, from the perspective of animal behaviour. He argues that the degree to which the cognitive system can be viewed as adapted to its environment cannot be decided by a priori reflection, but requires developing detailed case studies. Much research in animal behaviour attempts to provide such case studies—attempting to understand specific phenomena in terms of assumptions that the animal is optimizing food intake, minimizing the probability of being killed by a predator, and so on. This well-established research programme in the study of animal behaviour is strongly analogous to Anderson's programme of providing rational analyses for cognitive processes. Kacelnik considers two case studies: temporal discounting and risk sensitivity. Temporal discounting concerns the degree to which animals discount future utilities with respect to current utilities. The empirical data, obtained with both humans and animals, indicate a hyperbolic discounting function. Kacelnik suggests that, although there may be no direct normative justification for this discounting function, it may be viewed as a by-product of the fact that the discounting mechanism may be evolved to maximize

food intake in the context of foraging. The topic of risk sensitivity concerns how humans and animals choose between actions with uncertain outcomes. Specifically, how do the expected pay-offs from the action and the variance in expected pay-off interact to determine decision-making behaviour. For example, in foraging, it might be appropriate for a bird to choose a foraging site with lower expected mean pay-off, where that pay-off is relatively certain, so that the bird does not risk failing to obtain enough food to survive through a cold winter night. Kacelnik describes animal and human research which suggests that, at a qualitative level at least, a normative account of risk sensitivity accounts for the pattern of human choice behaviour in an experimental domain. Kacelnik concludes that a normative approach can have an important role in guiding research in human psychology, and that this normative approach may be usefully informed by evolutionary considerations. But he cautions that evolutionary considerations must be applied with caution: it is unlikely that humans or animals are fitness maximizers in all circumstances.

The second part of the book is on memory. This is the subject area where recent interest in rational analysis really began with the paper by Anderson and Milson which appeared in the *Psychological Review* in 1989. This section contains a variety of papers by leading researchers in this area. In Chapter 4, *The effectiveness of retrieval from memory*, Shiffrin and Steyvers focus on the question: given the constraints imposed by what has been stored in memory, is retrieval from memory optimal or close to optimal, and is retrieval in different tasks equally optimal or equally far from optimal. They suggest that these questions can be addressed in the context of models that are based on probabilities of matching probe cues to memory traces, and they couch their discussion in the context of one such model, REM (retrieving effectively from memory). Optimality is discussed first for explicit single item yes/no recognition. They show that relative optimality is less when extensions are made to tasks in which the probe consists of more than one item, such as paired recognition, and associative recognition. They argue that retrieval is probably less optimal in recall than in recognition. Extensions to generic and implicit memory are briefly considered.

In Chapter 5, *Predictions of a Bayesian recognition memory model (and a class of models including it)*, Chappell reports predictions from a Bayesian model of episodic recognition memory developed by McClelland and Chappell (1996). The key issue under consideration is the effect of the number of times that an item is presented in a learning phase for recognition judgements (e.g. was an item presented several times or just once). Moreover, to what extent is the recognition of an item which was, say, presented just once, affected by the number of presentations of other items in the list? Four cases are considered: PW (pure weak), where all items are presented once; PS (pure strong), where all items are presented several times; MW (mixed weak), where items are present once in a list where some items were presented once and others presented several times; and MS (mixed strong) which is defined analogously. He derives expressions relating hit rates and false alarm rates for recognition judgements in the four cases above. The hits rates and false alarm rates from 11 separate experiments taken from Ratcliffe *et al.* (1990) and Murnane and Shiffrin (1991) confirmed these relationships derived from a rational Bayesian analysis of memory performance.

In Chapter 6, *Cueing for context: an alternative to global matching models of recognition memory*, Dennis and Humphreys present a model that like Chappell (Chapter 5) and Shiffrin and Steyvers (Chapter 4) is framed as a rational Bayesian account of memory. However, as several chapters in this book argue, there may be evidence for which an adequate explanation requires making some assumptions about the algorithmic level. In this chapter Dennis and Humphreys make some algorithmic assumptions about cueing in single-item recognition that allows them to capture data other rational models cannot. Specifically, they assume that an item is recalled by cueing its bindings to other items in memory and then seeing if the current context occurs in any of those bindings. This contrasts with other approaches which cue by context and then check to see if the item occurs in any of the resulting bindings. Perhaps the most interesting effect this model accounts for is the enhanced recall of low frequency items. Because these occur in few contexts they would seem less confusable, and therefore easier to recognize. Dennis and Humphreys 'cueing for context' approach reveals how different approaches at the algorithmic level may lead to different predictions within a generally rational approach (see also, López *et al.*, Chapter 15).

