The Place of Cognitive Architectures in a Rational Analysis

John R. Anderson

Department of Psychology

Carnegie-Mellon University

This paper contains a summary of the main points that I will be making in my presentation at the Cognitive Science Meetings. Some more details will be in that presentation. All the details and formal derivations will be found in Anderson (in press).

This paper will consider the Soar architecture of Laird, Newell, and Rosenbloom (in press), my own ACT* architecture (Anderson, 1983), and the PDP architecture of McClelland and Rumelhart (Rumelhart & McClelland, 1986, McClelland & Rumelhart, 1986). Now that there are numerous candidates for cognitive architectures, one is naturally led to ask which might be the correct one or the most correct one. This is a particularly difficult question to answer because these architectures are often quite removed from the empirical phenomena which they are supposed to account for. In actual practice one sees proponents of a particular architecture arguing for that architecture by reference to what I call signature phenomena. These are empirical phenomena which are particularly clear manifestations of the purported underlying mechanisms. The claim is made that the architecture provides particularly natural accounts for these phenomena and that these phenomena are hard to account for in other architectures. In this paper I will argue that the purported signature phenomena tell us very little about what is inside the human head. Rather they tell us a lot about the world in which the human lives. The majority of this paper will be devoted to making this point with respect to examples from the SOAR, ACT*, and PDP architecture.

As a theorist who has been associated with the development of cognitive architectures for 15 years I should say a little about how I came to be advocating this position. I have been strongly influenced by David Marr's (1982) metatheoretical arguments in his book on vision which are nicely summarized in the following quote:

> An algorithm is likely to be understood more readily by understanding the nature of the problem being solved than by examing the mechanism (and the hardware) in which it is solved.

Marr made this point with respect to phenomena such as stereopsis where he argued that one will come to an understanding of the phenomena by focusing on the problem of how two two-dimensional views of the world contained enough information to enable one to extract a three-dimensional interpretation of the world and not by focusing on the mechanisms of stereopsis. He thought his viewpoint was appropriate to higher-level cognition although he did not develop it for that application. As recent as a few years ago I could not see how his viewpoint applied to higher level cognition (Anderson, 1987). However, in

the last couple of years I have come to see how it would apply and have realized its advantages. The basic point of Marr's was that if there is an optimal way to use the information at hand the system will use it. I have stated this as the following principle:

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

One can regard this principle as being handed to us from outside of psychology--as a consequence of basic evolutionary principles. However, I do not want to endorse this viewpoint on that principle because there are many cases where evolution does not optimize. Rather, I view it as an empirical hypothesis to be judged by how well theories that embody the principle of rationality do in predicting various cognitive phenomena. Developing a theory in a rational framework involves the following 6 steps:

1. Precisely specify what the goals of the cognitive system are.

2. Develop a formal model of the environment that the system is adapted to (almost certainly less structured than the experimental situation).

3. Make the minimal assumptions about computational costs.

4. Derive the optimal behavioral function given (1)-(3).

5. Examine the empirical literatures to see if the predictions of the behavioral function are confirmed.

6. If predictions are off, iterate.

The theory in a rational approach resides in the assumptions in (1) - (3) from which the predictions flow. I refer to these assumptions as the framing of the information processing problem. Note this is a mechanism-free casting of a psychological theory. It can be largely cast in terms of what is outside of the human head rather than inside. As such it enjoys another advantage which is that its assumptions are potentially capable of independent verification.

SOAR--Power Law Learning

The signature phenomenon I would like to consider for the SOAR theory is power-law learning which is referenced in many of the SOAR publications. This refers to the linear relationship that is obtained between the logarithm of the amount of practice and the logarithm of response time which implies that the performance measure is a power function of practice. In the Soar model the power law falls out of the chunking learning mechanism plus some critical auxiliary assumptions. Chunking refers to the collapsing of multiple production firings into a single production firing that does the work of the set. It is assumed that each chunk produces a performance enhancement proportional to the number

of productions eliminated. Chunks are formed at a constant rate--either on every opportunity or with equal probability on every opportunity. The final critical assumption is that as chunks span larger and larger units the number of potential chunks grows exponentially. As a consequence of the last assumption, learning will progress ever more slowly because it takes more experience to encounter all of the larger chunks.

I will offer a rational analysis of power law learning which will also explain the forgetting and massing functions. This will be part of a larger rational analysis of human memory which is the topic of the next section.

