John R. Anderson. 1990. The Adaptive Character of Thought. Hillsdale, NJ: Lawrence Erlbaum Associates.

1 Introduction

Contents

PRELIMINARIES

A person writes a research monograph such as this with the intention that it will be read. As a consumer of such monographs, I know it is not an easy decision to invest the time in reading and understanding one. Therefore, it is important to be up front about what this book has to offer. In a few words, this book describes a new methodology for doing research in cognitive psychology and applies it to produce some important developments. The new methodology is important, because it offers the promise of rapid progress on some of the agenda of cognitive psychology. This methodology concentrates on the adaptive character of cognition, in contrast to the typical emphasis on the mechanisms underlying cognition.

I have been associated with a number of theoretical monographs (Anderson, 1976, 1983; Anderson & Bower, 1973). The sustaining question in this succession of theories has been the nature of human knowledge. The 1973 book was an attempt to take human memory tradition as it had evolved in experimental psychology and use the new insights of artificial intelligence to relate the ideas of this tradition to fundamental issues of human knowledge. The theory developed in that book was called HAM, a propositional network theory of human memory. The 1976 book was largely motivated by the pressing need to make a distinction between procedural and declarative knowledge. This distinction was absent in the earlier book and in the then-current literature on human memory. The theory developed in that book was called ACT. In ACT, declarative knowledge was represented in a propositional network and procedural knowledge in a production system. The 1983 book was motivated both by breakthroughs in developing a learning theory that accounts for the acquisition of procedural knowledge and in identifying a neurally plausible basis for the overall implementation of the theory. It was called ACT* to denote that it was a completion of the theoretical development begun in the previous book.

In the 1983 book on the ACT* theory, I tried to characterize its relationship to the ACT series of theories and my future plans for research: ". . . my plan for future research is to try to apply this theory wide and far, to eventually gather enough evidence to permanently break the theory and to develop a better one. In its present stage of maturity the theory can be broadly applied, and such broad application has a good chance of uncovering fundamental flaws" (Anderson, 1983, p. 19).

My method of applying the theory has been to use it as a basis for detailed studies of knowledge acquisition in a number of well-defined domains such as high-school mathematics and college-level programming courses. In these studies, we have been concerned with how substantial bodies of knowledge are both acquired and evolve to the point where the learner has a very powerful set of problem-solving skills in the domain (Anderson, Boyle, Corbett, & Lewis, in press; Anderson, Boyle, & Reiser, 1985; Anderson, Conrad, & Corbett; in press).

The outcome of this research effort has certainly not been what I expected. Despite efforts to prove the theory wrong (I even created a new production system, called PUPS for Penultimate Production System, to replace ACT*—Anderson & Thompson, 1989), I failed to really shake the

old theory. As I have written elsewhere (Anderson, 1987c; Anderson, Conrad, & Corbett, in press), we have been truly surprised by the success of the ACT* theory in dealing with the data we have acquired about complex skill acquisition with our intelligent tutors. ACT* proved not to be vulnerable to a frontal assault, in which its predictions about skill acquisition are compared to the data. This book contains some theoretical ideas that are rather different than ACT*, produced by the new methodology that the book describes. These ideas do not so much contradict ACT* as they address the subject of human cognition in a different way.

The A in ACT* stands for Adaptive, and this book results from an effort to think through what it might mean for human cognition to be adaptive. However, this book is not cast as an update on the ACT* theory, but rather is an effort to develop some points about human cognition from an adaptive perspective. The majority of the book, the next four chapters, tries to develop theory from an adaptive perspective in four related fields of cognition. This chapter is devoted to setting the stage for that development.

To state up front where this chapter is going, the argument is that we can understand a lot about human cognition without considering in detail what is inside the human head. Rather, we can look in detail at what is outside the human head and try to determine what would be optimal behavior given the structure of the environment and the goals of the human. The claim is that we can predict behavior of humans by assuming that they will do what is optimal. This is a different level of analysis than the analysis of mental mechanisms that has dominated information-processing psychology. Having raised the possibility of levels of analysis, the questions arise as to just how many levels there are and why we would want to pursue one level rather than another. It turns out that there have been many ideas expressed on these topics in cognitive science. Rather than just present my position on this and pretend to have invented the wheel, it is appropriate to review the various positions and their interrelationships. However, if the reader is impatient with such discussion, it is possible to skip to the section in the chapter that presents "The New Theoretical Framework," where the discussion of rational analysis begins.

LEVELS OF A COGNITIVE THEORY

Table 1-1 is a reference for this section, in that it tries to relate the terminology of various writers. I start with the analysis of David Marr, which I have found to be particularly influential.

Marr's System of Levels

No sooner had I sent in the final draft of the ACT* book to the publisher than I turned to reading the recently published book by Marr (1982) on

TABLE 1-1
Levels of Cognitive Theory According to Various Cognitive Scientists

Marr	*Chomsky*	*Pylyshyn*	*Rumelhart and McClelland*	*Newell*	*Anderson*
Computational Theory	Competence	Semantic Level		Knowledge Level	Rational Level
Representation and Algorithm	Performance	Algorithm	Macrotheory/ Rules	Program Symbol Level	Algorithm
		Functional Architecture	Microtheory PDP models	Register Transfer Level	Implementation
Hardware Implementation		Biological Level		Device	Biological

vision. It contained a very compelling argument about how to do theory development. I read over and over again his prescription for how to proceed in developing a theory:

> We can summarize our discussion in something like the manner shown in Figure 1-4 [our Table 1-2], which illustrates the different levels at which an information-processing device must be understood before one can be said to have understood it completely. At one extreme, the top level, is the abstract computational theory of the device, in which the performance of the device is characterized as a mapping from one kind of information to another, the abstract properties of this mapping are defined precisely, and its appropriateness and adequacy for the task at hand are demonstrated. In the center is the choice of representation for the input and output and the algorithm to be used to transform one into the other. And at the other extreme are the details of how the algorithm and representation are realized physically – the detailed computer architecture, so to speak. These three levels are coupled, but only loosely. The choice of an algorithm is influenced, for example, by what it has to do and by the hardware in which it must run. But there is a wide choice available at each level, and the explication of each level involves issues that are rather independent of the other two (pp. 24–25).

Although algorithms and mechanisms are empirically more accessible, it is the top level, the level of computational theory, which is critically important from an information-processing point of view. The reason for this is that the nature of the computations that underlie perception depends more upon the computational problems that have to be solved than upon the particular hardware in which their solutions are implemented. To phrase the matter another way, an algorithm is likely to be understood more readily by understanding the nature of the problem being solved than by examining the mechanism (and the hardware) in which it is embodied.

In a similar vein, trying to understand perception by studying only neurons is like trying to understand bird flight by studying only feathers: It just cannot

TABLE 1-2
Marr's Description of the Three Levels at Which Any Machine Carrying out an Information-Processing Task Must be Understood

Computational Theory	*Representation and Algorithm*	*Hardware Implementation*
What is the goal of the computation, why is it appropriate, and what is the logic of the strategy by which it can be carried out?	How can this computational theory be implemented? In particular, what is the representation for the input and output, and what is the algorithm for the transformation?	How can the representation and algorithm be realized physically?

> be done. In order to understand bird flight, we have to understand aerodynamics; only then do the structure of feathers and the different shapes of birds' wings make sense. More to the point, as we shall see, we cannot understand why retinal ganglion cells and lateral geniculate neurons have the receptive fields they do just by studying their anatomy and physiology. We can understand how these cells and neurons behave as they do by studying their wiring and interactions, but in order to understand *why* the receptive fields are as they are—why they are circularly symmetrical and why their excitatory and inhibitory regions have characteristic shapes and distributions—we have to know a little of the theory of differential operators, band-pass channels, and the mathematics of the uncertainty principle (pp. 27–28).

Marr's terminology of "computational theory" is confusing and certainly did not help me appreciate his points. (Others have also found this terminology inappropriate—e.g., Arbib, 1987). His level of computational theory is not really about computation but rather about the goals of the computation. His basic point is that one should state these goals and understand their implications before one worries about their computation, which is really the concern of the lower levels of his theory.

Marr's levels can be understood with respect to stereopsis. At the computational level, there is the issue of how the pattern of light on each retina enables inferences about depth. The issue here is not how it is done, but what should be done. What external situations are likely to have given rise to the retinal patterns? Once one has a theory of this, one can then move to the level of representation and algorithm and specify a procedure for actually extracting the depth information. Having done this, one can finally inquire as to how this procedure is implemented in the hardware of the visual system.

Marr compared his computational theory to Gibson's (1966) ecological optics. Gibson claimed that there were certain properties of the stimulus which would invariantly signal features in the external world. In his terminology, the nervous system "resonates" to these invariants. Marr credited Gibson with recognizing that the critical question is to identify what in the stimulus array signals what in the real world. However, he criticized Gibson for not recognizing that, in answering this question, it is essential to precisely specify what that relationship is. The need for precision is apparent to someone, like Marr, working on computer vision. This need was not apparent to Gibson (see Shepard, 1984, for an extensive analysis of Gibson's theory).

I tried to see how to apply Marr's basic admonition to my own concern, which was higher-level cognition, but it just did not seem to apply. Although Marr's prescription seemed fine for vision, it seemed that the representation and algorithm level was the fundamental level for the study of human cognition. It was certainly the level that information-processing psychology had progressed along for the last 30 years.