In Chapter 7, *Sorting out core memory processes*, Schooler develops his work on the rational analysis of memory (J. Anderson, 1990), which proposes that memory's sensitivity to statistical structure in the environment enables it to estimate optimally the odds that a memory trace will be needed. J. Anderson and Schooler have analysed sources of informational demand on the environment: speech to children and word usage in the front page headlines of the *New York Times*. They showed that factors that govern memory performance, including recency, also predict the odds that an item (e.g. a word) will be encountered. In this chapter, Schooler develops the theory to make precise predictions about how the odds of encountering an item varies as a joint function of: (i) the statistical associations between the item and elements of the current context, and (ii) how long it has been since the item was last encountered. The prediction was confirmed environmentally for child-direct speech and *New York Times*' data. The corresponding behavioural prediction was tested using a cued recall task in which the cues were either strongly associated or not associated with the targets. In contrast to the environmental results, recall performance was more sensitive to the length of the retention interval in the presence of cues that are not associated, than in the presence of associated cues. Further modelling showed that incorporating estimates of the influence of non-retrievable processes (e.g. reading a word, deciding to respond, etc.) on overall performance reduces the discrepancy between the theoretical predictions and the observed data.

In Chapter 8, *Rational and non-rational aspects of forgetting*, R. Anderson reviews evidence for and against rational memory. He observes that existing analyses like those due to Anderson and Schooler, do not explicitly manipulate need-probability—the probability that an item will need to be recalled. R. Anderson cites evidence from his own recent work that captures need-probability operationally by varying the probability that an item is tested for recall on a particular trial—on some trials participants were told that there would be no recall test. This laboratory manipulation successfully confirmed that people's forgetting curves for an item

matched the probability that that item would be tested. R. Anderson goes on to suggest that certain aspects of the forgetting function are not captured by test probability, e.g. recency effects, and the effects of serial position within a list of to-be-remembered items. Finally, he makes some proposals about why apparent deviations from need probabilities may occur, e.g. the capacity of short-term memory and the ongoing process of learning the optimal solution.

In Chapter 9, *Adaptive analysis of sequential behaviour: oscillators as rational mechanisms*, Brown and Vousden argue that the brain implements various key memory functions using endogenous oscillators—nerve bundles that display an oscillatory dynamic. Using banks of units that oscillate at different frequencies allows the authors to model memory for serial order, i.e. the periods of different oscillators are associated with different positions in a sequence. Their OSCAR model can account for various effects in short-term memory for serial order and in speech production, e.g. temporal generalization—where items close to each other in a sequence are more confusable, and hierarchical representation where sequences may be chunked in to smaller sequential units. Oscillatory mechanisms are therefore clearly adequate to explain a variety of empirical phenomenon. Brown and Vousden also argue that such mechanisms are perfectly suited to adapting an organism successfully to its environment—they provide just the mechanism needed to adapt optimally an organism's periodic foraging behaviour to the periodic availability of food at different sites.

The third part of this book is on categorization and induction, which is another core area originally investigated by J. Anderson (1990). In Chapter 10, *The rational analysis of categorization and the ACT-R architecture*, J. Anderson and Matessa explore how a particular cognitive architecture, the ACT-R architecture, can be used to model categorization. The ACT-R architecture is a version of Anderson's ACT theory of cognition (Anderson, 1976, 1983), which was developed to incorporate the rational analyses of memory and choice in Anderson (1990). However, the ACT-R architecture does not embody Anderson's rational analysis of categorization. Anderson and Matessa show instead that the rationality of categorization behaviour emerges in the ACT-R framework. They discuss two radically different ways in which categorization can be implemented in ACT-R: an exemplar approach, where exemplars are stored in ACT-R's declarative memory, and a rule-based approach, where rules are part of ACT-R's procedural memory. Both approaches provide excellent fits with data from an important data set (Gluck and Bower, 1986). Anderson and Matessa argue that the rationality of categorization can be seen as derivative on the rationality of memory and choice. Crucially, this important generalization arises not purely at the level of rational analysis, but from the choice of a specific cognitive architecture. Thus, there may be valuable synergies between rational and architectural accounts of cognition.