A Rational Analysis of Human Memory

The claim that human memory is rationally designed might strike one at least as implausible as the general claim for the rationality of human cognition. Human memory is always disparaged in comparison to computer memory--it is thought of as slow both in storage and retrieval and terribly unreliable. However, such analyses of human memory fail both to understand the task faced by human memory and the goals of memory. I think human memory should be compared with information-retrieval systems such as the ones that exist in computer science. According to Salton and McGill (1983) a generic information retrieval system consists of four things:

(1) There is a data base of files such as book entries in a library system. In the human case these files are the various memories of things past.

(2) The files are indexed by terms. In a library system the indexing terms might be keywords in the book's abstract. In the human case the terms are presumably the concepts and elements united in the memory. Thus, if the memory is seeing Willie Stargell hit a home run the indexing terms might be Willie Stargell, home run, Three Rivers Stadium, etc.

(3) An information retrieval system is posed queries consisting of terms. In a library system these are suggested keywords by the user. In the case of the human situation it is whatever cues are presented by the environment such as when someone says to me "Think of a home run at Three Rivers Stadium".

(4) Finally there are a set of target files desired by which we can judge the success of the information retrieval.

One thing that is very clear in the literature on information retrieval systems is that they cannot know the right files to retrieve given a query. This is because the information in the queries does not completely determine what file is wanted. The best information retrieval systems can do is assign probabilities to various files given the query. Let us denote the probability that a particular file is a target by P[A]. In deciding what to do informational retrieval systems have to balance two costs. One is what Salton an McGill call the precision cost and which I will denote C_P. This is the cost associated with retrieving a file which is not a target. There must be a corresponding cost in the human system. This is the one place where we will see a computational cost appearing in our rational analysis of

memory. The other cost Salton and McGill call the recall cost and we will denote it C_R. It is the cost associated with failing to retrieve a target. Presumably in most cases it is much larger than the precision cost for a single file or memory.

Given this framing of the information processing problem we can now proceed to specify the optimal information-processing behavior. This is to consider memories (or files) in order of descending P[A] and stop when the expected cost associated with failing to consider the next item is greater than the cost associated with considering it or when

$$P[A]\, C_R < (1\text{-}P[A])\, C_P \tag{1}$$

We now have a complete theory of human memory except for one major issue--how should the system go about estimating P[A]. I propose that the system should use the item's past history of usage and the elements in the current context to come up with a Bayesian estimate of that probability. A particularly transparent way of stating this is with the Bayesian odds ration formula which we can state

$$\frac{P(A|H_A\&Q)}{P(\overline{A}|H_A\&Q)} = \frac{P(A|H_A)}{P(\overline{A}|H_A)} \cdot \prod_{i \in Q} \frac{P(i|A)}{P(i|\overline{A})} \tag{2}$$

where $P(A|H_A\&Q)$ is the posterior probability that the memory is needed given its past history and the cues in the current context, $P(\overline{A}|H_A\&Q)$ is $1\text{-}P(A|H_A\&Q)$, $P(A|H_A)$ is the posterior probability given just the history, $P(\overline{A}|H_A) = 1\text{-}P(A|H_A)$, $P(i|A)$ is the conditional probability that i would be in the current context if A is needed, and $P(i|\overline{A})$ is the conditional probability if A is not needed. This way of formulating the relationship nicely breaks up the need probability into the product of a history factor $P(A|H_A)/P(\overline{A}|H_A)$ plus a context factor the product involving the $P(i|A)/P(i/\overline{A})$. Note that in this context factor we are assuming the individual cues are independent of one another in order to obtain a product. I neither want to argue that this is really true nor that the human system actually acts as if it is. I am only using this as an approximation to get an indication of what the rational predictions are.

The History Factor

In investigating the implications of this rational analysis for the power-law learning function we need to focus on the history factor in the above equation. In particular we need to specify $P(A|H_A)$. To determine this we need to know how the past history of usage of a memory trace predicts whether it will be currently used. To determine this in a truly valid objective way we would have to follow people around, determine when they use particular facts, and induce what the empirical relationship is. It is nearly impossible to imagine collecting such objective statistics in the human case but such statististics are available for other information retrieval systems. For instance, there is data about how past borrowings from a library predict future borrowings(Burrell, 1980; Burrell & Cane, 1982). There is data

about how past accesses to a file predict future accesses(Stritter, 1977). The data for these different domains is quite similar in terms of the nature of the functional relationship between past use and current use. I propose that these relationships are true of all information retrieval systems including the human one.