What I initially failed to focus on was the essential but unstated adaptionist principle in Marr's argument. Vision could be understood by studying a problem only if (a) we assumed that vision was a solution to that problem, (b) we assumed that the solution to that problem was largely unique, and (c) we assumed that something forced vision to adopt that solution. For instance, in the case of stereopsis, we had to assume that vision solved the problem of extracting the three-dimensional structure from two two-dimensional arrays, and there was usually a single best interpretation of two two-dimensional arrays. Analysis of the visual environment of humans suggests that there is usually a best interpretation. To pursue Marr's agenda, it is not enough to argue that there is a unique best solution; we also have to believe that there are adaptive forces that created a visual system that would deliver this best solution. Perhaps other aspects of cognition deal with problems that have best solutions, and the organism is similarly adapted to achieve these best solutions. Once I cast what Marr was doing in these terms, I saw the relevance of his arguments to cognition in general.

Marr's hardware-implementation level may still be inapplicable to the study of cognition. It made sense in the case of vision, where the physiology is reasonably well understood. However, the details of the physical base of cognition are still unclear.

Marr's analysis of these levels is very much motivated by the issue of how to make progress in cognitive science. As he saw it, the key to progress is to start off at the level of computational theory. As he bemoaned about the practice of theory in vision:

> For far too long, a heuristic program for carrying out some task was held to be a theory of that task, and the distinction between what a program did and how it did it was not taken seriously. As a result, (1) a style of explanation evolved that invoked the use of special mechanisms to solve particular problems, (2) particular data structures, such as the lists of attribute value pairs called property lists in the LISP programming language, were held to amount to theories of the representation of knowledge, and (3) there was frequently no way to determine whether a program would deal with a particular case other than by running the program. (Marr, 1982, p. 28)

These problems certainly characterize the human information-processing approach to higher level cognition. We pull out of an infinite grab bag of mechanisms, bizarre creations whose only justification is that they predict the phenomena in a class of experiments. These mechanisms are becoming increasingly complex, and we wind up simulating them and trying to understand their behavior just as we try to understand the human. We almost never ask the question of why these mechanisms compute in the way they do.

Chomsky's Competence and Performance

Marr related his distinction to Chomsky's much earlier distinction between competence and performance in linguistics, identifying his computational theory with Chomsky's competence component. As Chomsky (1965) described the distinction:

> Linguistic theory is concerned primarily with an ideal speaker-listener, in a completely homogeneous speech-community, who knows its language perfectly and is unaffected by such grammatically irrelevant conditions as memory limitations, distractions, shifts of attention and interest, and errors (random or characteristic) in applying his knowledge of the language in actual performance. This seems to me to have been the position of the founders of modern general linguistics, and no cogent reason for modifying it has been offered. To study actual linguistic performance, we must consider the interaction of a variety of factors, of which the underlying competence of the speaker-hearer is only one. In this respect, study of language is no different from empirical investigation of other complex phenomena.
>
> We thus make a fundamental distinction between *competence* (the speaker-hearer's knowledge of his language) and *performance* (the actual use of language in concrete situations). Only under the idealization set forth in the preceding paragraph is performance a direct reflection of competence. In actual fact, it obviously could not directly reflect competence. A record of natural speech will show numerous false starts, deviations from rules, changes of plan in mid-course, and so on. The problem for the linguist, as well as for the child learning the language, is to determine from the data of performance the underlying system of rules that has been mastered by the speaker-hearer and that he puts to use in actual performance. (pp. 3–4)

The competence-performance distinction has been the source of a great deal of confusion and controversy. The relationship between competence and performance is really not the same as the relationship between Marr's level of computational theory and his lower levels.[1] In Marr's case, the lower levels achieve the goals of the computational level. Chomsky's competence level is a theory based on a certain subset of data that is thought to be a direct and reliable reflection of the person's linguistic knowledge. For instance, judgments of whether a sentence is grammatically well formed provide key data for a theory of competence, but time to understand a sentence is thought to be less stable and is consigned to a theory of performance. Performance is somehow constrained to reflect the competence, but it reflects other factors as well. Unlike Marr's case, performance is not just a matter of implementing the goals of competence. Indeed, unlike Marr's computational-level, Chomsky's competence is not concerned with the goals of the system. A computational-level theory of language would have to be concerned with the functionality of language—a concern that Chomsky explicitly rejected. In fact, in contrast to all the other proposals for a higher level, Chomsky's competence is unique, in that it explicitly eschews concerns with functionality. Therefore, it is removed from the other levels in Table 1-1. It is better not thought of as a level in the sense of Marr's levels.

Still, Chomsky used the competence level to serve the same role in theory building as Marr used computational theory. Under both analyses, the scientist should first work out the higher level. Both felt that this was a key to making progress. Also, the lower levels are constrained somehow to reflect the higher levels.

[1] Consequently, some people have questioned why I mentioned Chomsky's distinction at all. The answer is that Chomsky and Marr relate their distinctions to one another, and Chomsky's distinction is very well known.

Pylyshyn's Distinction Between Algorithm and Functional Architecture

Pylyshyn (1984) distinguished between three levels that are quite analogous to Marr's three levels. These he called the semantic level, the symbolic level, and the biological level. He developed the semantic level with respect to Newell's concept of a knowledge level, and we turn to reviewing that concept extensively at the end of this section. He had little to say about the biological level beyond making standard arguments for its inadequacy as the sole level of psychological explanation. Of major interest is a distinction he developed within the symbolic level, between mental algorithms and the functional architecture (Pylyshyn, 1980). These are two levels sandwiched between the biological and the semantic. They are very important levels, because they are the levels at which most cognitive psychologists have aimed their research.

The algorithm level is an abstract specification of the steps a system must go through to perform a process. This specification is abstracted away from the functional architecture that actually implements the steps of the algorithm. As Pylyshyn (1984) described the functional architecture:

> It includes the basic operations provided by the biological substrate, say, for storing and retrieving symbols, comparing them, treating them differently as a function of how they are stored, (hence, as a function of whether they

> represent beliefs or goals), and so on, as well as such basic resources and constraints of the system, as a limited memory. It also includes what computer scientists refer to as the "control structure", which selects which rules to apply at various times (p. 30).

Despite his reference to biology, Pylyshyn's functional architecture is an abstraction above the biological level. The analogy Pylyshyn used is that the distinction between the algorithm level and the functional architecture corresponds to the distinction between a canonical computer program and its machine implementation. Pylyshyn's assertion is that there is a particular distinguished algorithm level, rather than the situation in computers where there can be layers of languages each compiling into a lower level. As he wrote: "Rather than a series of levels, we have a distinguished level, the level at which interpretation of the symbols is in the intentional, or cognitive, domain or in the domain of the objects of thought" (Pylyshyn, 1984, p. 95).

One of the examples Pylyshyn used to illustrate the distinction between algorithm and functional architecture is the process of answering questions like "If John is taller than Mary and John is shorter than Fred, who is the tallest?" At the algorithm level, one could imagine a procedure for answering such questions in which each of the premises (e.g., John is taller than Mary) requires placing the terms in an ordered list, and answering the question involves reading off the person at one end of the list. Such a procedure could be implemented in one of many programming languages that would involve a set of instructions. Pylyshyn's interpretation of the algorithm level seems to be, really, the specific programming language rather than the general procedure. The functional architecture would be concerned with the implementation of the instructions of that programming language on a particular machine. Thus, the functional architecture might tell us how long it takes to insert an element into a list or how long it takes to read off the item at the end of the list.

As argued in Anderson (1987a), the assumptions of the ACT* theory can be sorted into assumptions about the algorithm level and assumptions about functional architecture. In ACT*, there is one set of production system principles for representing knowledge states and determining transitions among these knowledge states and another set of principles for computing activation levels for the knowledge structures that determine how these knowledge structures map onto the specifics of behavior like response time. The first set of assumptions is about the algorithm level, and the second is about the functional architecture. Amazingly, I did not realize that the assumptions of my theory were at two levels until 1984 (Anderson, 1984), when I began to think about the implications of the theory for intelligent tutoring. Curiously, it seemed that only the algorithm level had implications for intelligent tutoring, and intelligent tutoring only had implications for the algorithm level.

Pylyshyn introduced an interesting principle to distinguish what belongs to the functional architecture from what belongs to the algorithm level:

> ***Cognitive Impenetrability.*** **The operations at the functional architecture level are not affected by the organism's goals and beliefs.**

Thus, to call up the standard Sternberg (1969) model of memory scanning, although goals and beliefs determine what digits the subject will compare with what digits, the actual process of comparing one digit to another (the famous 35–40 msec) is not affected by goals and beliefs. Thus, the process of incorporating the experimenter's instructions is at the algorithm level, whereas the actual memory scan is at the level of functional architecture. Cognitive impenetrability gets at the essence of the difference between a symbolic and a subsymbolic level. Only the symbolic level should be influenced by the semantic contents of our knowledge.

McClelland and Rumelhart's PDP Level

A major new surge (to say the least) in cognitive science has been the appearance of connectionist theories, which try to model cognition in terms of units thought to reflect some aspects of neuronal functioning. Some question has arisen as to the level at which such connectionist theories are cast. It might seem obvious that they should be identified with the hardware level of Marr. However, Rumelhart and McClelland (1985), in their reply to Broadbent (1985), noted that their connectionist models (which they call PDP models – see McClelland & Rumelhart, 1986; Rumelhart & McClelland, 1986) are considerably abstracted from the hardware level and are really at the algorithm and representation level. As noted with respect to Pylyshyn's levels (see Table 1-1), Marr's representation and algorithm level can be broken into at least two levels, the levels that Pylyshyn called the algorithm level and the level of functional architecture. In Pylyshyn's terminology, connectionist models are theories at the level of the functional architecture.

