BEHAVIORAL AND BRAIN SCIENCES (1991) 14, 471–517
Printed in the United States of America

Is human cognition adaptive?

John R. Anderson
Psychology Department, Carnegie Mellon University, Pittsburgh, PA 15213-3890
Electronic mail: *anderson@psy.cmu.edu*

Abstract: Can the output of human cognition be predicted from the assumption that it is an optimal response to the information-processing demands of the environment? A methodology called rational analysis is described for deriving predictions about cognitive phenomena using optimization assumptions. The predictions flow from the statistical structure of the environment and not the assumed structure of the mind. Bayesian inference is used, assuming that people start with a weak prior model of the world which they integrate with experience to develop stronger models of specific aspects of the world. Cognitive performance maximizes the difference between the expected gain and cost of mental effort. (1) Memory performance can be predicted on the assumption that retrieval seeks a maximal trade-off between the probability of finding the relevant memories and the effort required to do so; in (2) categorization performance there is a similar trade-off between accuracy in predicting object features and the cost of hypothesis formation; in (3) casual inference the trade-off is between accuracy in predicting future events and the cost of hypothesis formation; and in (4) problem solving it is between the probability of achieving goals and the cost of both external and mental problem-solving search. The implemention of these rational prescriptions in neurally plausible architecture is also discussed.

Keywords: Bayes; categorization; causal inference; computation; memory; optimality; problem solving; rationality

1. A rational theory of cognition

There has been a long tradition of trying to understand human cognition as an adaptation to its environment. The writings of Brunswik (1956) and Gibson (1966; 1979) represent classic examples of this approach. More recently, Marr (1982) and Shepard (1987) have provided rigorous developments of this perspective. I (Anderson 1990a) have tried to develop what I call "rational analyses" of several broad categories of human cognition. A *rational analysis* is an explanation of an aspect of human behavior based on the assumption that it is optimized somehow to the structure of the environment. The term comes from an analogy with the "rational man" hypothesis that plays such a major role in economic explanation. As in economics, the term does not imply any actual logical deduction in choosing the optimal behavior, only that the behavior will be optimized.

This rationality thesis has often been supported with evolutionary arguments, but from my perspective the evolutionary connections are as much of a hindrance as a help. The idea that any aspect of an organism has evolved to an optimal form is quite problematical (e.g., see the readings edited by Dupré 1987). I will discuss these problems shortly; first I would like to state what draws me to the position in the absence of any strong encouragement from evolutionary considerations.

If there were some precise relationship between the structure of the environment and the structure of behavior, this would provide a much needed perspective in cognitive psychology. The principal goal of the field has been to find mental structures that explain behavior. This goal has been impeded by two related obstacles. First, there is a serious and perhaps intractable induction problem in inferring the structure of a black box from the structure of the behavior it produces. Second, there is a serious and intractable identifiability problem in that many different proposals about mental structure are equivalent in their behavioral consequences.

The rational approach helps rather directly with the induction problem. If we know that behavior is optimized to the structure of the environment and we also know what that optimal relationship is, then a constraint on mental mechanisms is that they must implement that optimal relationship. This helps suggest mechanisms, thus reducing the induction problem.

A rational theory can also provide help with the identifiability problem. It provides an explanation at a level of abstraction above specific mechanistic proposals. All mechanistic proposals which implement the same rational prescription are equivalent. The structure driving explanation in a rational theory is that of the environment, which is much easier to observe than the structure of the mind. One might take the view (and I have so argued in overenthusiastic moments, Anderson, in press) that we do not need a mechanistic theory, that a rational theory offers a more appropriate explanatory level for behavioral data. This creates an unnecessary dichotomy between alternative levels of explanation, however. It is more reasonable to adopt Marr's (1982) view that a rational theory (which he called the "computational level") helps define the issues in developing a mechanistic theory (which he called the level of "algorithm and representation"). In particular, a rational theory provides a precise characterization and justification of the behavior the mechanistic theory should achieve.

This target article discusses the general issues surrounding a rational analysis and reviews its applications to

 0140-525X/91 $5.00+.00

memory, categorization, causal analysis, and problem solving. It closes with an example of how rational analysis might be related to a mechanistic implementation.

1.1. The evolutionary perspective. A rational theory should stand on its own in accounting for data; it need not be derived from evolutionary considerations. Still, its connections with evolution are undeniable and have influenced the formulation of the rationalist program. It is accordingly worth discussing these evolutionary considerations explicitly.

Here is the simple view of evolution adopted here: At any stable point in evolution a species should display a range of variability in its traits. The differences in this range are not important enough in their adaptive value for any to have been selected. There may be some change in the species during this stable stage because of such things as genetic drift, in which the distribution of nonsignificant variability alters. Optimization might get involved if some truly novel genetic variation is created by some random mutation. Optimization is more often evoked, however, when the environment undergoes some significant change, after which the former range of traits is no longer equivalent in adaptive value. According to this view, changes in the environment are more significant than random changes in genetic code in driving evolutionary change.

If this view is approximately correct, evolution is a local optimizer. This can be understood as a kind of hill-climbing in which the set of possible traits defines the space and the adaptive value defines altitude. At a stable point in time the species is at some point or plateau of a local maximum. When there is an environmental change, the contours of the space change and the species may no longer be at a maximum. It will climb along the slope of steepest ascent to a new maximum and reside there. Extinction occurs when it is not possible for a species to adapt to the environmental changes. New species appear when different members of one species evolve to adapt to different environments. This means that the local optimum that any species achieves is a function of the accidents of its past. Maybe humans would be better adapted if they had the social structure of insects, but given our mammalian origins, there is no path to this hypothetical global maximum.

According to the hill-climbing metaphor, there are two major constraints on the predictions of an optimization analysis. One comes from the proximity structure of the space of traits and the other comes from the species' current location in that space. Only certain variations are reachable from where it is right now. Consider the case of the moths of Manchester, a standard illustration of evolutionary optimization (Kettlewell 1973): When pollution became a major factor in Manchester, the former peppered gray moth was no longer optimal in avoiding predators and a mutant black moth largely replaced it. Other conceivable morphological responses to predators are just as effective as changing color or more so. For example, one could imagine the development of offensive weapons like those of other insects. Moth mutants with offensive weapons do not occur, however, whereas color mutants do. Thus, color was a direction open for hill-climbing but offensive weaponry was not.

This means that any species or any aspect of a species is subject to constraints which depend on evolutionary history and can be arbitrary and complex. The more arbitrary and complex these constraints, the less explanatory an appeal to optimization will be. The general lesson we can take from such considerations is that in some cases much explanatory power is achieved by appealing to optimization and in other cases little. In optimal foraging theory (e.g., Stephens & Krebs 1986) we see a full range of explanatory outcomes from an optimization analysis. [See also: Fantino & Abarca: "Choice, Optimal Foraging, and the Delay-Reduction Hypothesis" *BBS* 8(3) 1985; and Houston & McNamara: "A Framework for the Functional Analysis of Behaviour" *BBS* 11(2) 1988.] My work on rational analysis is intended to explore explanatory power in human cognition.

The idea that cognition is optimized to the environment should be taken as a scientific hypothesis and evaluated by how well it organizes the data. One should not be surprised to find it successful in explaining some aspects of cognition but not others. In particular, I have held that the strong regularities in basic cognitive functions such as memory or categorization reflect a statistical optimization to the structure of the environment. Shepard has been a strong proponent of the idea that deep regularities of the mind reflect deep regularities of the world. "In view of the extended time base of biological evolution, I suppose that it would have been the most persuasive and enduring features, invariants, or constraints of the world in which we have evolved that would have become most deeply and thoroughly entrenched" (Shepard 1982, p. 50). It is interesting to ask whether these function-specific optimizations add up to an overall optimization of human cognition. I return to this question at the end of the paper.

As a final remark, it should be noted that developing a rational theory of an aspect of human cognition is a higher-risk, higher-gain enterprise than cognitive psychology's more normal endeavor of developing a mechanistic theory. It is high risk in that it may turn out that the given aspect of human cognition is not optimized in any interesting sense. In contrast, there is little doubt nowadays that human cognition is realized in a set of mechanisms. The rational approach is high gain in that we will have made a substantial discovery if we can show that certain aspects of human cognition are optimized in some interesting sense. This would be informative, whereas discovering that the mind is mechanistically realized is no longer news.

1.2. Applying the principle of rationality. So far the notion of a rational theory has been very general. I have followed a rather specific program for developing such a theory, one involving six steps (outlined in Table 1). Each step requires some discussion:

Step 1. The first step in developing a rational analysis is to specify the goals being optimized by the cognitive system. Any behavior can be seen as optimizing some imaginable goal. Thus, the mere fact that one can predict a behavior under the assumption of optimality is no evidence for a rational analysis. One must motivate the goals to be optimized. So far this has been easy because there is the strong constraint that these goals must be relevant to adaptation. If we were to analogize from other adaptionist applications, however, we would expect that,

Table 1 *Steps in developing a rational theory*

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 the minimal assumptions about computational limitations
4. Derive the optimal behavioral function given 1–3 above.
5. Examine the empirical literature to see whether the predictions of the behavioral function are confirmed
6. Repeat, iteratively refining the theory

as the cognitive analyses advance, circumstances will arise in which it becomes a significant research problem to determine what needs to be optimized. A good example of such an issue comes from optimal foraging theory where the question arises whether caloric intake should always be maximized and how to trade off caloric intake with other goals (Stephens & Krebs 1986) [See also Houston & MacNamara: "A Framework for the Functional Analysis of Behaviour" *BBS* 11(2) 1988] Although maximizing caloric intake is a good first-order rule of thumb, situations can arise which require more complicated formulations. Such considerations are a sign of maturity and success, indicating that we have reached a point where we can apply our theory to particular situations in such detail that we can raise very specific questions about what should be optimized.

Step 2. The second step is to specify the structure of the environment to which cognition is optimized. As in Step 1, any behavior can be seen as optimal in some possible environment. Thus, converging evidence is needed to support one's assumptions about the environment. It is here that the observability of the environment is a great potential strength of the rational analysis, although just how one capitalizes on this potential can vary. Sometimes one can appeal to existing scientific theory about the relevant environment, as in the rational analysis of categorization that is described later in this paper. Other times, one must resort to statistical studies of the environment, as in the rational analysis of memory. Occasionally, one must turn to plausibility arguments, as in the rational analysis of problem solving. The first approach yields the most compelling theory and the last the least compelling.

There are two problems in any attempt to characterize the environment. The first is that one's characterization is usually based on only a portion of the environment and it is uncertain whether this will generalize to other portions. The second problem concerns whether the environment that one has characterized is the right one. For example, if one takes an evolutionary perspective, one might argue that our current environment is quite different in its information-processing demands from the past environment that shaped our evolution. The solution to both of these problems should be to study the information-processing demands imposed by many aspects of the environment in many different cultures. To the extent that these studies tell the same story, generalization is justified. To the extent that different patterns appear, one must complicate one's environmental theory.

The theory of the environment that seems to arise in most situations is basically probabilistic. The information-processing implications of various environmental cues are not certain. This leads to the Bayesian character of the optimization analysis in Step 4. Many characterizations of human cognition as irrational make the error of treating the environment as being much more certain than it is. The worst and most common of these errors is to assume that the subject has a basis for knowing the information-processing demands of an experiment in as precise a way as the experimenter does. What is optimal in the microworld created in the laboratory can be far from optimal in the world at large.

An example is the matching law in operant conditioning (see Williams, 1988, for a review). Choices between two alternative behaviors generally tend to be proportional to the reinforcement allocated to each behavior. This can be derived for a momentary maximizing model (Shimp 1966) or a melioration model (Herrnstein & Vaughan 1980) in which the organism chooses the behavior with the greatest local probability of reinforcement. It turns out that such local optimization can in some circumstances lead to behavior which is globally nonoptimal (Herrnstein 1990). For example, there are experiments in which response *A* always is reinforced more than response *B*. However, the rate of reinforcement of both responses increases with the proportion of response *B*. The "optimal" behavior is to choose *B* often enough to keep the reinforcement level high but *A* often enough to enjoy its higher payoff. Organisms typically choose *A* too often to maximize this overall level of reinforcement. [See also: Rachlin et al.: "Maximization Theory in Behavioral Psychology" *BBS* 4(3) 1981.]

It has been called into question (e.g., Staddon 1987) whether organisms can be aware of such complex contingencies. It also seems inappropriate to call such behavior nonoptimal. It is nonoptimal in the experimenter's definition of the world but might actually be optimal in the organisms's definition (see Rachlin et al. 1981). To determine what is optimal from the organism's perspective, one must know the distribution of reinforcement contingencies in the organism's environment in general, the symptoms of these various contingencies, and what would be reasonable inferences about the current situation given these assumptions and the organism's current experience. In Bayesian terms, the distribution of contingencies is the prior probability, the symptoms given the contingencies are the conditional probabilities, and the inferences given the experience are the posterior probabilities. Characterizations of the organism as nonoptimal make the error of ignoring prior uncertainty and assume that the organism should have the same model of the situation as the experimenter. It is possible, given reasonable assumptions about priors, that local maximization has the highest expected value over possible contingencies in the environment. The hypothesis about the actual contingencies might have too low a posterior probability to influence behavior.

Step 3. The third step is to specify the computational constraints on achieving optimization. This is the true Achilles heel of the rationalist enterprise. It is here that we take into account the constraints that prevent the system from adopting a global optimum. As admitted earlier, these constraints could be complex and arbitrary. To the extent that this is the case, a rationalist theory would fail to achieve its goal, which is to predict behavior

from the structure of the environment rather than the structure of the mind

This potential danger has not caused problems in my own work on rational theory so far. It has only called for two global constraints on cognition. The first is that it costs the mind something to consider each alternative. A basic capacity limit on processing alternatives is quite reasonable given current conceptions of both cognition and computation. The second constraint is more uniquely human: There is a short-term memory limit on the number of things that can be simultaneously held in mind.

Step 4. The fourth step is to combine the assumptions in Steps 1–3 to determine what is optimal. Given the statistical characterization in Step 2, the optimization process becomes a Bayesian decision-making problem. Specifically, what behavior will maximize expected utility, defined as the expected goals from Step 1 minus the expected costs from Step 3? Expectations are defined in terms of the statistical structure of the environment from Step 2.

The major complication in this step is analytic tractability. While the idea of optimal behavior is precise enough, it is not always trivial to determine just what it is. In principle, it can always be determined by a Monte Carlo simulation in which we go through the consequences of different possible behavioral functions in the hypothesized environment. It would not be feasible to consider a full Monte Carlo simulation in practice, although we have done partial simulations (e.g., Anderson & Milson 1989). One consequence of the problem of analytic tractability is that many simplifying assumptions have to be made. This is a ubiquitous practice in science. Although I am fairly comfortable with the simplifying assumptions, the range of possible simplifying assumptions and their consequences should be explored. As will be noted in Step 6, most of my iterative theory construction activity has involved such exploration.

Step 5. The fifth step is to see whether subjects' behavior can be predicted from the optimal behavior. A noteworthy feature of my work to date is its reliance on existing experimental literature to test predictions of the rational theory. On the one hand, it is relatively impressive that such a scope of literature can be accommodated within the theory. On the other hand, this raises the question (worrisome to some) of whether the theory leads to novel predictions and experiments. There are many potential research directions. I have chosen to begin by using the wealth of the existing literature on human cognition. This paper will nonetheless describe a number of novel empirical tests (Phenomena 14, 16, and 21).

Step 6. The sixth step is to refine the theory iteratively. If one formulation does not work, we must be prepared to try another. Such iterations have often been seen as a sign that an adaptionist enterprise is fatally flawed (Gould & Lewontin 1979). As Mayr (1983) notes in response to Gould and Lewontin, however, iterative theory construction is the way of all science. In cognitive science we have certainly seen a long succession of mechanisms for explaining cognition. It is to be hoped that in cognitive science we understand that a theory should be evaluated by how well it does in organizing the data and not by whether it is the nth theory of that type that has been tried. Let me also add that my own experience with theory construction in the rationalist framework has been less iterative than my experience with theory construction in the mechanistic framework. This is what we would hope for – that rational considerations would provide more guidance in theory construction. The major point of iteration has been to play with approximating assumptions, trying to find a set that is reasonable and which leads to a modicum of analytic tractability. Thus, the iteration is being focused on trying alternative forms of the theory to fit the data.

1.3. Pre-empirical summary. The theory is created in Steps 1–3 in Table 1. The remaining points in that table are concerned with deriving and testing predictions from the theory. I refer to the theory created in Steps 1 through 3 as a *framing of the information-processing problem*. So far I have attempted framings for four separate aspects of cognition: memory, categorization, causal inference, and problem solving. Given that the evidence for rational analysis is based on how well we can predict behavior from the structure of the environment and not on rhetoric about the prior plausibility of an adaptionist analysis, this article reviews these four subtheories and 21 nontrivial behavioral phenomena which have been explained by rational analysis.[1] Table 2 provides a summary of these four subtheories by specifying their assumptions about the goals, the environments, and the costs. The following sections will review these subtheories.

2. A rational theory of memory

The goal of memory is to get access to needed information from the past (e.g., remembering where the car is parked in the airport parking lot). The relevant structure of the environment has to do with how the need for information tends to repeat itself in various situations. The simple cost function is that it costs something to retrieve a memory. These considerations all fit into a basic decision theory. The system can use its model of the environment and experience with any memory A to estimate a probability $p(A)$ that the memory is relevant at the current moment. If G is the value of achieving the current goal and C is the cost of considering any single memory, the optimal retrieval algorithm can be specified. A rationally designed information-retrieval system would retrieve memory structures ordered by their probabilities $p(A)$ and would stop retrieving when

$$p(A)G < C \qquad \text{(Equation 1)}$$

That is, the system would stop retrieving when the probabilities are so low that the expected gain from retrieving the target is less than its cost. This strategy guarantees retrieval episodes of maximum utility where this is defined as the probability of successful retrieval multiplied by G minus expected cost.

A basic assumption is that the probability of being needed, $p(A)$, is monotonically related to the latency and probability of recall, which are the two major dependent measures used in the memory literature. It is related to latency because memory structures are ordered according to $p(A)$, and it is related to accuracy because of the threshold on what items will be considered.

Table 2 *Summary of rational analyses of the four domains of cognition*

	Goal	Structure	Cost
Memory	Get access to needed experiences	Patterns by which need for information repeats	Retrieving a memory
Categorization	Predict features of new objects	How features of objects cluster together	Hypothesis formation
Causal inference	Predict future events	Statistical models of causal structure	Hypothesis formation
Problem solving	Achieve certain states in external world	How problems vary in difficulty and how similarity to the goal is related to distance to the goal	Internal effort in generating plans and external effort in executing plans

This means that a rational theory of memory is going to be a theory of how to estimate $p(A)$. It is assumed that, like computer information-retrieval systems (Salton & McGill 1983), human memory has two sources of information for deciding whether a memory A is going to be relevant in the current context. One is the past context-independent history of use of that memory. The other is the cues impinging on the person in the current context.[2]

Estimation can be characterized as a process of inferring a Bayesian posterior probability of being needed, conditional on these two sources of information. Formally, we are conditionalizing on a history H_A of being relevant in the past and a *set* of cues, denoted as Q. These cues will consist of a set of terms denoted by indices i. Characterized this way we can come up with the following Bayesian odds formula for estimating the conditional probability $P(A|H_A \,\&\, Q)$ which is $P(A)$ from Equation 1:

$$\frac{P(A|H_A \,\&\, Q)}{P(\bar{A}|H_A \,\&\, Q)} = \frac{P(A|H_A)}{P(\bar{A}|H_A)} \times \prod_{i \in Q} \frac{P(i|A)}{P(i|\bar{A})} \qquad \text{(Equation 2)}$$

That is, the odds ratio for item A is the product of the odds ratio for item A given history H_A multiplied by the product of the ratio of the conditional probabilities for each cue in the context. This equation makes certain assumptions about conditional independence, namely, that the degree to which A affects the probability of i in the context does not depend on A's past history or the other cues in the context. This assumption is typically made in the computer information-retrieval literature for purposes of tractability.[3]

The first term on the right side of Equation 2, $P(A|H_A)/P(\bar{A}|H_A)$, is basically a prior odds ratio for the item given its past history. This is the *history factor*. H_A will be a record of all the times A has been needed. As such, it reflects, among other things, frequency of use and how recently it was last used. The other quantities, the $P(i|A)/P(i|\bar{A})$ are the odds ratios of the conditional probabilities of the cues given that the memory is needed versus not needed. These ratios can be thought of as associative strengths between cues and memories. They constitute the *context factor*. This history and context factors will be considered in the next two subsections.

2.1. The history factor. To address the history factor, $P(A|H_A)$, we need to determine how the past history of a memory structure's use predicts its current use. To determine this in the most valid way, we would have to follow people about in their daily lives keeping a complete record of when they use various facts. Such an objective study of human information use is close to impossible. One possibility is to study records from nonhuman information-retrieval systems. Such studies have been done on borrowing books from libraries (Burrell 1980; Burrell & Cane 1982) and on accessing computer files (Satyanarayanan 1981; Stritter 1977). Both systems tend to yield rather similar statistics. It is also possible to look at various recorded subsets of the information-processing demands placed on human memory: Lael Schooler and I have examined the use of words in *New York Times* headlines and sources of messages in electronic mail. These display the same kinds of statistics as libraries or file systems. Thus, there appear to be certain universals in the structure of information presentation and use.

Burrell (1980; 1985) has developed a mathematical theory of usage for information-retrieval systems such as libraries (a similar model appears in Stritter, 1977, for file usage). His theory can be plausibly extended to human memory. It involves two layers of assumptions: First, the items (books, files, memory structures) in an information-retrieval system vary in terms of their desirability. These desirabilities vary as a gamma distribution. Second, desirabilities determine rate of use in a Poisson process. Burrell's model was augmented by Anderson and Milson (1989) with the assumption that there will be fluctuations in the desirability of items – an assumption that is true of books in libraries, words in the *New York Times*, or sources of electronic messages. The detailed application can be found in Anderson and Milson (1989) or Anderson (1990a), but, in summary, this augmented Burrell model predicts:

Phenomenon 1 Memory performance increases as a power function of the amount of practice (Newell & Rosenbloom 1981). Basically, if we look at these various sources (libraries, headlines, etc.) we find that the probability that an item will be used in the next unit of time increases as a power function of the number of times it has

been used. Thus, the memory function is a direct reflection of the environmental function.

Phenomenon 2. Memory performance decreases as a power function of delay between experience and test (Wickelgren 1976). Basically, if we look at these various sources we find that the probability that an item will be used in the next unit of time decreases as a power function of the time since last use. Again, the memory function is a direct reflection of the environmental function.

Phenomenon 3. Holding constant the number of exposures (Phenomenon 1) and the time since last exposure (Phenomenon 2), there is an effect of the spacing between the exposures (e.g., Bahrick 1979; Glenberg 1976); If an item has been used at time t, its probability of use at time $t + \Delta t$ is approximately maximal if its previous use was at time $t - \Delta t$. Again the environment shows the same relationship.

There is no other theory that can fit all three of these phenomena simultaneously. It needs to be stressed that these predictions derive from the structure of the environment and the definition of optimal memory performance. These empirical relationships can be seen directly in library borrowings, file access, *New York Times* articles, or electronic mail messages. Any system faced with the same statistics of information use and optimized in the sense defined will produce the basic human memory functions. *No additional assumptions are required.*

2.2. The context factor. The analysis of the contextual factor focuses on the quantities $P(i|A)/P(i|\bar{A})$, which are the cue strengths. Note that $P(i)$ and $P(i|\bar{A})$ are going to be nearly identical because conditionalizing on the nonrelevance of one memory structure out of millions cannot change the probability of any cue much. Thus, this discussion of cue strength will focus on the simpler form of $P(i|A)/P(i)$. Note that $P(i|A)/P(i) = P(A|i)/P(A)$. The cue strength (the ratio) thus reflects either the degree to which the context element (i) is more or less probable when a memory trace (A) is needed, or, equivalently, the degree to which it is more or less probable that a trace (A) is needed when a context element (i) is present. Intuitively, these cue strengths reflect the degree of association between the terms i and the memory structures A.

A critical question is how to estimate these cue strengths. The assumption of a rational analysis is that whatever factors should influence the estimation of $P(A|i)/P(A)$ should also show corresponding influences on memory. Thus, the associative effects in memory reflect a rational effort to adjust the estimation of need probability on the basis of statistical factors relating cues in the context to the probability that the memory will be needed. There are two obvious factors. One is based on direct experience and the other on inference. The direct experiential factor is the proportion of times that A has been needed in the past when cue i was in the context. The inferential factor is the proportion of times that memories similar to A have been needed when cue i was in the context. A memory model like SAM (Gillund & Shiffrin 1984) uses only the direct experiential factor but this is a poor basis when A is a recent memory and we have not had sufficient experience for estimating the true proportion. In that case we want to emphasize the inferential factor. In Bayesian terms, the inferential factor establishes a prior probability on the basis of similar memories which can be adjusted as the system gathers experience.

One question concerns how to measure similarity for the inferential factor. Following the lead of the work in information retrieval (Salton & McGill 1983), which is also reflected by a strong tradition in psychology, we decompose the memory trace A into a number of elements. Thus, for example, my memory of Ronald Reagan defeating Jimmy Carter might be decomposed into the elements *Ronald Reagan, defeated,* and *Jimmy Carter.* The similarity between a cue i and memory A would be defined in terms of the similarity of cue i to every element of A. Thus, if cue i were *George Bush,* its similarity to the Ronald Reagan memory could be defined in terms of the individual similarity of *George Bush* to *Ronald Reagan,* to *defeat,* and to *Jimmy Carter.* These element-to-element similarities should reflect the frequency with which a memory trace involving one element is needed when the other element is in the context. Operationally, these element-to-element similarities can be tapped by free association norms or direct ratings.

To summarize, the analysis of the contextual factor identifies direct cooccurrence of the target cue and the memory and similarity between the cue and memory components as factors which should be related to $P(A|i)/P(A)$. Given that identification, a number of phenomena from the experimental literature can be predicted, including the following:

Phenomenon 4. Recognition memories are generally poorer when cued by more frequent words (Gillund & Shiffrin 1984; Kintsch 1970). The basis for this can be seen in the ratio $P(i|A)/P(i)$. If $P(i|A)$ is held constant, as it usually is in memory experiments, there will be a negative effect of the frequency of i, which increases $P(i)$.

Phenomenon 5. Memories are poorer the more memories are associated to a particular term (Anderson 1983). The basis for this classic interferential effect is easy to see in the ratio $P(A|i)/P(A)$. If the probability of a particular memory is held constant as it usually is in a memory experiment, this ratio will vary with $P(A|i)$. The more things, A, associated with i, the lower $P(A|i)$ must be on average for any A.

Phenomenon 6. Memories are more accessible in the presence of related elements and less accessible in the presence of unrelated elements. This phenomenon occurs in many contexts but has been studied at great length in research on word priming, where words related to primes are processed more quickly than unprimed words but words unrelated to primes are processed more slowly (e.g., Meyer & Schvaneveldt 1971; Neely 1977). Because of the similarity factor, $P(A|i)/P(A)$ is greater than 1 for primes i that are related to the word. Since, the $P(A|i)$ must average to $P(A)$ and the $P(A|i)$ are greater than $P(A)$ for related primes, they must be lower than $P(A)$ for unrelated words. Thus, for unrelated words, $P(A|i)/P(A)$ will be a ratio that will lesson the estimated odds of A.

Many other phenomena are addressed in Anderson (1990a) and Anderson and Milson (1989), but the three examples above establish that frequency, interference, and priming effects all reflect adaptive estimation strategies on the part of memory. This contrasts with their typical treatment, according to which they are idiosyncrasies of the memory system. Interference (Phenomenon 5) is typically seen as a weakness. These effects,

as well as those associated with the history factor, however, reflect sensible procedures for estimating the potential relevance of a target memory rather than weaknesses or peculiarities of the memory system.

3. A rational theory of categorization

The goal assumed for categorization is to predict features of objects (e.g., predicting whether a creature will be dangerous). A category label is just another feature to be predicted. On this view, predicting that a creature will be called a tiger is no different from predicting that it will be dangerous. There has been a tendency in the empirical literature to assume that category labeling is the raison d'etre of a category. A number of investigators (e.g., Kahneman & O'Curry 1988; Mandler et al. 1988) have recently pointed out that the focus on predicting category labels may have distorted our understanding of categorization. Category formation is not equivalent to assigning labels, and it is possible to form categories in the absence of any labels (Brooks 1978; Fried & Holyoak 1984; Homa & Cultice 1984). A label is just another feature.

Having specified the goal of categorization, the next step in a rational analysis is to specify the relevant structure of the environment. Our rational theory of categorization rests on the structure of living objects produced by the phenomenon of speciation. Species form a nearly disjoint partitioning of the natural objects because they cannot interbreed. Within a species there is a common genetic pool, which means that individual members will display particular feature values with probabilities that reflect the proportion of those phenotypes in the population. Another useful feature of species structure is that the display of features within a freely interbreeding species is largely independent.[4] For example, there is little relationship between size and color in species where those two dimensions vary. Thus, the critical aspect of speciation is the disjoint partitioning of the object set and the independent probabilistic display of features within a species.

An interesting question concerns whether other types of objects display these same properties. Another common type of object is the artifact. Artifacts approximate a disjoint partitioning but there are occasional exceptions – for example, mobile homes are both homes and vehicles. Other types of objects (stones, geological formations, heavenly bodies, etc.) seem to approximate a disjoint partitioning but here it is hard to know whether this is just how we perceive them or whether there is any objective sense in which they are disjoint. One can use the understanding of speciation in the case of living objects and the manufacturer's intended function in the case of artifacts to objectively test disjointness.

I have taken this disjoint, probabilistic model of categories as the approximate structure of the environment for predicting object features. As discussed in Anderson (1990), it would be too costly computationally to calculate the exact probabilities with this model. Based on considerations of controlling computational cost, an iterative categorization algorithm (much like those of Fisher 1987 and Lebowitz 1987) has been developed that calculates approximations to the exact probabilities. The following is a formal specification of the iterative algorithm:

1. Before seeing any objects, the category partitioning of the objects is initialized to be the empty set of no categories.

2. Given a partitioning for the first m objects, calculate for each category k the probability P_k that the $m + 1st$ object comes from category k. Let P_o be the probability that the object comes from a completely new category.

3. Create a partitioning of the $m + 1$ objects with the $m + 1st$ object assigned to the category with maximum probability calculated in Step 2.

4. To predict the probability of value j on dimension i for the $n + 1st$ object, calculate:

$$Pred_{ij} = \sum_k P_k P(ij|k) \qquad \text{(Equation 3)}$$

where P_k is the probability the $n + 1st$ object comes from category k and $P(ij|k)$ is the probability of displaying value j on dimension i.

The basic algorithm is one in which the category structure is grown by assigning each incoming object to the category it is most likely to come from. Thus a specific partitioning of the objects is produced. Note, however, that the prediction for the new $n + 1st$ object is *not* calculated by determining its most likely category and the probability of j given that category. The calculation in Equation 3 is performed over all categories. This gives a much more accurate approximation to $Pred_{ij}$ because it handles situations where the new object is ambiguous among multiple categories. It will weight these competing categories approximately equally.

It remains to come up with a formula for calculating P_k and $P(ij|k)$. Since $P(ij|k)$ turns out to be involved in the definition of P_k, we will focus on P_k. In Bayesian terminology, P_k is a posterior probability $P(k|F)$ that the object belongs to category k given that it has feature structure F. Bayes' formula can be used to express this in terms of a prior probability $P(k)$ of coming from category k before the feature structure is inspected and a conditional probability $P(F|k)$ of displaying the feature structure F given that it comes from category k:

$$P_k = P(k|F) = \frac{P(k)P(F|k)}{\sum_k P(k)P(F|k)} \qquad \text{(Equation 4)}$$

where the summation in the denominator is over all categories k currently in the partitioning, including a potential new one. This then focuses our analysis on the derivation of a prior probability $P(k)$ and a conditional probability $P(F|k)$.[5]

3.1. The prior probability. The critical assumption with respect to the prior probability is that there is a fixed probability c that two objects come from the same category and this probability does not depend on the number of objects seen so far. This is called the *coupling probability*. If one takes this assumption about the coupling probability between two objects being independent of the other objects and generalizes it, one can derive a simple form for $P(k)$ (see Anderson, 1990a, for the derivation):

$$P(k) = \frac{cn_k}{(1 - c) + cn} \qquad \text{(Equation 5)}$$

where c is the coupling probability, n_k is the number of objects assigned to category k so far, and n is the total number of objects seen so far. Note that for large n this closely approximates n_k/n, which means that there is a strong base rate effect in these calculations with a bias to put new objects in large categories. The rational basis for this is presumably apparent.

We also need a formula for $P(0)$, which is the probability that the new object comes from an entirely new category. This is

$$P(0) = \frac{(1 - c)}{(1 - c) + cn} \qquad \text{(Equation 6)}$$

For a large n this closely approximates $(1 - c)/cn$, which is again a reasonable form – that is, the probability of a brand new category depends on the coupling probability and the number of objects seen. The greater the coupling probability and the more objects, the less likely it is that the new object comes from an entirely new category.

3.2. Conditional probability. In the case of the conditional probability, the critical assumption, based on our analysis of speciation, is that the probability of displaying features on various dimensions given category membership is independent of the probabilities on other dimensions. Thus we can write

$$P(F|k) = \prod_i P(ij|k) \qquad \text{(Equation 7)}$$

where $P(ij|k)$ is the probability of displaying value j on dimension i given that the object comes from category k. This is the same quantity that appeared in Equation 3. The importance of Equation 7 is that it allows us to analyze each dimension separately.

