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 "over-enthusiastic 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