Burrell developed a model for library borrowings which provides a good analytical starting point. There are three basic assumptions in Burrell's model. The first is that the items in a system vary in their desirability. Burrell assumes that the distribution of desirability is a gamma distribution with parameter b and index v. He is able to basically show such a distribution of borrowings in the case of a library system. The second assumption that Burrell makes is that there is an aging process such that items will decay in their borrowing rate with the passage of time. Again he can empirically validate that such an aging process does occur. This means that if we take an item from the gamma distribution with initial desirability λ its desirability after time t will be $\lambda r(t)$ where r(t) describes the rate of decay. Burrell uses a simple exponential decay in rate of the form. The third assumption of Burrell is that borrowings are a Poisson process and that times until next borrowing are exponentially distributed with rate $\lambda r(t)$.

With these assumptions we can derive what I call the recency-frequency function RF(n,t) which is the probability that an item introduced t time units ago and used n times over that period will be needed in the current time unit. It produces a linear relationship between number of uses, n, and need probability. This is a special case of a power function. When we consider plausible monotonic transformations from need probability to latency the linear relationship disappears but the power function relationship remains. Because of the aging factor r(t) we wind up predicting the forgetting function quite accurately as well.

Thus, we have shown that power law learning can be predicted from a rational perspective which sees human memory as adapting to the statistics of information use. Thus, it is what is outside the human head not what is inside that is controlling the memory performance. I should emphasize that this does not deny that chunking may be one of the mechanisms the mind uses to achieve this adaption. However, the argument is that the real explanation is in the outside world and not in the internal mechanisms.

ACT*--The Fan Effect

Now I would like to turn to the second architecture, ACT*, and consider a signature phenomenon which has played a key role in its development. This is the fan effect (Anderson,1983) . A typical experiment is focused on subjects' ability to recognize sentences that they have learned. According to ACT*, upon being presented with a sentence such as "The lawyer is in the park" the subject activates the concepts in the sentence such as lawyer, in, and park. Activation spreads from these concepts along various network paths. The time to recognize a sentence is a function of the amount of activation reaching the proposition node. The critical additional assumption in the ACT* theory is that the amount of activation that can spread out of a node is fixed and that the more paths emanating out of a concept the less activation can go to any one proposition and so the slower recognition will be. Fan refers to the number of such paths and is manipulated by manipulating the number of facts studied about a concept like lawyer.

We can extend our previous rational analysis of the fan fan effect to accomodate the fan effect. Here we will be interested in analyzing the context factor rather than the history factor since we are manipulating properties of the memory cues that we presented to subjects. That is we want to focus on the quantities $P(i|A)/P(i|\bar{A})$ where the i are concepts like lawyer and the A are the sentences to be recalled. We can rewrite these as

$$\frac{P(i|A)}{P(i|\bar{A})} = \frac{P(A|i)P(i)/P(A)}{P(\bar{A}|i)P(i)/P(\bar{A})} \qquad (3)$$

The P(i) drop out. Since $P(\bar{A})$ must be near one (there are millions of traces and no one can be very probable) it can also be ignored. To an approximation we can also ignore $P(\bar{A}|i)$. This is a good approximation to the extent that the probability of needing a trace remains low even in the presence of a predictive cue. If we allow this approximation we get the following which is very easy to analyze:

$$\frac{P(i|A)}{P(i|\bar{A})} \simeq \frac{P(A|i)}{P(A)} \qquad (4)$$

Our claims do not depend on making this approximation. It is just that they are a lot easier to see with the approximation. In our experiments P(A) is basically constant for all items and so the critical factor turns out to be the probability that the trace is relevent given a particular cue. This is precisely what is manipulated by fan in a typical experiment. The more facts associated with a particular concept the less likely any one is given the concept. Basically if the fan is n the probability is 1/n. Anderson (1976) did an experiment that decorrelated fan and probability by manipulating the probability of testing various facts associated with a particular concept. That experiment showed conclusively that the critical factor is probability and not fan.

Thus, the fan effect is a consequence of memory using the correlation between cues and a memory's relevence to predict when the memory is needed. It may be that spreading activation is one of the mechanisms that the mind uses to compute the correlation. However, for current purposes the critical fact is once again that the explanation of the phenomena lies in what is outside of the human head and not what is inside.

PDP -- Categorization

PDP models involve representing knowledge in a distributed form where specific experiences do not have specific encodings. On the other hand PDP models do learning locally such that changes in strengths of connection between specific elements must underlie these distributed encodings. This leads PDP models to naturally produce generalization phenomena such that they extract central tendencies out of the experience of specific instances. In introducing PDP models, McClelland, Rumelhart, & Hinton(1986) give a lot of play to categorization phenomena which is the identification of common categories in a set of tendencies. It receives more page space in their article than any other phenomena. There is a substantial literature in cognitive psychology on categorization behavior.