In Chapter 11, *Optimum performance and exemplar models of classification*, Nosofsky takes up the theme of exemplar models of categorization in more detail. The general claim embodied in such models is that people represent categories by storing individual exemplars of categories in memory, and classify objects according to their similarity with exemplars of different categories. Nosofsky's (1984, 1986) generalized context model (GCM) constitutes the most theoretically and experimentally

well-developed categorization model of this kind. This model assumes that exemplars are points in a psychological space, and that people may distribute their attention differentially over these psychological dimensions. The issue of 'rationality', in this context, concerns whether people distribute their attention in a way which is adaptive: to optimize classification performance. Nosofsky shows that the assumption that people do optimize their allocation of attention allows his model to account for performance in a variety of experimental paradigms. Moreover, Nosofsky shows that there is also a deep connection between exemplar-based models in general and rational considerations, pointing out how it relates to probabilistic views of classification. This provides an important link between rational analysis and processing architectures, as also discussed by McClelland, and Anderson and Matessa. He also reviews empirical data showing that people are often, although not invariably, able to learn quite complex likelihood-based decision boundaries for classification, noting that this 'rational' performance is consistent with an exemplar-based architecture for classification.

In Chapter 12, *A Bayesian analysis of some forms of inductive reasoning*, Heit considers how categorization relates to inductive inferences which allow people to make predictions. Specifically, he is concerned with the degree to which people evaluate inductive arguments such as: goldfish thrive in the sunlight; therefore, tuna thrive in the sunlight. He presents initial steps towards Bayesian rational analysis for this class of inferences. He argues that his model, which relies in essence on a single equation (Bayes' theorem) can account for a wide range of standard empirical results, concerning the effect of similarity between the two categories (e.g. goldfish and tuna), and more subtle effects of typicality and category diversity. Moreover, the Bayesian approach has the means to take account of the effects of prior knowledge, concerning, for example, beliefs about what Goodman (1955) terms the projectability of properties. Heit suggests that a fundamental challenge to this kind of account is to explain the origin of this kind of prior knowledge, and that this may be an important topic of future research.

In Chapter 13, *Dynamics of dimension weight distribution and flexibility in categorization*, Lamberts and Chong consider the exemplar framework for categorization discussed in Nosofsky (Chapter 11), and pay further attention to the way in which attentional weights are distributed in categorization performance. They consider how these weights, which appear to be adjusted to optimize classification performance, are adjusted in dealing with a new categorization task. One possibility is that differential weights reflect selective attention during learning, which may affect the precision with which different aspects of the stimulus are remembered. Another (and compatible) possibility is differential weights could result from an active decision process during the classification of new stimuli. To investigate this question, Lamberts and Chong studied very short-term changes in dimension weights. In their experiments, they attempted to get participants to employ an active decision process, by asking some subjects to pay attention to particular stimulus features of novel items, where other subjects are instructed to attend to all features. They found that people can flexibly adjust the degree to which they weight stimulus dimensions, and that this may influence their categorization decisions, favouring the view that weight changes are mediated by an active decision process. Moreover, this work suggests

that 'optimal' weights, as discussed by Nosofsky, may have the status of defaults, which can overridden by experimental instructions. They suggest that the flexibility of dimension weights may provide one important way in which general background knowledge can influence categorization.

