Rational Models of Cognition

Intermediate article

Nick Chater, University of Warwick, Coventry, UK
Mike Oaksford, Cardiff University, Cardiff, UK

CONTENTS

Rational models of cognition attempt to explain the function or purpose of cognitive processes.

CONSTRAINTS ON MODELS OF COGNITION

A scientific explanation of psychological, biological, or social phenomena can take one of two complementary forms. The first is mechanistic: phenomena are explained by analysing their internal causal structures. The second is purposive: phenomena are explained in terms of their purpose, what problems they solve.

In biology, purposive explanation concerns the function of biological structures and processes (e.g. the function of the heart is to pump blood). The same style of explanation is applied to animal behavior (e.g. the function of building nests is to provide a safe shelter for eggs). In the social sciences, 'rational choice' explanation views people as having the purpose of maximizing their 'utility', given the constraints imposed by their environment. Moreover, in everyday life, we explain each other's behavior by giving reasons for why this behavior 'makes sense', given our desires and our beliefs.

In cognitive science, however, mechanistic explanation has been predominant. Computational models, whether symbolic or connectionist, have focused on specifying architectures and algorithms for cognition; and experimental work has been oriented towards mechanistic questions, such as the limits of human memory, or the number of, and interconnections between, memory stores. The picture of the cognitive system that emerges from this focus on mechanistic explanation is as an assortment of apparently arbitrary mechanisms, subject to equally arbitrary limitations, with no apparent rationale or purpose.

By downplaying purposive explanation of cognition, cognitive science may have been missing an essential source of constraints on cognitive models: namely, that in many domains, cogntion appears to be extremely well adapted to the challenges that it faces. In perception, motor control, language processing, common-sense reasoning and decision-making, the cognitive system reliably (though not infallibly) handles perceptual and cognitive problems of great complexity, typically under conditions of uncertainty. The cognitive system can learn to deal with a remarkably broad range of challenges, both natural and artificial, from unicycling to backgammon to musical composition. And the cognitive system acquires, stores and retrieves a rich understanding of the everyday world. It seems plausible that, as for other biological structures, this success is not accidental. It seems more likely that the cognitive system is superbly adapted to serve practical and computational ends. Thus, cognitive models should, ideally, not just fit the empirical data, but also, where possible, make sense as solutions to adaptive problems that the cognitive system faces.

CHARACTERIZING RATIONALITY IN HUMAN COGNITION

Rational models of human cognition aim to explain the function or purpose of human behavior or the cognitive processes underlying it. An idealized methodology for providing such explanation is given in Anderson's (1990) notion of 'rational analysis'. This methodology has six steps:

1. *Goals*. Specify precisely the goals of the cognitive system.
2. *Environment*. Develop a formal model of the environment to which the system is adapted.
3. *Computational limitations*. Make minimal assumptions about computational limitations.
4. *Optimization*. Derive the optimal behavior function.

5. *Data*. Examine the empirical evidence to see whether the predictions of the behavior function are confirmed.
6. *Iteration*. Repeat, iteratively refining the theory.

The idea is that a rational model explains behavior as an optimal (or nearly optimal) attempt (step 4) to achieve certain goals (step 1), in the context of a particular environment (step 2), and with possibly limited computational resources (step 3). The project is empirical, in two senses. First, the goals, environment, and computational limitations can only be determined empirically. Second, the goal of a rational analysis is to explain patterns of empirical data. So, an optimal system for some aspect of categorization or reasoning is only of interest if it captures empirical data on how people do categorize or reason. As with any empirical scientific project, there may be a continuous adjustment of all the elements of the explanation, in order to obtain the most compelling relationship between theory and data (step 6).

How can this 'rational' style of explanation relate to, and potentially constrain, a mechanistic cognitive model? The answer is that the mechanistic cognitive model can implement the computations specified by the rational model (or, at least, some approximation to them). Thus, building a rational model complements, rather than displaces, traditional mechanistic modeling in cognitive science.

Rational models have been developed, more or less independently, in a number of contexts. One tradition, mentioned above, is 'rational choice' explanation, which, in its classical form, assumes that individuals make decisions in order to maximize their expected utility (or, in some biological contexts, to maximize their number of viable offspring). Rational choice explanation is the foundation of modern economics, and has applications in animal behavior, sociology, and political science. In cognitive science, rational models have been developed for specific cognitive processes in perception, categorization, reasoning, problem solving, memory, and language processing, rather than for the whole individual. We will discuss work in this tradition below; related approaches have also been developed independently in the study of vision (e.g. likelihood and simplicity models in perceptual organization, ideal observer models, and the computational level of explanation (Marr, 1982; Pomerantz and Kubovy, 1986)).