Rumelhart and McClelland (1986) argued that the algorithm level is not a real level, but rather that emergent properties of their functional architecture may approximate the rules that other theorists propose for the algorithm level. As they wrote:

> There is still another notion of levels which illustrates our view. This is the notion of levels implicit in the distinction between Newtonian mechanics on the one hand and quantum theory on the other. It might be argued that

> conventional symbol processing models are macroscopic accounts, analogous to Newtonian mechanics, whereas our models offer more microscopic accounts, analogous to quantum theory. Note, that over much of their range, these two theories make precisely the same predictions about behavior of objects in the world. Moreover, the Newtonian theory is often much simpler to compute with since it involves discussions of entire objects and ignores much of their internal structure. However, in some situations Newtonian theory breaks down. In these situations we must rely on the microstructural account of quantum theory. Through a thorough understanding of the relationship between the Newtonian mechanics and quantum theory we can understand that the macroscopic level of description may be *only an approximation* to the more microscopic theory. Moreover, in physics, we understand just when the macrotheory will fail and the microtheory must be invoked. We understand the macrotheory as a useful formal tool by virtue of its relationship to the microtheory. In this sense the objects of the macrotheory can be viewed as *emerging* from interactions of the particles described at the microlevel. (p. 125)

Under some interpretation, McClelland and Rumelhart must be right that the brain only approximates the symbolic rules we ascribe to it. However, the interesting question is, "Under what interpretation?". A computer only approximates the program it is implementing—there are failures of memory, interrupt device processes, overhead of operating systems, small surges of voltage, and so on. However, the approximation in the case of the computer is clearly both good and faithful. The Rumelhart and McClelland enterprise is based on the belief that the brain approximation of the algorithm level is neither good nor faithful. Thus, their point is not just that it is an approximation (surely their own PDP models are approximations in some sense) but that the algorithm level often can be a bad approximation that misses the essence of the behavior at hand.

Their basic reason for believing that the algorithm level is a bad approximation is their belief that learning is defined at the lower level. They used a compiler analogy to make their point. In their analogy, the algorithm level corresponds to a PASCAL program and the lower level to assembly code. Their view is that the mind programs itself at the assembly code level, and the assembly code can only be approximated by PASCAL code.

As they wrote in Rumelhart and McClelland (1985), "Because there is presumably no compiler to enforce the identity of our higher level and lower level descriptions in science, there is no reason to suppose there is a higher level description exactly equivalent to any particular lower level description" (pp. 195–196). Or, as they wrote in Rumelhart and McClelland (1986), "Since there is every reason to suppose that most of the programming that might be taking place in the brain is taking place at a 'lower level' rather than a 'higher level' it seems unlikely that some particular higher level description will be identical to some particular lower level description" (pp. 124–125).

I think they are wrong in this point and that their arguments in the quoted passages are so weak as to be vacuous. The reasoning implicit in both of these passages is basically, "We don't know how the brain does it, therefore it cannot" (i.e., an argument from ignorance). The psychological reality of the algorithm level is very much an empirical question to be decided by whether there are phenomena that can only be explained at this level. A great many phenomena of higher level cognition currently only have explanations at the algorithm level. This includes much of syntactic processing (Pinker & Prince, 1988), almost all of problem solving (Newell & Simon, 1972), learning of problem-solving skills (Anderson, 1981a), and human deduction (Johnson-Laird, 1983). Indeed, there is very little of what is conventionally called "thinking," which has been treated by connectionist models, let alone successfully treated. On the other hand, when we turn to the implementation of these thinking processes, such as memory for the facts being used, connectionist models enjoy great success.

The connectionists are focused on the level of functional architecture, because they believe that this level offers the key insights for making progress towards a successful scientific theory. Rather than the Marr-Chomsky approach of trying to guarantee some overall correctness or well-formedness of the computation, their concern is that the computation takes place in something at least approximating neural elements. As they wrote:

> We have found that information concerning *brain-style* processing has itself been very provocative in our model building efforts. Thus, we have, by and large, not focused on *neural modeling* (i.e., the modeling of neurons), but rather we have focused on *neurally inspired* modeling of cognitive processes. Our models have not depended strongly on the details of brain structure or on issues that are very controversial in neuroscience. Rather, we have discovered that if we take some of the most obvious characteristics of brain-style processing seriously we are led to postulate models which differ in a number of important ways from those postulated without regard for the hardware on which these algorithms are to be implemented. We have found that top-down considerations revolving about a need to postulate parallel, cooperative computational models (cf. Rumelhart, 1977) have meshed nicely with a number of more bottom-up considerations of brain style processing. (Rumelhart & McClelland, 1986, p. 130).

In writing the ACT* book, I was much concerned with this argument and tried to make that theory neurally realistic. I have since come to seriously question the force of the neural constraint for two reasons. As they noted in the preceding quote, our knowledge of neural mechanisms is weak. Thus, it is not clear what are "the most obvious characteristics of brain-style processing," and their neural assumptions may not correspond to what actually happens in the brain, as Crick and Asanuma (1986) complained with respect to PDP models. On the other hand, we may think certain things cannot happen that do and so unnecessarily restrict ourselves. So, we may very well be misguided by a premature insistence on neural fidelity.

However, my deeper concern is that it is not clear that the neural concerns provide much constraint, misguided or not. It is unclear what one cannot predict by a suitable arrangement of neural elements given that they are computationally universal. Even more disturbing, it seems that there are multiple arrangements of neural-like elements that will produce the same phenomena. That is, even if we restrict ourselves to some circumscribed class of neural models, like the PDP class, we will have the identifiability problems that have haunted all cognitive science theorizing at this level. Rather than too much constraint, it is likely to be, once again, a matter of too little.

Newell's Knowledge Level and the Principle of Rationality

As stated in the discussion of Pylyshyn and the PDP models, it is standard practice in cognitive psychology to work at what is variously called the symbol level (Pylyshyn), or the level of representation and algorithm (Marr). No one doubts the existence of a biological level below the symbol level, but we choose to work at the higher level, either because we believe we do not have adequate evidence about the biological level or because we believe we can make progress on psychological issues more rapidly by working at a higher level of abstraction. However, few until Marr and Newell had suggested that it was possible that there was a useful level of analysis above the symbol level. Newell formulated this as his knowledge level hypothesis.

> ***The Knowledge Level Hypothesis.*** **There exists a distinct computer systems level, lying immediately above the symbol level, which is characterized by knowledge as the medium and the principle of rationality as the law of behavior (Newell, 1982,p. 99).**

As the preceding quote indicates, Newell's development of the knowledge level was originally with respect to computer systems, but he extended it to the human situation. Newell saw a lot of similarity between levels of analysis for the computer and levels of analysis for the human. When speaking of computers, he used the terms *program* or *symbol level* to refer to what Pylyshyn called the *algorithm level,* the term *register-transfer level* to refer to *functional architecture,* and *device level* to refer to the *biological level.* However, our focus is on his *knowledge level,* which corresponds to Pylyshyn's *semantic level.* His statement of this level has proven to be quite influential in artificial intelligence.

The concept that gives precision to Newell's knowledge level is the principle of rationality. As Newell (1982) stated it,

> ***Principle of Rationality.*** **"If an agent has knowledge that one of its actions will lead to one of its goals, then the agent will select that action," (p. 102).**

There are a number of undefined terms in this specification, but he did develop what he meant by each:

Goals. The organism is assumed to want certain states of affairs to come to be.

Selection. The claim is that the organism will perform one of the actions it knows to achieve its goals. An important complication is that it is possible that multiple actions will achieve a goal or that multiple goals will conflict in the action they will call for. Here, the knowledge level is silent (although lower levels will not be). It just places constraints on actions, it does not uniquely prescribe them.

Implication. What does it mean to have knowledge that an action will lead to a goal? Newell had in mind the idea that the knowledge logically implies the goal, although he did not want to commit himself at this knowledge level to a particular symbol system to implement the logic.

Knowledge. An obvious definition for knowledge might be something like "whatever the person has encoded from experience," but Newell avoided this, perhaps because it seems impossible to know or even set bounds on what a person might encode from an experience or perhaps because he was writing about computers and not about people. Rather, he offered the following definition: "Whatever can be ascribed to an agent, such that its behavior can be computed according to the principle of rationality" (Newell, 1982, p. 105).

Many people suspect circularity upon reading these various assertions of Newell. Newell's basic position may be summarized thusly: "Knowledge and goals imply behavior." Basically, there are three terms related by the principle of rationality—goals, knowledge, and behavior. Given any two, one can infer or predict the third. Thus, if we know a person's goals and observe his behavior, we can infer his knowledge. Now that we have determined his knowledge, if we manipulate his goals, we should be able to predict his new behavior. Thus, we see that this implicational structure allows us to infer knowledge from one situation and make potentially disconfirmable predictions in a new situation. The appearance of circularity is only illusory.

The important feature of the knowledge level is that it allows an analysis of human behavior abstracted away from any assumptions about the symbols or processes in the human head. For instance, it does not matter whether the person knows *Spinoza was a human* and *Humans have color vision* and infers *Spinoza has color vision,* or whether the person has that fact directly represented. In either case, we can predict what the person will say to the question "Did Spinoza have color vision?" This is because we assume that we know the person's goal (to answer the question) and because, in either case, the principle of rationality predicts the same answer. The difference between these two knowledge representations is a nondistinction as far as the knowledge level is concerned. Note that the knowledge level only predicts the behavior, not how long it takes to calculate it.

Analysis at the knowledge level leads to considerable predictive force. Thus, we can predict that the thirsty person will drink the water offered to him quite independent of any psychological theory at the symbol level. Similarly, we can predict what answer my son will give to a subtraction problem. Indeed, what is amazing to me as a cognitive psychologist (perhaps not amazing to any other type of person) is just how much of human behavior can be predicted without recourse to any of the standard machinery of cognitive psychology. However, there are clear difficulties for the knowledge level. Their clarity is further testimony to the fact that the level has precision.