Part IV of this book is on reasoning. This is an area of cognition that had not seemed amenable to rational analysis. However, the chapters in this section reveal that rational analysis is proving vital to understanding the apparent errors and biases that have putatively been detected in this area of research over the last 40 years. In Chapter 14, *Causal mechanism and probability: a normative approach*, Glymour and Cheng consider the rational basis for people's judgements about causality. They argue that much recent research assumes that there is a fundamental divide between mechanistic and probabilistic analyses of causal inference. Mechanistic analyses describe the causal sequence that relates causes and effects. Probabilistic accounts analyse causal relationships in terms of the probabilities of effect given cause, effect in the absence of the cause, and similar notions. Glymour and Cheng argue that this is a false dichotomy, which causes researchers to overlook the nature of the evidence that supports the induction of mechanisms and to miss some important probabilistic implications of mechanisms. Moreover, they claim that this dichotomy has blocked the development of an alternative conception of how people learn the causal structure of their world: for discrete events, a central adaptive problem is to induce causal mechanisms in the environment from probabilistic data and prior knowledge. Viewed from this perspective, they show that the probabilistic norms assumed in the human causal judgement literature often do not map on to the mechanisms generating the probabilities. Their alternative conception of causal judgement is, they argue, more congruent with both scientific uses of the notion of causation and observed causal judgements of untutored reasoners. They illustrate some of the relevant variables under this conception, using a framework for causal representation now widely adopted in computer science and, increasingly, in statistics. They also review the formulation and evidence for a theory of human causal induction (Cheng, 1997) that adopts this alternative conception.

In Chapter 15, *The rational analysis of human contingency judgement*, López *et al.* present evidence that in causality judgements people are sensitive to factors that cannot be explained by a rational model. Learning such contingencies from sequences of discrete trials has been modelled using the Rescorla–Wagner (R-W) model which relies on gradual updating of associative strength between representations of the cause and the effect. Recently, Cheng (1997, see also Chapter 14) showed that, at asymptote, the R-W model computes the probabilistic contrast which is the normatively correct assessment of the contingency, i.e. human learning is rational (Shanks, 1995). The question remains whether the cognitive system uses something like R-W to compute the contrast, or whether it is computing this contrast directly. In this chapter López *et al.* address this question by observing people's pre-asymptotic behaviour under a variety of conditions. They observed that people's performance accorded well with the optimal model at asymptote. However, various trial-by-trial effects were also observed before asymptote that were not consistent with computing the probabilistic contrast. López *et al.* argue that these effects can only be explained if the cognitive system uses an algorithm like R-W to achieve

asymptotic behaviour that accords with the probabilistic contrast. López *et al.* also present experiments that seem to show that various probabilistic reasoning biases, the conjunction fallacy and cue competition effects, may be the result of the associative mechanisms underlying human learning.

Most of the chapters in this book concern the rationality of particular cognitive processes. In Chapter 16, *Rationality assumption of game theory and the backward induction paradox*, Colman, by contrast, considers rationality in the context of the interaction between individuals. 'Rational choice' theories of economic and social phenomena have been widely influential and highly controversial throughout the social sciences (see, e.g. Elster, 1986)—such theories are, in essence, rational analyses at the level of interpersonal or group phenomena. Indeed, the project of rational analysis in cognitive science may be viewed as an attempt to apply this style of explanation to explaining processes within the individual. Colman considers the fundamental issue of the viability of game theory as an appropriate normative and descriptive model of multiperson interactions. He considers a celebrated paradox of game theory—the backward induction paradox (induction here refers to the proof technique of mathematical induction, rather than inductive inference). Colman illustrates the paradox using what is known as the two-person centipede game (Rosenthal, 1981). On successive trials two players alternate in choosing whether to stop the game or to continue it. If a player stops the game, then there are no pay-offs to either player. But whenever a player chooses to continue, that player is fined £1 and the other player is rewarded £10. The game must finish in any case after a fixed number of moves. The principle of backward induction sanctions the following line of reasoning: suppose player A is scheduled to make the last move if the game were to run through to the end. On that final move, B is fined £1, so it is rational for B to stop the game on the move before. But this means that A, knowing that B will do this, should stop the game on the move before, and so on all the way back to the recommendation that the game should be stopped on the very first move. This seems paradoxical, because by following this 'rational' course of action both players receive no pay-off, whereas they could both have received a substantial sum of money by cooperatively continuing the game. Colman's proposed solution to the paradox is to argue that the cooperative strategy *is* rational, given the fact that the people making the decisions have imperfect information about each other, and must reason under uncertainty. He shows that, under these conditions, the backward induction argument can be blocked, and the cooperative strategy may be rational after all.