Many rational models, including those described below, use a particular theorem of probability, Bayes' theorem. Given two states A and B, the *joint* probability $P(A\,\&\,B)$ is the probability that both A and B are true; and the *conditional* probability of A given B, written $P(A\,|\,B)$, is the proportion of the probability associated with B that is also associated with A. So by definition, $P(A\,|\,B) = P(A\,\&\,B)/P(B)$ and $P(B\,|\,A) = P(A\,\&\,B)/P(A)$. Putting these together, and rearranging, we obtain Bayes' theorem:

$$P(B|A) = \frac{P(A|B)P(B)}{P(A)} \tag{1}$$

This simple theorem has considerable application, not only in building rational models of cognition, but also in statistics and the philosophy of science.

BAYESIAN MODELS OF CATEGORIZATION

Formulating a rational model of categorization requires specifying a goal or purpose which categorization is presumed to serve. Anderson (1991) makes the natural assumption that the goal of categorization (step 1) is to predict unknown features of objects from known features. He assumes, further, that the environment consists of classes characterized by a probabilistic relationship with a set of features (step 2). Specifically, given a class C_i, the assumption is that for each feature F_j, there is a probability $P(F_j\,|\,C_i)$ that a item of category C_i has the feature F_j; and, crucially, that this probability is *conditonally independent* of the other features that item has (formally, this means that $P(F_1\,\&\ldots\&\,F_n|C_i) = \prod_{j=1}^{n} P(F_j\,|\,C_i)$). Here, we shall assume that step 3 is null: no specific computational constraints are needed. Given these assumptions, what is the optimal way of predicting unknown features from known features (step 4)?

Rather than follow Anderson's precise formulation, for clarity we follow a simpler analysis. Suppose we know that an item possesses a set of features $F_1, \ldots F_n$, and want to know $P(C_i\,|\,F_1\,\&\ldots\&\,F_n)$ for each category C_i. That is, we want to know the probability that the item belongs to category C_i. Bayes' theorem gives:

$$P(C_i\,|\,F_1\,\&\,\ldots\,\&\,F_n) = \frac{P(F_1\,\&\,\ldots\,\&\,F_n\,|\,C_i)P(C_i)}{P(F_1\,\&\,\ldots\,\&\,F_n)} \tag{2}$$

$$= \frac{P(C_i)\prod_{j=1}^{n} P(F_j\,|\,C_i)}{P(F_1\,\&\,\ldots\,\&\,F_n)} \tag{3}$$

where the simplification follows because of Anderson's crucial assumption that features are conditionally independent. Finally, suppose we want to predict an unknown feature F_{n+1}. If we knew that the item belonged to category C_i, then the probability of F_{n+1} would simply be $P(F_{n+1} \mid C_i)$. But we know $F_1, \ldots F_n$, rather than the category, so we must predict F_{n+1} by summing these conditional probabilities, weighted by the probability of each C_i, given the known features $F_1, \ldots F_n$. Thus,

$$P(F_{n+1}|F_1 \,\&\ldots\&\, F_n) = \sum_i P(C_i|F_1 \,\&\,\ldots\,\&\, F_n)P(F_{n+1}|C_i) \qquad (4)$$

Bayesian models of categorization, of various forms, have been used to capture empirical data on categorization (rational analysis step 5) (Anderson, 1991), as well being widely applied in artificial intelligence and machine learning.

BAYESIAN MODELS OF BELIEF REVISION

Bayesian models are also widely used in understanding reasoning and belief revision. In artificial intelligence, there has been a substantial shift from logical to Bayesian views of how beliefs should be revised in the light of new knowledge. According to the logical viewpoint, knowledge is encoded as a set of axioms and their deductive consequences. New knowledge (for example, derived from perception or language) is encoded in new axioms; and the new knowledge state consists of the larger set of axioms and their deductive consequences. This approach runs into difficulties where new and old knowledge appear inconsistent, because, in most logical systems, all propositions (and their negations) follow from a contradiction, leading to potential inferential chaos. There have been numerous ingenious attempts to combat this difficulty. But the Bayesian approach aims to avoid it entirely, by assuming that 'knowledge' is only probabilistic – or more accurately, by modeling belief revision in terms of probability theory. In the probabilistic framework, outright contradictions need not occur (what was previously probable simply becomes much less probable). Pearl (1988) and others have shown how to build parallel distributed computational mechanisms for probabilistic reasoning for belief revision. These models depend, crucially, on making independence assumptions between pieces of information, in just the way that we assumed above that features were conditionally independent given the relevant category. For example, effects are typically viewed as conditionally independent given their causes.