Standard Bayesian models (Berger 1985) can be used to calculate a posterior density of probabilities and the mean of this density. There are different solutions for discrete and continuous dimensions. I will present the discrete case here. Anderson and Matessa (1990) can be consulted for the more complex continuous case. In the discrete case we have:

$$P(ij|k) = \frac{n_{ij} + \alpha_j}{n_k + \alpha_0} \qquad \text{(Equation 8)}$$

where the α_j are parameters reflecting our priors, $\alpha_0 = \Sigma \alpha_j$, n_k is the number of objects in category k which have a known value on dimension i, and n_{ij} is the number of objects in category k with the value j. α_j/α_0 reflects the prior probability of the value j and α_0 reflects the strength of belief in these priors. For large n_k $P(ij|k)$ approximates n_{ij}/n_k which one frequently sees promoted as the rational probability. It has to have this more complicated form to deal with problems of small samples, however. For example, if one had just seen one object in a category and it had the color red, one would not want to guess that all objects were red. If there were seven colors with the α_j for each equal to 1.0, the above formula would give $\frac{1}{4}$ as the posterior probability of red and $\frac{1}{8}$ for the other unseen six colors.

Basically, Equations 7 and 8 define a basis for judging how similar an object is to the category's central tendency.

3.3. Empirical phenomena. The algorithm and mathematical relationships just described have allowed us to simulate a large number of experimental results in the literature including:

Phenomenon 7. Subjects more reliably assign an object to a category the closer it is to the central tendency of the category (Hayes-Roth & Hayes-Roth 1977; Posner & Keele 1968; Reed 1972). This is in direct response to Equation 8 and variations on it which indicate that the conditional probabilities should be higher for closer objects.

Phenomenon 8. Although test objects tend to be more reliably categorized the closer they are to the central tendency of the category, there is also an effect of their distance from specific members of the category (Medin & Schaffer 1978). This is demonstrated by having subjects study instances that are distant from the central tendency. Subjects will do well in categorizing test items that are similar to the distant instance. The way the model handles this is to create two categories – one for the instances that define the central tendency and one for the oddball instance. This is done because it maximizes the overall predictive structure of the object set.

Phenomenon 9. Lately, much has been made of the fact that the probability of categorization is an exponentially decaying function of the number of features on which the test instance mismatches the category (Russell 1986; Shepard 1989). This is produced in the rational model by Equation 8, which makes conditional probability a product (not a sum) of the probabilities on individual dimensions.

Phenomenon 10. Subjects' categorization is sensitive to the number of objects in a category such that they have a greater tendency to assign objects to categories with larger membership (e.g., Homa & Cultice 1984). This sensible base rate effect is produced by Equation 5, which defines the prior probability of belonging to a category. It turns out that the model also predicts some of the subtle deviations from base rate effects reported by Medin and Edelson (1988).

Phenomenon 11. Subjects are sensitive to the correlation of features in an experimenter's category (Medin et al. 1982). The model predicts this despite the fact that it treats dimensions in its internal categories as independent. When there are correlated features it breaks out different categories to correspond to each combination of correlated values. This is just one example of many where the model does not observe the labeling conventions of the experimenter in deciding what category structure maximizes prediction.

Phenomenon 12. Subjects tend to identify a category structure in which the categories are sufficiently specific to pick up most of the predictive structure of the objects but are not unnecessarily narrow. Rosch (e.g., Rosch et al. 1976) called these the basic level categories. This model forms categories that correspond to the basic level categories subjects identify (Hoffman & Ziessler 1983; Murphy & Smith 1982) because this does maximize the predictive structure of the categories.

For details of these phenomena and others see Anderson (1990a; in press). The evidence seems quite compelling that subjects' categorization behavior can be seen as an attempt to optimize their predictions about the features of objects. Shepard (1987) also analyzes generalization phenomena as the consequence of trying to optimize

prediction. Although his method of analysis is quite different, he comes to very similar conclusions. This is comforting, because an optimization analysis should not depend on the details of the optimization methodology. What is optimal for achieving a goal in a particular environment should be the same under all carefully reasoned analyses.

4. A rational analysis of causal inference

The analysis of causal inference is quite parallel to the analysis of categorical inference. In the case of causal inference, the assumed goal of the system is to maximize the accuracy of its predictions about future events (e.g., whether a light will go on when a switch is flipped). To do this, the system must extract the laws that govern the world's predictive structure and recognize the situations in which these laws are apt to apply. These laws are serving the same basic role in making predictions about events as categories were in making predictions about objects.

There are a number of ways causal inference is more complicated and the rational analysis correspondingly less complete. For one, it is not entirely clear what the best conception is of a "causal law." I have opted for such rather traditional situation-action rules as, "If the forest is dry and lightning strikes, there will be a fire." Such laws are necessarily probabilistic. Another complication is that generalizing these laws requires generalizing over relational structures and the problem of relational generalization is notably more difficult than categorical generalization.

A situation can be conceived as presenting a set of cues, C, that might be relevant to predicting that an event, E, will occur (Einhorn & Hogarth 1986). The prediction task is to come up with a probability $P(E|C)$ of an event E conditional on the cues C. The relevant intervening constructs are the set of causal rules, i, that we have inferred. $P(E|C)$ can be calculated by the following rule:

$$P(E|C) = \sum_i P(i|C)P(E|i) \qquad \text{(Equation 9)}$$

where $P(i|C)$ is the probability of rule i applying in the presence of cues C and $P(E|i)$ is the probability of event E should rule i apply. This rule is the analog of Equation 3 for categorization.

Equation 9 focuses on $P(i|C)$, the probability of a causal law applying in a situation and $P(E|i)$, the probability of an event should a causal law apply. $P(E|i)$ will be basically derived from how often the event occurs when the rule is applicable. Analogous to Equation 3 the relevant equation for $P(i|C)$ is:

$$P(i|C) = \frac{Con(i)P(C|i)}{\sum_i Con(i)P(C|i)} \qquad \text{(Equation 10)}$$

where $Con(i)$ denotes the confidence that i exists and $P(C|i)$ is the probability of cues C if rule i did apply. Note $Con(i)$ plays the role of a prior probability but it is not a prior probability. Rather than reflecting a probability that the rule will apply it reflects a confidence in the existence of the rule.

Before turning to the empirical phenomena, a brief comment is required on the treatment of causality as probabilistic. There is a long standing tradition in philosophy, extending back to Hume and Kant, of treating causality as deterministic. I am relatively naive with respect to the underlying philosophical issues, but I am inclined to agree with Suppes's (1984) characterization of nonprobabilistic treatments of causality as fundamentally incoherent. More to the point, however, our enterprise is not directly concerned with philosophical issues but rather with how to maximize predictions about events in the future. Whatever one's conception of the nature of causality, I assume one would be hard pressed not to take a statistical and hence a probabilistic approach to the prediction task.

4.1. Empirical phenomena. This analysis identifies a number of quantities as relevant to causal inference – $Con(i)$, $P(C|i)$, and $P(E|i)$. $Con(i)$ is one's confidence in a causal rule such as "smoking causes lung cancer." $P(C|i)$ is the conditional probability of a cue C given that law i applies. An example might be the probability that one would see yellow teeth in someone to whom the law applies (i.e., a smoker). $P(E|i)$ is the probability of getting the effect when the law applies – that is, the probability that someone will have lung cancer because of smoking. We have been able to examine a number of empirical phenomena relevant to these quantities and to show their rational basis:

Phenomenon 13. With respect to $Con(i)$, there are a number of studies of human causal inference given 2×2 contingency data in which subjects experience an effect or the absence of an effect in the presence or absence of a purported cause (e.g., Arkes & Harkness 1983; Crocker 1981; Schustack & Sternberg 1981). Human inference has often been characterized as nonoptimal in such situations because people are more influenced by changes in the joint cooccurence of cause and effect than changes in the joint nonoccurrence of cause and effect. It is assumed that subjects should be symmetrically sensitive. This assumption depends, however, on subjects' prior beliefs about the frequency of an effect in the presence of a cause compared with its absence. If subjects are much more certain about what the frequency is in the absence of a cause than in its presence, an asymmetry is predicted that corresponds to what is empirically observed. It seems reasonable to assume that subjects will assume that any particular event should have a probability near zero in the absence of a cause whereas they should be uncertain just how strong the relationship is in the presence.[6] Anderson (1990a) reports very good data fits to human causal attributions given these assumptions. Thus, what has been characterized as nonrational may in fact be optimal given appropriate assumptions about prior probabilities.

Phenomenon 14. With respect to $P(C|i)$, there has been a long history of concern with the effect of cues about the temporal and spatial contiguity between cause and effect. At least since Hume's writings there has been the belief that humans reflexively see causality in cases of close spatial and temporal proximity. In joint research with Mike Matessa and Ross Thompson, however, we find that how subjects use these cues depends on the particular prior model they bring to the situation.

Because contiguity has been so closely tied to thinking about causality and since our research on this is not yet

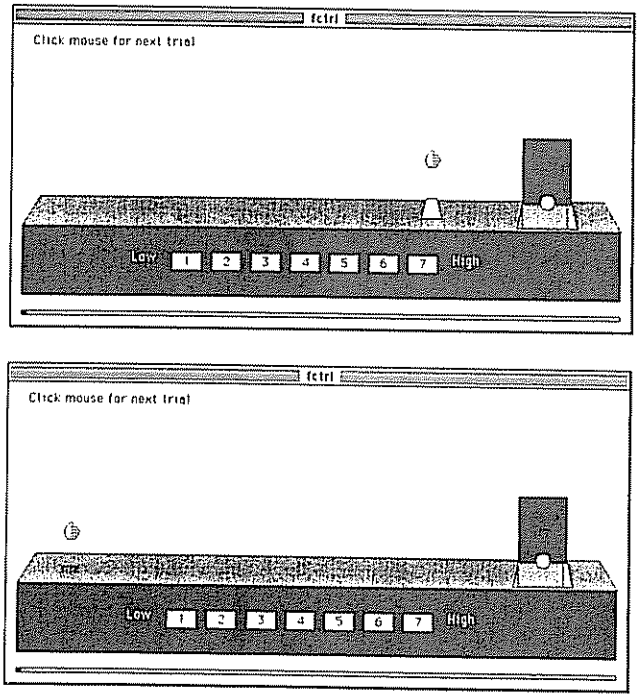

Figure 1 The stimulus situation for judging causality: (a) when a weight is dropped and a vibratory wave model is invoked and (b) when a ball is dropped and a projectile model is invoked

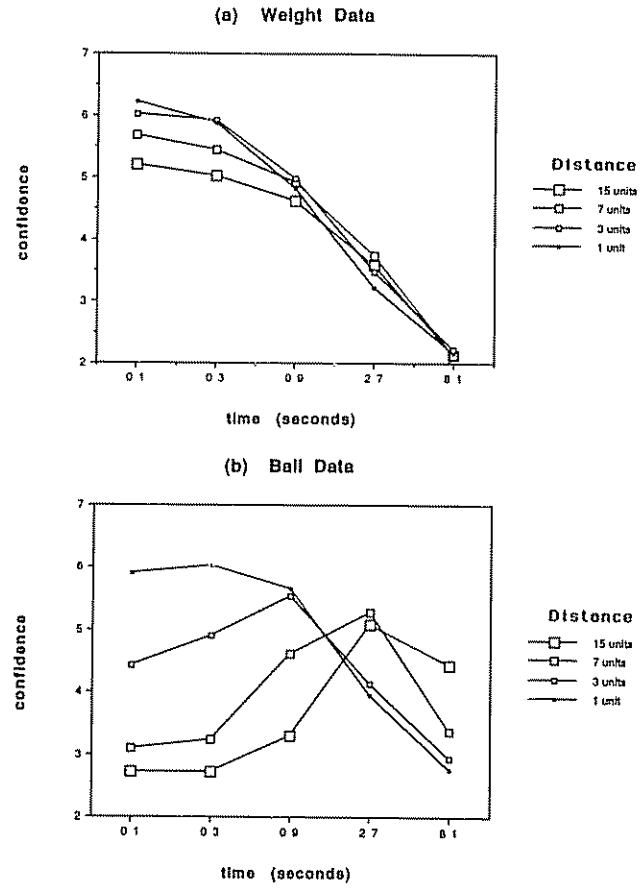

Figure 2 Strength of causal perception as a function of distance in space and time for (a) the weight and the vibratory wave model and (b) the ball and the projectile model

published, we will describe it in a little detail. Figure 1 shows the two experimental situations we used with our subjects: In all cases, the event to be explained is a trap door opening and a ball rising out of the hole at the right of the box. Prior to this, an event occurs to the left of the hole. The subject's task is to judge whether or not the first event appears to be the cause of the door opening and the ball coming out. In Figure 1a, a hand lets go of a weight and it drops onto the box. Subjects are told that it may have jarred loose a latch that opens the door and releases the ball. Subjects are asked to judge how likely it is that the weight is responsible. In Figure 1b, subjects see a hand drop a ball into a hole and are told that there might be a passage in the box through which it goes and comes out at the trap door. The time between the dropping of the weight or ball and the opening of the box is varied, as is the distance between where the weight or ball drops and the door.

Subjects were asked to rate on a 1–7 scale how strong a causal link they perceived between the two events (the dropping of either the weight or the ball and the ball rising from the hole). The first event occurred at various distances from the second and at various delays in time. Figure 2 shows how subjects' ratings varied as a function of these two dimensions. The results are in sharp contrast for the two conditions. In the case of the weight (Figure 2a), the results correspond approximately to the classical prescriptions. There is a large effect of temporal contiguity on the causal ascriptions. The effect of spatial contiguity is more complex. People ascribe more causality to the closer stimulus only in the case of short temporal contiguity. The effects of the two dimensions are not additive. This interaction is what would be predicted under the appropriate physical model: The weight dropped on the beam should set up a vibratory wave that travels through the beam nearly instantaneously (and presumably jars the mechanism that causes the door to open and the ball to spring forth). Any substantial delay between one event and the other should appear causally anomalous. In nonanomalous cases, distance becomes a valid cue because the force at a location should diminish with something between $1/d$ and $1/d^2$ where d is the distance from the source.

The case where the ball is dropped into the beam (Figure 2b) shows an even stronger interaction between time and distance. There is no favored time or distance. Rather, subjects are looking for a match between time and distance. The model they must be assuming is one in which the ball moves through the beam at a constant velocity and what they are looking for is cases where the ball would travel that distance in that time. Thus, in both parts *a* and *b* of Figure 2, we see that temporal and spatial contiguity are not used as mindless cues to causality, but rather are filtered through plausible models of the situation to make sensible inferences about causality. If closer temporal and spatial contiguity usually lead to causal ascriptions, it may reflect the fact that people tend to perceive situations in terms of something like the vibratory wave model. Schultz (1982) has emphasized the primacy of a prior model in people's causal attributions.

Phenomenon 15. People are also influenced by similarity between cause and effect in their causal attributions; there has been a long tradition of viewing this reliance on similarity as irrational (for a review, read

Nisbett & Ross 1980). We were able to show that under certain prior models similarity is a valid cue to causation – basically, when the cause transfers part of itself to the effect. For example, when all the wash turns out a brilliant purple it is reasonable to blame an item of clothing that has that color and people naturally use similarity in this case. On the other hand, they are very unlikely to blame a similarly colored wallpaper. Thus, similarity is not used mindlessly as a cue for causality but is within the framework of justifiable prior models. Such inferential use is statistical (or, in artificial intelligence terms, a heuristic) and may lead to incorrect conclusions on occasions. This does not make its use justified, however.

Phenomenon 16. When we see a causal law apply to an object it is reasonable to infer that it will apply again to that object but what other objects will it also apply to? It is an implication of the rational analysis of categorization and the rational analysis of causal inference that subjects ought to use category boundaries to define the range of generalization of their causal inferences. We showed experimentally (Anderson 1990a, Chapter 4) that subjects' causal inferences do take place in the manner predicted by the conjunction of the two models.

Phenomena 13–15 have a status somewhat different from that of the earlier ones. In the case of these three, there were certain aspects of human causal inference that had been commonly described as irrational. In each case there was a failure to consider reasonable prior models one might bring to situations of causal inference. When reasonable prior models are factored in, there is evidence for a great deal of rationality in human causal inference.

5. A rational analysis of problem solving

With respect to problem solving it was assumed that the goal of the system was to achieve certain states of affairs (e.g., getting from one location to another). Achieving a state of affairs has some value, G. The problem solver is characterized as searching for some plan which will achieve this state of affairs. Any plan is characterized as having a certain cost, C, and probability of success, P. The rational behavior is one of choosing among plans to find one that maximizes $PG - C$ and executing it, provided its expected value is positive.

This characterization of problem solving is unlike the characterization that exists in the problem-solving literature in that it considers plans as probabilistic and concerns itself explicitly with the trade-off of gains and costs. This seems more like the actual problem solving that people face daily, however. This characterization of problem solving is clearly connected to the decision-making tasks that have been studied and indeed such decision-making tasks are special cases of problem solving under this characterization. This leads to the first rational phenomenon under problem solving:

Phenomenon 17. Subjects appear to make decisions in which they combine subjective probabilities and subjective utilities in accordance with a standard economic logic (for a review, Kahneman & Tversky 1984). A rational analysis would predict the economic combination principle but it is often thought that rational subjects should combine the stated probabilities and utilities and not subjective quantities that systematically deviate from them. This only follows, however, if the stated quantities are the true values, and there is no justification for believing this is so. It is rare for probabilities that are stated to us to be accurate, particularly extreme probabilities (e.g., I don't know how many times programmers, students, friends, etc., said they were almost *certain* to have a task done by a particular date). Also, true adaptive utility (e.g., number of surviving offspring) could not be linearly related to money or other stated utilities over an unbounded region. Matching these discrepancies between stated and true quantities, subjective probability functions tend to discount extreme stated probabilities and subjective utility functions are nonlinear in money and wealth. Although there are still many second-order considerations (which would require a paper in themselves) it still appears to be more adaptive to use subjective probabilities and utilities than stated probabilities and utilities.

5.1. Combinatorial structure. Problem-solving tasks often have a combinatorial structure that most experimental decision-making tasks do not. A typical decision-making task involves taking a single action. A typical problem-solving task involves taking a sequence of actions or steps to achieve a goal. Therefore, there is a sequence of decisions to be made. For example, in the Tower of Hanoi problem (see Anderson 1990b, Chapter 8), no single allowable step will go from the start state to the goal state. There are logically more complex possibilities, but consider a pure serial step structure such as in an eight puzzle[7] or getting from an office to a new lecture hall.[8] In these cases each step causes a new state that "enables" the next step and finally results in the goal.

The probability that a sequence of steps intended to produce the goal will actually do so is the product of the probabilities that the individual steps will have their intended effects.[9] The cost is the sum of the costs of all the steps. This suggests that such step combinations can be treated as single macro-steps with the derived probabilities and costs, and that one can choose among them with the same expected utility logic (discussed under Phenomenon 17) used for choosing single steps.

The major difficulty is in applying this logic to discovering such sequences for the first time. This is the traditional domain of problem-solving search. In the problem-solving literature, the steps are called operators. The difficulty is searching the exponentially expanding space of operators leading from the current state: If a operators can apply to each state there are a^n chains of n operators extending from the start state.

The typical AI program searches through such sequences of steps using various heuristics, finds some sequence that leads to the goal state, and then executes that sequence. Thus, the typical scheme is to plan a complete solution and then act on it. There are two difficulties with this scheme. The first concerns the probabilistic nature of these steps. That is, the planned sequence of steps will lead to the goal only with a certain probability. With some probability they will fail and lead elsewhere. It would not always be rational to plan a long sequence of steps if the sequence is going to diverge from the intended path at an early point. The second difficulty is that limits on memory span results imply that one

cannot hold a long sequence of such steps in mind, let alone compare a large number of such sequences

Thus, one can at best plan a few steps before acting. This is in fact how human problem solving typically plays itself out. For example, when I put something in one side of the car I seldom consider whether it is the easier side to take it out from – only whether it is the easier side to put it in from. This can and does lead to overall problem-solving episodes that are nonoptimal. This iterative plan-and-act structure is an inevitable consequence, however, of the uncertainty of problem solving in these domains and the limitations of working memory.

The iterative plan-and-act structure in human problem solving is transparent in many situations. For example, people trying to solve the eight puzzle will plan short sequences of steps to get a piece in a desired position, execute them, plan, execute, and so forth. Their problem-solving episodes often consist of a pause when a plan of some sort is being hatched, a sequence of steps resulting in a piece in position, another pause, and so on.

The basic logic of such iterative plan-and-act schemes is to approximate optimality in global problem solving by achieving local optimality. This is reminiscent of the issues discussed with respect to the operant literature on the matching law. Just as local maximizing can lead to global nonoptimality, so too problem-solving sequences which are locally optimal might not be globally optimal. The argument here, as in the case of the matching law, is that such exceptions are too unpredictable statistically and too costly computationally to be considered. Such local optimization can have optimal expected value even if it produces a specific instance of behavior that is nonoptimal (ignoring computational cost) in a specific situation.

5.2. Rational analysis of partial plans. I use the phrase *plan* or *partial plan* to refer to a plan for achieving a goal which involves a move (a sequence of one or more steps) followed by the intention to complete the achievement of the goal after this move. An interesting question is how one chooses the partial plan to execute. One cannot use the criterion that the plan achieves the goal because the plan does not get one all the way to the goal.

Figure 3 is an objective analysis (i.e., not necessarily in the subject's head) of the state of affairs with respect to evaluating a plan i involving a move of one or more steps and then the intention to reach the goal. The diagram starts at state S and branches off from there. Each branch in that diagram is explained:

1. The move in plan i has some probability p_i of producing its intended state I and some cost a_i associated with that move.

2. Since I is not the goal state, there is some probability q_i that the goal state will be reached from I with a cost c_i.

3. With probability $(1 - q_i)$ the goal state cannot be reached from I. There will be some cost d_i associated with the moves before the goal is abandoned.

4. With probability $(1 - p_i)$ the move in plan i does not produce its intended state I but some unintended state U. There is a cost b_i associated with getting to U.

5. From state U there is still some probability r_i that the goal can be achieved with cost e_i.

6. Finally, there is a probability $(1 - r_i)$ that the goal

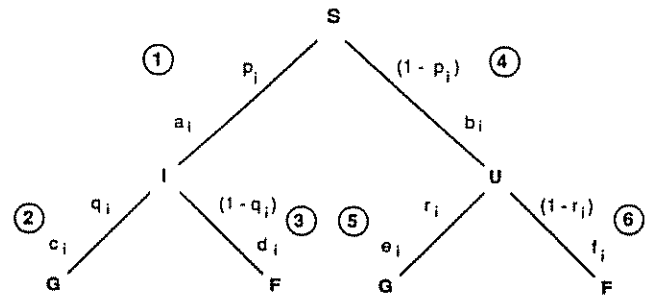

Figure 3. The state of affairs with respect to evaluating a move i. The first level in the tree represents whether the move achieves its intended state (I) or some unintended state (U). The second level reflects whether it is possible to get from these states to the goal (G) or not (F).

cannot be reached from state U and a cost f_i before the goal is abandoned.

The overall probability of success and cost associated with plan i is

$$P_i = p_i q_i + (1 - p_i) r_i \qquad \text{(Equation 11)}$$

$$C_i = p_i[a_i + q_i c_i + (1 - q_i) d_i] + (1 - p_i)[b_i + r_i e_i + (1 - r_i) f_i] \qquad \text{(Equation 12)}$$

Given this definition of P_i and C_i for a plan i, the system should choose among the plans just as it chose when there was only one step to the goal. The logic for choosing among plans is one in which one considers them mutually exclusive and chooses the one that maximizes $P_i G - C_i$ provided that quantity is positive. Indeed, multimove plans can be put into competition with single-move plans and a choice can be made among all of the possible plans according to this economic logic. In some cases, a multimove plan might be chosen over a single-move plan because the multimove plan is cheaper or more probable.

Anderson (1990a) goes through a series of mathematical simplifications to reduce Equations 11 and 12 into the following more easily analyzable forms:

$$P_i = p_i q_i / [1 - (1 - p_i) f] \qquad \text{(Equation 13)}$$

$$C_i = a_i + c_i \qquad \text{(Equation 14)}$$

where f is a parameter reflecting how much a failure to achieve an intended intermediate state affects the prospects of finally achieving the goal.

The upshot of these simplifications is to reduce the estimation problems to that of estimating p_i, the probability of the move in the partial plan producing the intended state; q_i, the probability of getting from that state to the goal; a_i, the cost of the partial plan; and c_i, the remaining cost if it succeeds. The decision procedure becomes simply to choose i with maximum expected $P_i G - C_i$ (as defined by Equations 13 and 14) or to give up if there is no i such that $P_i G > C_i$.[10]

The terms a_i and p_i can be estimated directly from knowledge of the steps in the move associated with the partial plan. The question becomes how to estimate q_i and c_i, since they are associated with unknown steps after the intermediate state has been reached. There are a number of possible bases for their estimation, but two have been analyzed in detail. One is the similarity between the goal and the state the move is intended to

result in. The second is the amount of effort expended so far.[11]

Bayesian techniques were used again to estimate q_i and c_i from similarity and effort. This involved developing prior models of how problems varied in the amount of effort they required for completion and how they varied in similarity. In the case of effort required for completion, the assumption was that problems varied according to a gamma distribution. With respect to similarity, a correlation was assumed between the number of differences between a state and a goal and the amount of effort required to reach the goal. If differences were independent there would be a perfect correlation – every difference would require an extra unit of effort to eliminate. Because of nonindependence of differences, this correlation is not perfect, and an exponential distribution of differences was assumed with the mean of the distribution proportional to the amount of effort.

A final issue that needs to be factored into the analysis of problem solving is an internal cost of generating moves. A satisficing model (Simon 1955) was produced in which the problem solver stops generating moves when the expected improvement from considering another move is less than the cost of generating it.

5.3. Empirical phenomena. Based on these assumptions a simulation of human problem solving was developed which reproduced the following phenomena (see Anderson, 1990a, Chapter 5, for details):

Phenomenon 18. The typical first attack on a problem in a novel domain is to do hill-climbing (e.g., Atwood & Polson 1976; Kotovsky et al. 1985). This falls out of the correlation between difference and amount of effort.

Phenomenon 19. When subjects have difficulty with a novel problem they switch from hill-climbing to means-end analysis (Kotovsky et al. 1985). This turns out to be the result of the fact that hill-climbing promises to minimize effort whereas means-ends analysis promises to maximize probability of success. Early in a problem sequence the subject is justified to believe in a high probability of success because most problems are solvable by this means. If the subject fails to succeed for longer than expected, he will switch to a method that promises to maximize probability of success.

Phenomenon 20. This analysis predicts that subjects are variable in their behavior and do not always make the optimal moves. This follows from the analysis of "satisficing" in which students stop seeking moves when the expected value of search decreases below a threshold. Moreover, their variability and nonoptimality is a function of both the value of the goal and the cost of considering moves.

The phenomena listed above do not really test the sophisticated interactions implied by the rational model of problem solving. This basically reflects a mismatch between this theory and the literature on problem solving. No literature deals with problem solving under conditions of uncertainty, real costs and gains, and complex problem-solving structure. There are decision-making tasks that involve uncertainty and real costs and gains. Unfortunately, they focus on simple one-step problems. There are problem-solving tasks that have a complex combinatorial structure, but they involve steps whose outcome is certain with unclear costs and gains. The rational model applies to these two types of tasks in a rather degenerate way.

In the case of decision making, the problem of managing search does not arise. The problem concerns how gradations in subjects' choice behaviors vary with gradations in the choice sets. In contrast, the problem-solving literature tends to treat the subject as totally deterministic, making all-or-none choices. There is no way to conceptualize the fact that subjects make different choices except to put a random function in the simulation. As a consequence, the typical problem-solving theories (including previous theories of my own) have great difficulty dealing with the statistics of group behavior and are content to simulate single subjects.

Phenomenon 21. Most everyday problem solving has a structure that differs significantly from the one studied in the decision-making and problem-solving literature. A paradigmatic example is route finding, where the maze of potential paths creates a very concrete search space in which the uncertainties of passage cause real variability and time and money are real costs. Recently, we have been studying route finding in the laboratory (Anderson & Kushmerick 1990), putting novel predictions of the theory to test. We show, for example, a linear relationship between decision time and the number of alternatives considered before our satisficing model selects an alternative. This is an example of a novel prediction of the rational model being put to test.

6. Implementation of a rational analysis

It is in the spirit of a rational analysis to prescribe what the behavior of a system should be rather than how to compute it. It is not our claim that the human system actually goes through the relatively complex Bayesian analysis used to establish what the optimal behavior was. Inevitably, however, the criticism is made that such rational models are unrealistic because they imply unrealistic mental computations. It is often quite easy to convert these rational prescriptions into plausible mechanisms, however. As one example, consider the following proposal for implementing the categorization prescription described earlier in a connectionist architecture.

Figure 4 shows a connectionist network designed to implement a category structure created by a rational model simulation of an experiment by Gluck and Bower (1988). Subjects in their experiment learned to associate two diseases, a rare and a common one, with the absence or presence of four symptoms (thus there are in effect eight symptoms). Our system in this simulation induced four categories to capture various patterns of symptom-disease correlation. It should be noted that the disease labels are features associated with the category and are no different from the symptoms. In the center of the figure are the four categories. Figure 4 illustrates associations from each of the 10 features (eight symptoms and two disease labels) to each category and from the category to each feature. Activation will spread from the input feature nodes to the category nodes and then to the output feature nodes.

We can use the rational analysis to prescribe the appropriate activation calculations. Let us first consider a calculation of the activation levels of the category nodes.

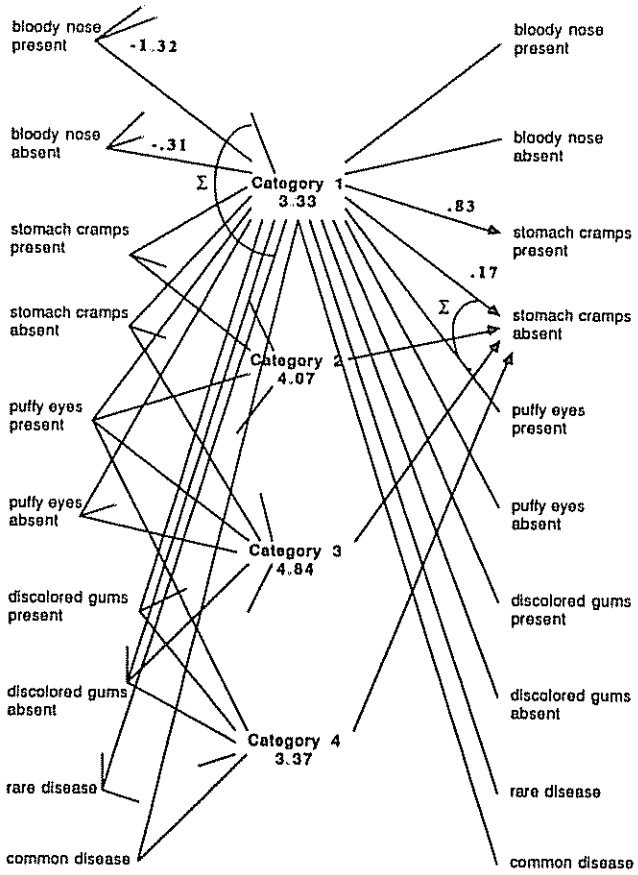

Figure 4. A schematic representation of how a category structure might be represented in an ACT declarative network

This activation calculation should implement Equations 4–8. The two relevant probabilities in this equation are $P(k)$ for each category and $P(ij|k)$ for each link from feature j to category k. Actually, we do not have to calculate Equation 4; rather, we need only calculate a quantity proportional to $P(k) \prod P(ij|k)$ which is what the numerator in Equation 4 will become. Continuing this logic of proportionality, we can substitute the numerators of Equations 5 and 8 and make it our goal to calculate a quantity proportional to $n_k \prod [(n_{ij} + \alpha_j)/(n_k + \alpha_0)]$. This is a multiplicative relationship whereas it is typical to think of activations from various sources as adding. We therefore take logarithms and make $\ln n_k$ the activation level of the category nodes and $\ln(n_{ij} + \alpha_j) - \ln(n_k + \alpha_0)$ the strengths of the j to k links. For example, for category 1, $n_k = 28$ and $\ln n_k = 3.33$, which is what is shown in Figure 4. Similarly, for category 1 and bloody nose $n_{ij} = 7$ and, assuming $\alpha_j = 1$, $\ln(n_{ij} + \alpha_j) - \ln(n_k + \alpha_0) = -1.32$, which is what is shown in Figure 4. Activation spreading from the prescribed features would calculate

$$\ln(n_k) + \sum_j [\ln(n_{ij} + \alpha_j) - \ln(n_k + \alpha_0)],$$

which is the log of a quantity proportional to the probability of the object coming from the target category.

The real purpose of categorization is not to assign objects to categories but rather to make predictions. This requires the spreading of activation out of the category nodes to the feature nodes and the accumulation of activation there. According to the rational analyses, we want to calculate Equation 3. This requires converting the log quantities, which is what the node activation represents, to probability, which is what the output activation should represent. Therefore, we need the following formula, which relates a node's activation, A, to the amount of activation output, O.

$$O = e^{bA},$$

which makes the output activation proportional to the probability. This output activation is multiplied by $P(ij|k)$, which represents outgoing link strength, to determine the amount of activation arriving at features j from category k.