Oaksford and Chater (1994) presented a rational analysis of human hypothesis testing in Wason's selection task. Illogical performance on this task has been taken to argue for human irrationality. Oaksford and Chater's analysis showed for the first time that behaviour on this task can be viewed as rational. However, Oaksford and Chater's optimal data selection account has been criticized for failing to model aspects of the data where people know that there are exceptions to a rule under test, e.g. when they know that there are some non-black ravens when testing the hypothesis that all ravens are black. More recently Oaksford *et al.* (1997) have shown that in the context of sequential sampling participants make some data selections that appear not to accord with the theory of optimal data selection. In Chapter 17, *A revised rational analysis of the selection task: exceptions and sequential sampling*, Oaksford and Chater clarify both issues. First, they show that allowing for the possibility of exceptions in their optimal data selection not only fails to alter the predictions of their model but also allows it to capture data previously thought to count against it. Second, they argue that the sequential sampling situation allows participants to update their estimates of the probabilities of categories used in the test rules on-line. Given the structure of the samples used in the experimental tasks they show that such on-line updating would have the effects observed by Oaksford *et al.* (1997).

In Chapter 18, *Rational analysis of causal conditionals and the selection task*, Over and Jessop make the important connection between recent research using the selection task (see Chapter 17) and causal reasoning (see Chapter 14), a connection first suggested by Oaksford and Chater (1994). Over and Jessop argue that apparent biases in causal judgements using 2×2 contingency tables such that information from different cells is differentially weighted correspond to apparent biases in selection task performance. Consequently, if selection task performance can be viewed as rational (see Chapter 17) and if data selection is operating according to similar principles in both tasks then causal inference too can be viewed as rational. They show that current measures of the informativeness of evidence converge on the intuitively correct data selections when a causal rule is used in the selection task. An important development is the use of the principle of maximum entropy in the construction of alternative hypotheses. This is a more satisfactory method than the minimal change strategy adopted by Oaksford and Chater (1994) even if it is well justified (see Oaksford and Chater, 1996).

Researchers have frequently bemoaned the artificiality of laboratory problem solving tasks because they do not appear to capture the open-ended, probabilistic character of real world problems. Anderson (1990) observed that much real world problem solving is guided by its environmental context. In Chapter 19, *The practice of mathematics and science: from calculus to the clothesline problem*, Kurz and Tweney make the crucial observation that in problem solving there is a two-way interaction between the environment and the problem solver because people can actively construct their own environments. For example, wheeled vehicles are only an adaptive solution to the problem of more rapid locomotion once we have built roads or railway tracks. Kurz and Tweney are principally concerned with scientific reasoning. They first review recent work on the cognitive psychology of science which often involves analyses of scientists' diaries detailing their experimental work. They point out that scientists' problem solving can be characterized as an interplay between a conceptual space containing the hypotheses and expectations under investigation and a perceptual space that may require active re-configuration in order to allow relevant observations. They illustrate these points with reference to Faraday's work on electromagnetic induction and on acoustics. Kurz and Tweney then investigate peoples' constructions of representations in order to solve a problem that requires setting up and solving a differential equation. Different problem solvers adopted different representations of the environment and hence different interpretations of the calculus. Kurz and Tweney conclude that actively structuring the environment either actually, as in experiment, or conceptually may alter the course of problem solving.

The fifth and last part of this book is on search. Chater *et al.*'s Chapter 20, *The rational analysis of inquiry: the case of parsing*, has two parts. The first provides a general framework for thinking about the rational analysis of inquiry—what is a rational way of devoting resources to obtaining information. The second describes an extended case study—a rational analysis for deciding on the order in which different possible readings of locally syntactic ambiguous structures should be considered by the parser.

The problem of deciding how to search for information arises, in different forms, in several chapters of this book (Oaksford and Chater's and Over and Jessop's chapters on the selection task, Young's and McGonigle and Chalmers' chapters on search) and is central to many areas of cognition. The rational analysis of inquiry is difficult because inquiry typically proceeds in the face of severe resource limitations. Therefore choices regarding which inquiries to make must be highly selective—but it is difficult to know which inquiries are likely to prove fruitful *before* they have be conducted. Chater *et al.* distinguish between disinterested inquiry, where the goal is finding out as much about the world as possible, and cases where the goal is provided by some external set of utilities, and inquiry is required simply to determine how to decide and how to act to maximize these utilities. They also distinguish between cases where the gathering of new information must be selective and cases where the computationally limited cognitive system must decide between different possible computations that it can make with existing information. They show how a simple set of tools from probability theory and information theory can be used to capture each kind of problem, and reveal the relationships between different classes of problem.