McClelland et al. do not actually simulate any specific experiment in this literature but rather offer a simulation of the extraction of the characteristics of the members of two gangs (the jets and the sharks) as a prototype of the experiments in the literature.

To develop a rational analysis of categorization behavior the first thing we need to ask is what are the goals of the cognitive system in forming categories. In much of the experimental literature on categorization one gets the feeling that the driving force behind categorization is some sort of social conformity--that we need to learn to use the same labels to describe objects as do other people. However, this clearly cannot be all of the picture, particularly because people can learn to identify categories in the absence of any labels. I think the real function of categorization is to maximize the system's ability to predict properties of objects including their labels. Clearly, a system that can make accurate predictions will be in a position to maximize its goals.

The reason people form categories to maximize prediction is because of the nature of objects in the external world. Formally, the following is the characterization that I will assume in my rational derivations. I will assume that the world seen so far has consisted of n objects which are partitioned into s disjoint sets or categories. Each object can be classified according to some r dimensions (for simplicity I will only consider cardinal dimensions) where each dimension i has some m_i values. The members of a category belong in that category by virtue of possessing theoretical probabilities p_{ij} that they will display value j on dimension i. These probabilities provide the intensional definition of a category in contrast to its extensional definition which can be gotten simply by listing the category members.

These assumptions are intended as descriptions of the external world not just of the perception of the world in the human head. One can ask why the objects in the world should partition themselves in disjoint partitions defined by conjunctions of features. I cannot say I know the total answer but there are some obvious things to point at. For instance there is the genetic phenomenon of species which enforces a disjoint (no crossbreeding) partitioning of conjunctively defined categories (the common genetic code within a species). Other types of objects like physical elements and tools tend to produce similar disjoint partitionings of conjunctively defined categories. One can also question the probabilistic definition of category membership since this is in contradiction to the tradition in the artificial intelligence work on categories. However, I think it is indisputable that category members do display their features with only certain probabilities. Most labradors are black and have four legs but neither feature is displayed universally.

From these assumptions one can derive a Bayesian algorithm to assign objects to categories and to estimate the theoretical probabilities p_{ij}. Again, I do not have the space to go into the details of the algorithm. I have applied the algorithm to the now classic data of Medin and Schaffer (1978) where it did better than their original model using only a single parameter rather than their many. I have also applied it to the long series of experiments involving the Posner and Keele (1968) stimuli using an encoding of these materials developed by Hintzman (1986). It accounts for all the phenomena that Hintzman lists for these materials. I have also successfully predicted the results of a complicated experiment of Elio and Anderson (1981) which no model before Hintzman's was able to

account for. Rather than discussing the specific experiments in detail it is worthwhile listing some of the major phenomena that are known about human categorization and explaining how the model accounts for each:

1. Clearly the research indicates that, to a degree, people extract the central tendency of a set of instances in that their behavior is a function of the distance from that central tendency. This simply reflects a sensitivity to the statistical correlation between features and category identity which amounts to using conditional probabilities in a Bayesian analysis.

2. In addition to distance from a central tendency the literature has found an effect of distance from specific examples(eg., Medin & Schaffer, 1978). This is produced by the tendency of the model to break diverse categories into subcategories where the features cluster together. The reason for this is that predictive power is gained by such decomposition.

3. It has shown that when a category has multiple central tendencies subjects can pick this up (Neumann, 1977). As with point (2) this is produced by the tendency to break a large diverse category into smaller categories that increase predictability.

4. There is an effect of category size as was discussed with respect to the Posner & Keele task. This is simply a sensitivity to base rates.

5. Rosch, Mervis, Gray, Johnson, Boyes-Braem (1976) has documented the many circumstances in which there appear to basic level categories. The existance of such categories in our framework is simply a consequence of the fact that these categories maximize the predictability of the world--which is basically Rosch's original point.

6. It is not necessary for feedback on category membership to be given in order for categories to emerge(Fried and Holyoak, 1984). Categories will emerge any time they increase in predictability of the universe. However, by applying category labels we increase the amount of structure that can be predicted and so enhance the value of category membership. So, labels should enhance categorization but are not essential.

7. The more things that can be predicted from category membership the more likely a category is to be formed even though this means one has to learn more about a category (Billman, 1983).

Thus it seems that categorization phenomena can be again explained from a rational perspective assuming that the controlling factor is the structure of the world and not the structure in the human head. Note again this analysis does not deny that PDP mechanisms may be the way that the mind implements this rational analysis. However, it denies that PDP models provide an adequate explanation of the phenomena.