The activation-based scheme just described will deliver a level of activation to node j which is proportional to the probability $Pred_{ij}$. If we assume that the probability of making a prediction is a function of this level of activation, we have an architecture that produces the categorization results we reviewed. There are such architectures. For example, the probability of pattern-matching in the ACT* (Anderson 1983) architecture is a function of level of activation.

Given this mapping of the categorization process into an architecture, one can specify the category learning algorithm. New category nodes have to be created every time an object cannot be assigned to an existing category (or, equivalently, each time it fails to get activation of a category node above threshold). Every time an object is assigned to an existing category the link strengths between the category and features have to be updated so that they are linearly related to $\ln(n_{ij} + \alpha_j) - \ln(n_k + \alpha_0)$ for incoming links and proportional to $n_{ij} + \alpha_j/(n_k + \alpha_0)$ for outgoing links. Also, the base level activation of the category node has to be updated to be linearly related to $\ln(n_k)$. Note that these are very trivial learning algorithms. We are not doing complex Bayesian mathematics; we are just updating strengths to reflect frequency.

There are striking similarities between this network and PDP networks with hidden units that are learned by the backprop algorithm (Rumelhart et al. 1986). The category nodes are like hidden units and the calculations in which they participate are similar if not identical. The real difference occurs in the learning algorithm. Whereas it takes huge numbers of iterations to get the backprop algorithm to learn a category structure, this algorithm basically gives asymptotic structure after each trial. The speed of learning of this procedure is gained by its strong priors concerning the structure of the environment whereas the backprop algorithm assumes, unrealistically I think, a much less structured environment.

Learning algorithms are not optimal in and of themselves. They are optimal with respect to environments. There are imaginable environments in which the backprop algorithm would do better. A rational approach encourages us to inquire about the structure of our actual environment and to design an algorithm optimal for it rather than designing algorithms which would only be optimal in some bizarre world.

7. Conclusions

This completes the review of the application of rational analysis. Although many details were necessarily omit-

ted, I hope it is clear from this review that many of the major characteristics of human cognition can be explained as an optimal response, in a Bayesian sense, to the informational structure in the world. We also see in examples like the network implementation of categorization, the potential to advance to the second stage of the program outlined by Marr, which is to implement these rational prescriptions in a cognitive architecture.

I am not advocating rational analysis to the exclusion of other approaches to cognition. I think we see evidence, however, that other approaches to human cognition would be more profitable if we took seriously the idea that there is a reason for the way the mind is. We are not trying to understand an arbitrary black box built out of billions of neurons.

Finally, to return to the title of this target article, is human cognition adaptive? Despite the compelling evidence that various components of it are optimized to the structure of the environment, it is unclear that we can leave with as positive an opinion about the functioning of the cognitive system as a whole. Consider memory, for example. Our memory performance can be relatively insensitive to our knowledge about our memory needs. Therefore, we may know we will need to remember a telephone number in an hour's time and will then be able to forget it. Memory does not respond to this knowledge and provide a momentary surge in the availability of the number in an hour's time, however. Rather, it responds to the general statistics of the number's use, oblivious to our "knowledge" about its future use. It is possible that the various components of cognition are optimized within their narrow bounds but that they are unable to pass information which allows a global optimum to be achieved. Just as we achieve only local optima in problem solving over time, we may only achieve local optima over components. If so, such local-by-component optimization may or may not be explainable by considering computational constraints and uncertainty in defining the global optimum.

ACKNOWLEDGMENT

I would like to thank Albert Corbett, Melvin Reder, and Lynne Reder for their comments on this paper. This research is supported by Grant 8705811 from the National Science Foundation and Contract N00014-90-J-1489 from the Office of Naval Research.

NOTES

1. This is not to imply that these are the only behavioral phenomena that can be explained or that there are contradictory phenomena. A fuller exposition of this approach can be found in Anderson (1990a).

2. Thus, for example, my memory for my locker combination has a history of being periodically useful and I may now be in context where the cues are those of the locker-room, such as lockers, showers, towels, etc.

3. Human memory may not be so constrained and it is interesting to ask which predictions might be upset by nonindependence.

4. This independence assumption is not perfect because more than one phenotypic feature can be controlled by the same gene. The most frequent form of nonindependence involves sex-linked characteristics. However, even here there is much less variation of most features (size) between sexes within a species than between species. As we will see (discussion of Phenomenon 11), in the presence of a strong set of sex-linked characteristics, the model would extract two categories – one for the male and one for the female members of the species.

5. A somewhat different Bayesian approach to categorization can be found in Cheeseman et al. (1988).

6. Consider one's confidence about the probability that a forest fire will start spontaneously at any moment when lightning does not strike versus the probability that it will start when lightning does strike. The probability in the former case is presumably near zero whereas I, at least, am very uncertain how to assign a probability in the latter case.

7. The eight puzzle is the tile-moving game (see Anderson 1990b) where the problem solver is given a 3 × 3 matrix containing eight tiles and one open space. The problem solver can move the tiles into the free space. The goal is usually to achieve some target configuration of the tiles.

8. Presumably a "step" in going from the office to the lecture hall is not an actual physical step but such a well-learned fragment of a route as getting from the office to the hall.

9. Conditionalized on the success of prior steps.

10. That is, if giving up is a costless option. Otherwise, one must consider the cost of giving up in competition with the other options and choose the minimum cost option.

11. Another possible basis we are currently studying is the history of past success with a method. Therefore, if I have had a history of success with trying to repair things by hand, I will be more inclined to try to repair something myself rather than call a repair shop.

Open Peer Commentary

Commentary submitted by the qualified professional readership of this journal will be considered for publication in a later issue as Continuing Commentary on this article. Integrative overviews and syntheses are especially encouraged.

Human cognition is an adaptive process

Gyan C. Agarwal

Department of Electrical Engineering and Computer Science (mc-154), University of Illinois at Chicago, Chicago, IL 60680

Electronic mail: *agarwal@uicbert.eecs.uic.edu*

Coming from an engineering background, I read this article with some amazement. My reason for writing this commentary is to point out that there are many parallels in the target article to what is common in control engineering practice. There are some very significant differences in approaching and even in defining the problem, however.

Anderson's basic premise is that human rational behavior can be adequately explained on the basis of optimization to the structure of the environment. It appears that the author has chosen to define the environment as the output of this black box (human behavior) because the field of cognitive psychology is unable, at least at present, to define the structure of the black box and its relation to the environment. The nature of the internal system and constraints because of the limited capability of any information-handling system is completely lost in such a view. It would be equivalent to posing an optimal control problem of predicting an optimal behavior of a system by assuming a performance index and the actual output but without any knowledge of the system itself or any of the prior inputs applied to it. In other words, the nature of the system and its history has nothing to do with its performance. In engineering analysis, such a statement could not be taken seriously.

The discussion of local and global optimality and iterative decision making may be more formally stated in terms of Bellman's Principle of Optimality: "An optimal policy has the property that whatever the initial state and initial decision are, the remaining decisions must constitute an optimal policy with regard to the state resulting from the first decision" (Bellman 1961, p. 57). The dynamic programming approach developed by Bellman has been extensively applied in deterministic as well as stochastic multi-step problems. It appears to me that Anderson is reinventing this approach limiting his analysis to output data only. The dynamic programming is an alternative to a variational approach in calculus and can be used to develop equations of Pontryagin's maximum principle. [Cf.: Clark: "Modeling Behavioral Adaptations" *BBS* 14(1) 1991; and Schoemaker "The Quest for Optimality: A Positive Heuristic of Science?" *BBS* 14(2) 1991.]

I take serious exception to the notion that a mechanistic theory is not needed and that a rational theory offers a more appropriate explanatory level for behavioral data. Anderson is asserting that a rational theory provides a precise characterization and justification of the behavior which the mechanistic theory should achieve. Would this not be the tail wagging the dog? The behavior of a physical system depends not only on the environment (either in the form of parameter dependence or external inputs) but more significantly on the structure and constraints of the system. I would think that similar conditions are also valid in biology.

Let me consider two examples to justify these concerns about Anderson's theory. First, for generating uniformly distributed random numbers on a digital machine, one can use one of several available algorithms. This problem has a well defined, discrete, deterministic system structure. The output represents a series of uniformly distributed random numbers with no apparent relationship between one and the other. The rational theory approach would not be able to predict the internal structure of such a mechanistic algorithm and the information derived from rational theory would be of little value.

A second example is the famous Fitts law for speed-accuracy tradeoff. Fitts (1954; 1964) had argued that speed-accuracy tradeoff can be accounted for by assuming that the channel capacity of the motor system is independent of task conditions. Fitts merely assumed that the average amplitude of movement is equivalent to average signal plus noise amplitude and that half the range of movement variability is equivalent to peak noise amplitude. Shannon's channel capacity theorem was then used to define the index of difficulty. Fitts's explanation neglected the principles of mass-limb system mechanics. His explanation would imply that an increase in movement time arising from an increase in movement distance results entirely from the consequent increase in information per symbol. This assertion neglects the fact that even with a perfect channel, movement time must increase with distance. Although Fitts's law has been shown to apply in many situations, several alternate hypotheses have been proposed (Gottlieb et al. 1989; Meyer et al. 1982; 1988). Fitts's equation is an empirical result and does not hold for movement velocity or muscle activations (Corcos et al. 1988). It also fails to provide any insight about the nature of motor control mechanisms and their learning and adaptation capabilities. (An alternative explanation of speed-accuracy tradeoff based on a mechanical model of limb and time optimal control theory has recently been provided by Logsdon [1990].) Fitts's explanation is equivalent to Anderson's rational theory, which considers only the behavioral data and an empirical fit.

I am also concerned about Anderson's definition of the principle of rationality and the steps necessary for developing such a theory. The first step is to specify the goals being optimized by the cognitive system. This is likely to be the most difficult problem and something that cannot be easily verified. I agree that any behavior can be seen as optimizing some imaginable goal. On the other hand, considering the nature of biological systems, there is no certainty of uniqueness between the goals and the cognitive performance. Anderson's assertion that characterizations of the organism are nonoptimal makes the error of ignoring prior uncertainty and assumes that the organism should have the same model of the situation as the experimenter. If Anderson's data and this theory do not match, how would one test the validity of the theory? In biology it is very difficult to develop criteria by which a theory is to stand or fall. As Wilkie wrote: "Even more suspect should be the theory which explains nothing because it can be adopted to explain anything. Such a theory can never be disproved, which gives it an illusory strength" (Wilkie 1954, p. 322).

Human cognition is based on the learning ability of biological neural nets. In the artificial neural nets of which Anderson is well aware, the learning process is an adaptive process. Although the biological neural nets are considerably more complex, there is no reason to doubt that these nets are also adaptive processes.

Some thinking is irrational

Jonathan Baron

Department of Psychology, University of Pennsylvania, Philadelphia, PA 19104-6196

Electronic mail: *baron@cattell.psych.upenn.edu*

Peterson and Beach (1967) reviewed a number of findings that suggested that human statistical judgments were roughly normative. Most of the data showed, as do Anderson's, that judgments were monotonic and sometimes linear functions of those predicted by a normative model. When all else is held constant, it is clear that people are often sensitive to the normatively relevant variables.

Since Peterson and Beach, however, investigators who have looked for nonnormative influences on performance have often found them. To be sure, as Anderson points out, some of these authors shot from the hip, using indefensible normative models. But other results cannot be so easily dismissed. For example: people are temporally impulsive, choosing a small immediate reward over a larger delayed reward repeatedly (Solnick et al. 1980); when they are asked to make repeated predictions of uncertain events, they often predict the less likely of two events, even when events are obviously independent (Gal 1990; Peterson & Uhela 1964); they are unable to ignore their knowledge of outcomes, even when they must ignore it in order to predict the behavior of others (Camerer et al. 1989); they fail to learn or unlearn correlations between cues and outcomes when prior beliefs are present (Chapman & Chapman 1967; 1979); and, in detecting correlations, they are subject to blocking effects and cue competition, which do not reflect reality either globally or locally (Chapman 1990; Wasserman 1990).

Anderson now tries to resurrect rationality as a basis for understanding human cognition. In some cases, this new approach leads to impressive results, reminiscent of the achievements of economics, which has taken a similar approach. But, like many economists, Anderson wants to go beyond these successes to claim that people are generally rational. To defend this theory, he has added: (1) a distinction between local and global optimization; (2) a claim that subjects do not understand the tasks as the experimenters do (sect. 1.2); (3) a claim that "what is optimal in . . . the laboratory can be far from optimal in the world at large" (sect. 1.2); and, (4) a claim that a full theory of rationality must account for the cost of thinking. I consider these in order.

(1) I interpret the global-local distinction (perhaps wrongly) to imply an admission that people can subvert their own goals as a whole, in the long run, even while they do their best to achieve some limited goal. If this interpretation is correct, then I would

agree with Anderson, except that I would like to see some evidence for the limited goal aside from the assumption that it must be present. I would also ask whether the limitation is remediable. If not, I would not want to call it irrational.

(2) As for Anderson's claim that subjects do not understand the tasks as the experimenters do, many experimenters indeed fail to check their subjects' understanding. In the studies cited earlier, however, subjects had an opportunity to learn the contingencies in effect even if they failed to understand the experimenters' instructions, and the biases survived the learning. In a great many other studies (see Baron 1988), experimenters test for misunderstanding before concluding that subjects are irrational.

Anderson's reporting of subjects' interpretations of decision making experiments (Proposition 17, sect. 5) does not account well for the results. He argues that the distortion of probabilities is reasonable given the fact that stated probabilities are usually inaccurate. This explanation cannot apply when probabilities are stated in terms of numbers of lottery tickets rather than subjective judgments, as they often are. Moreover, the type of inaccuracy found in probability judgment (Lichtenstein et al. 1982) cannot explain the distortion observed in decision making (Kahneman & Tversky 1979): In decision making, a probability of 1 is weighed more than a slightly lower probability, but in probability judgment, it is only a little more overconfident. Finally, irrationalities in decision making cannot be ascribed to probability distortion alone. Many are the result of framing effects, and these occur even when no uncertainty is present (e.g., Knetsch & Sinden 1984). (Anderson seems to imply in this section that nonlinear utility of money has been said to be irrational. This is puzzling. Anderson's equating "true adaptive utility" with number of offspring in a time of overpopulation is also puzzling.)

(3) Anderson's third claim – about the differences between the laboratory and the world – is a half truth. From the beginning of the study of "heuristics and biases" (e.g., Kahneman & Tversky 1972), irrational biases were thought to result from the overgeneralization of heuristics that are ordinarily useful. This is still the most promising general theory of irrational biases in my view (Baron 1990). At issue here is whether people can learn to discriminate conditions that control the usefulness of heuristics. Two arguments suggest that they can do so. First, unlike optical illusions (Funder 1987), most of the biases found in judgment and decision making are not universal. Moreover, those who do not show these biases seem no worse off than others. If anything, they are better off (Kuhn et al. 1988; Kuhn, in press; Larrick et al. 1990b). Second, many studies show that people can learn to improve their judgment and decision making (Larrick et al. 1990a; Nisbett et al. 1987).

(4) Finally, Anderson argues that the cost of thinking is a relevant consideration in determining rationality. Here I agree (Baron 1985). It is possible, however, that people are sometimes irrational in time allocation (Baron et al. 1986; Baron et al., in press). Moreover, many biases do not obviously save time or effort (e.g., the endowment effect found by Knetsch & Sinden 1984; the impulsiveness effect found by Solnick et al. 1980).

More generally, despite the value of the rationality assumption as a heuristic device for the development of a descriptive theory, a great deal of evidence indicates that people are sometimes systematically irrational, often in ways that are remediable. We must be careful not to miss (nor dismiss) the signs of such irrationality, because they provide us with a useful tool for the amelioration of the human condition.

The nonoptimality of Anderson's memory fits

Gordon M. Becker

Psychology Department, University of Nebraska at Omaha, Omaha, NB 68182

Electronic mail: *becker@unoma1.bitnet*

Anderson's "rational analysis" fits an impressive range of cognitive data; he has also been very resourceful in finding environments that make human memory appear optimal. He does not show, however, that these environments describe the "real environment" as conceived by the subject or by others, including other scientists or other sciences.

He erroneously describes (a) borrowings from libraries and accesses to computer files as "nonhuman information-retrieval systems," and (b) the use of words in *New York Times* headlines and the sources of messages in electronic mail as environmental demands placed on human memories. The underlying process in these systems is the human mind; and models (such as Burrell's 1980; 1985) that describe such data, are descriptions of human cognition. The Burrell parameters and distributions that Anderson uses in his memory studies were determined by data similar to that used later to "test" the model.

His definitions of parameters, and concepts like "optimal" and "environment," are often vague. For example, he defines G and C in Equation 1 as the relative gain and cost without specifying whether they are subjective (utility) or objective environmental measures. The payoff for *not* attaining the goal is not mentioned even though G is the (utility) difference between the payoffs for attaining and not attaining the current goal. The payoff when one does not attain the current goal is the expected gain from the next goal pursued, which depends on the expected gain from the following goal, and so on, which depends on all the possible goals one could have. Thus G in Equation 1 requires so much computation that Simon (1955) has argued that people satisfice instead of optimize. If the wrong G is used, the process will not be optimal. How does Anderson's subject know the correct value of G to use in Equation 1?

Anderson states that an optimal retrieval system "would stop retrieving when the probabilities are so low that the expected gain from retrieving the target is less than the cost of retrieving that item." Does he mean that the subject will continue retrieving memories and reaching the first goal repeatedly with the same gain G each time a retrieved memory structure achieves it? This would also require that utility be linear in both the C and G commodities. Repeating the same goal seems inconsistent with his example of remembering where he parked his car. If the gain from the first memory structure that correctly locates the car is G, wouldn't the gain from retrieving other "verifying" memory structures be less than G?

Equation 1 also assumes that retrieving a memory structure that does not attain the goal has no effect on the distance to the goal and no information about the relevance of other memories. A retrieved memory, however, can provide relevant (or misleading) cues not present in the initial set. After a retrieved memory structure failed to reach the goal, an optimal retrieval system would use this information, along with the joint probabilities to revise, in a Bayesian way, the probabilities of the other memories.

It should also be noted that Equation 1 does not describe the stopping problem for a parallel processing system that simultaneously retrieves n memories. The optimal rule for a parallel processor will replace the $p(A)$ in Equation 1 with $p'(A_1, \ldots, A_n)$, the probability that at least one of the n retrieved memory structures will reach the goal. This probability will also take into consideration the redundancy of different memory structures, because it is unlikely that they are independent. It should be noted, however, that this calculation is needed only when the cost depends on the number of memories retrieved, or when the

capacity of the parallel processor is such that it must limit its processing to less than the available memory structures. Thus the optimal stopping rule depends not only on the type of system, but on its specific limitations and costs, contrary to Anderson's statement that the optimal rule does not depend on the architecture.

Anderson's fits, however, are based on part of Equation 2; they ignore G and C and other parameters of Equation 2. The logic that allows him to claim that behavior is optimal when most of the equation is ignored is questionable. Moreover, he assumes that "the probability of being needed, $p(A)$, is monotonically related to the latency and probability of recall," and accuracy because of the threshold on what items will be considered. The reasoning here is: If $p(A_1) > p(A_2)$ then the latency and accuracy of $A_1 > A_2$. But, when he finds an estimate of $A_i > A_j$ he makes the *logical error* of concluding that $p(A_i) > p(A_j)$. Furthermore, the estimates of $p(A)$ are *not* obtained from an examination of the structure of the environment (as he suggests in Step 2 of Table 1) and no test is made to see whether the arbitrary equating of Burrell's parameter with $p(A)$ does in fact describe the "real" environment.

The Burrell parameter used to estimate $p(A/H_A)$, assumes that desirabilities "should be" distributed in the gamma distribution that fits the H_A responses of subjects and that these estimated desirabilities are those dictated by the environment. Although H_A is suppose to be the "record of all the times A has been needed," Burrell's estimate of $P(A/H_A)$ seems to be based on the assumptions that: (1) every time A was used it was relevant, (2) the subject always used an optimal strategy, and (3) the subject's responses are a direct reflection of the statistical structure of the environment and not the hypothesized structure of the mind. But H_A is a reflection of the subject and the estimate of $P(A/H_A)$ is based on the subject's responses not the environment. In fact we don't know how the fit relates to the environment . . . or optimality.

Thus even if Anderson's fits were optimal for the environments he describes, there is no evidence that they are optimal in the "real world" or even in the "perceived world" of his subject.

If human cognition is adaptive, can human knowledge consist of encodings?

Robert L. Campbell[a] and Mark H. Bickhard[b]

[a]IBM T. J. Watson Research Center, Yorktown Heights, NY 10598; and [b]Department of Psychology, Lehigh University, Bethlehem, PA 18015

Electronic mail: *[a]rlc@ibm.com or rlc@yktvmh.bitnet; [b]mhb0@lehigh.bitnet*

Have assumptions about mechanisms been avoided? Anderson asserts that a "rational" analysis of cognition can be separated from an "algorithmic" or "mechanistic" account (sect. 1.1), and that this amounts to "the framing of the information-processing problem . . . a nearly mechanism-free casting of a psychological theory." In his analyses of cognitive functions such as memory and categorization "the computational assumptions are indeed weak, involving claims that almost all information-processing theories would agree on" (Anderson 1990, p. 36). Weak though these assumptions may be, they still have consequences. Anderson's project is haunted by the ghosts of mechanisms that he has not yet exorcised from his "rational" level of analysis.

Is rational analysis committed to encodings? Anderson equates knowledge with *encodings*. By encodings, we mean objects, events, or structures of objects, in the mind that represent objects, or structures of objects, in the world and do so by correspondence (Bickhard & Campbell 1989; Bickhard & Richie 1983). In Anderson's (1983) ACT framework, declarative knowledge consists of hierarchical structures of encoding elements (the structures can be temporal, spatial, or propositional). Procedural knowledge consists of encoded production rules whose encoded conditions must be matched with symbols in working memory. Following Pylyshyn (1984) and Newell (1980), Anderson affirms that the algorithmic level is psychologically real and that what happens there is computations on symbols (Anderson 1990). Although Anderson does not mention it, he is also endorsing Fodor's (1975) "representational theory of mind," with consequences that we will explore below.

Anderson's specific rational analyses presume some obvious encoding atoms (such as "memory traces," sect. 2.2), or objects in the environment for such atoms to correspond to (such as discrete objects with discrete features, already clustered in predictively useful ways, sect. 3), so they too are hardly free of assumptions about mechanism.

In fact, Anderson takes encodingism for granted. Though he admits that "it has become apparent to me that this rational analysis has assumed the general ACT framework, if not the ACT* theory" (Anderson 1990, p. xi), the ACT assumptions, when finally enumerated, consist of such things as "a system in which memories are retrieved and tested for appropriateness" (p. 252). In such "weak" assumptions, encodingism is too deeply presupposed to be mentioned.

Is encodingism tenable? We have argued (Bickhard 1980; Bickhard & Campbell 1989; Bickhard & Richie 1983) that any framework that treats encodings as an irreducible, foundational form of representation is untenable. Encodings have to derive from some other form of representation, because they presuppose knowledge of what they are supposed to correspond to.

Foundational encodings are ubiquitous in the ACT framework. Two kinds of declarative encodings – temporal strings and spatial images – are hierarchical structures of elements that are held to encode the environment directly (preserving temporal or spatial sequence), whereas abstract propositions recode environmental structures by a process that has to be learned (Anderson 1983). Hence, for Anderson, temporal strings and spatial images appear to be foundational encodings, and their elements must be, along with the elements of abstract propositions.

Our argument against foundational encodings runs as follows: In the clear cases (Morse code, digital audio, etc.), encodings stand in for some other form of representation: X encodes Y means that X represents the same thing that Y represents. The encoding relationship presupposes that Y already represents something. Y might be an encoding itself, but if it is, it must stand in for another representation Z. The regress has to stop somewhere, and it cannot stop with an encoding. If Z is a foundational encoding, it must stand in for something already known, yet Z is supposed to be the means by which that thing is known. But "Z represents the same thing that Z represents" does not define an encoding. Hence encodings cannot be a foundational form of knowledge (Bickhard, in press a; in press b).

The incoherence of foundational encodings has been partially recognized by quite a few thinkers. For instance, Piaget (1970) argued that perception could not be a copy of structures in the world, but he did not extend the argument to concrete and formal operational structures, and Harnad (1990) has argued that digital encodings (symbols) cannot be a foundational form of knowledge, but he has not extended the argument to analog encodings (such as spatial images).

Can encodings develop? It follows directly from the incoherence argument that genuinely novel representations are impossible within an encoding framework. A fundamentally new kind of encoding cannot be acquired, because it would have to be defined in terms of the new kind of thing that it represents, yet that kind of thing supposedly cannot be known without the encoding (Bickhard, in press a; Campbell & Bickhard 1987).

The impossibility of novel encodings is best illustrated by the work of Fodor (1975; 1981). From Fodor's standpoint, all encoding approaches must posit an innate set of primitive encodings

All that distinguishes such approaches is the extent to which "complex" encodings can be defined as simple combinations of the primitive encodings

Fodor (1981) argues that, while "phrasal concepts" (e.g., sentences) may be built out of primitive encodings, "lexical concepts" cannot be, and must therefore be primitive encodings themselves. In Anderson's terms, this would be a claim that structures of encoding elements, such as phrase units, image units, or propositions, can be built of more basic encodings (it is not clear that they always are), but individual elements, – for example, words, basic subimages, and concepts – are not further reducible and are therefore primitive encodings.

Fodor's innate concepts cannot be learned; they must already be present to figure in any encoded hypothesis. Nor can they be products of any constructive developmental process (Fodor 1980). Fodor (1981) is forced to posit a process of "triggering," extrinsic to the passive built-in encodings, which elicits the activation of innate concepts through sensory conditions or the prior activation of other innate concepts.

Can encodings evolve? Positing innate encodings, however, just shifts the burden of their construction from the development of the individual to the evolution of the species. And Fodor's arguments imply that the acquisition of novel encodings through hypothesis testing is impossible in principle. Fodor's arguments lead to the conclusion that *if evolution is a variation and selection process, there is no way for encodings to evolve*. Though ambivalent about evolution, Anderson does treat it as a variation and selection process (sect. 1.1). He is occasionally willing to consider evolutionary constraints on the differentiation of "new representational types . . . there must reasonably have been time in our evolutionary history to create such a representation and an adaptive advantage to doing so" (1983, p. 46).

Given his commitment to encodingism and to evolution as variation and selection, Anderson is therefore obliged to (1) identify errors in Fodor's reasoning, (2) propose an alternative to encodingism, or (3) embrace Fodor's conclusions.

Preserving encodingism while refuting Fodor. To refute Fodor, Anderson would have to show that there is a process compatible with the rest of his theory that can generate emergent representation: representation constituted out of phenomena that are not themselves representational. This would be an uphill fight: A recent survey of production system models has concluded that the constructive processes invoked in such models are not even capable of generating new goals or radically reorganizing algorithms (Neches et al. 1987). Because Fodor's conclusions follow from the weakest assumptions of encodingism, Anderson would be hard pressed to avoid Fodor's *reductiones ad absurdum* while retaining anything like the ACT framework.

Replacing encodingism. Rejecting encodingism, in Anderson's case, would mean rejecting the physical symbol system hypothesis. Information-processing (IP) modelers would then have to venture into completely unexplored territory. There is not only the challenge of coming up with an alternative account of representation that avoids the stumbling blocks of encodingism and provides for the emergence of representation from something nonrepresentational. There is also the task of tracing its ramifications.

Convergently with some others (e.g., Brooks 1987), we have been engaged in this sort of effort for some time. An account of our alternative, *interactive representation*, would overflow this commentary. We would just like to point out that replacing encodings with interactive representation has forced changes throughout our conception of cognition, from learning to language to developmental stages to consciousness to psychopathology and beyond (Bickhard, in press b; Campbell & Bickhard 1986). For instance, traditional views of language as the recoding, transmission, and decoding of encoded messages cannot be maintained in an interactivist approach (Bickhard 1980a; 1987), hence language learning cannot be presumed to start with pairings of utterances and encoded meanings (as it is by Anderson 1983).

It would be most convenient if the changes that ensue from the adoption of nonencoded representations could be bottled up in a preprocessing stage, which converts everything into encodings, allowing computational business to go on as usual. Harnad (1990), for instance, proposes that perceptual categorization yields meaningful symbols, which can thenceforward be processed in the conventional fashion. But non-encoding-based sensory processing can't be divided this way (Bickhard & Richie 1983).

Embracing Fodor's nativism. The contemporary practice of IP modeling is already nativist, albeit unwittingly. Researchers simply introduce new elements of declarative representation whenever necessary to model a phenomenon, without considering their learnability. IP modelers could stipulate that any primitives introduced for modeling purposes are innately present and must be activated by triggering. Such a move, however, would wreak havoc on the empiricist allegiances usually professed by IP modelers. It would also restrict evolutionary constraints to operating on the generation and selection of combinations of encoding atoms, while leaving unsettled the question, ignored by Fodor, of the possible evolutionary origins of the primitive encodings.

Conclusion. It is certainly desirable to introduce evolutionary constraints, and constraints of optimization to the environment, into cognitive science, and we salute Anderson for doing so. But there is considerable irony in the introduction of such considerations within an encoding-based framework, which makes the evolution of mental representation impossible.

ACKNOWLEDGMENTS

Thanks to Jack Carroll, Steve Payne, and Kevin Singley for their comments.

Mechanistic and rationalistic explanations are complementary

B. Chandrasekaran

Laboratory for AI Research, Ohio State University, Columbus, OH 43210
Electronic mail: *chandra@cis.ohio-state.edu*

Evolution is more likely to be a "satisficer" than an optimizer. With that proviso, it does appear from Anderson's target article that a surprising number of the detail about human cognitive behavior can be explained as a satisficing, if not an optimal, response to the structure of the environment. What consequences follow for various research programs is not quite clear, however. I am an AI person, not a traditional cognitive psychologist. In addition to whatever explanatory powers psychologists want from their theories, I want the theories to have *design-prescriptive* powers as well; that is, I want them to tell me how to *create* mind-like entities. Historically, the route for this sort of progress has come from mechanistic explanations (ME) of mental phenomena. Thus, I read the target article from the perspective of what rational analysis (RA) has to say about ME.

Anderson's views on this range from his belief, in "overenthusiastic moments," that RA can supplant ME, to a more sober suggestion that RAs place constraints on MEs. Finally, he displays an ME for categorization that actually *implements* his RA, in the sense that it can be thought of as literally estimating the various probabilities involved.

Given a cognitive agent in an environment engaging in a certain behavior, how one should allocate an explanation of the behavior between the structure of the agent and the structure of the environment has been discussed before in psychology and AI. For example, consider Simon's (1982) ant: It produces a path

of great complexity on the beach, but a great deal of this complexity is explained by the properties of the environment, that is, the shape of the sandhills on the beach. In this case, roboticists charged with producing an artificial ant would be making a mistake if they thought that the ant had an internal structure that somehow had an encoding of the path. Not only would their explanations be wrong but they would find themselves constructing an ant that didn't work correctly.

But the problems Anderson is concerned with are not of this type. Here the explanations are not necessarily allocated between the internal structure of the agent and the structure of the external environment; *both* simultaneously account for the behavior, albeit in different ways. We need to define some notations to clarify this idea.

Let E stand for the environment and Si and bi for the structure of an agent and its behavior at time i. Let M stand for any body of mechanisms in the agent that takes as input Si, bi, and the response of the environment and produces as output $S(i + 1)$, that is, it is some sort of learning or structure-modifying function. Let us assume for the purposes of this discussion that the structure of E is invariant in time and that we are interested in the steady state properties of S and b, that is, for i = infinity.

M itself may be a complex collection of mechanisms with different time constants: one in the scale of biological evolution, perhaps another in the scale of cultural evolution, and a third in the scale of learning by an individual.

For the question, "Why is b the way it is?" we have two types of answers. One is that b is the way it is because S of such-and-such type produces it (traditional ME). The other is that b is the way it is because it is optimal for E (RA), but this story has a subplot: M modified S such that S was optimal for E, that is, it could produce an optimal b. Both answers involve S, sooner or later.

If our aim is to make agents that display behavior b, we either need to know S_{inf}, or we need to know M, S_{init} and have enough time to let M shape the S into S_{inf}. For the latter alternative, depending on whether S_{init} reflects the initial situation for the individual, the culture, or some point in biological evolution, we are talking of a more or less practical program.

In the above, I *have* accepted the RA hypothesis that b is optimal, but, as Anderson acknowledges in the concluding section of the target article, some b's may not be optimal after all. It seems to me that whether b is optimal depends on the following things:

(1) The presumed goal of the agent. If a behavior b is not optimal for goal g, perhaps it is optimal for goal g'. In some sense we can go shopping for goals. (Anderson is admirably careful about this issue in the examples he has studied; his statement of goals does not seem problematic, but it is not clear how long this fortunate state of affairs will last. For example, Marr assumes that a goal of the human visual system is to produce an account of 3-D shapes of the objects in the scene, but why is that not a reasonable goal for the frog's visual system as well? In general, why wouldn't any goals that we would ascribe to the human visual system not be appropriate goals for the frog's as well? To get RA off the ground, we will have to make additional assumptions, some of them about the *structure* of the respective visual systems.)

(2) The properties of S_{init} and M relative to the search space in which the specifications for optimal S_{inf} lie. Perhaps S will never get to the optimal S_{inf}. Thus, for a specific cognitive function, we will not know until after RA and data analysis are complete whether in fact b is optimal. In this sense, as a general research program, RA is asserting that b is optimal whenever it is.