The example Newell used to show the problems at the knowledge level is the fact that the knowledge level analysis would imply that someone who knows the rules of chess would play a perfect game, because such a game logically follows from this knowledge. To this, Newell acknowledged that the knowledge level is a "radical approximation," and, at many points, predictions derived from it would be overriden by considerations from a lower level, such as the impossibility of searching the game tree for chess in finite human time.

The knowledge level has much in common with Marr's computational level. Both are concerned with the issue of how the goals of the system constrain the behavior of the system. Newell's essential insight is that in the case of cognition, in contrast to vision, a major constraint takes place through the knowledge the person has acquired. However, in contrast to Marr, we do not find Newell making any claims that understanding the knowledge level is a prerequisite to doing research on other agenda in the study of cognition. Indeed, most of Newell's efforts have focused on the symbolic level and the constraint of having this system match the universal computability of the human system.[2] He has taken both biological and rational considerations simply as further constraints on his development of the symbol system.

As mentioned earlier, Newell avoided defining knowledge in terms of experience. As becomes apparent throughout the book, I think a principle of rationality should be defined in terms of a person's experience and not knowledge. The implications of experience are uncertain and fundamentally probabilistic, whereas the implications of knowledge are certain. Because it is experience that people have direct access to and not knowledge, our analysis of what behavior is rational has to face up to the probabilistic structure of uncertainty. As is argued in this chapter, some of the claims about human irrationality make the fallacy of treating uncertain experience as certain knowledge.

CURRENT FORMULATION OF THE LEVELS ISSUE

It is possible to amalgamate the ideas in the literature into a summary formulation that starts with the three levels of Marr's formulation and incorporates Pylyshyn's division of the second level into a level of algorithm and functional architecture. Thus, there are four levels of analysis that I call the rational level, the algorithm level, the implementation level, and the biological level. The highest level is called the rational level both because it is defined in terms of a principle of rationality (not quite Newell's; my principle of rationality is introduced shortly hereafter). The next level is called the algorithm level. The third level is called the implementation level, rather than Pylyshyn's level of functional architecture, because of difficulties with the use of the word architecture.[3] The lowest level is the biological level. Pylyshyn's term *biological level* is preferable to Marr's *hardware level* or Newell's *device level,* because it makes clear that we are talking about what is in the brain and not what is in the computer.

Having settled on names for these levels, however, does not settle the issue of their psychological reality. First, I begin with a bold assertion: When all is said and done and we know the truth about what is happening in the human brain, there will turn out to be only two levels of analysis that are psychologically real (i.e., in the brain). They are the algorithm level and the biological level. Marr had it right, and, despite practice in cognitive psychology, there is no intermediate implementation level except as an approximation useful in calculation. That is to say, Rumelhart and McClelland have it just wrong—it is not the algorithm level that has the

[2] Indeed, Marr (1982) was quite critical of Newell for this emphasis.

[3] As discussed in Anderson (1987a), architecture is better used to refer to the interface between the algorithm level and the implementation level.

status of Newtonian mechanics; it is the implementation level. The implementation level is an approximation to the biological level. We need it because we do not begin to know enough about the brain to specify the biological level. Thus, we need the implementation level as a computational approximation and as a holding position. Without it, as Pylyshyn has argued, we would not have any interpretation of the costs of the operations at the algorithmic level and would not be able to predict the temporal or reliability properties of various mental algorithms. However, it should be recognized as an approximation. There is no convincing evidence or argument that there is any real level between the biological and the algorithm level.

Not only does the implementation level lack true psychological reality, it has identifiability problems. That is to say, it is not possible to decide between many claims about the implementation level, such as whether processes are going on in parallel or serial (Townsend, 1974), the format of the knowledge representation (Anderson, 1978), or whether there is a distinct short-term memory (Crowder, 1982a). Of course, if implementation level theories are just crude approximations, there will be real limitations on our ability to discriminate among theories, because one cannot perform exacting tests of an approximate theory. However, the identifiability problem goes beyond this approximation limitation: The relationship of the implementation level to behavioral data is too indirect to allow identifiability, even if it were not an approximation. I expand on this issue of identifiability in the next section.

The Algorithm Level[4]

Given my pessimism about the reality and tractability of the implementation level, it might seem remarkable that I am optimistic in both senses about the algorithm level. The fundamental reason for my optimism about the reality of the algorithm level is my belief in Newell and Simon's (1976) physical symbol hypothesis, which Newell (1980a) stated as "The necessary and sufficient condition for a physical system to exhibit general intelligent action is that it be a physical symbol system" (p. 170). A physical symbol system is a system that manipulates symbols. Symbols, as they are used in the physical symbol hypothesis, are tokens that in essence are pointers to knowledge stored elsewhere. For instance, a variable, as it is used in a computer program, is a symbol. Symbols are basically the structures out of which the algorithm level is defined. This physical symbol hypothesis is basically a conjunction of two observations: (a) The only way we know how

[4] For a more complete discussion of the research issues at the algorithm level, read Anderson (1987a).

to achieve intelligence is by use of symbols, and (b) we now know a (growing) number of ways in which symbols can be implemented and manipulated in physical systems. From (a) and (b) the argument is made that the only way physical systems (including humans) can achieve their intelligence is by use of symbols.

The argument for the physical symbol hypothesis could be strengthened if we could argue that symbols are the only way to achieve intelligence, rather than the only *known* way, as just argued. The best argument that I have read to this effect is one that was recently made by Newell (in press):

> It is a law of nature that processing in the physical world is always local, that is, always takes places in a limited region of physical space. This is equivalent to there being no action at a distance, or, somewhat more specifically, that causal effects propagate with an upper velocity of c, the speed of light in vacuo. Consequently, any computational system ultimately does its work by localized processing within localized regions in space. What guides or determines this processing task must then also be local. If the task is small enough or simple enough, then the processes could have been assembled within the local region and the task accomplished. Ultimately, there is no alternative to doing it this way. However, with complex enough processing, additional structure from outside the local region will be required at some point during the processing. If it is required, it must be obtained. If it must be obtained, then some process must do it, using structure within the local region to determine when and how to go outside.
>
> The symbol token is the device in the medium that determines where to go outside the local region to obtain more structure. The process has two phases: first, the opening of *access* to the distal structure that is needed; and second, the *retrieval* (transport) of that structure from its distal location to the local site, so it can actually affect the processing. When to go outside is determined by when the processing encounters the symbol token in a fashion that requires the missing structure. . . .
>
> Hidden in this account is the basic proposition behind information theory, namely, that for a given technology there is a limit to the amount of structure that can be obtained in a given region of physical space. In information theory this is expressed as the channel or memory capacity, and it is measured in bits. But its foundation is the amount of variety in physical structure, given that there is a limit to the amount of energy that is available to detect it. It applies to all systems, discrete and continuous (Shannon, 1949). Thus, as the demands for variety increase—corresponding ultimately to the demands for variety in functions to be composed, the local capacity will be exceeded and distal access will be required to variety elsewhere. Thus, there must be something within the local site that indicates what additional knowledge is needed.

To summarize: All processing must be done locally. Only so much knowledge can be stored locally. True intelligence can require using unbounded knowledge. Hence, we need the access and retrieval functions that are the essence of symbols. Newell's argument leaves open the question of how much of human cognition involves symbols, because much of human cognition does not involve unbounded use of knowledge.

Given that the algorithm level is tied to symbols, it is the level at which Pylyshyn's principle of cognitive penetrability applies. It is here that knowledge is brought to bear with full force and so can influence cognition. In addition to cognitive penetrability, my conception of the algorithm level is distinguished by the fact that steps of cognition at the algorithm level are correlated with observable behaviors. It is this issue of relationship to behavioral data that we turn to next.

Behavioral Data

Under the ACT* theory and many other theories, steps of cognition at the algorithm level correspond to points of discrete changes in working memory. In ACT*, these discrete changes are produced by a production firing that enters new information into working memory. In contrast, a step of cognition at the implementation level in ACT* corresponds to a change in activation pattern. In ACT*, these changes in activation pattern are continuous, and even when simulated discretely there will be 10–100 of these steps at the implementation level before there is a step at the algorithm level (i.e., a production firing).

It is important that different states at the algorithmic level are correlated with changes in working memory states. A change in the state of working memory can result in external behaviors. Thus, one can use the steps of behavior of a subject to infer the steps of cognition at the algorithm level. Of course, steps at the algorithm level can pass by without any behavior, but much of the methodology of cognitive science can be aimed at bringing out behavioral indicants. I have used the term *protocol* to refer to any rich sequence of behaviors elicited by the experimenter to try to trace changes in working memory. Verbal protocols are the most common protocol methodology and succeed in situations where states of working memory are verbally reportable. However, other protocol methodologies include use of streams of terminal interactions (keystrokes, mouse clicks) or eye movements, and, in many situations, such methodologies might be preferable. The essence of a protocol is that it provide a running series of responses that can be used to infer the sequence of mental states.

Protocols, at their best, offer the prospect of providing a state-by-state description of the transitions at the algorithmic level, and the scientist is simply left with the task of inducing the rules that determine these transitions. Protocols in real life are never so fine-grained as to report every state, nor are the reports sufficiently rich to discriminate between all possible pairs of states; however, they are a major advance over the situation at the implementation level. There are problems with the use of protocols and many more incorrect criticisms of their use (see Ericsson & Simon, 1984, for a thorough discussion of the issues). Many of the false criticisms of protocols stem from the belief that they are taken as sources of psychological theory rather than as sources of data about states of the mind. To serve the latter function, one need not require that the subject accurately interpret his mental states, but only that the theorist can specify some mapping between his behavior and states of the mind.