Conclusions

In summary we have looked at three cognitive architectures For each we have have taken a signature phenomenon and developed a reasonable model of the world in which that phenomenon occurs and the goals of humans operating in that world. We have made a few assumptions about computational costs which are not at all mechanism specific. We have derived the signature phenomena as solutions to the optimization problems we defined. In each case this rational analysis led to an account that was as accurate or more accurate than the original mechanistic account.

Now we come to the hard question of what the implications are of these demonstrations. I am not really sure what the implications are but I will hazard a guess. This is that cognitive architectures should be viewed as notations for expressing the behavioral functions that emerge as the solutions to the optimization problems in a rational analysis. The real theory lies in the assumptions made in the statement of the optimization problem--i.e., the assumptions about the goals, the world, and the computational limitations. These assumptions do not have the same identifiability problems that the mechanistic models do and lead to a much deeper explanation of the phenomena at hand. However, something computationally powerful like a Turing-equivalent architecture is necessary if we are going to be able to express the solution to these optimization problems.

Thus the theory is in the framing of the information processing problem and the architectures provide notation for expressing the solutions to the optimization problems. I see a one-to-many mapping between framings and architectures. That is, one can take a single framing and for every architecture find some configuration of its mechanisms that enable the optimal behavior to be computed. Choice among architectures is then not to be determined by veracity of empirical predictions. Rather it is to be determined by how easy it is to work out the optimal behavior in that architecture. Ease of use is the classic criterion for selecting among notations. Empirical veracity is reserved for theories.

References

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

Anderson, J.R. (1983). *The Architecture of Cognition.* Cambridge, MA: Harvard University Press.

Anderson, J.R. (1983). Retrieval of information from long-term memory. *Science, 220,* 25-30.

Anderson, J. R. (1987). Methodologies for studying human knowledge. *The Behavioral and Brain Sciences, 10,* 467-505.

Anderson, J. R. (in press). *The Adaptive Character of Thought.* Hillsdale, NJ: Erlbaum.

Billman, D. (1983). *Inductive learning of syntactic categories.* Doctoral dissertation, University of Michigan, Ph.D. dissertation

Burrell, Q. L. (1980) A simple stochastic model for library loans. *Journal of Documentation, 36,* 115-132.

Burrell, Q. L (1985). A note on aging on a library circulation model. *Journal of Documentation, 41*, 100-115.

Burrell, Q L. & Cane V. R. (1982). The analysis of library data. *Journal of the Royal Statistical Society. Series A*(145), 439-471.

Elio, R. & Anderson, J. R. (1981). The effects of category generalizations and instance similarity on schema abstraction. *Journal of Experimental Psychology: Human Learning and Memory, 7*, 397-417

Fried, L. S., & Holyoak, K. J. (1984). Induction of category distributions: A framework for classification learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 10*, 234-257.

Hintzman, D. L. (1986). Schema Abstraction in a Multiple-Trace Memory Model. *Psychological Review, 93*, 411-428.

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

McClelland, J. L., Rumelhart, D. E., and the PDP research group. (1986). *Parallel distributed processing: Explorations in the microstructure of cognition.* Cambridge, MA: Bradford Books.

McClelland, J. L., Rumelhart D. E., & Hinton, G. E. (1986). *Parallel Distributed Processing.* Vol. 1: *The appeal of parallel distributed processing.* In D. E. Rumelhart & J. L. McClelland (Eds.).

Medin, D. L., & Schaffer, M. M. (1978). Context theory of classification learning. *Psychological Review, 85*, 207-238.

Neumann, P G. (1977). Visual prototype information with discontinuous representation of dimensions of variability. *Memory & Cognition, 5*, 187-197.

Posner, M. I. & Keele, S. W. (1968). On the genesis of abstract ideas. *Journal of Experimental Psychology, 77*, 353-363.

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

Rosenbloom, P S., Laird, J. E., & Newell, A. (In Press). *Working Models of Human Perception. The chunking of skill in knowledge.* London: Academic Press. In H. Buoma & B. A. G. Elsendoorn (Eds.).

Rumelhart, D. E., McClelland J L., and the PDP research group (1986) *Parallel distributed processing Explorations in the microstructure of cognition.* Cambridge, MA: Bradford Books.

Salton, G., & McGill, M. J. (1983). *Introduction to modern infromation retrieval.* New York: McGraw-Hill.

Stritter, E. P. (1977). *File migration.* Doctoral dissertation, Stanford University, Stanford: Stanford Linear Accelerator Center.