My points in the above have been that ME and RA are complementary analyses, not alternatives, and that the RA program is not unambiguous in its methodology. Now I want to examine the claim that RA places strong constraints on ME.

The optimality of an agent's behavior does not imply that the agent is using explicit optimization to produce the behavior. This has been a pet peeve of mine about quite a bit of work in AI which assumes (1) that the job of an intelligence is to produce correct or optimal answers, and (2) the mechanisms for production of intelligent behavior should implement normative methods of producing optimal answers, most commonly some form of logic or Bayesian analysis. On the contrary, it seems to me that it is neither necessary nor desirable that the mechanisms of behavior production be explicit implementations of normative methods.

I need to clarify some terms before I proceed. The idea of the "structure" of cognition is a bit too vague. We can assume that what is meant by that word is, in information-processing language, two things: a mechanism and some content that has been put into the mechanism. To use some concrete examples, one proposal for a cognitive mechanism is a search engine, the latest example of the proposal being the SOAR architecture of Rosenbloom et al. (1987). Such a mechanism corresponds to a language in which specific programs with specific content can be written. Thus, SOAR can be programmed to have knowledge about some domain and methods. A SOAR machine so programmed can actually work on problems in that domain. The metaphor of a programming language does not restrict the above idea to symbolic mechanisms. The PDP-style connectionist research similarly specifies an abstract mechanism, but that mechanism itself needs to be given content to solve specific problems. For example, when one designs a PDP-style network to solve word recognition, one has in fact used the abstract "programming language" of a PDP-style connectionist mechanism to produce a specific "program" of that type.

In my view, what is interesting about both these (and many other) mechanisms that have been proposed for cognition is that they can be used to implement both optimal and nonoptimal methods for specific goals. In fact, this property seems to be especially desirable: The agent can adapt itself to changes in the environment without changing the basic mechanism because it is neutral with respect to optimal and nonoptimal algorithms. I want to give two examples of this.

I am told that frogs' visual systems are so organized that sensing danger they jump toward blue and away from green. It has been proposed that this is optimal behavior: Blue represents a body of water, safer for the frog, and green represents land, full of predators. The neural mechanisms that implement this strategy could just as easily implement some other strategy of color preference. The RA that suggests that the optimal behavior is "jump to blue" is putting no constraints whatsoever on the basic neural mechanisms. Of course, such an RA *is* placing a constraint on the *content* of the mechanism, namely, that it should be programmed to prefer blue to green.

Similarly, I am told that during the plague in medieval Europe, some villages, on hearing of the breakout of the disease in a nearby village, engaged in a ritual of dancing at night near the village dump, making loud noises with pots and pans. As it happened, such villages had a smaller chance of catching the infection. A modern-day RA would show that this was actually optimal behavior, because the ritual kept the rats away from the dumps and consequently from the village. Many cognitive mechanisms can implement this strategy, including the following pair: a more or less random behavior-generating mechanism that first conceived of some version of the ritual, and a cognitive mechanism that in some way remembered and passed on the ritual. Villages that survived were more likely to pass the ritual on. The same mechanisms could be used to implement relatively ineffective strategies.

Depending on what is meant by the word "structure," RA can be thought of as giving clues about the structure of cognition. If structure means abstract mechanisms, the case is in general less compelling. If structure means the totality of abstract mechanism plus content, it seems quite reasonable to say that RA can give clues about structure, since, as seen in the above examples, it gives clues about content.

This brings me to the relation of Anderson-style RA to Marr's approach. Anderson proposes that RA is similar to Marr's computational level. It is true that Marr, in his work on vision, proposes that we should start by asking "What are the goals of computation?" and, "What is available in the image?" The latter question, in its generalization, can be construed as "What is the nature of the environment?" Although Marr uses his analysis of what is in an image to constrain what he could plausibly expect the visual system to be computing, however, his computational level account of vision is really a proposal about the *structure* of the visual system. Marr did not infer the existence of a level called the "2½-D sketch" purely from the properties of the image and the hypothesized goal of the visual system. It was a *hypothesis*, inspired by the analysis of the image no doubt, but nevertheless a hypothesis. The computational level, to use Newell's (1982) term, avoids symbol-level commitment to how the computation is implemented, but it nevertheless remains a partial specification of a structure. The spirit behind Marr's levels is close to that behind Newell's distinctions between knowledge and symbol levels, in that both are attempts to develop a way of talking about structure without being tied down to the incidental aspects of implementation, but neither is an attempt to avoid specifying the structure needed for explanation.

I have tried to clarify the relationship between Marr's computational level and Anderson's RA, because I think the former still hews to the ME program. I can actually try to implement Marr's three stages (using additional commitments) to make a vision machine. I can't in general implement an RA to make the corresponding cognitive machine. Anderson's classification net is not really a counterexample, because, as I have argued, we do not in general want to be committed to literal implementations of optimizing methods to achieve optimizing behavior. In summary, I have supported that side of Anderson that believes that RA and ME are complementary. I have also argued that RA may give general guidance, not about the abstract mechanisms of cognition, but about their content.

Before concluding, I'd like to express my admiration for Anderson's piece as a tour de force of analysis and writing that illuminates the relation among behavior, the structure of the agent, and the environment. I think RA also helps provide arguments for why AI should worry about natural (i.e., human) intelligence. Often AI people make a fairly strong distinction between human and machine intelligence claiming that there is no reason to base our mechanically intelligent agents on the structure of human cognition. If we want our machines to share our goals and operate intelligently in the sorts of environments we operate in, we had better look to the structure of human intelligence for inspiration, because according to RA, it is probably pretty optimal for the task.

ACKNOWLEDGMENT

Support provided by AFOSR Grant 89-0250 in the preparation of this commentary is gratefully acknowledged.

Normative theories of categorization

James E. Corter

Teachers College, Columbia University, New York, NY 10027
Electronic mail: *jecorter@cutcv2.bitnet*

"Rational" or normative accounts of categorization are not new. At least since Rosch's seminal work on "basic level" categories (Rosch et al. 1976), it has been proposed that the categories we tend to use most are those that optimize some useful characteristic. For Rosch, one potential optimization criterion was the "informativeness" of categories. Medin (1983) and Jones (1983) suggested that it is desirable to maximize the certainty of inferences that can be made about the features of instances of a category ("category validity"). Gluck and Corter (1985) pointed out that the expected number of correct inferences that can be made actually involves two factors. The first is the category validity, which is indexed for an individual value of a single feature by $P(f_i|c)$, the conditional probability of the feature value given that the instance is a member of category c. The second factor is $P(c)$, the overall relative frequency or base rate of the category. Anderson's account of categorization begins with these same assumptions.

A relatively novel aspect of Anderson's theory concerns the role of prior expectations in affecting what categorizations are made. A Bayesian approach seems perfectly suited to providing a rapprochement between "similarity-based" and "theory-based" views of categorization. A "theory-based" view of categorization (e.g., Murphy & Medin 1985) holds that the categories people form are largely determined by their causal theories about why the categories should exist, rather than by statistical criteria measuring category-feature associations. For example, rats and humans alike find it more natural to associate nausea with a recently experienced food substance than with a recently experienced light or tone stimulus. Normatively, it seems difficult to argue against a Bayesian solution, in which the influence of prior knowledge and beliefs on inferences is represented by prior probabilities on propositions, and Bayesian methods are used to incorporate these priors with the evidence at hand, producing posterior probabilities for the various propositions or inferences. Indeed, there has been much interest in Bayesian methods in the machine learning community. In Anderson's model, Bayesian priors are used to model the influence of the base rate of a category, $P(c)$. The influence of these base rates, however, is assumed to be moderated by a free parameter of the model, the "coupling parameter." This parameter is interpreted as the subject's prior tendency to put any two objects into the same category and is introduced to improve the performance of the category-learning model during the early stages, when only a few instances have been seen. But the need for this parameter seems to weaken the case for Anderson's claim that category-learning mechanisms operate so as to maximize the normative criterion. This parameter does not seem to represent a cognitive limitation, which is allowed in Anderson's rational analysis, but rather a fairly arbitrary processing assumption.

Anderson's larger endeavor, to provide a framework for developing and evaluating normative accounts of cognitive phenomena, is a valuable contribution. Anderson proposes that a rational analysis should begin with a specification of the goals the organism should maximize. The second step, to describe fully the environment in which the person is to operate, is a difficult one, but undeniably important. Indeed, Rosch et al. (1976) described this step as the major goal of their work on basic level categories. One might argue with some of Anderson's results here, however. For example, Anderson concludes that categories are almost always disjoint. This seems to ignore the possibility of classifying things at varying levels of abstractness – after all, a thing can be a computer, an object, a word processor, a possession, and a Zenith simultaneously. Another controversial conclusion about the nature of the environment concerns the assumption of independence of features, which is necessary to make the computational problem tractable and to give the theory more predictive value. Anderson points out that within a species, many features seem to vary independently (i.e., size and coloration). But between species, features are not independent. For animals, the features *sings*, *flies*, *lives in trees*, and *has feathers* are correlated. It is this correlation that makes the category *bird* so salient. Step 3 of Anderson's methodology is to make some set of (minimal) assumptions about cognitive limitations. The assumptions must be minimal because very stringent limitations might result in a complete obliteration of any bias toward "rational" performance, thus negating the value and severely limiting the testability of the rationality hypothesis.

The truly bold hypothesis at the heart of Anderson's theory is

that human cognitive mechanisms might have adapted so as to maximize normative goals. It seems somewhat unclear whether Anderson means that the human mind incorporates a flexible set of abilities, allowing people to "learn to learn" in the most rewarding way, or whether he means that "locally" better ways of thinking and learning have been bred into humans by natural selection. The unresolved question here is whether these "optimal" cognitive tendencies are encoded genetically in the individual, or in plastic (but very fundamental) learning mechanisms. A third possibility, not discussed by Anderson, is that superior outcomes might reinforce and encode optimal categorization performance not at the individual level at all, but rather at the level of a society of category users. Freyd's (1982) shareability hypothesis suggests that the categories that survive in a society of cognitive beings are those that can be readily explained and communicated among individuals. Corter and Gluck (1985; submitted) have suggested that the categories that survive in a culture and language tend to be those that that are most informative, that is, that allow the maximal number of inferences to be made concerning the values of features. Thus, optimal categories can be selected for at a societal-cultural level without postulating any optimizing cognitive tendencies on the part of the individual.

Rational analysis: Too rational for comfort?

Ronald de Sousa

Department of Philosophy, University of Toronto, Toronto, Ontario M5S 1A1, Canada
Electronic mail: *sousa@vm.epas.utoronto.ca*

Philosophers define the induction problem in such a way as to guarantee that it will be insoluble. Perhaps this is merely a matter of not wanting to drive ourselves out of business. Anderson's approach, by contrast, promises to take account of the fact that in an evolutionary context the problem is heavily constrained by the range of specific problems organisms face. This results in an exciting research program, holding out hope of understanding what types of inferences need to be privileged to succeed at the business of life.

But in the promising aspect of this project also lies its danger. To be adequately constraining and realistic, our conception of that "structure of the environment" to which our inferential and categorizing strategies are to be adapted must be relativized to the interests and capacities of some existing organism. For psychological purposes, there is no objective "structure of the world" independent of what we need to find in our specific environment. This is not to take sides on the issue of scientific realism. We could easily believe that the structure of the world as described by the hard sciences is in some strong sense the real, objective structure of the world; but that structure is notoriously difficult for organisms like ourselves to visualize and manipulate: It is counterintuitive to the point of unintelligibility to common sense. The level at which we form the sort of predictive and categorial common sense with which psychology must be concerned may be relatively far removed from anything we can call the "objective structure of the environment." But how exactly are we to define that level? It must presumably lie somewhere between the level of physics (a level largely irrelevant to our concrete experience) and that level at which inferences are made on the basis of highly specific knowledge-drive rules, constrained by the demands of a particular ecological niche.

Evolution is "local optimizer" (sect. 1.1, para. 3). But our conception of where local maxima were actually to be found in the history of evolution must remain largely a matter for speculation. Because some evolutionary changes result from drift and not from selective pressure, as Anderson points out, we are unlikely ever to discover whether a particular case of less-than-perfect adaptation results from the limitations of a particular hilltop or from some episode of random drift resulting in a changed landscape of available hills to climb.

On the other hand, the most favorable cases of apparent adaptation may partly be the product of a certain way of conceptualizing the problem. Rational analysis tells us how things would be if they were optimally designed for solving a range of problems understood in a certain way. One obvious danger is that the crucial features of "the environment" may be fabricated on the basis of particular capacities of a given species to detect and manipulate just those features. This would not diminish the interest of Anderson's descriptive work, but it would introduce a circularity into his explanatory project. It could also confirm the skepticism that some have expressed about the prospect of finding any differences in intelligence between different species (MacPhail 1987).

The real pitfall, however, lies in the opposite direction. If the mechanisms are conceived too abstractly, the promised constraints will be weak or nonexistent. One is almost bound to come up with the result that our cognitive powers are well adapted up to a point, and that beyond that point our irrationality is just brute fact. For example: I find it hard to see how the rational analysis of problem solving tells us anything about "the informational structure of the world" in any empirical sense, as opposed to merely working out some plausible but a priori Bayesian principles of choice under uncertainty. Another example is provided by the "prior models" used to supplement the Humean principles of causal inference by contiguity. The ball and projectile experiment (target article, Figure 2b) gives us just what we would expect from the principle of contiguity, supplemented by the belief that the ball must travel through the mechanism before it can release the projectile. The "prior model" is just an application of the principle of contiguity itself.

In summary, there are actually three poles of attraction, with their attendant dangers, not two. The third pole – contrasting both with objectifying science and with speculation about the informational structure of excessively specific niches – is a merely a priori analysis of the demands of life. I suspect, to give one last example, that Anderson's "environmental constraint" on memory is of this kind. If one were to invent a principle governing the availability of recalled information, what else would we devise but a formula that specifies that the information most often needed is the information most often used? This seems to be the content of Anderson's remark that the "relevant structure of the environment has to do with how the need for information tends to repeat itself in various situations." As Anderson himself points out, the matching of our memory function to that "structural" fact is actually rather inconvenient in all those cases in which we know something more about the future need for some specific item of information. But is his theory able to provide any explanation for this inconvenient feature? In reply, Anderson appeals to the fact that our cognitive capacities are limited by local optima, but it's not clear what more is being said here than that some of our cognitive capacities are not as satisfactory as a perfect designer would have wished, because it somehow happened that way. Although I find Anderson's work seductive and impressive, I am tempted to conclude that he is, after all, too much like a philosopher.

Adaptivity and rational analysis

Bradley W. Dickinson

Department of Electrical Engineering, Princeton University, Princeton, NJ 08544-5263
Electronic mail: *bradley@princeton.edu*

Anderson has selected a rather provocative title for his target article, particularly in view of the carefully delimited scope of

his discussion explaining and illustrating the methodology of rational analysis. In my view, adaptivity is a process by which changes in system properties, either structural or parametric, produce changes in system response that improve, or even optimize, some intrinsic system performance measure. The system performance measure incorporates relevant environmental influences and uncertainties, and the changes in system properties result from a feedback process that may take into account observed performance effects of prior response changes. These changes occur over time, and adaptive phenomena may exist at multiple time scales.

From this perspective, rational analysis as described by Anderson involves models for instantaneous "snapshots" of cognitive processes where adaptivity over long time scales may be an underlying mechanism. The probabilistic models that are formulated involve plausible relationships between variables, and to the extent that accurate conclusions about experimental outcomes are reached, they provide a framework for explanations. The capability of using a variety of environmental models, from theoretical to statistical to ad hoc, makes rational analysis a very flexible tool.

Because categorization was included as the example of where a theoretical model of the environment could be used, I think it would have been appropriate to describe the probability model based on speciation more explicitly. The reader is left without any insight into the statement that the model leads to an excessively complicated computational task. Is this a case of fundamental constraint complexity (corresponding to Step 3 in target article Table 1), or is this a case in which simplification is needed only for efficient simulation of the model (corresponding to Step 4 in Table 1)? Without knowing the theoretical model, the reader also has no basis for understanding the nature of the approximations used to obtain the empirical estimates of the probabilities used in the categorization algorithm.

For an extension of the ideas of rational analysis to models where optimization is carried out over time, dynamic programming models may be employed. The computational constraints and optimality issues to be addressed, however (again corresponding to Steps 3 and 4 of Table 1), are much more severe. In the context of behavioral adaptation processes, some dynamic programming models are discussed in *BBS* by Clark (1991).

Adaptive cognition: The question is how

Jonathan St. B. T. Evans

Department of Psychology, Polytechnic South West, Plymouth, Devon PL4 8AA, England

Anderson's title poses the wrong question. Surely we would all agree that intelligent behaviour is a function of the organism's goal, the environmental structure and the cognitive mechanisms available. We would further agree that the cognitive acts we study are, as Newell and Simon (1972) put it, of *intendedly* rational behaviour on the part of our subjects. The real questions concern not whether cognition is adaptive but in what ways people attempt to achieve their goals and what kinds of mechanisms have been developed for these purposes. In fact, those psychologists like myself who earn their living by studying cognitive biases can do so only on the assumption that our subjects intend to be rational. If behaviour were random or unmotivated then the errors we observe would be of no theoretical or practical interest.

Anderson's thesis should be seen not so much as a theory that cognition is rational, but rather as a theory of what rational cognition looks like. Actually, it breaks down into two components that can be assessed separately. The first is a particular theory of rationality. The second is the contention that this theory is reflected in the behaviour of human beings. This approach has been tried before – with rather limited success.

Two of the most notable previous attempts to apply a theory of rationality as a descriptive psychological theory illustrate two broadly different concepts of rationality. The first is the notion that logic describes human deductive reasoning. This approach, sponsored by many philosophers and psychologists – including Piaget – defines rationality by *mechanism*. Roughly speaking, the argument runs that people are intelligent, intelligence requires accurate reasoning, logic describes how correct deductions are made, so people must reason by logic. This theory has had a hard time of it in the past 20 years or so, however, with much experimental evidence of logical errors and content-dependent reasoning (see Evans 1989), as well as the postulation of a rival method of deductive competence in the form of mental-models theory (see Johnson-Laird & Byrne, in press).

The second major attempt concerns the application of economic decision theory to the study of behavioural decision making. This approach defines rationality by *purpose*. Thus, decision behaviour is rational if it maximises the organism's expectation of gain or minimises its expectation of loss. This rationalist approach also underwent a rapid fall from grace. Within a decade of man's being declared a good intuitive statistician by Peterson and Beach (1967), the economic model was judged a descriptive failure (Slovic et al. 1977). In parallel with the study of deductive reasoning, the decision and judgement literature had thrown up a plethora of experimental findings – including much evidence of bias – that appeared to require a different approach, such as the assumption that reasoning and judgement is based on the widespread use of heuristics (Kahneman et al. 1982).

It is clear that Anderson's theory is of the second kind, in which rationality of purpose – that is, achieving goals within a defined environmental structure – takes priority over considerations of mechanism. Indeed, his theory can be seen as an extension of the economic decision model to a range of cognitive activities that have not traditionally been thought of as involving decision tasks. No such theory can be complete, however, without assumptions about mechanism. For example, on what basis does the chooser decide that one act is preferable to another? Does one rely on past experience or engage in mental modelling of possible future worlds?

Consider two sample problems: (1) how to choose a move in a given chess position and (2) how to decide which candidate to vote for in an election. In either case there are two fundamentally different ways one might suppose that the decision is made. One is to assume that the chooser attempts to calculate possible future consequences of the choices by imagining the possible moves and counter-moves of the game or by running a mental simulation of the likely future behaviour of the political candidates. At some arbitrary point in the projected future the tree is terminated and the current choice is determined by the optimal world state that can be reached. This is certainly one model of rational choice – favoured, for example, by decision analysts (see von Winterfeldt & Edwards 1986). The second approach assumes that little or no forward search occurs and that one instead tries to match the features of the current situation to one previous experience and to choose on that basis. Thus I recognise that this chess position is similar to one in a previous game in which a particular move was effective, or I rely on a belief that one candidate's political party has a better track record.

Anderson's discussion of problem solving, in particular, makes it clear that his theory favours the second notion. But is this because of considerations about mechanisms (he refers to working memory constraints) or because it maps more conveniently onto the model of rationality he favours? Does it not in fact demonstrate that his cognitive theory – like everyone else's – is determined both by considering the function that the system serves and by the plausibility of the postulated mechanisms?

In conclusion, Anderson presents us with an interesting and powerful cognitive theory, but he confuses us with his use of the concept of rationality. He does not distinguish himself from

other cognitive scientists by his belief that cognition is adaptive, but rather by his proposal of a particular theory of rationality. Furthermore, his particular applications of the theory are – in common with other cognitive theorists – constrained by his view of the plausibility of the cognitive mechanisms involved.

Rational analysis and illogical inference

Edmund Fantino and Stephanie Stolarz-Fantino

Psychology Department, University of California San Diego, La Jolla, CA 92093-0109[1]

Electronic mail: *ps28%sdcc12@ucsd.edu*

Anderson's view of human cognition as an optimal response to environmental demands is appealing and potentially productive. If it attracts the attention it deserves in terms of research activity designed to elaborate it, this approach may well revolutionize the theoretical and empirical future of cognition. In addition, rational analysis has the potential to help bridge the unfortunate gap between cognitive psychology and more behaviorally oriented approaches, which we have lamented elsewhere (e.g., Stolarz-Fantino & Fantino 1990; see also Rachlin 1989; White et al. 1989). [See also Fantino & Abarca: "Choice, Optimal Foraging and the Delay-Reduction Hypothesis" *BBS* 8(2) 1985.]

We applaud rational analysis as an eminently plausible approach to categorization, causal inference, problem solving, and memory, but as the target article makes clear, the actual application in any given case – however straightforward in principle – may be difficult or impossible in practice. The four steps required to develop the optimal behavioral function for any given situation may be extremely difficult to traverse. Moreover, as Anderson notes, what "is optimal in the micro-world created in the laboratory can be far from optimal in the world at large" (sect. 1.2, para. 5). As an example, the target article discusses operant research in which local optimization may lead to behavior that is globally nonoptimal. In one series of such experiments, Heyman and Herrnstein (1986) showed that pigeons would match their choice responses to the rates of reinforcement obtained on two available schedules even when matching resulted in much lower rates of reinforcement than a maximizing strategy. Michael LaFiette and the first author are completing a comparable experiment with college students. Somewhat to our surprise, our students did no better than Heyman and Herrnstein's pigeons: Subjects displayed a matching strategy even when it resulted in lower monetary earnings. Anderson raises the question of whether organisms can be sensitive to such complex contingencies (citing Staddon 1987). In support of this caveat are the results of a second portion of our study: When cues in the experimental room made it clearer to the subject that more money could be earned with a maximizing strategy, our college students did adopt one.

Unfortunately, resolving some other examples of ostensibly nonoptimal behavior may be more difficult. For example, subjects demonstrating the conjunction effect found by Tversky and Kahneman (1982) report that the conjunction of two events is (more rather than less) likely to occur than one of the events alone. We have found this effect even when our instructions to college students were manipulated to facilitate logical thinking (Stolarz-Fantino & Fantino 1990). It is possible that when subjects display the conjunction effect, they are applying strategies that have proved adaptive in other contexts – categorization tasks, for example. If this is true, subjects may rate the sometime, "Linda is a bank teller and is active in the feminist movement," as more likely than, "Linda is a bank teller and may or may not be active in the feminist movement," because it seems more informative about Linda. In any event, it would be instructive to show how rational analysis might account for this type of illogical inference.

ACKNOWLEDGMENT

Preparation of commentary was supported by NIMH Grant MH-20752 to the University of California at San Diego.

NOTE

1. The authors are also affiliated with San Diego State University, San Diego, CA 92182.

Beyond Helmholtz, or why not include inner determinants from the beginning?

Hans-Georg Geissler

Department of Psychology, University of Leipzig, 0-7030 Leipzig, Germany

I am struck by the similarity between rational analysis (RA) and a methodology called the "adaptive view" (AV), which I suggested several years ago (Geissler 1983; cf. also Geissler 1976). There is agreement about the basic role of optimality assumptions as well as the stress put on the objective, environmental determination of cognition (Geissler 1983, p. 88). At the same time, overall closeness brings important specific differences into prominence. Most strikingly, AV differs from RA (a) by the emphasis it puts on inner codeterminants and on indirect validation of perception and (b) by the specification of optimality (p. 89). To some extent the differences between RA and AV reflect differences in the domains for which they are devised. There is a considerable overlap between both methodologies, however, which strongly suggests the need for a synthesis of divergent features. Specifically, I maintain that RA might benefit from explicitly taking into account inner constraints of cognition on which AV focuses. I will discuss this point in three steps.

Parameter constraints and cross-task analysis. Anderson takes care to state assumptions of RA as a basis for deductive procedures so that they do not refer to inner constraints (the "structure of mind") with the exception of short-term memory limitations. This position represents a favorable alternative to inductive curve fitting strategies which notoriously suffer from arbitrariness of parametrization. Note, however, that the parametrization objection does not apply to inductive cross-task comparison techniques that are based on parameter invariances among task-specific families of functions (for a brief survey see Geissler & Buffart 1985). This rationale has been successfully applied in visual space orientation (Geissler 1980; 1970), recognition of serial structures (Geissler et al. 1978) and multiple categorization (Geissler & Puffe 1983). Cross-task comparison procedures simultaneously validate parameter constraints of the processing systems and assumptions on processing strategies, whereas RA considers only the latter. Hence a potential drawback of RA seems to be that by excluding parameter constraints from explicit consideration it has to dispense with important theoretical tools. To illustrate within the context of the target article consider the function:

$$y = 4.93\,\{1 - [abs(t - t_0)/8.1]^{0.38}\}\,\{abs[0.2\,abs(t - t_0)] - 0.24\,d\} + 2.21 \qquad \text{(Equation 1)}$$

Equation 1 provides a rough approximation (obtained by a handheld calculator) of the confidence judgments of both Figures 2a and 2b of the target article, where d denotes distance, t time and t_0 time at the relative confidence maximum. To satisfy the weight data it suffices to assume $t_0 = 0$. The ball data require separate estimates of t_0 for each value of d. The result (Equation 1) is compatible with Anderson's notion that temporal and spatial contiguity are filtered through different models of the situation. However, in detail, Equation 1 is at variance with a straightforward rationality assumption: t_0 can be reasonably

interpreted as the time the spread of a vibratory wave or a ball movement takes to match the given distance. But then the content of the second product term of Equation 1 cannot be reduced to matching. This becomes obvious from the fact that the *d* range enters for the ball data the same way as for the weight data. Furthermore, there is no effect of absolute time. Rather, it enters into the first (distance independent) *and* into the second term relative to matching time.

The limited data base does not warrant a detailed discussion of the parameter constraints per se. A possible exception is the decay parameter of 8.1 seconds, which obviously depends on autonomous characteristics of the processing systems. It may be noted that there is a significant body of evidence from various paradigms of a universal critical time period of about 9 seconds (cf. Geissler 1987; 1991).

Rule constraints and optimality. In addition to parametric constraints the example illustrates the presumptive existence of a complementary type of inner constraint on cognition that may be called a rule constraint. This involves limitations of task adaptivity caused by the format of potential rules. As causality judgments are a particular case of information integration in the sense defined by N. H. Anderson, rule (1) can be looked upon as a complex "mental algebra" (cf., e.g., N. H. Anderson 1981). A first argument suggesting that this classification is useful comes from the empirical evidence that complex rules result from combining a small number of elementary rules. Thus a major goal of RA, the derivation of cognitive rules, could be related to an inner iterative process of rule construction.

A second argument relates rule constraints to optimality. An example bearing on sensation is given in Geissler and Puffe (1983). The general argument is as follows. Suppose some set of constraints among subjective variables is modelled by a set of rule constraints among subjective variables. Then, in general, both sets of constraints do not uniquely determine the mappings of objective variables onto subjective ones. Uniqueness can be attained by assuming extremality principles as a basis for selection rules. For sensory attributes these principles take the form of invariance or constancy predicting power laws of mapping for multiplicative, and logarithmic laws for additive rule constraints. In the realm of perceptual organization the same basic principle takes the form of the minimum principle of structural information (cf. Leeuwenberg & Buffart 1983).

Extremality principles of this type are forms of optimality criteria that seem to have nothing to do with those advocated by Anderson. On a deeper level, however, at least two relationships between both types of criteria can be relevant within a broader theoretical framework: (1) Cognitive performance can become optimal in the sense of an absolute structural minimum in a stationary state after the application of Anderson's rule. This may be trivial in some problem-solving tasks where the solver proceeds along a trajectory of minimal length from the start to the goal after recognizing the general solution. Minimality reached in this way can be nontrivial in other cases, for example, if the task of cognition is to find a shortest cognitive code of a category. (2) A cognitive code may involve a hierarchy, with both types of optimal criteria operating. A memory search task, for example, may be accomplished using the optimal memory code at hand.

Anderson's theory of categorization presupposes well-defined and fixed features. (2) will become relevant in situations in which features must be considered the result of structure formation processes that become part of adaptive behavior. This expanded notion of adaptivity may also apply to the assimilation of category structures. Geissler and Puffe (1983) and Buffart and Geissler (1983) provide evidence that the phenomenon of basic level categories, which is considered primary by Rosch and coworkers, Hoffmann and Ziessler (1983) and others, can be derived from a generalized minimum principle.

The place of Bayesian statistics. The above considerations can be summarized as follows: (a) The deductive something of RA and more inductive strategies of analysis may represent complementary exploratory paths rather than mutually exclusive ones. (b) It seems useful to complement the rationale of RA by techniques which take into account inner constraints on brain activity or, if you like, the "structure of mind." (c) RA and AV become related theoretical points of view as soon as the representation and generation of information become mutually dependent.

If this is true, what about Bayes' theorem? To me it seems a very forceful implementation of Helmholtz's principle of unconscious or inductive inference. It has turned out to be of great heuristic value in the prediction of visual illusions within the indirect validation framework I used in early work (Geissler 1970). Still, I suspect it needs modifying if inner constraints are implied. If a guess is permitted, it might take the form of a self-consistency relation established within neuronal networks of the types suggested by the principles of adaptive resonance (Grossberg 1988) or reentry (Edelman 1987).

Does the environment have the same structure as Bayes' theorem?

Gerd Gigerenzer

Institut für Psychologie, Universität Salzburg, 5020 Salzburg, Hellbrunnerstraße 34, Austria

Electronic mail: *1gigerenz@edvz uni-salzburg ada.at*

Cognition should not be divorced from its environment, argued Egon Brunswik (1964), comparing the two to a married couple who have to come to terms with one another by mutual adaptation. His "ratiomorphic explication" of cognition started with analyzing the statistical texture of the natural environments (the ecological validities) and the degree to which perception is adapted to that texture. Anderson's program of "rational analysis" is quite similar: To specify the statistical structure of the environment, and, on the assumption that cognition is adapted to that structure, to infer the structure of cognition – or, at least, to infer constraints imposed on cognition by the environment. Both Brunswik and Anderson study the coming-to-terms of the married couple as an adaptation of only one partner (cognition) to the other, and both view the mind as an intuitive statistician. But here the similarities end.

Does Anderson pursue his own program? The crucial Step 2 of the "rational analysis" is "to specify the structure of the environment to which cognition is optimized," which "is much easier to observe than the structure of the mind." How do we observe that environmental structure? Among three approaches, Anderson proposes the "appeal to existing scientific theory" as the most compelling, to be illustrated with a rational analysis of categorization. So let us look at that: What is the structure of the environment that is reflected in the structure of category formation? Anderson proposes two structural components, the disjoint partitioning of the object set, and the independence of features of objects. Both are necessary assumptions for his Bayesian modelling of categorization and other cognitive functions. In the case of categorization, the evolutionary rationale Anderson gives is twofold: (1) that species cannot interbreed (disjoint partitioning of the object set), and (2) that features within species are displayed largely independently of one another.

Even if these two structural components were characteristic for the evolutionary context – Anderson himself admits that independence does not hold when features are controlled by the same gene – the question is whether they are characteristic for other contexts, too, as Anderson assumes. Conditional independence is a mathematically convenient assumption in standard

Bayesian models, but not necessarily valid in natural environments. Brunswik in fact focused on the *dependencies* between features of objects in natural environments, which for him defined the texture of an environment. Similarly, physicians look for clusters of dependent symptoms to arrive at a disease classification. In general, conditional nonindependence among testimonial evidence, clinical symptoms, and other features poses a well-known problem in the sequential application of Bayes' rule, as it does in Anderson's "rational theory." Dependence between the prior probability and the likelihood ratio (which measures the impact of new information) in Bayes' theorem poses another problem (e.g., Birnbaum 1983).

This may be sufficient to illustrate why I do not think that Anderson pursues his own program: to analyze the structure of the environment. Rather, he seems to have started with Bayes' theorem as a model of rationality and to have assumed that the structural assumptions underlying Bayes' theorem specify the structure of the environment as well. This is a legitimate heuristic: to start with some statistical model of inference – Fisher's analysis of variance, Neyman-Pearson decision theory, multiple regression, Bayes' theorem – and to investigate the hypothesis that the mind is an intuitive statistician of that kind or school (Gigerenzer 1991). And this is what I understand Anderson to be doing. But this is not Anderson's program according to his own lights.