Note that the argument is not that the algorithm level has behavioral consequences and the implementation level does not. Rather, it is that the transitions at the algorithm level have direct behavioral indicants, whereas the transitions at the implementation level are only indirectly inferable from the behavioral indicants at the algorithm level. As an example of this point, consider the Sternberg task. In this task, subjects are shown a small study set of digits and then asked to decide whether a particular test digit is in that set. At the algorithm level, this can be modeled as a rather trivial task (e.g., see the ACT* model in Anderson, 1983): A single step is involved in deciding whether the test digit is in the memory set or not. A unique behavioral indicant of this is given—the subject says "yes" or "no." The Sternberg task has been of interest as a domain in which to study the implementation level. Many theories (Baddeley & Ecob, 1973; Glass, 1984; Sternberg, 1969, 1975; Theios, Smith, Haviland, Troupmann, & Moy, 1973) have been proposed in which detailed comparisons are being carried out between the test digit and the items in memory set either in serial or parallel. The behavioral data that has principally been used to decide among such theories is reaction time—the time for subjects to make this judgment. This does not give us a behavioral indicant for each step in the process but rather only a final datum. As a consequence of the poverty of the data relative to the complexity of the implementation level theories, it is generally regarded as impossible to decide such issues as whether the comparisons are being performed in serial or parallel.

This identification of the algorithm level with the behavioral function of the cognitive system is more abstract than the interpretation advocated by Anderson (1987a) or by Pylyshyn (1984), who basically identified it with a programming language. A programming language comes with a commitment to a particular syntax and potentially has implementation constraints. Rather, our use of algorithm is at a level of abstraction more like its typical use in computer science where it is an abstract specification of computation that can be realized in many computer languages. Issues of syntax are issues of architecture, namely the notation that interfaces the algorithm level and the implementation level.

The Rational Level

So far, I have discussed three levels of analysis: a biological level, which is real but almost inaccessible to cognitive theorizing, the approximate but essential implementation level, and the real and accessible algorithmic level. Is there a higher level where we should begin our inquiry, as Marr and Chomsky have advocated? As indicated in Table 1-1, I think there is a higher level, called the rational level, which is close in character to Marr's computational level. This book is mainly devoted to developing theory at the rational level, although it contains some speculations about how this relates to issues at other levels.

The rational level of analysis offers a different cut at human behavior. It is not an attempt to propose an information-processing analysis of mind at some level of aggregation from the molecular to the behavioral. It is not "psychologically real," in the sense that it does not assert that any specific computation is occurring in the human head. Rather, it is an attempt to do an analysis of the criteria that these computations must achieve to assure the rationality of the system. This turns out to be an important level at which to develop psychological theory, but a theory at this level is not directly about what the mechanisms of the mind are. Rather, it is about constraints on the behavior of the system in order for that behavior to be optimal. If we assume that cognition is optimized, these behavioral constraints are constraints on the mechanisms. This level of analysis is important, because it can tell us a lot about human behavior and the mechanisms of the mind. The function of this book is to demonstrate the usefulness of the rational level of analysis.

The idea that we might understand human behavior by assuming it is adapted to the environment is hardly new. It started with the functionalist school in the beginnings of American psychology (e.g., Dewey, 1910; James, 1892). More recently, it has been associated with psychologists such as Brunswik (1956), Campbell (1974), and Gibson (1966). We have already discussed at length the adaptionist basis of Marr's contribution to vision. Neisser's (1976; 1982) emphasis on understanding cognition in ecologically valid situations has adaptionist components to its motivation. Cosmides (1989) and Shepard (1987) represent recent efforts to develop analyses with very explicit evolutionary connections. Shepard's work is particularly close to the material in Chapter 3 and we will include there some specific discussion of his theory.

Thus, in proposing a rational analysis of human cognition I can hardly claim to have invented the wheel. However, there is something quite different about the wheel that is being described in this book. This is a result of both how rational analysis is related to the various levels of a cognitive theory and because of the particular research program attached to rational analysis. It is more in keeping with the spirit of the rational-man analysis in economics (from which it borrowed its name) than with most other applications in psychology (the spirit is similar, however, to Marr and Shepard). In particular, it leads us to formal models of the environment from which we derive behavior. Thus, its spirit is one which focuses us on what is outside the head rather than what is inside and one which demands mathematical precision. The next section of this chapter describes the new theoretical framework associated with rational analysis.

THE NEW THEORETICAL FRAMEWORK

Although the proposals in this book differ in some details from other proposals advanced in the ACT* book, the more substantial difference concerns the philosophy from which they are developed. This philosophy comes from merging a negative conclusion about the goals of cognitive science with a positive conclusion about its prospects. On the negative side, I have come to appreciate the profound lack of identifiability in the enterprise of cognitive science. On the positive side, I have come to realize the considerable guidance that rational considerations provide. I first consider the negative point and then the positive.

Lack of Identifiability at the Implementation Level

Cognitive psychology would be a rather unreal science if we worked only at the algorithm level. Our minds are not abstract algorithms left to compute away but have significant temporal and reliability properties. Thus, one has to consider implementation-level issues, but it is rather dissatisfying to pursue implementation-level theories in face of their identifiability problems. Because the rational level offers a different cut at cognition (rather than a higher level of abstraction), it allows one to pursue issues of the temporal and reliability properties of human cognition in a way that is free from problems of approximation and identifiability. It also allows one to view these properties as design features of the human mind, rather than as design flaws. Thus, the issue of identifiability proves to be a substantial part of the motivation for the rational analyses of this book. This subsection is devoted to making that point in more detail.

If we confine ourselves to behavioral data, then by definition all we can see are the steps of mind at the algorithm level. Even here we do not really see the steps of the mind, but rather their behavioral consequences. Any theory of cognition that confines itself to behavioral data has to be judged by how well it does at predicting the specific behaviors that occur in response to specific experiences. That is, we are limited to what goes into the system and what comes out—where "what comes out" includes things like response latency or intensity. A large fraction of cognitive psychologists—myself included—have taken as our goal to induce what is happening in the mind at the implementation level from this information. Recall that the implementation level is concerned with a model of the mental steps that take place between overt behaviors.

I have tip-toed around the feasibility of this enterprise for the last 10 years, because it would be unpopular to say it could not be done. However, the pretense can no longer be maintained, 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. A number of people have argued elaborate special cases of this non-identifiability (Anderson, 1978; Townsend, 1974). The general case can be argued so simply that it is hard to believe that the field has not accepted the conclusion long ago.

Basically, what we are trying to induce is the function that maps input to output. We choose to specify this function as a set of mechanisms, but this should not obscure the fact that these mechanisms compute an input-output function, and it is this function that we can empirically test. Said another way, if two different sets of mechanisms compute the same input-output function, there is no way to discriminate among them. Now, one of the simple things we know from work on formal machine theory is that there is an infinite number of mechanisms that compute the same input-output functions. That is, there is a many-to-one mapping from mechanisms to behavioral functions, and, consequently, identifying the behavioral function will not identify the mechanism. So, behavioral data will never tell us what is in the mind at the implementation level. It is time we stopped fooling ourselves.[5]

Responses to Lack of Identifiability. There are three standard responses to this dilemma. One is to appeal to parsimony and assume that the simplest set of mechanisms is correct. This might offer some hope of deciding within a circumscribed class of machines like Turing machines, but parsimony is meaningless when we compare different classes of computing mechanisms, such as PDP models and production systems. It is not possible to define an objective and acceptable metric to compare the parsimony of theories in such different formalisms. The inability of advocates of either class to make headway in arguing against the other should convince us of that. Second, as has been argued elsewhere (Anderson, 1983), it stretches credulity beyond any reasonable bounds to assume that nature chose the most parsimonious design for the mind. Thus, even if parsimony were capable of settling scientific disputes, it has no chance of telling us what is in the human head.

The second response is to argue optimistically that if we have enough data from enough phenomena of sufficient complexity, then the identifiability problem would go away. The supposed insight is that if we have enough behavioral constraints, there will be only one mechanism that satisfies them. However, identifiability problems do not go away with behavioral complexity. Again, this is easy to see in formal function and machine theory. There are lots of equivalent versions of complex Turing machines. All the complexity does is make it harder to choose. In general, identifiability problems are simpler when the behavior is simpler. There are a lot fewer programs that might be reasonably written[6] to write "hello" than to parse a sentence. Indeed, my reason for optimism about identifiability at the algorithm level is that at that level of abstraction we have to account for simple one-step transitions between reportable states rather than complex sequences of unobserved computations.

The third response to the identifiability problem is to appeal to physiological data to help tell us what is going on. The advantage of physiological data is that it offers the potential of providing a one-to-one tracing of the implementation level, just as protocols provide the potential for that kind of tracing of the algorithm level. Our knowledge of the mechanisms of early vision has developed because of physiological data. The right kind of physiological data to obtain is that which traces out the states of computation of the brain. Although there is still far to go, there has been considerable recent progress on this score (Dawson & Schell, 1982; Donchen, McCarthy, Kutas, & Ritter, 1983; Farah, 1988; Phelps & Massiotta, 1985; Posner, Peterson, Fox, & Raichle, 1988; Roland & Friberg, 1985). The wrong kind of physiological constraint is to make arguments based on things like speed of neural processing. As witness that this is the wrong kind of constraint, three very different theories (ACT*—Anderson, 1983; SOAR—Newell, in press; and PDP—McClelland & Rumelhart, 1986; Rumelhart & McClelland, 1986) have been proposed and defended in terms of neural timing.

Conclusion About the Identifiability Issue. The study of cognitive behavior is an interesting and worthwhile endeavor, despite the identifiability problem at the implementation level. We are making important progress in the absence of physiological data. For many purposes, such as application to education, such a physiological base would be excess

[5] In the Appendix, I deal with counter claims based on considerations of computational complexity and processing time.