The rational analysis concerns how a serial parser can deal with the massive local syntactic ambiguity in natural language. A serial parser can only explore one option at a time—and presumably the more options that are explored erroneously the more likely the system is to 'crash' irretrievably. The rational analysis aims to find the optimal way of exploring parses to minimize the probability of a 'crash' and thus maximize the probability of a successful parse of the whole sentence. The analysis shows that, under certain assumptions, the parser should select hypotheses in order of their 'specificity' (roughly, how specific their predictions are about future context) multiplied by their prior probability. This recommendation differs from many traditional accounts of parsing, which assume that parsing decisions are based on linguistic principles such as 'minimal attachment' (Frazier, 1979), and also appears to differ from constraint-based approaches to parsing (e.g. MacDonald *et al.*, 1994), which only take account of prior information. Testing this rational analysis empirically is an interesting project for future research.

In Chapter 21, *Rational analysis of exploratory choice*, Young considers the problem of sequential search from a different perspective—starting from Anderson's (1990) rational analysis of problem solving. He extends this style of analysis to a class of exploratory search situations which involve selecting one of a number of possible options, where little is known about the options before exploration begins. He formulates the situation as one of single-move, multistage search, where the probabilities of success of the different options are initially unknown, but where a range of assessment methods is available to provide information about each option.

The assessment methods differ in their costs and in the quality of the information they deliver. The analysis defines an optimal strategy, which is applied to an experimentally studied task where a subject has to use an unfamiliar computer package. Simulation of the optimal strategy shows that it exhibits a number of features characteristic of the empirical data, such as repeated scanning of the menus, progressive focusing on a subset of the options, and iterative deepening of attention.

The core of McGonigle and Chalmers' Chapter 22, *Rationality as optimized cognitive self-regulation*, is an account of their recent comparative and developmental work on search behaviour. They argue that organisms can exploit the structure of their environment in crucial ways in the attempt to minimize the cognitive and physical costs of search. This is a particularly pressing problem for an organism when foraging for food—it does not want to return to previous locations, i.e. it wants to search non-reiteratively, in order to minimize costs. Planning non-reiterative search can also be combinatorially explosive, as the number of possible search paths increases exponentially with the number of locations to be searched. McGonigle and Chalmers show that monkeys and human neonates adopt very similar strategies: they use the features of search locations to minimize these costs. For example, if search locations are ordered by size or brightness they are capable of exploiting these properties to regulate their search behaviour, e.g. search the smallest, then the next smallest and so on. McGonigle and Chalmers argue that such skills provide the primitive operations of higher level cognitive skills, for example, this simple ordering behaviour may provide the learned primitives for transitive reasoning: if $A > B$ and $B > C$, then $A > C$.

This book covers a broad spectrum of the core research areas of cognitive psychology. In each area, the work reported here reveals the wealth and depth of current research adopting the rational analysis approach. This approach has illuminated research in all these areas revealing that to a close approximation human cognition seems well, if not always optimally, adapted to the environment. Several chapters have also shown that the algorithms that implement these rational models also need to be invoked in explaining the detailed patterns of performance. We believe that continued research on rational models of cognition represents the most promising and fruitful approach currently available in cognitive psychology and hope that this collection will prove an encouragement to others to develop their own rational analyses of cognitive phenomena.

Acknowledgements

We would first like to thank the authors of the chapters in this volume. All have been immensely helpful and responsive in getting their contributions to us promptly so that we have had little trouble meeting publisher's deadlines. We would also like to thank the editorial team at Oxford University Press for their support while producing this book. We would particularly like to thank Gordon D. A. Brown for his support and assistance.

Many of the chapters included in this volume were presented in preliminary form at the Conference on Rational Models of Cognition organized by the editors of this

book at the University of Warwick, UK, in July 1996. We acknowledge grants received in support of this conference from the Cognitive Psychology Section of the British Psychological Society, the University of Warwick Research Fund, and the British Academy. We also thank Greg Jones of the Department of Psychology, University of Warwick, for his support and help, and Chris Richley and Becki Grainger for their help in organizing and running the conference.

We would also like to thank our respective families for their forbearance and understanding during the production of this book, MO: Julia, Joanne and David Oaksford; NC: Louie and Maya Fooks.