Is Bayesian statistics adaptive? "The information-processing implications of various environmental cues are not certain. This fact leads to the Bayesian character of the optimization analysis. . . ." Why? The same fact leads Brunswik to the multiple-regression character of optimization. Neither Brunswik nor Anderson explains why they believe that their respective statistics would be adaptive. Bayesianism seems to be flexible enough to apply to any environment, even to those commonly seen to contradict it, such as Allais's and Ellsberg's paradoxes (e.g., Jeffrey 1987). But insofar as there is a specific Bayesian model of some cognitive function, I believe that the question whether the model applies to a given environment can be answered only empirically, not a priori. This can be done by checking whether a *structural isomorphism* exists between a given environment (or task) and the specific Bayesian model (see Gigerenzer & Murray 1987, pp. 162–74). Similarly, if we want to see the mind as a rational intuitive statistician (Bayesian or otherwise), then we need to postulate not only a statistical algorithm, but in addition some heuristics (or a second-order algorithm) that check whether the structural assumptions of the algorithm hold in the given environment over time and space.

Toward domain-specific theories of cognition. One direction for revising the rational analysis would be to change the singular form "to specify the structure of the environment" (Step 2) into the plural form "to specify the structures of environments." Different environments may have different structures, and these may also change over time. Thus, the program would need an extra step before Step 2 to obtain a categorization of various environments. Let us call the product of this categorization a set of domains. Domains may correspond with respect to level of abstraction and predictive power to Rosch's basic level objects (e.g., Rosch 1978). For example, the recent proposal of domain-specific theories of reasoning has greatly advanced the potential to predict people's information search in the Wason selection task (e.g., Cheng & Holyoak 1985; Cosmides 1989). Proposed domains (of human interaction) include social contracts, threats, permissions, obligations. In a social contract, for example, a decisive structural component seems to be that a participant can be cheated and that subjects consistently search for information that can reveal potential cheaters. [See Maynard Smith: "Game Theory and the Evolution of Behaviour" *BBS* 7(1) 1984; and Caporael et al.: "Selfishness Examined: Cooperation in the Absence of Egoistic Incentives" *BBS* 12(4) 1989.]

Bayesian models can indeed be very useful in suggesting a conceptual language for talking about differences in structures across domains. But domains also have surplus structures, such as cheating options, which go beyond standard statistical structures. If we take Anderson's program seriously and start with a theory of environments (as opposed to starting with Bayes' theorem) then we might indeed make the "substantial discovery" that Anderson promises. But we might also discover that cognition is more flexible and does not always rely on Bayes' theorem and strong assumptions such as independence. A highly adaptive intuitive statistician of the mind might even work with exploratory data analysis.

Optimality and psychological explanation

Peter Godfrey-Smith

Museum of Comparative Zoology, Harvard University, Cambridge, MA 02138[1]

Anderson seeks to apply optimality methods to cognition without appealing to a biological justification for this approach, claiming that the optimality framework should simply "stand on its own feet in accounting for data." The program that results, according to Anderson, is a "high-risk, high-gain enterprise" compared to standard approaches to cognitive psychology. That this strategy is high-risk is quite certain; its promise of high gains is another matter. The main issue I discuss is whether the optimality approach, removed from its biological context, is in principle able to deliver the explanatory payoff Anderson seeks.

We must distinguish two kinds of explanatory success that any approach to cognition might deliver. First, there is what Anderson refers to as "organizing the data." A theory might produce impressive generalizations and predictions about actual cognitive phenomena, leading us to believe that it gives us an accurate account of the actual structure of the mind. But once we have a good understanding of *how* the mind is wired up, we should be led to a second question: *Why* is the mind wired the way it is? These questions are quite distinct, though it is possible for them to be investigated at the same time and with similar methods.

Optimality theory is controversial and high risk, but it may appear that the risks are offset by the possibility that this approach can yield both kinds of explanatory success. Very possibly, this is how things appear to Anderson, who thinks that as well as organizing the data, his approach takes seriously "the idea that there is reason for the way the mind is." The problem with Anderson's approach is that severing optimality from biology robs the optimality approach of its ability to yield insight into this mental raison d'être. This is because it is only as a component within a more general biological approach that optimality has this kind of explanatory potential.

Before I enlarge on this, it is important to recognize that Anderson might choose to claim only the first kind of explanatory role for his "rational analysis." That is, the sole purpose might be to describe the actual structure of cognition, making no claims to explain why we are structured the way we are. I do not discuss this view of Anderson's target article in detail. I am skeptical about his specific models but provide motivation for this skepticism with one brief comment only. It is unlikely that an analysis of some aspect of cognition in terms of costs and benefits will succeed if the only costs considered are those "internal" costs deriving from the mental effort involved in retrieving memories, forming hypotheses, and the like (Table 2). Such costs must surely pale beside practical, external costs resulting from bad behavioral choices. To some extent, internal costs are related to the external, practical ones. Expending "internal effort" is practically costly if it also spends time, for example, and the mental operations Anderson considers may monopolize the resources of attention in critical decision-making situations. But if Anderson thinks internal effort has signifi-

cant costs of its own, unrelated to the ways cognition affects behavior, this surely needs to be argued

The more important problem in Anderson's account, however, does not concern the details of his rational analyses, but his attitude toward optimality as an explanatory tool The real home of optimality-based, adaptationist explanation is within the second theoretical project outlined above, the project of explaining why cognitive mechanisms are structured the way they are Anderson seeks to explain the structure of the mind in terms of the statistical structure of the environment But even if we show that some aspect of cognition is an optimal response to a specific environmental situation, that on its own does not explain why this aspect of cognition exists in the form it does. For why should our brains contain mechanisms optimally suited to environmental problems? Optimality can be used to explain how we are wired only if this explanation also cites a process though which optimality is *generated* This is the role played by evolution by natural selection It could also be played by any mechanism of individual learning and cultural transmission which can produce approximations to optimal traits (Boyd & Richerson 1985) Links, similarities, and correspondences between the structure of mind and the structure of the environment explain nothing if we have no idea of a process generating and maintaining those relations

Now, as Anderson says, showing that an evolved structure is optimal is a very difficult business There is a range of evolutionary forces and factors that can thwart the optimizing pressure of selection (Gould & Lewontin 1979) These constraints and random elements are important components in an evolutionary perspective on cognition The explanatory power of optimality in psychology is only as great as the role of optimizing forces within the range of phylogenetic and ontogenetic factors that produce human cognitive mechanisms If other factors are more important than selection, either generally or in some specific domain, then we should look to these for an explanation of why cognition is the way it is It does not do to be skeptical about the known forces generating optimality, but then to adopt a strongly optimality-based approach to cognition, without proposing other factors by which this alleged optimality can be produced Showing that a certain relation exists between the mind and the world is not explaining much about cognition, unless we are able to show that the mind is the way it is *because* the world is the way it is

NOTE

1. Author is affiliated with the Philosophy Department (B-002), University of California at San Diego, La Jolla, CA 92093

Bayes in the context of suboptimality

Robert A. M. Gregson

Department of Psychology, University of New England, Armidale 2351, Australia

Electronic mail: *rgregson@gara une oz au*

As a committed Bayesian, I welcome Anderson's attempt to give coherence and rigour to a theory of cognition I think we should address the question of whether or not the formalism is sufficiently complete to be testable as a model of real behaviour It is Step 3 (the "Achilles heel" of section 1 2) that raises interesting problems for me

A Bayesian expression such as his Equation 2 follows from the axioms of probability if each of its component probability terms independently satisfies those axioms The $p(i|A)$ terms may initially be degrees of belief but a special sort of belief that is coherent has to be involved If a Bayesian foundation is to be adopted, then either we use well-behaved probability measures, or we use some psychological probabilities – call them ψp's – that fuzzily conform to some other axioms

Human observable behaviour is sometimes noted to be infraBayesian, in the sense that revisions of subjective relative likelihoods $\Lambda_{1,2}$, where $\Lambda_{1,2} = p(H_1|E)/p(H_2|E)$ do not occur in the face of new evidence to an appropriate degree

There are thus two distinguishable senses in which a Bayesian model of human cognition might depart from the purely mathematical expression Either I might write (in the simplest two-hypothesis case)

$$p(H_1|E) = \frac{\theta \cdot p(E|H_1) \cdot p(H_1)}{[\theta \cdot p(E|H_1) \cdot p(H_1) + \theta\, p(E|H_2) \cdot p(H_2)]} \quad \text{(Equation A)}$$

where θ is an operator which modifies the weight of E but at the same time the probabilities $p(E_1|E)$ are well-behaved; or I might rewrite the whole expression as

$$\psi p(H_1|E) = \frac{\psi p(E|H_1) \cdot \psi p(H_1)}{[\psi p(E|H_1) \cdot \psi p(H_1) + \psi p(E|H_2) \cdot \psi p(H_2)]} \quad \text{(Equation B)}$$

and wish to advance the notion that ψp functions also satisfy such expressions as Equation B The whole, however, may not be optimal in the sense of Equation A when θ is an identity operation It is not clear which way Anderson is going, and optimality in the second case may be hard to define (de Groot 1970) Note that θ is not a cost risk factor in the decision theoretic sense Anderson concatenates other such subjective probabilities as his *Con* terms, which is admissible but doesn't get round the problem implicit in Equation B

My two other concerns relate to the fact that real decisions are sequential in real time A Bayesian decision is a best decision in the long run, and if we wish to use the rule recursively then we need expectations about the future stability of the conditional probabilities $p(E|H)$ holding in the future Nonstationary processes can be treated with the modern variants of Kalman filtering (Spall 1988) but to assume that human judges can revise simultaneously, from incomplete data, the $p(E|H)$ or $\psi p(E|H)$ terms in their repertoire at a sufficient rate to get near optimising the revision of $p(H|E)$ is to assume a bit much

I can see that a symbiote of a computer and a human operator could function as Anderson suggests, but with classes of coupled probabilities, as in his section 3 1, and the class size itself a variable over time, I think we have a situation in which it is beyond unaided human intelligence to behave anything like rationally. Multivariate time series models have been constructed from a Bayesian perspective (Broemeling 1985), but I am not happy with notions that human neural networks are going to function that way

What Anderson is describing is a protocol for a ceiling on effective cognition; a sort of AI cognition without thought about the biological substrate which has to run the process It is refreshing to have some Bayesian arguments instead of the Neyman-Pearson inference that ruins much of experimental psychology, but Anderson's appeal to a connectionist net is an appeal to accept my Equation B as a network description. Networks are not, in this day and age, adequately summarised as lines and boxes; their properties are subtle and diverse A network can, for example, transmit chaotic patterns (Mpitsos et al 1989) but losing the parameter values involved in so doing If tight adherence to a Bayesian formalism is desired, the parameters have to get through, or the loss of precise parameter transmission through the cognitive apparatus has to be presented as a cause of suboptimality

I don't think that appeals to "general statistics" (sect 7) really say anything; the strength of a Bayesian approach is to use degrees of belief in a formalized sort of reasoning, as input to a recursive algorithm that subsequently is about probabilities, and about what are, in effect, recommendations for action It wasn't advanced (posthumously) by Thomas Bayes as a model of

imperfect human cognition, but as a prescription Anderson's program still has a long way to run

Rational analysis and the Lens model

Reid Hastie and Kenneth R Hammond

Center for Research on Judgment and Policy, University of Colorado, Boulder, CO 80309-0344
Electronic mail: *rhastie@clipr colorado edu*

We are impressed by the close correspondence between the rational analysis framework developed by Anderson and the "Lens" model for the analysis of judgment (Brehmer & Joyce 1988; Brunswik 1934; 1943; 1956; Hammond 1955) Briefly, the Lens model organizes the study of judgment phenomena according to the elements and relationships summarized in the diagram that gives the approach its name (Figure 1):

the judgments of the organism being studied (Y_s);

the attributes or events being judged in the environment (Y_e);

the cues that afford the judgment (x_i);

the judge's policy for cue utilization (a regression equation that summarizes the policy as weights and function forms that take the cue values as "input" and produce a predicted judgment ($\hat{Y}_s$));

the structure of the environment (a second regression equation that summarizes the structure as weights and function forms that take the cue values as "input" and produce a prediction of the environmental state ($\hat{Y}_e$))

Anderson's rational analysis can be mapped onto the Lens model: The model of the organism corresponds to Anderson's cognitive algorithmic model (represented in the ACT* architecture); the model of the environment corresponds to Anderson's rational analysis (a Bayesian algorithm for optimal performance) There are some infelicities in the analogy: Anderson uses Bayesian probability theory to formulate a model of adaptive performance in the environment, Lens modelers use statistical linear regression models (however, neither preference is a necessary one); Anderson develops his rational analysis by making a priori assumptions about cognitive computational limitations; the Lens modelers do not motivate their acceptance of the regression model's representation of the environment with explicit references to cognitive limitations Most of Anderson's applications of rational analysis are to hypothetical environments, but his framework suggests an implicit commitment to Brunswik's "representative design" methodology for identifying the structure of the actual environment; whereas the Lens model researchers' commitment to "representative design" is explicit (and enthusiastic)

The analogy between the two frameworks is inviting, and we would like to make some predictions about the general conclusions that will emerge from research within the rational analysis framework, based on the findings from Lens model research in the domain of judgment and decision making tasks (Brehmer & Joyce 1988; Slovic & Lichtenstein 1971) We should emphasize that the Lens model is typically applied to to-be-judged events that are sampled in a manner that is "representative" or reflects the actual pattern of occurrences of the events in a natural environment Obviously, if the structure of the stimulus set and its relations to the judged outcomes are altered (e g , to produce a uniform distribution of experimental stimuli across a factorial design), these general conclusions might change

(1) Accuracy ("achievement," r_a): Moderate to low levels of correspondence between judgments and environmental criteria in tasks that include ecologically typical levels of "irreducible uncertainty" (e g , clinical judgment, financial, and meteorological forecasting)

(2) Model-Model Correspondence (*C*): The strongest relationship among components of the framework holds between the two models constructed by the researcher

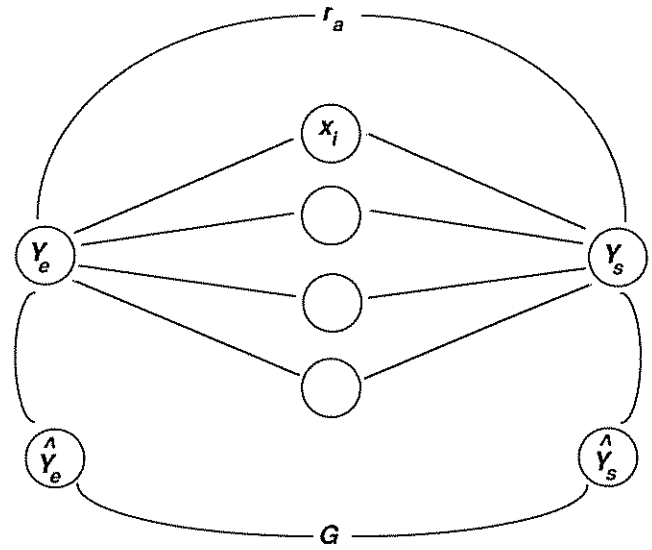

Figure 1 (Hastie & Hammond) Brunswik's Lens model: Environmental terms on the lefthand side (Y_e, $\hat{Y}_e$), cues for judgment in the center (x_i), and judgment responses on the righthand side (Y_s, $\hat{Y}_s$)

(3) Bootstrapping: The *model* of the judgment policy predicts environmental events more accurately than the judge's own judgments (due to a lack of consistent application of the policy by the judge)

(4) The model of the environment predicts behavior well – about as well as the model of the judge This is the observation that appears to have contributed most to Anderson's enthusiasm for rational analysis.

(5) Model-Judge Fit: Not surprisingly, regression models (and cognitive models) of human behavior, tailored to the specific task under consideration, describe judges' cue-utilization policies well.

(6) Model-Environment Fit: Here there is great variation in findings implying that we need to develop higher order principles to prescribe the types of models that are most appropriate for alternate types of environments This presupposes that we develop a useful theory of environments or tasks – we believe that this is the preeminent problem for psychologists who would respond to Anderson's challenge to pay more attention to the environment Lens model researchers have taken some initial steps in this direction (e g , Hammond 1988)

We believe that the parallel emphasis on the significance of environmental structure and the necessity for research design that reflects such structure by both a leader in the field of cognitive psychology and a widely known approach to the field of judgment and decision making signifies an important event. Strong similarities combined with different subject matters will produce complementarities in knowledge that advance both fields

Probing the "Achilles' heel" of rational analysis

Keith J Holyoak

Department of Psychology, University of California, Los Angeles, CA 90024
Electronic mail: *holyoak@cognet ucla edu*

Anderson's systematic attempt to explain basic aspects of human cognition by optimization assumptions provides a welcome

supplement to more familiar investigations of information-processing mechanisms. He is careful to avoid taking the stance of "methodological imperialism" (Thagard 1989), which would elevate one privileged methodology above all others. Nonetheless, as a proponent of rational analysis he would clearly like to minimize the dependence of rational theories on assumptions about computational mechanisms. Indeed, the third step in Anderson's stated procedure for developing a rational theory is to "Make the minimal assumptions about computational limitations" (Table 1) that are needed "to specify the computational constraints on achieving optimization" (sect. 1.2, para. 8). Anderson is aware that if these constraints were "complex and arbitrary," this step could prove to be "the true Achilles' heel of the rationalist enterprise" (sect. 1.2, para. 8). Anderson is optimistic, however, that two simple constraints will suffice: Considering alternatives entails cognitive costs, and short-term memory has limited capacity. Before accepting this sanguine assessment, it seems that a more skeptical examination of the potential Achilles' heel is warranted. I focus on two points.

First, an additional and fundamental class of constraints involves limitations on the types of information the system can encode or represent for use in cognitive tasks. There is no point, for example, in applying the rationalist framework to predict human responses to ultrasonic waves; rational analysis is only appropriate for creatures with the relevant representational capabilities, such as (in this case) bats. Representational constraints are by no means limited to peripheral sensory encoding. For example, in his analysis of memory for sentences, Anderson assumes that the memory trace of a simple transitive sentence can be decomposed into elements consisting of the concepts corresponding to the verb, the subject, and the object. It is somewhat unclear whether or not it is assumed that these constituent elements preserve information about role bindings (e.g., whether or not the elements of *The dog bit the boy* differ from those of *The boy bit the dog*). Whether the representations used in human memory retrieval encode such structural relations as role bindings is an empirical question (about which there has been some debate); and the answers to such questions about representational capabilities will surely have implications for a rational theory of memory retrieval. A related illustration is provided by the work of Kunda and Nisbett (1986), who found that people were much better able to estimate covariation between events that were highly codable (e.g., scoring basketball points) than events that people find difficult to encode (e.g., friendly behavior). In general, then, optimization is constrained by the representational capabilities of cognitive mechanisms.

Second, it can be argued that the cost of considering alternatives, one of the "minimal assumptions about computational limitations" acknowledged by Anderson, cannot be clearly defined except in relation to more specific assumptions about computational mechanisms. The time costs of considering alternatives will depend on such factors as whether the underlying algorithm is serial or parallel (Rumelhart et al. 1986). It is sometimes the case, for example, that the number of processing cycles required for a relaxation algorithm to reach a stable state proves to be independent of the size of the network (e.g., Thagard et al. 1990), so that given an adequate number of processors, solution time will be roughly constant over wide variations in the numbers of units and links. Given that each unit may represent an alternative hypothesis about the interpretation of a constituent of the input, it follows that how much additional time (if any) it costs the mind to consider each alternative may depend on the computational mechanisms available to make the choices. Although some cognitive decisions depend on the serial and costly evaluation of alternatives, others most likely do not.

There is therefore reason for rationalists to guard the Achilles' heel of their framework with some care. The constraints on information processing imposed by the computational mechanisms provided by evolution are undoubtedly complex, although unlikely to be arbitrary, and the progress of rational analysis may be more dependent on our prior understanding of these mechanisms than Anderson would hope. He may take cheer, however, in reflecting that Achilles ran far and swiftly before anyone could bring him down.

Adaptive rationality and identifiability of psychological processes

Dominic W. Massaro[a] and Daniel Friedman[b]

[a]Program in Experimental Psychology and [b]Department of Economics, University of California, Santa Cruz, CA 95064

Electronic mail: *massaro@fuzzy.ucsc.edu*

This commentary addresses identifiability – a single, but central theme of Anderson's important treatise on adaptive rationality. Anderson states that his research, the research of other psychologists, and psychological science in general is hamstrung by an identifiability problem: "And so I will bluntly say that it is just not possible to use behavioral data to develop a theory of the implementation level in the concrete and specific terms to which we have aspired" (Anderson 1990). (Anderson is using implementation level to refer to the functional level of theorizing that is dear to the hearts of most psychologists.)

Identifiability is related to the important theorems of E. F. Moore and subsequent work in formal automata theory. Moore was concerned with the behavior of sequential machines. As observers, we have access only to the machine's inputs and outputs. It is not possible to look inside the machine's black box. The question is, to what extent can one machine be distinguished from another, given a set of input-output observations? One of Moore's (1956, p. 140) theorems states that "given any machine S and any multiple experiment performed on S, there exist other machines experimentally distinguishable from S for which the original experiment would have had the same outcome." In other words, Moore proved that a given input-output function can be exactly mimicked by some other different input-output function.

As an example of a lack of identifiability, there are two well-known methods of multiplication: successive addition and logarithms. In the first case 5×7 is computed by adding 7 to 0 five times. In the second case, the logarithm of 5 is added to the logarithm of 7, and then the anti-log of the sum is taken.

These two methods can correspond to two different machines or to two different psychological processes. Moore's theorem states that input-output functions would not allow us to distinguish between these two machines. The same input-output functions are observed for both machines. For psychology, this means that some model of an experimental result is not unique. Some other model can be manufactured to give exactly the same predictions. For Anderson, this identifiability problem places an enormous constraint on "traditional" psychological research concerned with mechanism or process. How can we converge on a given process or mechanism when we can always compose another set of processes to make the same prediction? Anderson's solution is adaptive rationality. He states: "A rational theory can also provide help with the identifiability problem." Somehow, by analyzing behavior within this framework, we can bypass the identifiability limitation. For Anderson, rationality is a normative concept that postulates that behavior is optimal given the environment and given computational constraints.

Our commentary addresses several points related to this claim. First, we are sympathetic to aspects of the overall approach. It is productive for psychologists to use normative models as benchmarks for actual performance, and many have done so recently. For example, Massaro and Friedman (1990) used Bayes' theorem as a benchmark for assessing the predictive properties of psychological models of pattern recognition. In

addition, the predictions of the models were tested against actual results. The conclusion was that behavior appeared to follow the normative predictions of Bayes' theorem – with the additional assumption that the probability of a response reflected the posterior probabilistic prediction of Bayes' theorem. Psychological models equivalent to Bayes' theorem were therefore supported, while those making different predictions were falsified. Even though we used a normative analysis, identifiability problems remained. Cohen and Massaro (in press) modified the falsified models to bring them in line with normative predictions and actual results. The latter effort illustrates the reality of Moore's theorem and Anderson's concern with identifiability. The falsified models can be modified to make the same target predictions. Thus, identifiability remains a barrier, even when a normative analysis is used.

Just as several psychological models can be used to predict a given input-output function, several different normative models could be implemented for a given situation. The normative models might make the same predictions or different ones. In either case, it is necessary to test among them. If the different normative models might make different predictions, then it is somewhat easier to determine which one is correct (see also Gigerenzer et al. 1988). For example, there are two normative models in the Linda-is-a-feminist situation – straight probability theory and Bayes' theorem. Tversky and Kahneman (1983) assumed that the first model is the one that should be used to perform optimally. Massaro (1987) has argued, on the other hand, that Bayes' theorem is more appropriate for optimality. We might assume that subjects could use either of these algorithms to derive their answer. The first algorithm might be deemed appropriate if the subject interprets the question as one involving the likelihood of single or multiple events (e.g., Is Linda a bank teller? or Is Linda a bank teller and a feminist?). The second algorithm might be deemed appropriate if the subject interprets the question as one involving the likelihood of something given a single source or given two sources of information. That is, what is the likelihood of Linda given bankteller? or the likelihood of Linda given bankteller and feminist? The scientist needs to determine which of the two models actually corresponds to the processes (mechanisms) engaged by the task of interest. This question about rationality is identical in form to a question about process.

The major reason Anderson's rational analysis does not solve the identifiability problem, however, is his assumptions about processing constraints in the implementation of the models. For example, he assumes that there is a processing cost for retrieving a memory. Thus, tests of this description of behavior have the same identifiability constraint as process-oriented research. Without some sort of independent evidence for the processes (mechanisms) assumed by the implementation of the two models, we cannot determine whether behavior is rational. By permitting computational constraints (Step 3 of Anderson's Table 1), a normative model could always be implemented in several ways with different predictions in the different implementations. We need some independent evidence, other than the "normativeness" of the model, that allows us to prefer one implementation over another. This evidence is exactly the goal of process-oriented research that Anderson claims is plagued with identifiability problems. Perhaps Anderson is right when he states that specifying the computational constraints on achieving optimization "is the true Achilles heel of the rationalist enterprise."

To state it baldly, it should now be obvious that even research within the perspective of rationality faces problems of identifiability. We have demonstrated that the framework of adaptive rationality does not lessen the identifiability problem. The primary reason is that rationality analysis cannot be carried out independently of process concerns. This brings the investigator right back within the constraints of process-oriented research. In addition, any criterion insisting on rational behavior in every domain is unrealistic. As stressed by Gould (1986), and acknowledged by Anderson, we can be certain that not all behaviors are adaptive. Thus, the investigator will have to distinguish among both rational and irrational models of behavior.

Psychologists are hence left with choosing among a multiplicity of sufficient models – consistent with Anderson's claim of the futility of process-oriented research, but inconsistent with his remedy. Anderson argues that the search for psychological processes or mechanisms is plagued with identifiability problems. We agree with this assessment. We believe that the barrier of identifiability can be overcome, however; scientific inquiry can potentially choose among equally accurate models by extending the empirical data base, evaluating the models on the basis of parsimony, and testing among viable models using the principles of falsification and strong inference (Massaro 1987; 1989; see also multiple book review, *BBS* 12(4) 1989). In Massaro & Friedman's (1990) analysis, some of the unidentifiable models in a task with just two response alternatives made different predictions for four responses. In Cohen and Massaro (in press), some of the indistinguishable models required more free parameters than others. Extending the data base is a valuable strategy for distinguishing models that make identical input-output predictions.

To return to the multiplication example, reaction times (RT) are a valuable dependent variable. There is an illustrative series of experiments on how children add two numbers. Experiments have been able to distinguish between two viable models of addition by measuring reaction times to different problems. The results indicate that 6 + 3 takes about the same amount of time as the problem 4 + 3. These problems take longer than 7 + 1, however. In total, the results indicate that, at one stage of development, the child recognizes the numbers, chooses the larger one, and then adds the smaller number by counting from the larger to the smaller in steps of one. One could cast doubt on the previously mentioned log-anti-log model by looking at reaction times and accuracy for larger multiplicands. Thus, the identifiability problem is not as insurmountable as Anderson claims.

In closing, Anderson's thesis might be illuminated by noting an important distinction made in evolutionary biology between proximate and ultimate causes of (or influences on) behavior (Alcock 1989). Proximate causes address psychological processes that influence behavior. For example, we might ask what environmental information the gannet (a large seabird) uses to signal closing its wings when landing on water. Ultimate causes might concern why the gannet closes its wings when landing – what evolutionary significance it might have. As psychologists, we have usually been concerned with proximate (immediate or close in time) influences. For example, what are the visual features actually used in letter recognition? How are these features combined? and How is a decision made given this information? Ultimate causes, such as the evolution of the visual system to detect edges and other properties of letters, are of less interest. The psychologist's concern with proximate causes, in many respects, makes the framework of evolution less applicable.

Anderson's rational approach works on the level of ultimate causation. To the extent that proximate and ultimate causation interact, a rationality framework helps lessen the identifiability problem. Indeed, normative analysis is often a valuable guide to the study of behavior, and Anderson makes a useful contribution in forcefully stating its importance and illustrating its uses. However, the psychologist's primary charge remains one of uncovering proximate causation – how behavior is actually achieved, not simply how it is optimal or how it reflects its evolutionary history. As a framework for proximate causation rational theory appears to offer no free lunch.

ACKNOWLEDGMENT

The writing of this paper was supported, in part, by grants from the

Public Health Service (PHS R01 NS 20314) and from the National Science Foundation (BNS 8812728)

The cognitive laboratory, the library and the Skinner box

Howard Rachlin

Psychology Department. State University of New York. Stony Brook. NY 11794-2500

Those of us who have been trying to use economic models to explain the overt behavior of whole animals can only applaud such attempts as Anderson's to do the same in cognition Anderson's theories of memory, categorization, and causal inference throw new light on these phenomena, thereby broadening the power of economic analysis in psychology

There is a danger in using this approach, however, one that becomes particularly critical in the analysis of cognitive processes – the danger of taking a concept that has an operational meaning in a money economy and applying it in a vaguely metaphorical sense to a biological process Thus, a squirrel storing nuts may have only surface similarity to a person depositing money in a bank; the dynamics of the two processes (in response to variation of environmental constraints) may fundamentally differ

Consider "cost " In a money economy cost may be reckoned in terms of money; in a barter economy cost may be reckoned in terms of the amount of a commodity given up But what corresponds to cost in a cognitive system? The list in the last column of Table 2 is not helpful The concept, "internal effort" has some intuitive relation to economic cost but is not defined operationally The problem is not with the rationalist enterprise as such but in the application of the rationalist mode of analysis to wholly covert cognitive mechanisms Such an application is not only metaphorical but the metaphors are inevitably mixed

To see what I mean, consider how you might talk about the functional characteristics of an automobile as a whole; you might refer to acceleration, handling, braking, comfort, appearance, and so on, using the sorts of terms you see in *Consumer Reports* These terms refer to functions of the car as a whole But these terms (and the functions they imply) are completely inappropriate in discussing the operation of an internal mechanism like the carburetor A functional analysis of the carburetor is certainly possible but it would have to be made in terms of the carburetor's own environment (the thermodynamics of the air-gas mixture), not that of the car as a whole

Anderson applies his economic analysis to cognitive processes to discover an underlying internal mechanism (as in sect. 6) The difficulty is that an internal cognitive mechanism (unlike the carburetor of an automobile) cannot be isolated and studied in relation to *its* environment Instead, Anderson talks about cognitive mechanisms as if their environments were those of the person as a whole This is like talking about the acceleration, handling, comfort and appearance of a carburetor

As an illustration of both the power of Anderson's economic analysis and the problems involved in applying it to internal mechanisms, let us consider section 2, "A Rational Theory of Memory " It begins: "The goal of memory is to get access to needed information from the past (e g , remembering where the car is parked in the airport parking lot) " But surely the goal is to *find* the car; even remembering where it is would seem to require imagination and thought as well as memory – and then you would still not have found the car It is unclear how you would get from the retrieval of a memory (a picture of the parking spot? a number like E7? a view of a region? a cognitive map?) to the behavior, or even whether the former is a necessary step in the latter

In discussing "The History Factor" (sect 2 1), Anderson says that the best way to get from history to current use would be to "follow people about in their daily lives keeping a complete record of when they use various facts " But even if you did follow people around, how would you separate the facts making contact with their memory mechanisms from the facts affecting their behavior directly, from the facts remembered and forgotten, from the facts ignored? This is an especially acute problem when molar characteristics of behavior (e.g , response rate over a time period) are direct functions of molar characteristics of the environment (e g., reinforcement rate over a corresponding period). Instead, Anderson quite sensibly observes such activities as book borrowings from libraries and access to computer files These interesting data are *examples* of memory processes (with entirely overt costs and benefits) but Anderson uses them as *analogies* to a completely different memory process, an occult one, going on in a person's head

Anderson shows that people's behavior in cognitive laboratories may be derived from their behavior in libraries (assuming that the predictability of Phenomena 1–3 from the library and computer access data is not based merely on the statistical properties of the group data usually obtained in cognitive studies). This discovery is no inconsiderable accomplishment in itself But I question whether the postulation of an internal library-like process (a library in the mind) is a necessary intermediate step

Ignoring that intermediate step, Anderson might have noted that the correspondence between people in the library and people in the cognitive laboratory extends even to a pigeon in a Skinner box; Phenomena 1 and 2 exemplify generalized matching (Baum 1974); Phenomenon 3 is straightforward temporal discrimination, Phenomena 4 and 5 correspond to those discovered by Rescorla, Jenkins and Gibbon (Gibbon 1986) in the control of behavior (A) by conditional and discriminative stimuli (i's) The present article is some evidence that (as I have always suspected) these phenomena are fundamentally economic in nature

The same sort of comments apply to categorization, inference, and problem solving. As with memory, these processes are easily definable as purely overt behavioral processes Skinner notwithstanding, there is no need to stop using mental terms when talking about behavior The relation between mental events and physiological events may be characterized as one between patterns of overt behavior and underlying physiology rather than (or as well as) between computer flow diagrams and underlying physiology (Rachlin 1989) The data Anderson uses to construct his models are (necessarily) behavioral and the conclusions apply not only to cognitive studies of humans but also to ethological and operant studies Anderson's insistence that yet-to-be-discovered internal cognitive mechanisms will also work that way cannot of course be disproven but is, so far, gratuitous

The demonstration that, given *environmental* constraints as they are, people process information as efficiently as possible (i e , that they extract reinforcement efficiently from a varying environment) would be a great contribution to behavioral psychology I think Anderson thinks this is what his research does show I just wish he'd say it directly

Rational analysis will not throw off the yoke of the precision-importance trade-off function

Wolfgang Schwarz

Freie Universität Berlin. Psychologisches Institut. D-1000 Berlin 33. Germany

In cognitive psychology it is generally recognized that the importance (or scope) of a theory on the one hand and the

precision of its formulation (and thus the degree to which it is testable) on the other hand are inversely related. A successful theoretical approach will be uniformly superior to its competitors, that is, for any given precision, its importance will be greater.