[6] By "reasonable" I mean to exclude programs with useless steps. If we were to allow them, the identifiability problem would only be worse.

baggage. However, when we are inducing a scientific theory from behavioral data, we should not lose track of what we are doing. We are inducing an *abstract* function that maps input to output. We need a notation for codifying that function so we can communicate it to others, reason about it, and derive predictions. This is what our cognitive architectures and implementation theories provide us with—a system of notation for specifying the function. We should not ascribe any more ontological significance to the mechanisms of that architecture than we do to an integral sign in a calculus expression. If two theorists propose two sets of mechanisms in two architectures that compute the same function, then they are proposing the same theory. There are still bases for choosing among notations such as simplicity and tractability, but we are not choosing among scientific theories when we do so; we are choosing among notations for the same theory according to their convenience. To summarize, the argument is not that we should abandon developing implementation theory, but rather that their scientific claims should be read as the abstract behavioral functions they compute, not the specific mechanisms proposed. Part of the attraction of a rational approach is that it provides a way of specifying these functions without commitment to mechanism.

It should be clear how this position resembles behaviorism and how it differs. Behaviorism was correct in its usually unstated assumption that you cannot infer the mechanisms in the black box from what goes in and what comes out. It was incorrect in going from that to its claims that there should be restrictions on the notation with which theories are stated. The consequences of these restrictions was to produce theories that were incapable of computing the complex cognitive functions that people could. Said another way, the inadequacy of behaviorism was not its claim that a scientific theory was a mapping from stimulus to response but in the unnecessary restrictions it placed on the computational power of the mapping. All successful criticisms of behaviorist theories have focused on their computational power. The success of modern cognitive psychology stems from the computational power of our theories.

The Principle of Rationality

One of the consequences of our excessive concern with mechanism is that we often act as if God created the mind more or less arbitrarily, out of bits and pieces of cognitive mechanisms, and our induction task is to identify an arbitrary configuration of mechanisms. Of course, this is not the modern scientific understanding of human nature. The human is not a random construction but a construction that has been, to some degree, optimized to its environment by evolution. The behavior computed by our cognitive mechanisms must be optimized to some degree and in some sense. If we could only specify that degree and sense, we would be in a position to place enormous constraints on our proposals for cognitive mechanisms, both at the implementation level and the algorithm level. This is the strongest appeal of a rational approach.

Evolutionary Optimization. It is a hard issue to specify to what degree and in what sense we would expect to see human cognition optimized. I have tried to work through current ideas about evolutionary optimization (a very controversial area—see Dupre, 1987, for a representative set of readings). Here is my summary of the consensus (such as there is one) cast in terms familiar to a cognitive scientist rather than the terms from that literature: At any stable point in evolution, a species should display a range of variability in traits. The differences in this range are not sufficiently important in their adaptive value that any have been selected out. There may be some change in the species during this stable stage because of a phenomenon known as genetic drift, in which the distribution of nonsignificant variability changes. The optimization process might get called on if some truly novel genetic variation is created by some random mutation. However, it is generally thought that optimization is more generally called in when the environment undergoes some significant change after which the former range of traits is no longer equivalent in terms of adaptive value. This is the view that sees changes in the environment as more significant in driving evolutionary history than are random changes in genetic code.

The significance of this viewpoint is that it characterizes evolution as a local optimizer. I understand this in terms of a hill-climbing metaphor in which the set of possible traits defines the space, and the adaptive value defines height. 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 of a species occurs when it is not possible 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 optimum that any species achieves is a function of the constraints of its past. Maybe humans would be better adapted with the social structure of insects, but given our mammaliam origins, there is no path of hill climbing from where we are to this hypothetical global maximum.

Within the hill-climbing metaphor, there are two major constraints on prediction by optimization analyses. One is the proximity structure on the space of traits, and the other is where the species currently is in that space. Only certain variations are reachable from its current location. So consider the case of the moths of Manchester that serve as a standard illustration of evolutionary optimization. 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. There are other conceivable morphological responses to predators as effective—or more so—than changing color. For instance, one could imagine the development of offensive weapons such as possessed by other insects. However, moth mutants with offensive weapons do not occur, but color mutants do. Thus, color was a direction that was open for hill-climbing, but offensive weaponry was not.[7]This means that any species or aspect of a species is optimized, subject to some constraints that depend on evolutionary history and that can be pretty arbitrary and complex. The more arbitrary and complex these constraints, the less explanation there will be in appealing to optimization. The general lesson we can take from optimization explanations is that, in some cases, much explanatory power is achieved by appealing to optimization, and, in other cases, little explanatory power is achieved. Optimal foraging theory (e.g., Stephens & Krebs, 1986) is a field where we see a full range of explanatory outcomes from an optimization analysis. My book explores the question of how much explanatory power can be achieved in the case of human cognition. In particular, this book is an exploration of the following hypothesis.

General Principle of Rationality. **The cognitive system operates at all times to optimize the adaptation of the behavior of the organism.**

I have called this the principle of rationality, because it has a lot in common with the economist's position that people are rational agents and their economic behavior can simply be predicted on the assumption that they optimize their economic self-interests. This is a controversial position and one that it seems most people view as wrong, at least in detail (Hogarth & Reder, 1986). It is also generally viewed in psychology that people are anything but rational creatures and that their intellectual functions are shot through with intellectual fallacies. I try to reconcile the current position with this general wisdom in psychology in the last section of this chapter.

Part of the problem is with the term *rationality*. It has evolved two senses in social science. Perhaps the more obvious sense (which is close to Newell's sense) is that humans explicitly engage in logically correct reasoning in deciding what to do. Criticisms of human rationality are often arguments that humans do not do this. The second sense is that human behavior is optimal in terms of achieving human goals. This is the position in economics and the position advanced in this book. It explicitly disavows any claims about the mechanisms in the human head that achieve this optimization—they certainly do not have to involve logical calculation. (Only the theorists' predictions about human behavior require logical calculations.) It would have, perhaps, been less contentious and also perhaps clearer if I had chosen to call my principle the "principle of adaptation." However, I chose the terminology by analogy to economics before I appreciated its unfortunate consequences. Now I am stuck with it. At least I have tried to choose the right title for the book.

The principle should be taken as a scientific hypothesis to be judged by how well it does in organizing the data. One should not be surprised to find it doing well at explaining some aspects of cognition and not others. Obviously, I would not be writing this book if I did not believe I had achieved some success. My own sense is that cognition is likely to be one of the aspects of the human species that is most completely optimized and optimized in a clean, simple way so that it will yield to scientific analysis. This is because cognition seems one of the more malleable of human traits and, hence, more easily optimized and not as much subject to the constraints of evolutionary history. However, this is merely bias. The proof or disproof of the conjecture should not come from a priori considerations, but from how well the principle of rationality does in leading to successful theory.

[7]My 4-year-old son, who is enamored with "Teenage Mutant Ninja Turtles," has a different view about the plausibility of mutating offensive weaponry.

Applying the Principle of Rationality

How does one use the principle of rationality to develop a theory of cognition? Developing a theory in a rationality framework involves the following six steps:

1. Precisely specify what are the goals of the cognitive system.
2. Develop a formal model of the environment to which the system is adapted (almost certainly less structured than the standard experimental situation).
3. Make the minimal assumptions about computational limitations. This is where one specifies the constraints of evolutionary history. To the extent that these assumptions are minimal, the analysis is powerful.
4. Derive the optimal behavioral function given items 1 through 3.
5. Examine the empirical literature to see if the predictions of the behavioral function are confirmed.
6. If the predictions are off, iterate. In my own experience, my problems have been with the mathematical analyses required in step 4, which can often be quite complex.

The theory in a rational approach resides in the assumptions in items 1 through 3, from which the predictions flow. I refer to these assumptions as the *framing of the information-processing problem.* Note that this is a nearly mechanism-free casting of a psychological theory. Ideally, most of the interesting assumptions in this theory come in step 2, because the structure of the environment is what is easiest to verify. One can, in principle, look and see if these assumptions are objectively true of the world. To the extent that assumptions in step 3 play a significant role, this ideal is only approximated. The reader will find in subsequent chapters that the computational assumptions are indeed weak, involving claims that almost all information-processing theories would agree on (such as a short-term memory limitation of some sort or that it takes time to process an alternative.)

It is worth commenting on the fact that this process of theory building is iterative. If one framing does not work, we have to be prepared to try another. Such iterative behavior has often been seen as a sign that an adaptionist enterprise is fatally flawed (Gould & Lewontin, 1979). However, as Mayr (1983) noted in response to Gould & Lewontin, iterative theory construction is the way of all science. Certainly, in cognitive science, we have seen a long iteration of mechanisms to explain cognition. Hopefully, we understand in cognitive science that a theory is to be evaluated by how well it does in organizing the data and not by whether it is the *n*th theory that has been tried. Having acknowledged this, I must note that my own experience with theory construction in the rationalist framework is 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.

Advantages of Rational Theory

In summary, let me list the advantages of the rational approach in order of increasing importance:

1. It offers a way to avoid the indentifiability problems of the mechanistic approach. One has a theory that depends on the structure of an observable environment and not on the unobservable structure in the head.
2. It offers an explanation for why the mechanisms compute the way they do. We do not have to view the human mind as a random set of postulates let loose on the world.
3. It offers real guidance to theory construction. If the mind is not a random set of mechanisms, but is structured to optimize its adaptation, one can use the hypothesis of optimization to guide the search for a scientific theory. Otherwise, one has to rely on very weak methods to search a very large space of psychological hypotheses.

As Marr stressed, a special case of the third point is the role that a rational analysis can have in guiding the design of mechanistic theories at the algorithm and implementation levels. I hope to be able to follow up the analysis in this book with a new theory within the ACT framework to update the ACT* theory (see Anderson, 1983, for a discussion of the distinctions between frameworks and theories). Such a new theory will take advantage of the guidance of a rational theory.