References

Anderson, J. R. (1976). *Language, memory, and thought*. Lawrence Erlbaum, Hillsdale, NJ.

Anderson, J. R. (1983). *The architecture of cognition*. Harvard University Press, Cambridge, MA.

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

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

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

Chater, N. (1995). Neural networks: the new statistical models of mind. In *Connectionist models of memory and language* (ed. J. P. Levy, D. Bairaktaris, J. A. Bullinaria, and P. Cairns). UCL Press, London.

Chater, N. and Oaksford, M. (1996). The falsity of folk theories: implications for psychology and philosophy. In *The philosophy of psychology*, (ed. W. O'Donohue and R. Kitchener) pp. 244–56. Sage Publications, London.

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

Elster, J. (ed.) (1986). *Rational choice*. Basil Blackwell, Oxford.

Frazier, L. (1979). On comprehending sentences: syntactic parsing strategies. Ph.D. Thesis, University of Connecticut. Indiana University Linguistics Club, West Bend, IN.

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

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

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

Gluck, M. A. and Bower, G. H. (1986). From conditioning to category learning: an adaptive network model. *Journal of Experimental Psychology: General*, **117**, 227–47.

Goodman, N. (1955). *Fact, fiction, and forcast*. Harvard University Press, Cambridge, MA.

Hilbert, D. (1925). Über das unendliche. *Mathematische Annalen*, **95**, 161–90.

Kahneman, D. and Tversky, A. (1973). On the psychology of prediction. *Psychological Review*, **80**, 237–51.

Kahneman, D., Slovic, P., and Tversky, A. (eds.) (1982). *Judgement under uncertainty: heuristics and biases*. Cambridge University Press, Cambridge.

MacDonald, M. C., Pearlmutter, N. J., and Seidenberg, M. S. (1994). Lexical nature of syntactic ambiguity resolution. *Psychological Review*, **101**, 676–703.

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

McClelland, J. L. and Chappell, M. (1996). Familiarity breeds differentiation: a Bayesian approach to the effects of experience in recognition memory. Unpublished Manuscript. Department of Psychology, Carnegie Mellon University.

Murnane, K. and Shiffrin, R. (1991). Interference and the representation of events in memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, **17**, 855–74.

Nosofksy, R. M. (1984). Choice, similarity, and the context theory of classification. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, **10**, 104–14.
Nosofksy, R. M. (1986). Attention, similarity, and the identification-categorization relationship. *Journal of Experimental Psychology: General*, **115**, 39–57.
Oaksford, M. and Chater, N. (1994). A rational analysis of the selection task as optimal data selection. *Psychological Review*, **101**, 608–31.
Oaksford, M. and Chater, N. (1996). Rational explanation of the selection task. *Psychological Review*, **103**, 381–91.
Oaksford, M., Chater, N., Grainger, B., and Larkin, J. (1997). Optimal data selection in the reduced array selection task (RAST). *Journal of Experimental Psychology: Learning, Memory and Cognition*, **23**, 441–58.
Ratcliffe, R. Clark, S., and Shiffrin, R. (1990). The list strength effect: I. Data and discussion. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, **16**, 163–78.
Rosenthal, R. W. (1981). Games of perfect information, predatory pricing and the chain-store paradox. *Journal of Economic Theory*, **25**, 92–100.
Shanks, D. (1995). Is human learning rational? *Quarterly Journal & Experimental Psychology*, 48A, 257–79.
Simon, H. A. (1955). A behavioral model of rational choice. *Quarterly Journal of Economics*, **69**, 99–118.
Simon, H. A. (1956). Rational choice and the structure of the environment. *Psychological Review*, **63**, 129–38.
Stich, S. (1985). Could man be an irrational animal? *Synthese*, **64**, 115–35.
Stich, S. (1990). *The fragmentation of reason*. Cambridge, Cambridge, MA: MIT Press.
Sutherland, S. (1992). *Irrationality*. London: Constable.
Wason, P. C. (1966). Reasoning. In *New horizons in psychology*, (ed. B. Foss) pp. 135–51, Penguin, Harmondsworth, Middlesex.
Wason, P. C. and Johnson-Laird, P. N. (1972). *The psychology of reasoning: Structure and content*. Harvard University Press, Cambridge, MA.

Part I

General issues