"Why not *both* – precision *and* importance?" – to me, this seems to be Anderson's basic motivation for his "rational analysis" of cognition. To this end he has borrowed some approved elements (such as sequential decision-making or process-oriented stochastic modeling) from precise, but narrow paradigmatic approaches and used them to represent human memory on a large scale. As always, his presentation is alluringly many-sided, impressively far-reaching, and enviably well written. I do not think, however, that Anderson's rational analysis throws off the yoke of the precision/importance trade-off (PIT) function. Rather, I find his arguments about cognitive adaptation and optimization stimulating from a substantive point of view, but relatively inaccurate with regard to theory formulation and testability.

A first significant hint in this direction is Anderson's complete omission of the two books – and large parts of their contents – which essentially define the current state of the art of psychological modeling, namely Townsend and Ashby (1983) and Luce (1986).

A second point is that most of Anderson's formal representations are only qualitative statements, with no precise model-theoretical background. To give just the simplest example: On page 13 of his classical treatise, Feller (1968) admonishes us never to "speak of probabilities except in relation to a given sample space." Anderson (sect. 2, para. 2) repeatedly uses rather complex joint probability measures like $P\{A, H_A, Q\}$. What exactly is the sample space associated with these measures? From my point of view (e.g., Schwarz 1989; 1990), this omission is not just technical carelessness – rather, it points to the principal problem that such things as "historic factors" and "sets of cues" (sect. 2, para. 5–6) are simply too complex to be useful for the definition of a formal measurable set on which probability statements can be based. The exact relation of Equation 2 to the developments in sections 3–5 or, in fact, to any observable event is not clear to me: Precisely how does the subject use the odds defined on the left side of Equation 2? Also, it is well known (e.g., Chow & Schechner 1985) that Anderson's (Equation 1) one-step look-ahead decision rule will not necessarily be optimal under a lot of very plausible conditions. That the need probability $P(A)$ is monotonically related to the latency and probability of recall should ideally not be a "basic assumption" (sect. 2, para. 2), but rather a consequence of a given model.

Bayesian analysis (sect. 2, para. 4) is based on first (and purely formal) probabilistic principles; it goes nowhere beyond "conventional" probabilistic analysis, to which it properly belongs as a useful corollary; and by itself it does not constitute any substantive insight (cf. Feller's, 1968, p. 124, penetrating "Note on Bayes's Rule"). Actual applications require additional assumptions about the specific representation and retrieval processes involved in a given task.

Thus, Anderson and Anderson & Milson (1989) elaborate on a stochastic theory of storage and retrieval processes from Burrell (1980), which is precisely the kind of modeling (e.g., queueing theory, hazard rate) that we lack in cognitive psychology. Indeed, how much we actually lack it is perhaps best illustrated by Anderson and Milson's (1989, pp. 705; 718) surprising "discovery" that the mean of a posterior density of a rate parameter behaves for all practical purposes like a probability – all the more surprising as Anderson and Milson (Equation 2, p. 704) are concerned with likelihood ratios anyway, which may in general exceed 1, not only for densities, but also for "true" probabilities.

The information-retrieval model considered by Anderson (sect. 2.1) corresponds to the often-used library metaphor and is evidently plausible. A problem here is that there are many stochastic process models which are no less plausible – a quick glance through (e.g.) the *Journal of Applied Probability* invites us to compare human memory not only to a library, but also to proofreading (Chow & Schechner 1985), oil exploration (Benkherouf & Bather 1988), the selection of the best secretary (Petruccelli 1981), and so on – activities that all bear some structural similarity to fundamental memory functions. What we really need is a general and formal methodology to compare competing models and to identify their critical and relevant parts. Excellent examples are the works of Townsend and Vorberg on the identifiability of serial versus parallel processing (Townsend 1971; Vorberg 1977) or on general continuous-time Markov accounts of memory (Vorberg & Ulrich 1987).

Exploring the behavior of a model by simulation (target article, sect. 1.2, para. 11; Anderson & Milson 1989), though occasionally informative at an early stage, tends to produce an unsuitable overconfidence in our understanding of the simulated model and it often introduces an unpleasant degree of subjectivism, favoring parameter settings that are purely ad hoc (Anderson & Milson 1989, p. 707) and not leading to a sound and established estimational and hence evaluative procedure. Also, simulation may lead to a premature abandonment of efforts to derive closed-form analytic expressions. For example, models very similar to Anderson's (sect. 2.1) "augmented Burrell model" have been (or can be) solved analytically. In contrast, the general "multinomial" approach to the modeling of cognitive processes recently introduced in an exceptionally clear and important article by Riefer & Batchelder (1988) offers a general theoretical frame in which it is possible to formulate, estimate, and evaluate most current approaches to cognition (including, for instance, Anderson's rational analysis of problem solving, sect. 5) in an explicit and rigorous way.

On the nonapplicability of a rational analysis to human cognition

Eldar Shafir

Department of Psychology, Princeton University, Princeton, NJ 08544
Electronic mail: *eldar@clarity.princeton.edu*

People are complex organisms who interact with their environment in diverse ways. Like other animals, we are very good at certain of these interactions, and less impressive at others. How good animals are at particular tasks depends on many things, among them the structure of their environment and the nature of their mental apparatus. Anderson urges us to "take seriously the idea that there is a reason for the way the mind is." He then suggests the reason: "Behavior is optimized to the structure of the environment." Although most people would find the first suggestion unobjectionable, the second is less convincing. After all, there is a reason for the way telephones are, and the human spinal cord, and international borders, and all these are probably not optimized to the structure of the environment.

Although he acknowledges in passing the "problematical" nature of evolutionary considerations, Anderson then goes on to explore "how much explanatory power can be achieved in the case of human cognition." This, as others have argued before, is pure speculation, which we may never be able to investigate properly (see Lewontin, 1990, for an excellent discussion). For one thing, the forces of natural selection come into play only when there is difference in the probability of survival and reproduction. And it is not at all obvious that the presence or absence of some detail in cognitive function will have a significant influence on reproductive rates. Furthermore, even if there are differences in reproductive rates, those differences cannot be the cause of evolution unless they are transfered genetically. And it is not clear that our particular ways of solving

problems, for example, are genetically determined, as is only emphasized by Anderson's insistence on the shaping role of the environment

Given that the optimizing nature of behavior is not immediately obvious, nor entailed by evolutionary theory, it makes most sense to consider some empirical evidence that sheds light on the issues In fact, a large amount of evidence has been collected that indicates less than optimal behavior in many situations Let us touch on the issues of mental effort, local versus global optimization, and problem specification, all of which figure prominently in the target article.

According to Anderson, one of the key predictors of human performance in a rational theory is the cost of mental effort But numerous studies of behavior indicate that people are often willing to expend more mental effort on heuristics that are in fact less ideal for the task at hand For example, Kahneman and Tversky (1984) show that people who are about to pay $15 for a calculator are more willing to drive to another store to buy the calculator for $5 less than is a second group of people who are about to buy a calculator for $125 Although the full account of this behavior is beyond our present purpose (for a discussion, see Kahneman & Tversky 1984; Shafir et al 1989; Thaler 1980), suffice it to note that, in this and other problems of its kind, people deviate from the standard rational theory of consumer behavior by working harder to evaluate gains and losses in relative rather than absolute terms By basing their decision on the same $5 difference rather than the ratio it forms of the total price, subjects would save on mental effort and conform with the rational model, and this is what they do not do In a similar vein, when asked to estimate the likelihood that a totally uninformative description of a person is of an engineer, subjects use the more effortful representativeness heuristic, rather than simply relying on the base rates that are given and that would lead to a normatively more adequate response (Kahneman & Tversky 1973) Although in Anderson's rational analysis "cognitive performance maximizes the difference between the expected gain and cost of mental effort," subjects in the experiments above do not seem confined to such considerations

Anderson argues that whereas local optimality is normally achieved, global optimality is sometimes foregone for reasons such as limited memory But a number of studies show that the discrepancy between local and global strategies really is not so easy to resolve Redelmeier and Tversky (1990; see also Slovic et al 1978), for example, asked hundreds of clinicians to make treatment choices pertaining either to a single patient or to a group of comparable patients The clinicians weighted certain criteria (such as personal concerns of the patient and cost effectiveness) differently in the two cases and, as a result, exhibited different preferences between treatments They preferred one treatment when dealing (locally) with one patient, and the other treatment when contemplating (globally) the entire group Although their local and global preferences are clearly discrepant, it is not at all clear that one is "right" and the other "wrong": It may just be that we give relatively more weight to the human dimension when patients are in our office, and less when we contemplate the finances of a public health policy Although not licensed by Anderson's rational approach, conflicting local and global strategies may just be a natural outcome of the way we process information

A precise description of the environment is a critical part of the rational theory Work on "framing" has shown, however, that the same description of an environment leads to different behaviors when framed differently Thus, when asked to choose between two alternative treatments of a disease, people prefer one treatment if the outcomes are framed in terms of lives lost, and the other treatment if the problem is framed in terms of lives saved (see Kahneman & Tversky 1984; McNeil et al. 1982) Such violations of "description-invariance" as well as of "procedure invariance" (where people express discrepant preferences depending on the particular elicitation procedure that is used; see Slovic et al , 1982, for a discussion) have now been documented in numerous domains, in hypothetical as well as real world situations, with both high and low stakes, and both with and without monetary incentives These behaviors are not going to go away, and they do not seem to point us in the same direction as Anderson's conclusion "that many of the major characteristics of human cognition can be explained as an optimal response, in a Bayesian sense, to the informational structure in the world."

The rationality of causal inference

Thomas R. Shultz

Department of Psychology, McGill University, Montréal, Québec H3A 1B1, Canada

Electronic mail: *ints@musicb.mcgill.ca*

Rational analysis could be viewed as an attempt to rise above some of the major debates in contemporary cognitive science concerning the best mechanistic explanations of cognition It does this by showing that a wide variety of cognitive phenomena are optimal responses to goal satisfaction in particular environments given certain minimal computational limitations.

The breadth of intended coverage for rational analysis is impressive At the current rate of application, we may soon be asking whether it qualifies as a candidate for a unified theory of cognition One reason the approach seems so general is that it operates at a more abstract level than the more mechanistic candidates for unified theories

Although Anderson admits that ordinary humans do not explicitly undertake Bayesian computation, he claims that rational analysis can easily be converted into plausible reasoning mechanisms It is doubtful, however, that the constraints supplied by rational analysis are sufficient to favor particular mechanisms Much current debate centers around the fact that a fair number of distinct algorithms account for many of the phenomena listed by Anderson Rational analysis will satisfy researchers used to mechanistic accounts only insofar as it supplies sufficient mechanistic constraints

One of the four main areas of application in the target article is causal inference, an area that optimizes my own interest and expertise Anderson specifies his analysis of causal inference in two concise Bayesian equations (9 and 10). As with many Bayesian analyses, it is difficult to understand how the probabilities on the right sides of these equations are any more fundamental than those on the left sides For example, in Equation 10, the probability of a rule applying in the presence of cues seems no less fundamental than the probability of the cues if the rule did apply What evidence would support Anderson's view that the reverse is true? What sort of reasoning mechanism would conform to this arrangement of conditional probabilities?

Anderson makes a very useful point when he argues that cues to causality depend on underlying causal models (cf Doyle, 1989, for a more elaborated version of this). He applies this notion a bit too narrowly, however For example, in Phenomenon 15, Anderson argues that similarity is used when the subject holds a causal model specifying that a cause transfers part of itself to the effect, as in coloring phenomena But we have found that similarity is also used in contexts of number and size (Shultz & Ravinsky 1977). These cases too could be understood as using similarity only when it conforms to a causal model, but this would require a somewhat wider range of underlying causal models.

Anderson stresses that in a rational analysis the goal of causal inference is to predict future events But causal inference traditionally involves explanation and planning as well as prediction People may be interested in explaining an event by finding its cause even if the need for predicting the event never arises Once causal relations are known, they can also be used in the

construction of plans to satisfy goals. As Anderson (1990) realizes, this brings us into the related realm of problem solving. My point is that an unwarranted theoretical emphasis on prediction leads to the unfortunate theoretical neglect of explanation and planning.

Indeed, the only experiment Anderson reports here in much detail (Phenomenon 14, temporal and spatial contiguity) is one that involves explanation (or attribution, as it is more commonly known). Phenomena 13 (contingency) and 15 (similarity) also involve causal attribution rather than prediction. With respect to Phenomenon 16 (generalization), Anderson (1990) does present some new and interesting experiments on prediction, but it seems to me that the goals for causal inference ought to include explanation at least, and probably planning, as well.

Phenomenon 16, that a causal law generalizes to objects of the same category, seems obvious from many perspectives and so is not a unique prediction of a Bayesian analysis. It is not a well studied problem, and Anderson's efforts may stimulate more research.

Phenomenon 13, ignoring joint nonoccurrence information in contingency analysis, could likewise be explained by people's tendency to avoid negative information, presumably because of its complexity (Schustack & Sternberg 1981), or to ignore nonevents, presumably because they cannot ordinarily be detected. Anderson's Bayesian explanation of this phenomenon appears to work less well than the foregoing explanations, first, because it has subjects assuming the causal relation they are trying to detect, and second, because it contradicts the well documented tendency of people to ignore prior probabilities (Tversky & Kahneman 1980). The tendency to ignore priors could bode ill for any account that postulates Bayesian mechanisms.

The fact that many of these phenomena have various mechanistic explanations is symptomatic of the fact that a rational analysis under-constrains mechanistic accounts. Although it may always be possible to design probability values that conform both to Bayesian formulas and to behavioral data, the exercise is not terribly satisfying unless it distinguishes a mechanistic account. If the mechanistic account turns out to involve computations on probabilities, then it behooves the researcher to use behavioral data to constrain the choice of probabilities more directly, rather than using probabilities that simply make the equation fit the data to be explained.

So far, Anderson provides no clear analysis of the time course of causal knowledge. Is novel behavior typically as optimal as well-practiced behavior? Eventually, this may get cashed out in terms of the progressive refinement of prior probabilities, but it would have been interesting to see this done on some well known developmental phenomena.

Anderson identifies all of his causal inference phenomena (13–16) as having previously been characterized as irrational. It is extremely useful to show how such phenomena can be alternatively described as rational under differing assumptions. Anderson could well have included research by Kuhn (1989) on the *errors* people make in using the covariation heuristic in causal reasoning.

But how does the brain think? (or wasn't that the question?)

Steven L. Small

Department of Neurology, University of Pittsburgh, Pittsburgh, PA 15261
Electronic mail: *sls@dsl.pitt.edu*

Introduction. Anderson describes a way to analyze cognition in terms of human behaviors that are optimized to particular features of the external environment. He refers to the method as "rational analysis"; I refer to the underlying principle as either the "adaptation" or "optimization" assumption. Using rational analysis, he investigates several aspects of human cognition. I believe that the formal statement of the adaptation assumption adds a useful perspective to our thinking about cognition, but that the formal application of the method in rational analysis does not explain the kinds of things that need explaining.

Adaptation and cognitive science. Human cognition is as likely a priori to be adaptive as other human brain processes, such as regulating blood pressure or seeing in color. That it can be shown to be adaptive is interesting, but does not lead to an understanding of how it works. Cognitive scientists have always operated under some form of the optimization assumption, yet they believed that their mission was to explain not how the behaviors were optimized, but how the mind actually produced the behaviors. Similarly, cognitive neuroscientists focus on the phylogenetic and ontogenetic development of the brain and how the resulting brain structures lead to cognitive behavior.

Definitions and theory application. Anderson begins by stating that the goal of cognitive science is to predict the "output of human cognition." After presenting his method of analysis of cognitive behavior, he applies it to four areas in cognition. A number of questions can be raised about the definition and the areas chosen for analysis.

First, cognitive behaviors are not exclusively "outputs"; in fact, many of the tasks studied by cognitive scientists, such as reading (Just & Carpenter 1980), listening to music, understanding visual scenes (Marr & Poggio 1977), and inventing new scientific theories (Langley et al. 1983), cannot readily be described as outputs. But there is no doubt that they are cognitive tasks.

Second, four aspects of cognition were chosen for an analysis according to the adaptive theory: (a) recall memory; (b) categorization; (c) causal inference; and (d) problem solving. These tasks have something in common that makes them easily analyzed in terms of optimization; they are (generally) conscious, symbolic (propositional), and controlled processes.

Third, the data best applied to the optimization theory also have idiosyncrasies in common: They come from carefully controlled studies of goal-directed, rewarded, rational, propositional behaviors. Anderson thus provides an interesting account of some statistical regularities of certain conscious goal-directed behaviors. Although this might be useful as a descriptive theory, it does not help in uncovering the explanatory theory sought by a majority of researchers.

By defining cognition in a narrow way, excluding automatic aspects of visual and linguistic processing, for example, some of these objections can be averted. It is undesirable, however, to restrict the notion of cognition in this way, as it excludes some of the best understood aspects of overall behaviors that are universally considered cognitive. To continue the previous examples, the understanding of spoken language or of visual scenes requires both so-called "low level" and "high level" cognitive abilities. The interdependence of such processing "levels" requires their coordinated consideration in the development of cognitive theories.

Neuroscience. Although some cognitive researchers are reluctant to include basic neurobiological concepts in the development of cognitive theories, there is increasing interest in using data from neurobiology to constrain theory construction in cognitive science. The two sides of this coin are (a) the use of basic neuroscientific data from researchers in anatomy, physiology, and pharmacology; and (b) the consideration of clinical neuroscientific data (from neurology and psychiatry) about people with obviously abnormal cognitive systems and specific neurobiological defects. Must cognitive scientists restrict themselves to the brain as a black box? The optimization assumption side-steps the entire issue of how the mind/brain works, but that is in fact the main issue for a large proportion of cognitive scientists.

In this regard, do maladaptive and pathological cognitive

behaviors need explanation in a theory of cognition? My view is that they do, and that instead of assuming adaptation, one should make the mechanisms of adaptation one of the principal foci of inquiry. There are constraints on cognitive behavior within the brain that are at least as important as those in the environment. A brain lesion in the left inferior frontal gyrus (Broca 1861) or a deficiency of dopamine in the prefrontal cortex (Meltzer & Stahl 1976) can lead to tremendous changes in cognitive behavior. Aphasia, disinhibition, auditory and visual hallucinations, self-mutilation, stuttering, and dyslexia are all cognitive behaviors, they all need to be explained by a cognitive theory and are best understood in terms of internal mechanisms rather than adaptations to the external environment. In fact, the combination of basic neurochemical research and the failure of clinical techniques based on external adaptation has led clinical psychiatry to drop the adaptation assumption as an explanatory tool in favor of internal, mechanistic (pharmacological) explanations.

Connectionist modelling and David Marr. Anderson describes a connectionist network that he suggests represents an "algorithmic level" description of (part of) his "computational level" conception, thus relating his proposal to the levels of theory construction postulated by David Marr (1982). Like many others, Anderson assumes an independence of Marr's levels, which may not be correct. I share the view advocated by some that there is a fundamental inseparability (a mutual interdependence) among these levels: One cannot know the correct theory without paying attention from the outset to the algorithmic and implementational aspects, as they themselves provide interactive constraints on the theory construction.

The architecture and processing of the connectionist network constitute the real theory: it is not a better theory than any other network architecture unless it better meets known processing constraints. Learning the regularities in the environment faster does not make the theory better, unless the character of this learning models human learning more accurately. Likewise, learning in an asymptotic fashion is only interesting if that is how people do it. *Any* network should learn some regularities of the environment faster when given some initial knowledge (Rumelhart & McClelland 1986). Anderson's arguments about his network ignore the fact that (except for models of learning per se) it is the acquired (internal) representations, and not details about the learning, that constitute the most interesting aspect of network models.

Conclusion. Anderson demonstrates formally that conscious recall memory, categorization, causation, and symbolic problem solving can be viewed as optimized (adapted) behaviors in response to inherent aspects of the external environment. That the functions of the human organism represent (in part) adaptations to the *milieu extérieur* has been widely held for over a century (Bernard 1865). The question for cognitive science is not whether cognition is adaptive – it is – or even which are the optimized aspects of the environment. The question is how the brain performs this adaptation, both developmentally and in everyday cognitive activities. Until we know something about this, we do not know anything about cognition.

A Bayesian theory of thought

Howard Smokler

Department of Philosophy, University of Colorado, Boulder, CO 80309

Anderson's work is quite interesting to one of a philosophical bent. Not only does he deal with a number of mental processes that are of interest to philosophers but he attempts to explain them under the rubric of a Bayesian theory.

Most philosophers consider the Bayesian model of expected utility the unique normative one for judging conscious human action and belief as rational. From their perspective the work of psychologists, even if it were to show that a significant proportion of human beings did not act consciously as if they were Bayesian maximizers, would not be of any significance. It would only reveal the fact that many of them already believe: that the majority of people are rational only at times, if at all.

From their perspective, it is encouraging that Anderson's work points in a direction opposite to the one indicated by the work of so many other psychologists. But this work is in no sense of real importance to philosophers. The program for naturalizing epistemology – of making the theoretical results of sciences, including the social sciences, bear on philosophical problems – is undercut in most cases by the distinction between the normative and the empirical. Psychological explanations of categorization, causal inference, and problem solving, on this philosophical view, provide a basis for normative prescriptions for belief or action. If the project of naturalizing epistemology is to succeed, some fresh thinking must be forthcoming on this subject; I wonder what suggestions Anderson might have.

Psychologists like Anderson have explored the theoretical possibility that rationality functions as a maximizing and therefore adaptive process in human beings, making an explicit analogy with the evolution of biological traits. But as others have pointed out, many maximizing processes are not conscious and the irony in Anderson's studies seems to be that even some of the conscious processes he studies are treated as if they were unconscious. Consider a conscious process explained in terms of nonconscious mechanisms: In his application of an optimizing model to causal inference, Anderson treats it as if it were unconscious, as if one could make a causal judgement without knowing anything about the meaning of the representations that allow such a judgement to be made. I concede that there are aspects of all conscious processes that are not conscious, but, in principle, should a conscious process be explained almost completely as the outcome of a preconscious or unconscious optimization? Some element of following a rule must surely be present if the notion of inferring is to remain true to the intuition that it involves conscious patterns of mental activity. Otherwise this leads without justification to a form of reductionism that is contrary to our intuitions about what it is to think rationally. Some unspecified mechanism is being assumed without further warrant. [See also Searle: "Consciousness, Explanatory Inversion and Cognitive Science" *BBS* 13(4) 1990.]

Another problem is that the specification of costs/benefits that allows for the operation of the optimizing mechanism is only metaphoric. What does it mean that the cost of making a causal inference affects the formation of a hypothesis? I do not grasp the underlying mechanism.

Rationality and irrationality: Still fighting words

Paul Snow

Department of Computer Science, Plymouth State College, Plymouth, NH 03264

Electronic mail: *paulsnow@oz.plymouth.edu*

Cohen (1981) occasioned a debate in these pages about rationality. He wished especially to rebut the idea that research into how humans perform deductive and probabilistic inference had "bleak implications for human rationality" (p. 317). Kyburg's (1983) *BBS* paper and the resulting commentaries freshened that debate, emphasizing domains of reasoning generally similar to Cohen's. Professor Anderson, in the current target article, looks for rationality in four specific cognitive domains.

Ironically, some of Anderson's explanations depend on subjects' proper handling of base rates (or prior probabilities)

Cohen, on the other hand, was concerned with explaining subjects' apparent failures to use base rates properly (as in the famous "cab color" inference problem, pp 328–29)

Of Anderson's four cognitive domains, three involve commonplace, learnable, and purposeful behaviors: classification, causal inference, and planning If the environment affords any strategies for these tasks that frequently produce respectable results at low computational cost, then research of the sort described by Anderson would seem to have a high a priori prospect of success

People will presumably do something when asked by experimenters to perform in these task domains It is plausible that what the subjects might choose to do would involve some trade-off between effort and efficacy

A search for a descriptive model of those subjects' trade-offs, like the one outlined in Table 1 of the target article, would thus seem to have bright prospects. There are plenty of degrees of freedom available to model makers What shall we designate as the measure of success or of cost? Which environmental constraints shall we recognize in the model? Which simplifying assumptions shall we adopt?

Success in coming up with a model seems unsurprising Lack of potential surprise, rather than criticism about iterative hypothesis refinement that Anderson easily deflects, dilutes the evidence's probative power; for as probabilists (Polya 1954) and nonprobabilists (Shackle 1949) alike remind us, we learn little from unsurprising events That is, our prior beliefs about the rationality of human cognition are apt to be unchanged by modeling success that was never much in doubt in the first place

The properties of memory are presumably not as open to subjects' choices as the other behaviors. Yet, here too, the modeler's jacket is generously cut We find in the conclusions of the target article yet another degree of freedom: Which "components" of the "cognitive system" are to be modeled as locally optimal? (By the way, in the example task of phone number recall discussed in the conclusion, might not the "cognitive system" learn a few mnemonic tricks if remembering phone numbers were important enough?)

And yet, of 21 phenomena studied by Anderson, only seven are accompanied by a remark that suggests that other workers interpret the phenomenon as evidence of irrationality. In the other 14, the rational analysis illuminates issues other than the "big question": Is cognition adaptive?

In the three phenomena cited in connection with the "history factor" of memory, a beautiful, nonobvious similarity is established among diverse engineering problems of designing information-retrieval systems. The six categorization phenomena support a critique of existing research in the field: a misplaced (in Anderson's view) preoccupation with predicting category labels Phenomenon 16 links results from the causal inference and the categorization models to suggest new hypotheses for investigation The final four phenomena support principled criticism of the kinds of tasks often studied in the planning and problem-solving literature

In addition, for all four domains, Anderson motivates interesting Bayesian models of how to approach a task Quite apart from whether or not people are "rational," these models are surely interesting in themselves, and useful to the artificial intelligence enterprise, among others

Anderson's rational perspective, then, is a successful point of departure for science. Existing interpretations of experiments are challenged, new hypotheses are suggested for further experimentation, and similarities are disclosed among problems that otherwise appear distinct A powerful general-purpose modeling tool, Bayesian analysis, is artfully adapted to specific applications

Does the worth of Anderson's perspective depend much on whether or not human cognition is really rational? It had better not, since on the evidence of the Cohen & Kyburg commentary, the experimental record is equivocal People don't even always agree on what being rational means

As a springboard for successful science, the opposing perspective is fruitful, too Systematic inquiry into how behavior departs from selected normative accounts of rational behavior has been a fertile source of inspiration for hypotheses (in the work of Kahneman & Tversky, 1979, for instance) and cognitive engineering design (such as the extensive work in the nonprobabilistic management of uncertainty surveyed by Prade 1985)

Big questions fascinate us; for many scientists, interest in such matters is surely part of the explanation of how they came to be scientists in the first place Having a suspicion about how the answer will turn out can be a strong motivating force We all know good scientists whose work is informed by deeply held religious convictions, and other good scientists whose belief in a universe without any hint of the supernatural is just as strong

If science is roomy enough for both views on that particular big question, then surely rationalists and their critics can both be accommodated. Professor Anderson succeeds in demonstrating that his beliefs are respectable On the other hand, it is probably too much to hope that science will resolve the rationality question, any more than it will settle questions of religion There is other work to be done, on questions that are smaller, but whose answers are more nearly within our cognitive grasp

Computational resources *do* constrain behavior

John K. Tsotsos[1]

Department of Computer Science, University of Toronto, Toronto, Ontario M5S 1A4, Canada

Electronic mail: *tsotsos@ai.toronto.edu*

Anderson says: "A rational approach encourages us to inquire about the structure of our actual environment and to design an algorithm optimal for it rather than to design algorithms which would only be optimal in some bizarre world " The rational world Anderson proposes is as bizarre as the approaches he is criticizing. Behavior is not only a function of environment; behavior develops as a satisficing function constrained by environmental conditions as well as computational resources (and perhaps other things) In fact, given our current understanding, the limits on computation imposed by our brains may play the largest role in shaping behavior

Anderson says that in his work he has yet to find computational limitations posing danger to his scheme. Unfortunately, I don't think he has looked hard enough The combinatorial problems are very apparent, and in fact in most (if not all) natural problems, optimal solutions are computationally intractable in any implementation, machine or neural. A few examples are in order

(1) Vision
- Unbounded visual search, using a passive sensor system is NP-complete (Tsotsos 1989; 1990a)
- Unbounded visual search, using an active sensor system is NP-complete (Tsotsos, submitted)
- Polyhedral line-labelling is NP-complete (Kirousis & Papadimitriou 1985)

(2) Reasoning
- Finding the optimal satisficing strategy for simple and-or graphs is NP-hard (This refers to the task of deciding which operator to use to reduce a goal to its subgoals) (Greiner 1990)
- Finding the best explanation for a class of independent problems using probability theory (and several other

forms of abduction) is NP-hard (Bylander et al. 1989)
Abductive reasoning for all but the simplest theories is NP-complete (Selman & Levesque 1989)
Many forms of default reasoning are NP-hard (Kautz & Selman, in press; Selman & Kautz 1990)
Many of the strategies for defeasible inheritance in taxonomic hierarchies are intractable (Selman & Levesque 1989).

(3) Neural networks
For directed Hopfield nets, determining whether a stable configuration can be found is NP-complete (Godbeer 1987)

This listing only scratches the surface of the literature on the topic; there are many more examples. As should be clear, the problem-solving type of task, where a sequence of actions is required as opposed to single actions, is not the only one that has a combinatorial nature, as Anderson claims. Neither are the problems above obscure and isolated; rather, they are quite broad and natural. All "interesting" intelligent problems appear to be susceptible to combinatorial explosion. It is important to stress, however, that the examples given above do not by themselves "prove" that these problems or cognition in general are computationally intractable. They simply constitute evidence that the computational issues are real and may place severe constraints on algorithms proposed for the problems of cognition.

What does a computer scientist do when confronted with such a potentially intractable problem? A variety of approaches are possible.

(1) Develop an algorithm that is fast enough for small problems, but would take too long with larger problems. This approach is often used when the anticipated problems are small.

(2) Develop a fast algorithm that solves a special case of the problem, but does not solve the general problem. This approach is often used when the special case is of practical importance.

(3) Develop an algorithm that quickly solves a large proportion of the cases that come up in practice, but in the worst case may run for a long time. This approach is often used when the problems occurring in practice tend to have special features that can be exploited to speed up the computation.

(4) For an optimization problem, develop an algorithm which always runs quickly but produces an answer that is not necessarily optimal. Sometimes a worst case bound can be obtained on how much the answer produced may differ from the optimum, so that a reasonably close answer is assured. This is an area of active research, with suboptimal algorithms for a variety of important problems being developed and analyzed.

(5) Use natural parameters to guide the search for approximate algorithms. There are a number of ways a problem can be exponential. Consider the natural parameters of a problem rather than a constructed problem length and first attempt to reduce the exponential effect of the largest valued parameters.

NP-completeness effectively eliminates the possibility of developing a totally satisfactory algorithm. Once a problem is seen to be NP-complete, it is appropriate to direct efforts towards a more achievable goal. In most cases, a direct understanding of the size of the problems of interest and the size of the processing machinery is of tremendous help in determining which are the appropriate approximations. Could evolution have discovered this through millennia of experimentation?

It would be an extreme and untenable position to claim that behavior is made up of a large number of side-effects due to approximations; however, how much of our behavior can be legitimately put into this class? This is currently unknown and seems to me an interesting question for further study.

NOTE

1. The author is also affiliated with the Canadian Institute for Advanced Research.

Human and nonhuman systems are adaptive in a different sense

Tamás Zétényi

Department of General Psychology, L. Eötvös University, H-1378 Budapest, Pf 4. Hungary

In recent treatments of conceptual phenomena by cognitive science, three levels of analysis have been distinguished: a computational, algorithmic and an implementational level (see Marr 1982). In his target article Anderson undertakes to perform all three.

Anderson argues that if we describe the statistical structure of the environment, we can predict human cognition, which is supposed to be an optimal response to it, and is therefore adaptive.

He attempts to support the above claim by analyzing the behavior of nonhuman information-retrieval systems such as libraries or computer data bases. The results of the analysis of these systems are indeed correctly predicted by the carefully selected list of phenomena. What we need is an algorithm to represent the possible transformations between input and output. Bayesian theory is undoubtedly suitable for this purpose, so if we try to implement the resulting algorithm into another physical system, namely a connectionist(like) network, it is likely to run without a single error message.

What I find problematic in this is the claim that the memory functions of a nonhuman information-retrieval system can be plausibly extended to human memory. In spite of the cases provided by Anderson I wonder whether the algorithm can be implemented into human cognition.

Libraries and psychological experiments are indeed similar to each other in some respect. They have a limited set of items to handle. There are similar tasks too: A book or an item of a nonsense trigram list must be remembered upon request. It is true in both cases that memory performance will increase with practice and decrease with duration; and frequency also has an effect on it.

But let us assume that items in a library or files in a data-base are not independent entities; they communicate just as they do in cartoons. They can retrieve certain ideas, form categories, make decisions, and solve problems without any assistance. If this were the case, their behavior would indeed be similar to that of humans. This is the type of performance we call cognitive.