A similar shift from logic to probability theory has been advocated in the psychology of reasoning. It has been argued that various apparent experimental demonstrations of irrationality can be reinterpreted. For example, Oaksford and Chater (1994) have argued that searching instances which confirm a conditional rule 'if *A* then *B*' is rational from a probabilistic perspective, because a confirming instance can substantially raise the probability that the statement is true. Yet on a traditional viewpoint in the psychology of reasoning, searching for confirmatory evidence is misguided, because general statements cannot be logically derived from their instances – the next observation could always be a refutation. Thus, the human tendency to seek confirming evidence may appear irrational from a logical perspective, but entirely rational according to a Bayesian rational analysis. (*See* **Reasoning**)

EMPIRICAL EVIDENCE FOR AND AGAINST RATIONALITY

The case noted above highlights the difficulty of interpreting empirical evidence for or against rationality: the interpretation depends on the theoretical perspective adopted. But it might seem that rational models of cognition do not usefully contribute to the debate on whether people are rational, because they seem to assume the idea of the rationality of cognition from the outset. The approach seems to presuppose rationality, regardless of any empirical evidence that might be collected. The picture is, however, not so straightforward.

First, the dictates of a rational cognitive model will typically only be implemented approximately. These approximations will result in irrational behaviour. For example, Chater and Oaksford have given a Bayesian rational model of how people reason with syllogisms (e.g., 'all *X* are *Y*, all *Y* are *Z*, therefore, all *X* are *Z*'). Where there is a probabilistically valid conclusion for a syllogism, the heuristics generally generate it successfully; but they also generate other conclusions, giving 'irrational' answers for syllogisms where no conclusion follows.

Second, there is an important distinction between the rationality of specific cognitive processes and the rationality of the whole person, which is comprised of the interaction of innumerable cognitive processes. For example, the tendency of the cognitive system to pay attention to relative rather than absolute magnitudes may be highly adaptive in encoding information about the external world (because many aspects of the world are 'scale-invariant' (Chater and Brown, 1999)). But this may give rise to irrationality in risky decision-making, where, for example, the difference between prizes of $0 and $10 may be viewed as far less significant than the difference between prizes of $90 and $100 (Kahneman *et al*., 1982) – even though the differences are objectively the same. In general, we might conjecture that specialized cognitive processes might exhibit greater 'rationality' than the whole individual. This is because specialized processes need only be adapted to some relatively narrow class of tasks (e.g. interpreting stereoscopic disparities between the two eyes, segmenting the visual field) which has been encountered throughout an individual's life, and perhaps also through millions of years of evolutionary history. The whole person, on the other hand, must cope with an endless variety of tasks (e.g. making financial decisions), for which neither experience nor evolution may provide much guidance. If this is the case, then rational choice explanation, as described above, may seek support from human rationality just where it is weakest – a disturbing reflection from the point of view of the foundations of economics.

Third, the attempt to apply rational models of cognition can be viewed as a way of measuring the degree of rationality of the cognitive system. The rationality of thought and behavior can only be assessed against a standard of 'correct' performance. But to choose an appropriate standard of correct performance, we need to have decided what computational function the cognitive system is attempting to perform – and this is the goal of rational analysis. We cannot merely stipulate the standards against which cognition should be measured. If we do so, we run the risk of, for example, condemning people as irrational because they fail to reason logically, when they are reasoning quite rationally according to the dictates of probability, as noted above. Thus, far from presupposing human rationality, the project of building rational models of cognition should provide a test for when and to what degree people are rational.

References

Anderson JR (1990) *The Adaptive Character of Thought*. Hillsdale, NJ: Erlbaum.

Anderson JR (1991) The adaptive nature of human categorization. *Psychological Review* **98**: 409–429.

Chater N and Brown GDA (1999) Scale invariance as a unifying psychological principle. *Cognition* **69**: B17–B24.

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

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

Oaksford M and Chater N (1994) A rational analysis of the selection task as optimal data selection. *Psychological Review* **101**: 608–631.

Pearl J (1988) *Probabilistic Reasoning in Intelligent Systems*. Palo Alto, CA: Morgan Kaufman.

Pomerantz JR and Kubovy M (1986) Theoretical approaches to perceptual organization: simplicity and likelihood principles. In: Boff KR, Kaufman L and Thomas JP (eds) *Handbook of Perception and Human Performance*, vol. II 'Cognitive Processes and Performance'. New York, NY: Wiley.

Further Reading

Anderson JR (1991) Is human cognition adaptive? *Behavioral and Brain Sciences* **14**: 471–517.

Anderson JR (1994) *Rules of the Mind*. Hillsdale, NJ: Erlbaum.

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

Oaksford M and Chater N (1998) *Rationality in an Uncertain World*. Hove, UK: Psychology Press.

Oaksford M and Chater N (eds) (1998) *Rational Models of Cognition*. Oxford, UK: Oxford University Press.

Shanks DR (1995) Is human learning rational? *Quarterly Journal of Experimental Psychology* **48A**: 257–279.

Shepard RN (1987) Towards a universal law of generalization for psychological science. *Science* **237**: 1317–1323.