Although a rational explanation is more satisfying than a mechanistic explanation in terms of point number 2 in the preceding list, there is another sense in which mechanistic explanations are more satisfying: For whatever reason, we enjoy having an image of what might be happening in the head. Thus, even if a rational theory could predict all the relevant data, we would want to pursue a mechanistic theory. Rational and mechanistic approaches need not be in conflict. We can take Marr's view that the rational analysis provides the needed guidance for the mechanistic approach. We can emerge from our scientific endeavor with both an answer to what is happening (modulo identifiability limitations) and why it is happening (modulo the relativism of adaptation because of evolutionary history and biological constraint).

IS HUMAN COGNITION RATIONAL?

As indicated earlier, it is common wisdom in psychology that humans are irrational, and this seems to go to the heart of the proposal in this book. Many a person has, in effect, said to me, "Your analyses are interesting, but they must be wrong, because human thought has been shown not to be rational." Most of these demonstrations of human irrationality come from the fields of decision making and social judgment and are not from the more basic cognitive domains that are the focus of this book. The basic resolution of this apparent contradiction between the results of these other fields and the current book is that rationality is being used in two senses and that rationality in the adaptive sense, which is used here, is not rationality in the normative sense that is used in studies of decision making and social judgment. For an extensive discussion of these two views of human rationality and their relationship to evolution, see Stich (in press).

It is possible that humans are rational in the adaptive sense in the domains of cognition studied here but not in decision making and social judgment. However, in this section, I argue that many of the purported demonstrations of human irrationality are demonstrations in the normative sense and not the adaptive sense. I enumerate in the following subsections some of the ways in which criteria of normative rationality can deviate from criteria of adaptive rationality.

Computational Cost

One problem with normative definitions of rationality is that they ignore computational cost. This is nicely illustrated in the application of Newell's principle of rationality to chess. There, the observation was that knowledge of the rules of chess plus Newell's principle of rationality implied playing a perfect game of chess. As Newell noted, this ignores the astronomical cost of searching the entire game-tree of moves. In an adaptive analysis, one has to place the cost of performing the computation into the equation to be optimized (step 3 in our prescription for developing a theory on a rational framework). This makes the principle of rationality developed here more like Simon's (1972) theory of bounded rationality, although Simon has insisted that there is a difference (see Simon, in press, and the discussion in Chapter 6).

This observation is one of the potential Achilles' heels of a rational approach. If we have to know computational cost to know what is rational, we may have to specify the mechanisms of cognition in advance of rational analysis. This is just what we are trying to avoid. However, I hope to be able to illustrate that we can progress with very weak assumptions about computational cost. This is certainly the case in the analysis of the chess dilemma.

Is There an Adaptive Cost?

A question that is rarely asked is whether there is really a cost associated with the purported irrationality. If a person prefers A to B, B to C, and C to A, but there are no differences among A, B, and C in their adaptive value, then the intransitivity does not violate the adaptive principle of rationality. It is important to stress that adaptation in the genetic sense is measured in number of surviving offspring (which is what controls evolutionary selection) and not money, power, or happiness. Thus, the gambler's fallacy may lead someone to lose money in Las Vegas, but if it leads him or her to try for a third child after two boys (because a girl is due), then it is quite adaptive.

Nisbett and Ross (1980), after documenting the abundance of experimental evidence for a number of intellectual fallacies, noted that some of them may have little cost. A good example is the primacy effect, where people give too much weight to initial evidence and discount later evidence. Suppose, for example, that primitive man is trying to decide which of two fishing spots yields a better chance of catching a fish. Suppose he samples one and succeeds. The primacy effect means that he is going to tend to discount later evidence about the efficiency of the two fishing holes and continue with the first. Conversely, if his first experience is bad, he will tend to avoid that fishing hole, irrespective of later experiences. Such behavior is not rational in a normative sense.

The interesting question is "How irrational is it in an adaptive sense?". On careful analysis, Nisbett and Ross concluded that it is not very costly at all. If the fishing holes yield a very similar probability of a fish, then it would not matter which hole was chosen. If one hole had a near-one probability of yielding a fish and the other had a near-zero probability, primitive man would choose the right one, despite the primacy effect, because it is very unlikely that his first experience would be misleading, and even if it was it would be quickly overwhelmed by subsequent experience. (The primacy effect is not so strong that we totally ignore all subsequent experience).[8]

Heuristics Need to Be Evaluated in Terms of Expected Value

People act according to principles that cannot be guaranteed to be correct and can fail in specific cases. Such principles are called heuristics, and there is no reason why normatively irrational heuristics cannot be adaptive. For instance, many people are likely not to believe an argument if they perceive that the arguer does not believe it. According to normative models, the validity of an argument is a function of the argument and not the beliefs of the arguer. However, it is an open question whether, given the fallibility of validity judgment, people are more likely to come to erroneous beliefs behaving in accord with this heuristic.

[8] To explore this more systematically, one must make some assumptions about the distribution of successful fishing holes. Suppose that there is a uniform distribution from zero to one of successful fishing holes in terms of probability of catching a fish on a given day. That is to say, the chances that a new fishing hole will yield a probability p of catch each day is the same for all p. This means that if our primitive man chose randomly which hole to fish at, his expected probability of catching a fish any day would be .50. On the other hand, if he were omniscient and knew which of the two holes was best, he could expect to catch a fish 2/3 of the time. Said another way, the omniscient primitive man would catch 1/3 more fish than the random primitive man, giving him a considerable survival advantage. Of course, primitive man could not be omniscient. But let us suppose he was rational, took a modest sample, and went with the evidence of that sample. Suppose he tried the first hole three times and the second hole twice, and went with whichever hole yielded the most fish, choosing the second hole if there was a tie. This would yield him an expected .625 chance of a fish per day, or 94% of the omniscient maximum. However, primitive man is not rational. Suppose he showed such a strong primacy effect that he would choose the second hole after a successful first catch only if he failed on his other two samples of the first hole and succeeded with his two samples of the second. Similarly, he would only choose the first hole after a failure to get a first catch if his next two tries at the first hole were successful and his two tries at the second hole both failed. It turns out that his expected catch per day would be .603, or 96% of the rational man's catch. Presumably, this does not convey much of a survival disadvantage.

A case that Nisbett and Ross discussed in detail is the fallacy in human judgment that causes should resemble their effects. J.S. Mill (1843/1974) wrote, "The most deeply-rooted fallacy . . . is that the conditions of a phenomenon must, or at least probably will, resemble the phenomenon itself" (p. 765). Nisbett and Ross (1980) documented some of the mispractices of medicine that derive from this. For instance, in medieval times the lungs of foxes were prescribed as a cure for asthma, because the animal was regarded as remarkable for its strong power of respiration. People ridiculed the hypothesis that yellow fever might be caused by mosquitoes. Much human suffering has been created or prolonged by the insistence that causes must resemble their effects.

However, use of similarity is, on the whole, rational—as is expanded on in chapter 4. We can reject many spurious correlations as noncausal because of total lack of similarity between purported cause and effect. We do not want to believe that roosters' crowing causes the sun to rise, that lying in bed causes one to vomit, or that being homosexual causes one to have AIDS. As Nisbett and Ross conceded, the similarity heuristic has probably guided medical discovery such as vaccination, the use of cold compacts to treat burns, and the relationship of smoking to lung cancer. The use of similarity is a heuristic, and any heuristic can sometimes lead one far astray. The only claim is that one will do better on average if one follows it than if one ignores it.

Applications of Normative Models Often Treat Uncertain Information as Certain

If I were to take issue with the validity of any demonstrations of human irrationality in the normative sense, it would be with certain applications of normative models in defining rational behavior. Normative prescriptions take the form of "If situation *X* holds, then action *Y* is prescribed." Many applications overestimate the certainty of knowing whether situation *X* holds. For instance, consider a recent set of medical decisions I had to make. I was told by an internist, and then by a surgeon, that I had an indirect hernia (it had no symptoms that I could detect), that there was a 5% chance it would become strangulated, and that strangulated hernias are fatal 30% of the time and result in serious complications 50% or more of the time. I was told that a hernia operation was nearly totally free of danger and complication. These facts were basically confirmed as common medical knowledge by a number of nonsurgeon physician friends who had no interest in the surgeon's fees and presumably some interest in my welfare. Simple mathematics showed that surgery reduced my chance of premature death by 1.5% (actually a bit more complicated) and a serious medical complication by 2.5% at little cost. These are small probabilities, but if we keep encountering such risks we are playing Russian roulette, according to decision theorists (Dawes, 1987), and assuring our early demise. I have to admit, I was ill at ease with this analysis, but I forced myself to behave rationally. Later, I learned that what was diagnosed as a hernia was not a hernia but rather what hernia experts call a "weakness"; that if it ever developed into a hernia it would have become a direct hernia; that strangulation is rare in the case of direct hernias; that there is a high probability (5-15%) that hernias will reoccur after an operation; that the probability is even higher for someone operated on for a weakness; and that there are substantially higher probabilities of strangulation with reoccurring hernias. I only found this out upon further research when my hernia operation did fail and I had a real symptomatic hernia. Thus, what had happened was I had read too many articles on rational decision making, treated the premises provided to me by physicians as certain, and proceeded to act in the rational manner.

The case in the literature that best illustrates this overemphasis on the certainty of the premises is the famous Kahneman and Tversky (1973) demonstrations concerning humans' failure to take into account base rates. For instance, they ask subjects to read descriptions of individuals and judge whether these individuals are engineers or lawyers. They also tell subjects information about the base rates of engineers and lawyers in the population (for instance, 70% of the population are lawyers). Subjects completely ignore this base rate information and make their judgment on how well the description matches a prototypical engineer or lawyer. However, the prescriptively normative Bayes Theorem says base rates should have strong effects.