Nonhuman information-retrieval systems, as far as I know, never make changes in the entire material without human assistance. Subjects in a psychology experiment try to answer according to the instructions provided by the experimenter. Nevertheless, they show a strong tendency to retrieve items which are not included in the list. The types of errors are covered in detail by textbooks on human memory. Retrieving out-of-list items is excluded in the case of nonhuman systems. For example, we will not find on a library shelf a monk copying some codex, although these two are commonly associated with the concept "book" in memory. Similarly, it does not happen that a volume not satisfied with its borrowing index changes the color of its cover. It does not occur, either, that two different files of a data-base exchange strings without assistance. Books have never moved to certain places because they decided to form a new club (categorization), or because they found the shelf wet (reasoning). These would be instances of adaptation; these answers are certainly optimal with respect to environmental circumstances by some standard. The basic structure of a library or data-base will not change over time, so there is room for rational analysis. In contrast, a human retrieval system is an ever changing "environment."

Note that there is something strange in the argumentation above, namely, personification. I have personified the book or

the computer data-base file; I assumed a book was able to do something without any outer force. I think Anderson makes the same mistake. He personifies the library or data-base system in order to help explanation; then he forgets that the example is a mere metaphor, and takes it seriously.

My point is that the computational conditions and algorithm established for a particular system may not be valid for a different system. Furthermore, this may only become obvious after we have attempted to implement it. That is, I do not think rational theory in its present form is suitable to answer the question of whether or not human cognition is adaptive.

Author's Response

More on rational analysis

John R. Anderson

Department of Psychology, Carnegie Mellon University, Pittsburgh, PA 15213

Electronic mail: *anderson@psy.cmu.edu*

One thing you can count on in writing a *BBS* article is that there will be a wide range of points made in the commentaries. Indeed, this is the principal motivation for writing such an article – to gain new perspectives on one's research. Many of the comments I had not anticipated. One thing I had anticipated was the wide range of opinions about the plausibility of the optimization of cognition. Some thought it was obvious that cognition had to be highly adapted and others thought it highly implausible. This suggests that it is a profitable issue to pursue.

My approach has been to try to understand the adaptiveness of certain areas of human cognition that are close to my research interests in problem solving and learning. Many of the commentaries are cast at this empirical level. Some of the commentators discuss various empirical analyses in the target article; others point to more phenomena in defense of cognitive rationality; still others point to phenomena that contradict rationality. This response will first go through these three categories of comments. Then it will consider the more theoretical points that were made about the role of mechanism in a rational analysis, and close with a discussion of methodological points.

Memory. Both **Becker** and **Zetenyi** question whether the environment to which the model is optimized is really the one that humans face. There is certainly reason for raising this question about the original work of Anderson and Milson (1989) that was based on library borrowings and file accesses. However, we have found similar statistics in true sources of information available to humans such as *New York Times* headlines. Becker wonders whether it isn't the human mind that is determining the structure of domains like word use in the *New York Times* rather than the reverse, which is the claim of the rational analysis. This is an interesting hypothesis but it does not stand up to detailed scrutiny. For example, when there were a spate of articles on the Challenger accident it certainly was not human memory that caused the Challenger to explode. It is also known (Simon 1955) that a wide range of domains show the same general statistics as word usage. This includes the frequency with which animals of various species appear in the environment. Again, it seems implausible to claim that our memories control what animals we run into. The correct model of these regularities is like Burrell's (1980), which identifies the abstract features that unite these diverse domains.

Becker claims that the assumptions underlying the memory analysis are vague. This is to confuse generality and abstractness with vagueness. The assumptions are just cast at a very general level so that they might describe many systems. To be sure, systems can be found that do not satisfy them, but they do cast a wide net. I am not sure I understand all of Becker's specific assertions and questions about the memory assumptions but Anderson and Milson (1989) did produce running simulation models that reproduced data. So there is no problem with preciseness of specification and no apparent difficulty with the accuracy of the predictions.

Schwarz takes issue with some of the technical developments in Anderson and Milson (1989), criticizing us for mathematical infelicities as well as for providing simulations rather than closed-form solutions. On the first charge, I plead guilty and with each succeeding year I see how to make the mathematical points both more elegant and easier to understand. With respect to the second charge, however, computer solutions are the wave of the future in Bayesian applications (e.g., Gelfand & Smith 1990; Tanner & Wong 1987; Tierney & Kadane 1986). In the past, Bayesian applications have been limited by the need for closed-form solutions. One can now investigate more directly the nature of optimum decisions under reasonable sets of assumptions rather than artificially constrained assumptions.

Zetenyi raises the question of whether one can produce intrusion errors in nonhuman systems. Such errors are a legend in computer-based information retrieval systems (Salton & McGill 1983). It is true that the Anderson and Milson model does not really give any careful analysis of intrusion errors in human memory. That would require further analysis of the process that decides whether a retrieved memory satisfied the information-retrieval goal.

Categorization. The rational model of categorization also received a number of comments. Although he provides a nice summary of the categorization model, **Corter** questions some of the specifics. First, he criticizes the coupling parameter, c, as an arbitrary free parameter. He might well question some of the other parameters in the model as well. These are typically set to be what are standards for noninformative priors in the Bayesian literature, so they might not seem to be free parameters. In an adapted system, all these parameters, including c, should be set to reflect their true values in the environment. This means that the use of noninformative priors for the others, and of parameter estimation for c, are temporary holding patterns until we can do the environmental studies to determine what values these should have. This is what is so powerful about a rational analysis – the parameters have meaning in the external world. In the case of c, this refers to how often two objects we encounter are members of the same species. We are now

beginning to study published sources like the Audubon Society's annual Christmas bird count to get an estimate of this quantity.

Corter and **Gigerenzer** also question the assumptions that categories are disjoint and that features are independent within categories. Again, these are assumptions that we can begin to seriously explore in the environment. For example, we have been studying how well our categorization model applies to the various machine-learning data bases that describe aspects of the environment (Anderson & Matessa, submitted). The disjoint assumption seems to be working in domains as diverse as predicting diseases and classifying irises. The independence assumption requires some more elaborate comment. It should be understood that the assumption is not that features are independent globally but that they are independent only within a category. Thus, there is nothing problematical about Corter's bird example. Also, it is not the case, as Gigerenzer implies, that the categorical model makes it hard to capture dependence among features. In fact, it creates separate categories specifically to capture such dependences and, as Anderson and Matessa report, the model has great success in utilizing such dependencies for prediction.

Nonetheless, there is evidence for within-category dependence as well. For example, in applying the model to the classical iris data base of Fisher (1936) we find that within specific species of irises there is a correlation between sepal length, sepal width, petal length, and petal width. This is presumably produced because we are measuring four quantities that reflect one overall genetic trait of size. Similarly, we would expect length of left and right arms to be correlated in humans. Thus, the problem is that even if the "true" features are independent, the measurements we record for classification may not be. Of course, none of this causes difficulties for the general proposal of optimal categorization. Rather, it simply means that we need to change our model of the environment. Note also that this is not a case of changing the model of the environment to fit the behavioral data but rather changing it to fit the environment. The rational theory of human categorization is a function of the correct theory of the environment.

Dickinson wonders about whether the iterative categorization algorithm reflects a simplifying assumption for purposes of prediction or whether it is taken as a serious claim about a limitation on human categorization. It did start out as a simplifying assumption to allow us to make approximate estimates of the ideal quantity. However, the more I have worked with it, the more I have become convinced that it is a fundamental property of human categorization and in fact optimal given the assumption that there is a cost associated with each hypothesis considered. Right now I have no proof of optimality, but the human data do show a key feature predicted by the iterative algorithm and not the ideal: As discussed in Anderson (1990), the categories that human subjects form depend on the order in which they study the instances.

Causal inference. The treatment of causal inference also came in for some comment. **Schultz** asks about the coherence of the overall analysis. His remark about equation 10 that "the probability of a rule applying in the presence of cues seems no less fundamental than the probability of cues if the rule did apply" corresponds precisely to my sentiments about the matter. Equation 10 is a rather awkward attempt to cast the basic Bayesian insight that one's posterior confidence in a cause occurring should be a combination of one's prior confidence that the rule exists and how well the current situation fits the rule. Perhaps a better way to cast it would be in terms of an odds ratio formula:

$$\frac{P(i|C)}{P(\bar{i}|C)} \approx \frac{Con(i)}{Con(\bar{i})} \times \frac{P(C|i)}{P(C)} = \frac{Con(i)}{Con(\bar{i})} \times \frac{P(i|C)}{P(i)}$$

Here $P(i|C)/P(\bar{i}|C)$ is the posterior confidence, $Con(i)/Con(\bar{i})$ is the prior confidence, and either $P(C|i)/P(C)$ or $P(i|C)/P(i)$ are measures of how well the cues and the cause match up.[1] $P(C|i)/P(C)$ is a measure of how much more probable the cues become over their base frequency when the cause is operative and $P(i|C)/P(i)$ is a measure of how much more probable the cause is over base frequency when the cues are present. Both ratios are equivalent measures of how well the cues match the rules. Cast in this form the relationship shows that $P(i|C)$ and $P(C|i)$ can be thought of symmetrically. It better reflects my own sense of what is going on and perhaps also Shultz's.

Shultz argues that human's demonstrated neglect of prior probabilities bodes poorly for a number of our treatments, including the analysis of 2 × 2 contingency tables. However, as argued in Anderson (1990), these examples of neglecting the base rate actually reflect a neglect of the stated probabilities and not of the experienced proportions. People are in fact very sensitive to experienced frequencies, as demonstrated in many domains, including causal reasoning.

Shultz also points to a major incompleteness in the rational treatment of causal inference which is that there is no analysis of the development of the causal models which underlie causal inference. When does the knowledge develop that allows the differential analysis of the situations in Figure 1? And can its development be given a rational analysis? I wish I knew.

Smokler questions the treatment of causal inference as unconscious. [See also Searle: "Consciousness, Explanatory Inversion, and Cognitive Science" *BBS* 13(4) 1990; Velmans: "Is Human Information Processing Conscious?" *BBS* 14(4) 1991; and Dennett & Kinsbourne: "Time and the Observer: The Where and When of Consciousness in the Brain" *BBS* 15(2) 1992.] Rational analysis has no commitments on the issue of consciousness; it is concerned with constraints on the system's output whether conscious or not. I have actually been struck by demonstrations such as the classic ones of Michotte (1946) in which causal inference does seem automatic and unconscious. I suspect, however, that in other cases it is quite reflective. A virtue of the current approach is that it allows us to analyze both situations in terms of constraints on output. This emphasis on output leads to the focus on prediction to which both Smokler and Shultz object. The analysis of the attribution with respect to Figure 1 derives from a predictive analysis. The basic assumption is that unpredictable things will appear causally anomalous. It is possible, as Shultz suggests, that we will have to treat attribution and prediction differently. It is hard to identify what the adaptive func-

tion of causal attribution might be if it is not in service of prediction. With respect to Shultz's request that one extend causal inference to planning: I think that is a major accomplishment of my rational analysis of problem solving

I appreciate **Geissler**'s equation for fitting the data in Figure 2 and I think I understand most of it. Its point seems to be that since the same parametric form of equation fits the two situations, it is not that subjects are adopting the optimal model in both cases. This is not the only equation that can fit the data; I think different ones can be produced for the two cases of causal attribution which are justified by the different physical models. (This will be further developed in later publications.)

More rational domains. I have certainly not exhausted the domains where rationality can be demonstrated, and a number of commentators have added examples of their own: **Fantino & Stolarz-Fantino** show that given adequate information subjects will not adopt a maximizing strategy when it is nonoptimal. **Massaro & Friedman** point out that human pattern recognition appears to be optimal, as we would expect of such a deeply entrenched cognitive function. Massaro & Friedman even point to an adaptive treatment of the bothersome Linda-is-a-feminist example.

Irrational domains? Next we come to the question of whether there are domains of cognition where humans behave irrationally. **Baron, Evans, Fantino & Stolarz-Fantino, Massaro & Friedman, Shafir,** and **Snow** all raise this as central points in their commentaries. They all agree in pointing to the psychological domain of judgment and decision making as the place to show human irrationality. A number of the phenomena they mention were addressed in Anderson (1990) but the topic deserves some more discussion here. Before responding to some of the more specific points it is worthwhile to make three general points relevant to all of the commentaries.

(1) As noted by **Evans,** there are two senses of the term "rational." One refers to cognition that involves what may be called "rational thought," where this is defined by some normative model. This is not the sense of the term in a rational analysis. Here, the term is used in the economist's sense, which is that the output of the system is optimal and no claim is made about the mental processes by which this output is computed. These two senses need not be congruent; in particular, cognitive processes which are judged irrational by some normative model might still lead to the right behavior.

Many of the purported demonstrations of irrationality pertain to the first sense of the term. It is no accident that these studies tend to come from domains of judgment and decision making where such normative models, often of dubious applicability, abound. [See Cohen: "Can Human Irrationality Be Demonstrated" *BBS* 4(3) 1981; and Kyburg: "Rational Belief" *BBS* 6(2) 1983.]

Anderson (1990) enumerates some of the ways in which rationality in the first sense might not be rationality in the second sense. The central point is that we cannot know whether a behavior in the laboratory is adaptive in the second sense by applying a priori normative criteria. The only way to decide is to determine what the consequences of that behavior are in the real world (Funder 1987). A person may choose to do something for the wrong reason but it is possible that this will turn out to lead to the best consequences. It is an empirical question not a logical one, whether an action leads to good or bad consequences. My favorite example involves demonstrations that subjects do not pay attention to stated base rates: A minimal condition for this to be irrational is that the stated probabilities be veridical in the real world. Stated probabilities are definitely not veridical in the real world. For example, it was only a few years ago that we were told that there was a 1 in 100 chance of developing full blown AIDS if one tested positive for the HIV virus. It is also interesting that, in contrast to stated probabilities, categorization and causal inference task performance shows considerable sensitivity to the base rates with which the categories and causes are experienced.

(2) We have to factor in computational costs before deciding what is and is not rational. A perfect game of chess is the rational behavior given the rules of chess but it is clearly impossible given the impossibility of searching the whole game tree in real time. More will be said about mechanism in a later section.

(3) Finally, there clearly has to be something irrational in human behavior by the adaptive definition. No one can look at our current predicament – with huge stockpiles of nuclear weapons and so forth – and say that all human behavior is currently adaptive in the biological sense. It is quite a different matter, however, ascertaining where the difficulty lies. I am torn among three types of explanations. The first is that our behavior is adaptive for an environment other than the one we find ourselves in. This may be an explanation of the problems brought on by modern technology. In the target article I also adopted this as an explanation for some of the nonoptimal behavior associated with the matching law. A second explanation is that, while many of our basic cognitive functions are optimized from their own myopic view of the world, the system as a whole might not be optimized. This may either be because it is too expensive computationally to achieve such global coordination or because we simply have this failure to optimize globally as a weakness. A third explanation is that our cognitive machinery may just have some nonoptimalities. I view the absence of an adequate short term buffer as a prime candidate for such a nonoptimality. The rational analysis of memory concerned what would be optimal in the absence of such a buffer. As noted in the target article, rational analysis asks what is optimal given certain constraints. It is interesting to the extent that these constraints are few and simple.

Baron's commentary is a nice presentation of some of the issues involved in deciding rationality. Just as he finds himself half agreeing with me, I find myself half agreeing with him. For reasons enumerated above, I do not agree that most of the experiments he cites adequately establish the conditions to demonstrate irrationality in the adaptive sense. A brief comment is needed on his questioning the measure of adaptation in number of offspring: I am not advancing this equation as a moral imperative, only as an acknowledgement of the underlying mechanism that is supposed to produce evolutionary optimization. There may be better personal arguments for optimizing money, the happiness of oneself and others, or any other goal. It is just that these goals do not produce optimization of the

species. (Of course, there are serious questions about just how much optimization of the species even this produces.)

Evans raises the question of how the system can perform the complex computations required to determine the optimal behavior. He also provides the answer: The system can have stored the solutions to these problems based on personal or evolutionary experience

It is not reasonable to go through every purported example of irrationality cited in the commentaries, but it is worthwhile to address a couple – discussions of others can be found in Anderson (1990). **Fantino & Stolarz-Fantino** and **Massaro & Friedman** both bring up Kahneman and Tversky's feminist bank teller as an example of irrational behavior. The basic result is that people think representative conjunctions of events are more probable than any of the the components being conjoined. Although I do not claim to have any insights into the phenomenon, nor do I want to deny its importance, it seems a paradigm case of a frequently cited example that is irrelevant to the thesis of rationality as developed here. What does the verbal exercise of assigning probability values have to do with adaptive behavior? People may assign a higher probability to Becker's winning a tennis tournament than to his winning a particular match in the tournament but I haven't heard of any betting establishment getting customers to take poorer odds on the conjunction than the individual events.

As a second example of a behavior that might seem to be of more direct adaptive significance, consider the Kahneman and Tversky (1984) example cited by **Shafir** where people are more willing to drive to a store to save $5 on a $15 calculator than on a $125 calculator. As an aside, I note that this experiment did not actually require subjects to drive but rather took hypothetical verbal responses, but let us assume that the result actually reflects what people will do. Although this is an inconsistency, it is far from obvious it is maladaptive. Is it maladaptive to drive for the $5 savings in the $15 case? Or is it maladaptive not to drive for $5 savings in the $125 case? I assume it is not clear to the reader whether either decision is right or wrong since it is not clear whether $5 is worth the drive. Inconsistent choices among decisions that do not have a clear consequence cannot have maladaptive consequences. As discussed in Anderson (1990) at length, this seems characteristic of all the demonstrations of an effect of "framing" on choice (these effects occur when there is not a choice that is clearly correct). This is not to say that this research is not important; it is just irrelevant to rationality of cognition in the sense of rational analysis.

Mechanism and rational analysis. Rational analysis is offered as a theoretical approach to cognition that contrasts with the typical mechanistic approaches of cognitive science. However, as noted in the target article, it does require some assumptions about mechanism to establish the costs and define the constraints under which optimization takes place. The relationship between mechanism and rational analysis came in for considerable comment. I did not recognize any single recurrent theme, but there were a number of useful points:

Agarwal asserts that in engineering, any attempt to understand behavior that did not start with mechanism would be dismissed. Yet there are many successful efforts to understand behavior in natural systems that do start from the environment and largely ignore mechanism (Stephens & Krebs 1986). So perhaps we have a difference between disciplines. Part of the reason may be that evolution has had a lot longer to optimize the mechanisms than engineers have had. Also, modeling random number generation is not one of the tasks that psychologists find themselves called upon to perform. Note, however, that I am not "ignoring the nature and history of the system" as Agarwal suggests. The history of the system is the central story in my analyses. The nature of the system is downplayed but not ignored.

Campbell & Bickard argue that the analysis depends on a lot of assumptions in the ACT* theory about how knowledge is encoded. Their discussion about "encodingism" raises enough issues to constitute an article in itself, but for present purposes it is enough to note that their basic premise seems to be off the mark. The rational analysis, while it does implicitly assume some things from ACT*, does not require anything like the detail of commitment to ACT* representational assumptions these commentators imply. Indeed, one of the motivations for the rational analysis was to retreat from the representational assumptions in ACT*. Elsewhere (Anderson 1978; 1987) I have discussed my dissatisfaction with these representational assumptions. My problems with representation are problems of identifiability, not with the problem Campbell & Bickard raise. It is unclear whether there is anything in the rational analysis which would be inconsistent with their alternative proposal of "interactive representation."

Holyoak notes that choices about how to represent an event can determine what the predictions are. In particular, predictions of the memory, categorization, and causal models are subject to how the events are decomposed into elements or features. Note that such decomposition assumptions are very weak relative to the detail of the ACT* representational assumptions. I suspect that they are less vulnerable to problems of identifiability. Evidence for this lies in Holyoak's observation that differences in representational choices at this level have natural behavioral consequences – specifically, you cannot predict features you do not represent.

Such weak representational assumptions exemplify what I would like to find throughout with respect to mechanism. It is not that I have some aversion to assumptions of mechanism. It is rather that I have never been totally comfortable about mechanistic assumptions that suffer identifiability problems. What I would like to do is to find a level of mechanistic abstraction that does not have this difficulty. Thus, I do not agree with **Massaro & Friedman** that process assumptions such as "it costs something to consider a memory" have identifiability problems within a rational framework. That is to say, I do not think that, given a fixed environmental model, one can get the same predictions by optimizing within some alternative mechanistic assumptions that deny the existence of a cost associated with considering a memory. I have no proof of this conjecture and it would only take a single counter-example to prove me wrong. However, my belief is that rational analysis encourages a level of abstraction in mechanistic assumption that avoids identifiability problems. [See also Schoemaker: "The Quest for

Optimality: A Positive Heuristic of Science?" *BBS* 14(2) 1991.]

In a somewhat different vein, **Godfrey-Smith, Rachlin,** and **Smokler** all complain about the use of internal costs as vague and pale in comparison to external costs. Internal costs are quite undeniable, however, and potentially quite dramatic. The perfect example is searching the chess game tree to find the best move. No one can do this, no matter what external reward hangs on their chess play – there are just not enough seconds in the universe, let alone a lifetime. In this situation, the internal costs can be overwhelming. Chess is not an isolated example. Whenever we are dealing with the knowledge built up in a lifetime (as we are in memory, categorization, causal inference, and problem solving), the costs of exhaustively considering the implications of that knowledge must be overwhelming and we can be sure any optimal system would pay serious heed to the costs of combing through this knowledge base. It is true that the rational analysis does not go into great detail about how these internal costs make themselves felt, but I consider such abstraction a virtue.

There is indeed something special about external cost and reward and in this sense the rational analysis of problem solving is critical. The considerations about memory, categorization, and causal inference all have the goal of getting successful problem-solving behavior where all these mounting internal costs get paid back with some reward. Again, I prefer to think of such external rewards abstractly in terms of utility. **Rachlin** is wrong in claiming that we need some observable quantity like money. An abstract quantity like utility can be given a rigorous analysis (e.g., DeGroot 1970).

Chandrasekaran is interested in what constraints a rational analysis may impose on mechanism. He argues that it only implies constraints on the content and not on the structure of the mechanism. I am inclined to agree, and see the classification net as just the realization of a particular content given the constraint of a connectionist net. The rational analysis specified the content, not the use of the net. I also agree with **Schultz** that rational analysis underconstrains mechanistic accounts, but this is not necessarily a liability.

Tsotsos is a computer scientist who seems to be making a claim that is opposite to **Chandrasekaran**'s. He claims that there are all kinds of mechanistic constraints on optimization. An inspection of his list, however, reveals that these limitations are very general and would apply to a large variety of machines. For example, going to connectionist systems would offer no help here. Rather, his computational limits are constraints imposed by the logical content of the problem. I can only assume that Tsotsos has misunderstood my point, as I find myself quite in agreement with his. I did not deny that computational limitations pose a problem; rather, I said assumptions about mechanism (e.g., serial versus parallel processing) have not determined the predictions of the theory. With respect to his remarks about how NP-completeness can have a dramatic effect, the rational theory of categorization is a response to what I suspect is an NP-complete problem. Again, discussions of algorithms at this level of abstraction is more comfortable than the excessive specificity that has dominated cognitive science. As a final comment, **Tsotsos**'s examples are all counters to **Holyoak**'s claim that different assumptions about mechanism can change what is difficult.

Small asserts that an enterprise in cognitive science is to be judged by whether it leads us to understand mechanism. Here I simply disagree. One should be willing to take insight from wherever it comes. Small argues that cognitive scientists are not just interested in outputs of cognition and then cites a number of examples. His notion of output is very different from mine because I would consider all his examples as studies of the output of the cognitive system. How is interpreting a scene different from retrieving a memory in this respect? In both cases we are studying the output of an internal process and in both cases an adaptionist approach would analyze this process in terms of its ultimate contribution to external adaptation.

Small further claims that a rational analysis cannot explain maladaptive and pathological behaviors produced by lesions or neurochemical imbalance. Those, he argues, are data that require mechanistic explanations. I agree that pathology cannot be given a useful adaptionist explanation. This was never the goal of a rational analysis. Neuroscience is the way to go there. I have also claimed (Anderson 1978; 1987) that neuroscience offers the prospect of avoiding the identifiability problems that haunt behavioral data. However, the rate of progress by the neural route seems slow and behavioral data are fascinating in their own right.

Methodology. Finally there are the commentaries concerned with various aspects of the theoretical methodology:

Snow wonders about the strength of my tests of the rationality thesis. He thinks the tasks were made for a rational analysis and the assumption of local optimization allows a convenient escape hatch should things get difficult. All I can do in response is to point to the commentaries which were impressed by the range of data that the model fit. Snow also concludes that issues of optimization may be unresolvable. All I can say is that I hope not, and time will tell.

Massaro & Friedman suggest that gathering more data may eliminate identifiability problems, but, as developed at length in Anderson (1990), as long as we restrict ourselves to behavioral data, no amount of data will eliminate these problems.

Godfrey-Smith takes issue with my refusal to commit myself to the biological processes by which the optimality is generated. In his view, such an analysis is not explanatory. In my view, this is a paradigm case of the arbitrary use of the term "explanatory": To think one could establish to everyone's satisfaction that these cognitive functions were optimized to the environment and then have such an amazing result simply shrugged off as not "explanatory"! To be sure, it would be better to have a detailed analysis of the genesis of this optimization. Perhaps when we better determine where optimization exists and what form it takes we can begin to explore the question of its genesis. Trying to reconstruct the evolution of our cognitive faculties is extremely difficult and perhaps impossible. This should not stop us from trying to determine whether they have reached optimal form. [See

also Clark: "Modeling Behavioral Adaptation" *BBS* 14(1) 1991.]

Gigerenzer and **Hastie & Hammond** comment on the similarities and differences between this program and that of Brunswik. I think the similarities are quite strong and am inclined to agree with Hastie & Hammond's assessment that rational analysis is quite in keeping with the spirit of Brunswik's work. They are a little bothered (and Gigerenzer is very much bothered) by the difference between the linear regression methodology associated with Brunswik and the Bayesian statistical inference I used. The differences are more apparent than real, however. First, general linear models have a very coherent development within the Bayesian framework (Box & Tiao 1973). Bayesian methodology was chosen because it is designed to determine optimal decisions under uncertainty. As noted in my comments under categorization, the Bayesian models do not require independence to work, nor does the rational model of categorization produce independence. A possible difference between rational analysis and Brunswik's analysis is the potential role a Bayesian model allows for priors. If one had only weak priors then Bayesian analysis would proceed indistinguishably from conventional linear regression. However, in the presence of strong priors the conclusions can be quite different. Another difference is that rational analysis places no premium on representative design; it only requires that when you design an experiment you do not delude yourself into believing it is representative of the world.

My view of the priors in the Bayesian model is that they incorporate the experience from the evolutionary and personal history of the individual. Thus, if the system has a prior that a new species of mammal will be brown this is basically a reflection of frequency in experience. The basis of the model is hence more frequentist than one might assume. Indeed, one might consider my application of Bayesian models to exemplify a methodology known as "empirical Bayes" which some Bayesians dismiss as not truly Bayesian (Lee 1989). Although it uses the tools of Bayesian analysis, it is not committed to interpreting the probabilities as subjective. I use Bayesian methods for their technical virtues; I do not feel prepared to judge the statistical debates that separate the Bayesians and frequentists. I think this places in a somewhat different light the comments of **Gregson, Schwarz,** and **Smokler,** which are addressed at the Bayesian aspects of my work.

With respect to **Gregson**'s comments, it is worth emphasizing again that I regard the question of how people process stated probabilities as totally irrelevant to the question of whether their cognitive system can be understood as responding optimally in a Bayesian sense to experienced probabilities. Thus, the evidence about the non-Bayesian character of probability judgments is irrelevant. One may or may not be able to give a rational analysis of these probability judgments. To do so would require specifying the goal people are trying to optimize in making these judgments, the relevant statistical structure of the environment in which they make these judgments, and the potential computational costs. Such an analysis would apparently be very different from typical normative analyses of probability judgments: People's goals are surely not to appear as good Bayesians; stated probabilities cannot be taken at face value in the real world, and there are surely some significant computational costs.

Schwarz's comments on the probability developments are those of an unrepentant frequentist. There is much more to Bayesian analysis than he suggests. Specifically, in contrast to his quote from Feller (1968), one can speak of probabilities without specifying a sample space. As to whether I am a frequentist or a Bayesian, I leave that for the future to decide.

In a rather different vein, **de Sousa**'s comments on actually pinning down the relevant structure of the environment identify some important issues. He points out that the environment in which an individual finds itself is somewhat a function of the individual – because only certain parts can be sensed, because only certain aspects are relevant, and because we in part design our environment. This does not introduce circularity into the project but it could certainly make it quite a bit more difficult. Whether or not it actually does remains to be seen. Contrary to de Sousa, I regard our treatment of the environment as an analytic success that has avoided such pitfalls.

I would like to conclude this Response with a final word on **Rachlin**'s curious commentary, in which he seems to be asking me to admit to being a behaviorist and throw off my mentalist shackles. I have always felt that something was lost when the cognitive revolution abandoned behaviorism; my work on rational analysis can be viewed as an attempt to recover that. In doing this, however, I do not want to lose the cognitive insight that there is a mind between the environment and behavior. What I am trying to do is to regain the insight that the mind exists to support adaptive behavior.

NOTE

1. This formula is only an approximation because it involves replacing $P(C|\bar{i})$ with $P(C)$. The argument is that the base probability of the cues C does not change much when we conditionalize on the nonapplicability of a single rule $\bar{i}$.