Let us assume that subjects understood what they were told and took as their task the official task. Why should they believe the abstract information about base rates? As my medical experience testifies, information about base rates is typically unreliable. The instances of inaccurate base rate information abound. There were the famous polls that guaranteed Dewey's election. As another instance, a few years ago we were told that 1 in 100 people who tested positive for the AIDS virus would develop full-blown AIDS. Now, that estimate is up by more than a factor of 10. As a more humorous example, our local Pittsburgh magazine does a poll of local residents to get information about restaurants. Every year at least a plurality of people claim that Pizza Hut makes the best pizza. Should I really ignore personal experience and testimony of others to the contrary and accept this abstract base rate information gathered by a "reliable" source?

Believers in the need to use base rates are a die-hard crew, and whatever example I bring up they always say, "But, of course, you should have known that the information in that case was invalid. What you should pay attention to is *valid* information about base rates." Ignoring the problematical issue of whether there was any valid abstract information about base rates in our evolutionary history (from the village priests?), let me ask where is such valid information today, and how do we know it is valid?

I have asked bright colleagues where they think good information about base rates might be found. The most common answer is information like in *Consumer Reports* on things like repair statistics. Then I asked them how they know this is valid information in contrast to the AIDS information or the Dewey poll. It may well be that *Consumer Reports* does provide valid information, but none of my colleagues are in possession of reasons for believing so beyond an interesting "Well, if their information was bad, we would have heard about it," or "It proved reliable for me when I bought such-and-such a car."

Another comment I receive from colleagues who are less die-hard believers in abstract base rates goes something like this: "Well, all right, such base rate information is often invalid, but surely you cannot be arguing that one would be better off, on the average, to ignore such information." I am not arguing this, but, on the other hand, I can see no basis to argue that one would be better off to pay attention to abstract information about base rates. It is very much an open question that requires further analysis of how often base rate information is misleading and the costs and benefits of using frequently flawed information.

The only time it is clear that we should heed base rate information is for domains where we have personal proof that the information is valid. Thus, if one has personal experience that the reports from *Consumer Reports* have proven valid, then one should be influenced by them. On the other hand, when some high authority (medical, religious, political, or academic) has a pronouncement to make about something for which we have no personal experience or contradictory personal experience, we should be very suspicious. I wish I had been.

As a final observation, there is evidence that when our experience with base rates is concrete and not abstract (seeing is believing), and our behavior involves responding to the object in question, not engaging in a verbal exercise, people are extremely sensitive to base rates. A good example of this is the accuracy with which people probability match (see Kintsch, 1970b, for a review) when they are trying to predict an event in a random sequence. Interestingly, this probability matching has often been described as nonrational. If the probability is higher of a 1 in a random sequence of 0's and 1's, the subject should always predict 1 to maximize correct prediction. Upon close inspection, it turned out that subjects were not accepting the authoritative reports of experiments that these were random sequences, and they were searching out sequential patterns. Subsequent chapters in this book describe other instances in which people are very sensitive to concrete base rates, although, again, we often mischaracterize such sensitivity as irrational.

The Situation Assumed Is not the Environment of Adaptation

Adaptation is defined with respect to a specific environment. Often, the normative model assumes a different situation. Thus, human memory is often criticized because it cannot easily perform simple tasks, like storing a list of 20 words. However, human memory did not evolve to manage a list of 20 words but rather to manage a data base of millions of facts and experiences. In chapter 2, when we view memory in light of this situation, its behavior in a memory experiment appears quite adaptive.

Another example concerns the constant demonstration of human fallacies in experiments on deductive reasoning (Anderson, 1985, chapter 10). Deductive reasoning enables one to go from certain premises to certain conclusions. However, as discussed earlier, certain premises are rare or nonexistent in situations of adaptive importance. As a consequence, there is no reason why humans should have evolved to engage in correct deductive reasoning? Cosmides (1989) argues that we can understand the pattern of success and failure in reasoning about the Wason card-sorting task according to what situations are adaptively important.

An extension of this line of argument can be used to explain the apparent irrationality of modern life. Although it is important to avoid exaggerated doom-saying, we probably all agree that current human behavior is harming the prospects for human survival by creating huge nuclear arsenals and environmental disasters. A possible explanation is that human tendencies, adaptive in other earlier environments, are playing themselves out disastrously in the current modern technological age. One must be cautious of such rational explanations that make reference to past environments, rather than the current, because it is always possible to invent environments in which any behavior would be adaptive. This is not to say that one cannot make explanations by appealing to the past; however, they require independent evidence about what the past environments were really like.

Conclusions

Research comparing human behavior to normative models has been extremely useful; however, one must be careful in understanding its implications for adaptive rationality. It may well be that certain aspects of human cognition cannot be understood profitably in the framework I am advocating. However, I think we have grossly overestimated the irrationality of human cognition in this sense. Moreover, as I have begun to discover, there are aspects of human cognition that can be very profitably understood within the rationality thesis.

THE REST OF THIS BOOK

The next four chapters of this book are devoted to extensive analyses of a number of aspects of human cognition from the perspective of this rational framework. The fruitfulness of these analyses is the real evidence for the principle of rationality. Throughout the book, I briefly speculate on what the implications of these analyses might be for the ACT architecture, which is concerned with the algorithm and implementation levels. However, I have left working out these implications for another day. The fact that rational analyses can stand on their own is evidence that this is a level of theoretical analysis that can be pursued independently. I have not worked out the detailed architectural implementation, because it would be premature until I have fully worked out these rational derivations.

The book ends with a short chapter of general discussion. I wrote the chapter with some reluctance in response to those who felt the need to have general questions addressed after four chapters of detail. It should not be read without first reading the four contentful chapters.

APPENDIX: NON-IDENTIFIABILITY AND RESPONSE TIME

In the main body of this chapter, the claim was made that behavioral data will not allow us to determine the underlying mechanisms at the implementation level. The reason for this is that there are many different sets of mechanisms that can mimic one another. This is a well-established fact in formal automata theory. However, in the conventional understanding of formal automata theory, the fact that two systems display the same input–output behavior does not guarantee that they will show the same timing behavior. That is, although the two systems may produce the same output, they may take different times to compute it. Indeed, a lot of work in automata theory is concerned with studying machines that compute the same behavior but with different temporal functions.

Suppose we had two implementation theories that agreed in the input–output behavior of the system. Both would claim that the mind went through some set of mental steps (largely unseen) to perform some task. They would be different because they claimed that the mind went through different sets of steps to achieve the same end state. The claim that might be

advanced is that, even though the end states were the same, the time would be different because the two systems performed different internal steps.

Suppose system 1 takes n_1 steps to perform a task and system 2 takes n_2 steps, $n_1 < n_2$. It might seem simple to get timing mimicry despite the different number of steps. Let us just have system 2 take a fraction n_1/n_2 of the time to make its steps. This is the basic speed-up argument.

However, there is a well-known objection to such a speed-up proposal in automata theory. This objection involves the concept of computational complexity. System 1 and system 2 may take differing number of steps to perform their tasks, depending on the complexity of the problem. A simple example is that time to parse a sentence should vary with its length. Now if system 1 displays some function f_1 of complexity, and system 2 displays another function f_2 of complexity, it may not be the case for any constant speed-up factor a, $af_2(n) \leq f_1(n)$ for all n where n measures complexity (e.g., length of sentence). For instance, let f_2 be a squaring function and f_1 a linear function. For no $a > 0$ is it the case that $an^2 < n$ for all n. Basically, if n gets large enough, the system with the worse complexity function will start to lag behind the system with the better complexity function.

However, the whole problem with such arguments is that they depend on unbounded complexity, and people never deal with problems of unbounded complexity. As long as there is a bound on complexity, the argument vanishes. There are real and very sharp limitations on the complexity of human behavior. For instance, we can parse longish sentences only if they are basically linear concatenations of small phrases that we can parse separately. As another example, we can process in detail only a small part of a visual array at a time (that around the fovea). Again, we process a large array by a sequence of glances.

This complexity-bound argument is particularly forceful when one realizes that working-memory limitations and chunk-size limitations place very severe limitations on the number of elements that can be processed and, hence, on complexity functions. Any psychologically accurate model is going to have to involve linear concatenations of the processing of these limited-size chunks. With such severe limitations on complexity, mimicry of processing times would be particularly easy to achieve.

This observation has been discussed in Anderson (1979) and in Berwick & Weinberg (1984). Thus, the point of this argument is to say that, although the formal results are real problems for simulation of formal machines solving formal problems, they are not real problems for simulation of humans solving human problems.

I have been asked what happens to such speed-up arguments if one has physiological evidence that the operations of the brain can be performed only so fast. In principle, such considerations can serve to eliminate certain speed-up proposals. However, in practice, there are plenty of theories that do not push the brain beyond its limit. Typically, neurophysiological timing arguments, when they are invoked, are quite questionable. For example, J. A. Anderson (1973) has argued that the brain cannot do serial searches with 35 msec. comparison times per item, as was proposed by Sternberg (1969). Although I am inclined to believe that memory sets in the Sternberg task are not processed serially, there are no strong reasons for proclaiming 35 msec serial processing impossible. Perhaps certain schemes for processing an item could not be implemented in 35 msec, but Sternberg never (to my knowledge) made a commitment to a particular neural implementation.

2 Memory

Contents

PRELIMINARIES

I begin my detailed application of rational analysis with human memory for a number of reasons. Human memory is the field of cognitive psychology with which I have had the longest association, going back over 20 years. It is also the area of cognition where I got my first glimmers of how a rational