References

Alcock, J. (1989) *Animal behavior.* Sinauer. [DWM]

Anderson, J. R. (1978) Arguments concerning representations for mental imagery. *Psychological Review* 85:249–77. [rJRA]

(1983) *The architecture of cognition.* Harvard University Press. [aJRA, RLC]

(1987) Methodologies for studying human knowledge. *Behavioral and Brain Sciences* 10:467–505. [rJRA]

(1990a) *The adaptive character of thought.* Erlbaum. [arJRA, RLC, DWM, TRS]

(1990b) *Cognitive psychology and its implications. 3rd ed.* W. H. Freeman. [aJRA]

(in press) The place of cognitive architectures in a rational analysis. In: *Architectures for intelligence,* ed. K. vanLehn. Erlbaum. [aJRA]

Anderson, J. R. & Kushmerick, N. (1990) A rational analysis of production system architecture. Paper presented at the 31st Annual Meeting of the Psychonomic Society, November 16–18. [aJRA]

Anderson, J. R. & Matessa, M. (1990) A rational analysis of categorization. In: Proceedings of the Seventh International Machine Learning Conference. Palo Alto, CA. ed. M. Kaufmann. [aJRA]

(submitted) Explorations of an incremental, Bayesian algorithm for categorization. *Machine Learning.* [rJRA]

Anderson, J. R. & Milson, R. (1989) Human memory: An adaptive perspective. *Psychological Review* 96(4):703–19. [aJRA, WS]

Anderson, N. H. (1981) *Information integration theory.* Academic Press. [H-GC]

Arkes, H. R. & Harkness, A. R. (1983) Estimates of contingency between two dichotomous variables. *Journal of Experimental Psychology: General* 112:117–35. [aJRA]

Atwood, M. E. & Polson, P. G. (1976) A process model for water jug problems. *Cognitive Psychology* 8:191–216. [aJRA]

Bahrick, H. P. (1979) Maintenance of knowledge: Questions about memory we forget to ask. *Journal of Experimental Psychology: General* 108:296–308. [aJRA]

Baron, J. (1985) *Rationality and intelligence.* Cambridge University Press. [JB]

(1988) *Thinking and deciding.* Cambridge University Press. [JB]

(1990) Harmful heuristics and the improvement of thinking. In: *Developmental perspectives on teaching and learning thinking skills*, ed. D. Kuhn. Karger. [JB]

Baron, J., Badgio, P. & Gaskins, I. W. (1986) Cognitive style and its improvement: A normative approach. In: *Advances in the psychology of human intelligence*, vol. 3, ed. R. J. Sternberg. Erlbaum. [JB]

Baron, J., Badgio, P. & Ritov, Y. (in press) Departures from optimal stopping in an anagram task. *Journal of Mathematical Psychology.* [JB]

Baum, W. M. (1974) On two types of deviation from the matching law: Bias and undermatching. *Journal of the Experimental Analysis of Behavior* 22:231–42. [HR]

Bellman, R. (1961) *Adaptive control processes: A guided tour.* Princeton University Press. [GCA]

Benkherouf, L. & Bather, J. A. (1988) Oil exploration: Sequential decisions in the face of uncertainty. *Journal of Applied Probability* 25:529–43. [WS]

Berger, J. O. (1985) *Statistical decision theory and Bayesian analyses.* Springer-Verlag. [aJRA]

Bernard, C. (1865) *Introduction a l'etude de la medicine experimentale.* (2nd ed., 1867) Emile Martinet. [SLS]

Bickhard, M. H. (1980) *Cognition, convention, and communication.* Praeger. [RLC]

(1987) The social nature of the functional nature of language. In: *Social and functional approaches to language and thought*, ed. M. Hickmann. Academic Press. [RLC]

(in press a) The import of Fodor's anticonstructivist arguments. In: *Epistemological foundations of mathematical experience*, ed. L. P. Steffe. Springer. [RLC]

(in press b) How does the environment affect the person? In: *Children's development within social contexts*, ed. L. T. Winegar & J. Valsiner. Erlbaum. [RLC]

Bickhard, M. H. & Campbell, R. L. (1989) Interactivism and genetic epistemology. *Archives de Psychologie* 57:99–121. [RLC]

Bickhard, M. H. & Richie, D. M. (1983) *On the nature of representation: A case study of James J. Gibson's theory of perception.* Praeger. [RLC]

Birnbaum, M. H. (1983) Base rates in Bayesian inference: Signal detection analysis of the cab problem. *American Journal of Psychology* 96:85–94. [GC]

Box, G. E. P. & Tiao, G. C. (1973) *Bayesian inference in statistical analysis.* Addison-Wesley. [rJRA]

Boyd, R. & Richerson, P. J. (1985) *Culture and the evolutionary process.* University of Chicago Press. [PG-S]

Brehmer, B. & Joyce, C. R. B., eds. (1988) *Human judgment: The SJT view.* North-Holland. [RH]

Broca, P. P. (1861) Nouvelle observation d'aphemie produite par une lesion de la partie posterieure des deuxieme et troisieme circonvolutions frontales. *Bulletin de la Societé Anatomique de Paris* 6:398–407. [SLS]

Broemeling, L. D. (1985) *Bayesian analysis of linear models.* Marcel Dekker. [RAMG]

Brooks, L. (1978) Nonanalytic concept formation and memory for instances. In: *Cognition and categorization*, ed. E. Rosch & B. B. Lloyd. Erlbaum. [aJRA]

Brooks, R. (1987) Intelligence without representation. MIT AI Report. [RLC]

Brunswik, E. (1934) *Wahrnehmung und Gegenstandswelt: Grundlegung einer Psychologie vom Gegenstand her* (Perception and the world of objects: The foundations of a psychology in terms of objects). Deuticke. [RH]

(1943) Organismic achievement and environmental probability. *Psychological Review* 50:255–72. [RH]

(1956) *Perception and the representative design of psychological experiments.* University of California Press. [aJRA, RH]

(1964) Scope and aspects of the cognitive problem. In: *Contemporary approaches to cognition*, ed. J. S. Bruner et al. Harvard University Press. [GG]

Buffart, H. F. J. M. & Geissler, H.-G. (1983) Task-dependent representation of categories and memory-guided inference during classification. In: *Trends in mathematical psychology*, ed. E. Degreef & J. van Buggenhaut. North Holland. [H-GC]

Burrell, Q. L. (1980) A simple stochastic model for library loans. *Journal of Documentation* 36:115–32. [aJRA, GMB, WS]

(1985) A note on aging on a library circulation model. *Journal of Documentation* 41:100–15. [GMB]

Burrell, Q. L. & Cane, V. R. (1982) The analysis of library data. *Journal of the Royal Statistical Society* Series A(145):439–71. [aJRA]

Bylander, T., Allemang, D., Tanner, M. & Josephson, J. (1989) Some results concerning the computational complexity of abduction. Proceedings of the First International Conference on Principles of Knowledge Representation and Reasoning, Toronto. [JKT]

Camerer, C., Loewenstein, G. & Weber, M. (1989) The curse of knowledge in economic settings: An experimental analysis. *Journal of Political Economy* 97:1232–54. [JB]

Campbell, R. L. & Bickhard, M. H. (1986) *Knowing levels and developmental stages.* Karger. [RLC]

(1987) A deconstruction of Fodor's anticonstructivism. *Human Development* 30:48–59. [RLC]

Chapman, G. B. (1990) Models of contingency judgment. Ph.D. thesis, Department of Psychology, University of Pennsylvania. [JB]

Chapman, L. J. & Chapman, J. P. (1967) Genesis of popular but erroneous psychodiagnostic observations. *Journal of Abnormal Psychology* 72:193–204. [JB]

(1969) Illusory correlation as an obstacle to the use of valid psychodiagnostic signs. *Journal of Abnormal Psychology* 74:271–80. [JB]

Cheeseman, P., Kelly, J., Self, M., Stutz, J., Taylor, W. & Freeman, D. (1988) A Bayesian classification system. In: Proceedings of the Fifth International Conference on Machine Learning. San Mateo, CA. [aJRA]

Cheng, P. & Holyoak, K. (1985) Pragmatic reasoning schemas. *Cognitive Psychology* 17:391–416. [GG]

Chow, C. & Schechner, Z. (1985) On stopping rules in proofreading. *Journal of Applied Probability* 22:971–77. [WS]

Clark, C. (1991) Modeling behavioral adaptations. *Behavioral and Brain Sciences* 14(1):85–118. [BWD]

Cohen, L. J. (1981) Can human irrationality be experimentally demonstrated? *Behavioral and Brain Sciences* 4:317–70. [PS]

Cohen, M. M. & Massaro, D. W. (in press) On the similarity of categorization models. In: *Probabilistic multidimensional models of perception and cognition*, ed. F. G. Ashby. Erlbaum. [DWM]

Corcos, D. M., Gottlieb, G. L. & Agarwal, G. C. (1988) Accuracy constraints upon rapid elbow movements. *Journal of Motor Behavior* 20:255–72. [GCA]

Corter, J. E. & Gluck, M. A. (1985) The pragmatics of semantics: Information, communication and levels of abstraction. Paper presented at meeting of the Society for Philosophy and Psychology, University of Toronto. [JEC]

(submitted) Explaining basic categories: Feature predictability and information. [JEC]

Cosmides, L. (1989) The logic of social exchange: Has natural selection shaped how humans reason? *Cognition* 31:187–276. [CG]

Crocker, J. (1981) Judgment of covariation by social perceivers. *Psychological Bulletin* 90:272–92. [aJRA]

de Groot, M. H. (1970) *Optimal statistical decisions.* McGraw-Hill. [rJRA, RAMG]

Doyle, R. (1989) Reasoning about hidden mechanisms. Proceedings of the International Joint Conference on Artificial Intelligence. [TRS]

Dupré, J. (1987) *The latest on the best.* MIT Press. [aJRA]

Edelman, G. M. (1987) *Neural Darwinism.* Basic Books. [H-GC]

Einhorn, H. J. & Hogarth, R. M. (1986) Judging probable cause. *Psychological Bulletin* 99:3–19. [aJRA]

Evans, J. St. B. T. (1989) *Bias in human reasoning: Causes and consequences.* Erlbaum. [JSBTE]

Fantino, E. & Abarca, N. (1985) Choice, optimal foraging and the delay-reduction hypothesis. *Behavioral and Brain Sciences* 8:315–30. [aJRA]

Feller, W. (1968) *An introduction to probability theory and its applications, 3rd ed.*, vol. 1. Wiley. [WS]

Fisher, D. H. (1987) Knowledge acquisition via incremental conceptual clustering. *Machine Learning* 2:139–72. [aJRA]

Fisher, K. A. (1936) Multiple measurements of taxonomic problems. *Annals of Eugenics* 7:179–88. [rJRA]

Fitts, P. M. (1954) The information capacity of the human motor system in controlling the amplitude of movement. *Journal of Experimental Psychology* 47:381–91. [GCA]

Fitts, P. M. & Peterson, J. R. (1964) Information capacity of discrete motor responses. *Journal of Experimental Psychology* 67:102–12. [GCA]

Fodor, J (1975) *The language of mind* Crowell [RLC]
(1980) Fixation of belief and concept acquisition In: *Language and learning. The debate between Jean Piaget and Noam Chomsky*. ed M Piattelli-Palmarini Harvard University Press [RLC]
(1981) The present status of the innateness controversy In: *RePresentations*, ed J Fodor MIT Press [RLC]
Freyd, J J (1982) Shareability: The social psychology of epistemology *Cognitive Science* 7:191–210 [JEC]
Fried, L S & Holyoak, K J (1984) Induction of category distributions: A framework for classification learning *Journal of Experimental Psychology Learning, Memory, and Cognition* 10:234–57 [aJRA]
Funder, D C (1987) Errors and mistakes: Evaluating the accuracy of social judgment *Psychological Bulletin* 101:75–90 [rJRA, JB]
Gal, I (1990) Understanding repeated choices under uncertainty Ph D thesis, Department of Psychology, University of Pennsylvania [JB]
Geissler, H -G (1970) Calibration law and space orientation (in German) Unpub doctoral dissertation, Humboldt University, Berlin [H-GG]
(1976) Internal representation of external states: Aspects of an indirect validation approach to psychophysics In: *Advances in psychophysics*, ed H -G Geissler & Yu M Zabrodin Deutscher Verlag der Wissenschaften [H-GG]
(1980) Perceptual representation of information: Dynamic frames of references in judgment and recognition In: *Psychological research, Humboldt University 1960–1980*, ed F Klix & B Krause Deutscher Verlag der Wissenschaften [H-GG]
(1983) Physical correlate theory versus indirect calibration approach *Behavioral and Brain Sciences* 2:316–18 [H-GG]
(1987) The temporal architecture of central information processing: Evidence for a tentative time-quantum model *Psychological Research* 49:99–106 [H-GG]
(1991) New magical numbers in mental activity? On a taxonomic system for critical time periods In: *Cognition, information processing and psychophysics: Basic issues*, ed H -G Geissler, S W Link & J T Townsend Erlbaum [H-GG]
Geissler, H -G & Buffart, H F M J (1985) Task dependency and quantized processing in classification In: *Cognition, information processing and motivation, selected revised papers*, vol 3, ed G d'Ydewalle North-Holland [H-GG]
Geissler, H -G & Puffe, M (1983) The inferential basis of classification: From perceptual to memory code systems In: *Modern issues in perception*, ed H -G Geissler, H F M J Buffart, E L J Leeuwenberg & V Sarris Deutscher Verlag der Wissenschaften North-Holland [H-GG]
Geissler, H -G , Klox, F & Scheidereiter, U (1978) Visual recognition of serial structure: Evidence of a two-stage scanning model In: *Formal theories of perception*, ed E L J Leeuwenberg & H F J M Buffart Wiley [H-GG]
Gelfand, A E & Smith, A F M (1990) Sampling-based approaches to calculating marginal densities *Journal of the American Statistical Association* 85:398–409 [rJRA]
Gibbon, J (1986) The structure of subjective time: How time flies In: *The psychology of learning and motivation*, vol 20, ed G H Bower Academic Press [HR]
Gibson, J J (1966) *The senses considered as perceptual systems* Houghton Mifflin [aJRA]
(1979) *The ecological approach to visual perception* Houghton Mifflin [aJRA]
Gigerenzer, G (1991) From tools to theories A heuristic of discovery in cognitive psychology *Psychological Review* 98 (in press) [GG]
Gigerenzer, G & Murray, D J (1987) *Cognition as intuitive statistics* Erlbaum [GG]
Gigerenzer, G , Hell, W & Blank, H (1988) Presentation and content: The use of base rates as a continuous variable *Journal of Experimental Psychology: Human Perception and Performance* 14:513–25 [DWM]
Gilland, G & Shiffrin, R M (1984) A retrieval model for both recognition and recall *Psychological Review* 91:1–67 [aJRA]
Glenberg, A M (1976) Monotonic and nonmonotonic lag effects in paired-associate and recognition memory paradigms *Journal of Verbal Learning and Verbal Behavior* 15:1–16 [aJRA]
Gluck, M A & Bower, G H (1988) From conditioning to category learning: An adaptive network model *Journal of Experimental Psychology: General* 8:37–50 [aJRA]
Gluck, M A & Corter, J E (1985) Information, uncertainty, and the utility of categories In: *Proceedings of the Seventh Annual Conference of the Cognitive Science Society* Erlbaum [JEC]
Godbeer, G (1987) The computational complexity of the stable configuration problem for connectionist models M S thesis, Department of Computer Science, University of Toronto, September (Tech Report 208/88) [JKT]
Gottlieb, G L , Corcos, D M & Agarwal, G C (1989) Strategies for the control of voluntary movements with one mechanical degree of freedom *Behavioral and Brain Sciences* 12:189–250 [GCA]
Gould, S J (1986) Evolution and the triumph of homology, or why history matters *American Scientist* 74:60–69 [DWM]
Gould, S J & Lewontin, R C (1979) The spandrels of San Marco and the Panglossian paradigm: A critique of the adaptionist program *Proceedings of the Royal Society of London* 205:581–98 [aJRA]
Gould, S J & Lewontin, R C (1979) The spandrels of San Marco and the Panglossian paradigm: A critique of the adaptionist program Reprinted In: *Conceptual issues in evolutionary biology. An anthology*, ed E Sober MIT Press (1984) [PG-S]
Greiner, R (1990) Finding the optimal derivation strategy in a redundant knowledge base *Artificial Intelligence* (in press) [JKT]
Grossberg, S , ed (1988) Competitive learning: From interactive activation to adaptive resonance In: *Neural networks and natural intelligence* MIT Press [H-GG]
Hammond, K R (1955) Probabilistic functioning and the clinical method *Psychological Review* 62:255–62 [RH]
(1988) Judgement and decision making in dynamic tasks *Information and Decision Technologies* 14:3–14 [RH]
Harnad, S (1989) The symbol grounding problem *Physica D* 42:335–46 [RLC]
Hayes-Roth, B & Hayes-Roth, F (1977) Concept learning and the recognition and classification of exemplars *Journal of Verbal Learning and Verbal Behavior* 16:321–38 [aJRA]
Herrnstein, R J (1990) Rational choice theory: Necessary but not sufficient *American Psychologist* 45(3):356–67 [aJRA]
Herrnstein, R J & Vaughan, W , Jr (1980) Melioration and behavioral allocation In: *Limits to action: The allocation of individual behavior*, ed J E R Staddon Academic Press [aJRA]
Heyman, G M & Herrnstein, R J (1986) More on concurrent interval-ratio schedules: A replication and review *Journal of the Experimental Analysis of Behavior* 46:331–51 [EF]
Hoffman, J & Ziessler, C (1983) Objectidentifikation in kunstlichen Begriffshierarchien *Zeitschrift für Psychologie* 194:135–67 [aJRA, H-GG]
Homa, D & Cultice, J (1984) Role of feedback, category size, and stimulus distortion in the acquisition and utilization of ill-defined categories *Journal of Experimental Psychology: Learning, Memory, and Cognition* 10:83–94 [aJRA]
Houston, A I & McNamara, J M (1988) A framework for the functional analyses of behaviour. *Behavioral and Brain Sciences* 11:117–63 [aJRA]
Jeffrey, R (1987) Risk and human rationality *Monist* 70:223–36 [GG]
Johnson-Laird, P N & Byrne, R (in press) *Deduction* Hove & London: Erlbaum [JSBTE]
Jones, G V (1983) Identifying basic categories *Psychological Bulletin* 92:174–77 [JEC]
Just, M A & Carpenter, P A (1980) A theory of reading: Eye fixations to comprehension *Psychological Review* 87:329–54 [SLS]
Kahneman, D & O'Curry, S (1988) Surprise as an indication for spontaneous categorization Presented at the 29th Annual Meeting of the Psychonomic Society, November 10–12 [aJRA]
Kahneman, D & Tversky, A (1972) Subjective probability: A judgment of representativeness *Cognitive Psychology* 3:430–54 [JB]
(1973) On the psychology of prediction *Psychological Review* 80:237–51 [ES]
(1979) Prospect theory: An analysis of decision under risk *Econometrica* 47:263–91 [JB, PS]
(1984) Choices, values and frames *American Psychologist* 39:341–50 [aJRA, ES]
Kahneman, D , Slovic, P & Tversky, A (1982) *Judgment under uncertainty: Heuristics and biases* Cambridge University Press [JSBTE]
Kautz, H & Selman, B (in press) Hard problems for simple default logics *Artificial Intelligence* [JKT]
Kettlewell, P (1973) *The evolution of melanism* Oxford University Press (Oxford) [aJRA]
Kintsch, W (1970) Models for free recall and recognition In: *Models of human memory*, ed D A Norman Academic Press [aJRA]
Kirousis, L & Papadimitriou, C (1985) The complexity of recognizing polyhedral scenes Proceedings of the 26th Annual Symposium on Foundations of Computer Science, October, Portland, OR [JKT]
Knetsch, J L & Sinden, J A (1984) Willingness to pay and compensation: Experimental evidence of an unexpected disparity in measures of value *Quarterly Journal of Economics* 99:508–22 [JB]

Kotovsky, K., Hayes, J. R. & Simon, H. A. (1985) Why are some problems hard? Evidence from Tower of Hanoi. *Cognitive Psychology* 17:248–94. [aJRA]

Kuhn, D. (1989) Children and adults as intuitive scientists. *Psychological Review* 96:674–89. [TRS]

(in press) *The skills of argument*. Cambridge University Press. [JB]

Kuhn, D., Amsel, E. & O'Loughlin, M. (1988) *The development of scientific thinking skills*. Academic Press. [JB]

Kunda, Z., & Nisbett, R. E. (1986) The psychometrics of everyday life. *Cognitive Psychology* 18:195–224. [KJH]

Kyburg, H. E., Jr. (1983) Rational belief. *Behavioral and Brain Sciences* 6:231–73. [PS]

Langley, P., Bradshaw, G. L. & Simon, H. A. (1983) Rediscovering chemistry with the Bacon system. In: *Machine learning: An artificial approach*, ed. R. S. Michalski, J. G. Carbonell & T. M. Mitchell. Tioga. [SLS]

Larrick, R. P., Morgan, J. N. & Nisbett, R. E. (1990) Teaching the use of cost-benefit reasoning in everyday life. *Psychological Science* 1:362–70. [JB]

Larrick, R. P., Nisbett, R. E. & Morgan, J. N. (1990) Who uses cost-benefit reasoning? Unpub. manuscript. University of Michigan, Ann Arbor, MI. [JB]

Lebowitz, M. (1987) Experiments with incremental concept formation: UNIMEM. *Machine Learning* 2:103–38. [aJRA]

Lee, P. M. (1989) *Bayesian statistics*. Oxford University Press. [rJRA]

Leeuwenberg, E. L. J. & Buffart, H. F. M. J. (1983) An outline of coding theory: Summary of some related experiments. In: *Modern issues in perception*, ed. H.-G. Geissler, H. F. M. J. Buffart, E. L. J. Leeuwenberg & V. Sarris. Deutscher Verlag der Wissenschaften/North-Holland. [H-GG]

Lewontin, R. C. (1990) The evolution of cognition. In: *An invitation to cognitive science*, vol. 3, ed. D. Osherson & E. Smith. MIT Press. [ES]

Lichtenstein, S., Fischhoff, B. & Phillips, B. (1982) Calibration of probabilities: The state of the art to 1980. In: *Judgment under uncertainty: Heuristics and biases*, ed. D. Kahneman, P. Slovic & A. Tversky. Cambridge University Press. [JB]

Logsdon, J. B. (1990) Principles of optimum control of rapid human limb movements. M.S. thesis, Department of Electrical Engineering and Computer Science, University of Illinois at Chicago. [GCA]

Luce, R. D. (1986) *Response times*. Oxford University Press. [WS]

Macphail, E. M. (1987) The comparative psychology of intelligence. *Behavioral and Brain Sciences* 10:645–95. [RdS]

Mandler, J. M., Bauer, P. J. & McDonough, L. (1988) Differentiating global categories. Presented at the 29th Annual Meeting of the Psychonomic Society, November 10–12. [aJRA]

Marr, D. (1982) *Vision*. W. H. Freeman. [aJRA, SLS, TZ]

Marr, D. & Poggio, T. (1977) A theory of human stereo vision. Artificial Intelligence Laboratory, Massachusetts Institute of Technology. Technical Report AI Memo 451. [SLS]

Massaro, D. W. (1987) *Speech perception by ear and eye: A paradigm for psychological inquiry*. Erlbaum. [DWM]

(1989) Testing between the TRACE model and the fuzzy logical model of perception. *Cognitive Psychology* 21:398–421. [DWM]

Massaro, D. W. & Friedman, D. (1990) Models of integration given multiple sources of information. *Psychological Review* 97:225–52. [DWM]

Mayr, E. (1983) How to carry out the adaptionist program? *American Naturalist* 121:324–34. [aJRA]

McNeil, B. J., Pauker, S. G., Sox, H. C. & Tversky, A. (1982) On the elicitation of preferences for alternative therapies. *New England Journal of Medicine* 306:1259–62. [ES]

Medin, D. L. (1983) Structural principles of categorization. In: *Interaction: Perception, development and cognition*, ed. B. Shepp & T. Tighe. Erlbaum. [JEC]

Medin, D. L. & Edelson, S. M. (1988) Problem structure and the use of base-rate information from experience. *Journal of Experimental Psychology: General* 117:68–85. [aJRA]

Medin, D. L. & Schaffer, M. M. (1978) Context theory of classification learning. *Psychological Review* 85:207–38. [aJRA]

Medin, D. L., Altom, M. W., Edelson, S. M. & Freko, D. (1982) Correlated symptoms and simulated medical classification. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 8:37–50. [aJRA]

Meltzer, H. Y. & Stahl, S. M. (1976) The dopamine hypothesis of schizophrenia: A review. *Schizophrenia Bulletin* 2:19–76. [SLS]

Meyer, D. E. & Schvaneveldt, R. W. (1971) Facilitation in recognizing pairs of words: Evidence of a dependence between retrieval operations. *Journal of Experimental Psychology* 90:227–34. [aJRA]

Meyer, D. E., Smith, J. E. K. & Wright, C. E. (1982) Models for the speed and accuracy of aimed movements. *Psychological Review* 89:449–82. [GCA]

Meyer, D. E., Abrams, R. A., Kornblum, S., Wright, C. E. & Smith, J. E. K. (1988) Optimality in human motor performance: Ideal control of rapid aimed movements. *Psychological Review* 95:340–70. [GCA]

Michotte, A. (1946) *La perception de la causalité*. Vrin. [rJRA]

Moore, E. F. (1956) Gedanken-experiments on sequential machines. In: *Automata studies*, ed. C. E. Shannon & J. McCarthy. Princeton University Press. [DWM]

Mpitsos, G. J., Burton, R. M. & Creech, H. C. (1988) Connectionist networks learn to transmit chaos. *Brain Research Bulletin* 21:539–46. [RAMG]

Murphy, G. L. & Medin, D. L. (1985) The role of theories in conceptual coherence. *Psychological Review* 92:293–316. [JEC]

Murphy, G. L. & Smith, E. E. (1982) Basic level superiority in picture categorization. *Journal of Verbal Learning and Verbal Behavior* 21:1–20. [aJRA]

Neches, R., Langley, P. & Klahr, D. (1987) Production systems, learning, and development. In: *Production system models of learning and development*, ed. D. Klahr, P. Langley & R. Neches. MIT Press. [RLC]

Neely, J. H. (1977) Semantic priming and retrieval from lexical memory: Roles of inhibitionless spreading activation and limited-capacity attention. *Journal of Experimental Psychology: General* 106:226–54. [aJRA]

Newell, A. (1980) Physical symbol systems. *Cognitive Science* 4:135–83. [RLC]

(1982) The knowledge level. *Artificial Intelligence* 18:87–127. [BC]

Newell, A. & Rosenbloom, P. (1981) Mechanisms of skill acquisition and the law of practice. In: *Cognitive skills and their acquisition*, ed. J. R. Anderson. Erlbaum. [aJRA]

Newell, A. & Simon, H. A. (1972) *Human problem solving*. Prentice-Hall. [JSBTE]

Nisbett, R. E. & Ross, L. (1980) *Human inference: Strategies and shortcomings of social judgment*. Prentice-Hall. [aJRA]

Nisbett, R. E., Fong, G. T., Lehman, D. R. & Cheng, P. W. (1987) Teaching reasoning. *Science* 238:625–31. [JB]

Peterson, C. R. & Beach, L. R. (1967) Man as an intuitive statistician. *Psychological Bulletin* 68:29–46. [JB, JSBTE]

Peterson, C. R. & Uhlaner, J. J. (1964) Uncertainty, inference difficulty, and probability learning. *Journal of Experimental Psychology* 67:523–30. [JB]

Petruccelli, J. D. (1981) Best-choice problems involving uncertainty of selection and recall of observations. *Journal of Applied Probability* 18:415–25. [WS]

Piaget, J. (1970) *Genetic epistemology*. Columbia University Press. [RLC]

Polya, G. (1954) *Mathematics and plausible reasoning*. Vol. 2: *Patterns of plausible inference*. Princeton University Press. [PS]

Posner, M. I. & Keele, S. W. (1968) On the genesis of abstract ideas. *Journal of Experimental Psychology* 77:353–63. [aJRA]

Prade, H. (1985) A computational approach to approximate and plausible reasoning with applications to expert systems. *IEEE Transactions on Pattern Analysis & Machine Intelligence* 7:260–82. [PS]

Pylyshyn, Z. W. (1984) *Computation and cognition*. MIT Press. [RLC]

Rachlin, H. (1989) *Judgment, decision, and choice: A cognitive/behavioral synthesis*. W. H. Freeman. [EF, HR]

Rachlin, H., Battalio, R. C., Kagel, J. H. & Green, L. (1981) Maximization theory in behavioral psychology. *Behavioral and Brain Sciences* 4:371–88. [aJRA]

Redelmeier, D. A. & Tversky, A. (1990) The aggregate/individual discrepancy: Contrasting perspectives in medical care. *New England Journal of Medicine* (in press). [ES]

Reed, S. K. (1972) Pattern recognition and categorization. *Cognitive Psychology* 3:382–407. [aJRA]

Riefer, D. M. & Batchelder, W. H. (1988) Multinomial modeling and the measurement of cognitive processes. *Psychological Review* 95:318–39. [WS]

Rosch, E. (1978) Principles of categorization. In: *Cognition and categorization*, ed. E. Rosch & B. B. Lloyd. Erlbaum. [GG]

Rosch, E., Mervis, C. B., Gray, W., Johnson, D. & Boyes-Braem, P. (1976) Basic objects in natural categories. *Cognitive Psychology* 7:573–605. [aJRA, JEC]

Rosenbloom, P. S., Laird, J. E. & Newell, A. (1987) SOAR: An architecture for general intelligence. *Artificial Intelligence* 33:1–64. [BC]

Rumelhart, D. E. & McClelland, J. L. (1986) *Parallel distributed processing: Explorations in the microstructure of cognition*. Vol. 1: *Foundations*. MIT Press. [SLS]

Rumelhart, D. E., Hinton, G. E. & Williams, R. J. (1986) Learning internal representations by error propagation. In: *Parallel distributed processing*, vol. 1, ed. D. E. Rumelhart & J. L. McClelland. MIT Press. [aJRA]

Rumelhart, D. E., McClelland, J. L. & the PDP Research Group (1986)

Parallel distributed processing: Explorations in the microstructure of cognition. vols 1–2 MIT Press [KJH]

Russell, S J (1986) A quantitative analysis of analogy by similarity In: Proceedings of the National Conference on Artificial Intelligence. Philadelphia, PA American Association for Artificial Intelligence (AAAI) [aJRA]

Salton, G & McGill, M J (1983) *Introduction to modern information retrieval* McGraw-Hill [aJRA]

Satyanarayanan, M (1981) A study of file sizes and functional lifetimes In: Proceedings of the Eighth Symposium on Operating Systems Principles. Asilomar, CA, December [aJRA]

Schultz, T R (1982) Rules for causal attribution *Monographs of the Society for Research in Child Development* 47(1, Serial No 194) [aJRA]

Schustack, M W & Sternberg, R J (1981) Evaluation of evidence in causal inference *Journal of Experimental Psychology: General* 110:101–20 [aJRA, TRS]

Schwarz, W (1989) The generalized Bouman-van der Velden quantum coincidence detector *Biological Cybernetics* 61:315–18 [WS]

(1990) Stochastic accumulation of information in discrete time: Comparing exact results and Wald approximations *Journal of Mathematical Psychology* 34:229–36 [WS]

Selman, B & Kautz, H (1990) Model-preference default theories *Artificial Intelligence* 45:287–322 [JKT]

Selman, B & Levesque, H (1989a) Abductive and default reasoning: A computational core Proceedings of the Eighth National Conference on Artificial Intelligence, Boston, MA [JKT]

(1989b) The tractability of path-based inheritance Proceedings of the Eleventh International Joint Conference on Artificial Intelligence. Detroit, August [JKT]

Shackle, G L S (1949) *Expectation in economics* Cambridge University Press (Cambridge) [PS]

Shafir, E B, Osherson, D N & Smith, E E (1989) An advantage model of choice *Journal of Behavioral Decision Making* 2:1–23 [ES]

Shepard, R N (1987) Towards a universal law of generalization for psychological science *Science* 237:1317–23 [aJRA]

(1989) A law of generalization and connectionist learning Plenary address to the Cognitive Science Society, Ann Arbor, MI. August 18 [aJRA]

Shimp, C P (1966) Probabilistically reinforced choice behavior in pigeons *Journal of the Experimental Analysis of Behavior* 9:443–55 [aJRA]

Shultz, T R & Ravinsky, F (1977) Similarity as a principle of causal inference *Child Development* 48:1552–58 [TRS]

Simon, H A (1955a) A behavioral model of rational choice *Quarterly Journal of Economics* 69:99–118 [aJRA, GMB]

(1955b) On a class of skew distribution functions *Biometrika* 52:425–40 [rJRA]

(1982) *The sciences of the artificial, 2nd ed* MIT Press [BC]

Slovic, P & Lichtenstein, S (1971) Comparison of Bayesian and regression approaches to the study of information processing in judgment *Organizational Behavior and Human Performance* 6:649–744 [RH]

Slovic, P, Fischhoff, B & Lichtenstein, S (1977) Behavioral decision theory *Annual Review of Psychology* 228:1–39 [JSBTE]

(1978) Accident probabilities and seat belt usage: A psychological perspective *Accident Analysis and Prevention* 10:281–85 [ES]

(1982) Response mode, framing and information-processing effects in risk assessment In: *New directions for methodology of social and behavioral science. No 11 Question framing and response consistency*. ed R Hogarth Jossey-Bass [ES]

Solnick, J V, Kannenberg, C H, Eckerman, D A & Waller, M B (1980) An experimental analysis of impulsivity and impulse control in humans *Learning and Motivation* 11:61–77 [JB]

Spall, J C (1988) *Bayesian analysis of time series and dynamic models* Marcel Dekker [RAMG]

Staddon, J E R (1987) Optimality theory and behavior In: *The latest on the best: Essays on evolution and optimality*. ed J Dupré MIT Press [aJRA, EF]

Stephens, D W & Krebs, J R (1986) *Foraging theory* Princeton University Press [arJRA]

Stolarz-Fantino, S & Fantino, E (1990) Cognition and behavior analysis: A review of Rachlin's *Judgment, decision and choice* *Journal of the Experimental Analysis of Behavior* 54:317–22 [EF]

Stritter, E (1977) File migration Unpub Ph D thesis (STAN-CS-77-594). Stanford University [aJRA]

Suppes, P (1984) *Probabilistic metaphysics* Basil Blackwell [aJRA]

Tanner, M A & Wong, W H (1987) The calculation of posterior distributions by data augmentation *Journal of the American Statistical Association* 82:528–40 [rJRA]

Thagard, P (1989) Explanatory coherence *Behavioral and Brain Sciences* 12:435–67 [KJH]

Thagard, P, Holyoak, K J, Nelson, G & Gochfeld, D (1990) Analog retrieval by constraint satisfaction *Artificial Intelligence* 46:259–310 [KJH]

Thaler, R H (1980) Toward a positive theory of consumer choice *Journal of Economic Behavior and Organization* 1:39–60 [ES]

Tierney, L & Kadane, J B (1986) Accurate approximations for posterior moments and marginal destinies *Journal of the American Statistical Association* 81:82–86 [rJRA]

Townsend, J T (1971) A note on the identifiability of parallel and serial processes *Perception & Psychophysics* 10:161–63 [WS]

Townsend, J T & Ashby, F G (1983) *Stochastic modeling of elementary psychological processes* Cambridge University Press (Cambridge) [WS]

Tsotsos, J (1989) The complexity of perceptual search tasks Proceedings of the International Joint Conference on Artificial Intelligence. Detroit, August [JKT]

(1990) Analyzing vision at the complexity level *Behavioral and Brain Sciences* 13(3):423–45 [JKT]

(submitted) Active vs passive visual search: Which is more efficient? *International Journal of Computer Vision* [JKT]

Tversky, A & Kahneman, D (1980) Causal schemas in judgment under uncertainty In: *Progress in social psychology*, vol 1. ed M Fishbein Erlbaum [TRS]

(1982) Evidential impact of base rates In: *Judgment under uncertainty: Heuristics and biases*. ed D Kahneman, P Slovic & A Tversky Cambridge University Press [EF]

(1983) Extension versus intuitive reasoning: The conjunction fallacy in probability judgment *Psychological Review* 90:293–315 [DWM]

Von Winterfeldt, D & Edwards, W (1986) *Decision analysis and behavioural research* Cambridge University Press (Cambridge) [JSBTE]

Vorberg, D (1977) On the equivalence of parallel and serial models of information processing Paper presented at the 10th Annual Mathematical Psychology Meeting. Los Angeles, CA [WS]

Vorberg, D & Ulrich, R (1987) Random search with unequal search rates: Serial and parallel generalizations of McGill's model *Journal of Mathematical Psychology* 31:1–23 [WS]

Wasserman, E A (1990) Attribution of causality to common and distinctive elements of compound stimuli *Psychological Science* 1:298–302 [JB]

White, K G, McCarthy, D & Fantino, E (1989) Cognition and behavior analysis *Journal of the Experimental Analysis of Behavior* 52:197–98 [EF]

Wickelgren, W A (1976) Memory storage dynamics In: *Handbook of learning and cognitive processes*, ed W K Estes Erlbaum [aJRA]

Wilkie, D R (1954) Facts and theories about muscle *Progress in Biophysics and Biophysical Chemistry* 4:288–324 [GCA]

Williams, B A (1988) Reinforcement, choice, and response strength In: *Stevens's handbook of experimental psychology*. ed R C Atkinson, R J Herrnstein, G Lindzey & R D Luce Wiley [aJRA]