

Unified Theories of Cognition

Allen Newell

For multiple book review of
BEHAVIORAL & BRAIN SCIENCES

Psychology has arrived at the possibility of unified theories of cognition—theories that gain their power by positing a single system of mechanisms that operate together to produce the full range of human cognition.

I do not say they are here. But they are within reach and we should strive to attain them.

Book published by:

Harvard University Press
Cambridge, Massachusetts
London, England
1990

current cognitive theories. But the conclusion to be drawn is quite a different one, namely, one consistent with random-walk models being a harbinger for unified theories. The details of the models are not critically important, though an appropriate set of details is necessary to get out detailed predictions. Any mechanism with the common properties epitomized in Figure 1-11 will do all right. This means that we may be able to settle on an important schematic characterization of an elementary mechanism of the mind. And we can trust the incorporation of this mechanism into unified theories of cognition that may appear to have quite different structure.

To sum up this whole section, I see a number of harbingers of unified cognitive theories. None spans the whole of cognition, though Act* comes close, but each covers an important aspect of what a unified theory requires. And what is more important, these harbingers all announce fundamentally the same theory. Of course, they differ in detail and each has the stamp of the individual theorists. But where they address the same issues they yield approximately the same answers and use variants of the same mechanisms. And there are other examples as well that I have not taken the space to mention (MacKay, 1987). It is time to attempt to push this idea forward.

1.4. The Task of the Book

My goal for this book is to convince you that unified theories of cognition are really worth striving for—now, as we move into the 1990s. The way not to do this is to talk about it, although I have been doing exactly that in this initial chapter. The only way I know how to make the case is to put forth a candidate, so that we can see what a unified theory of cognition means and why it is a plausible goal for cognitive science. Thus, throughout the rest of the book I will focus on a single candidate.

1.4.1. The Risks

Presenting a single candidate is not without serious risks, though it seems to me the only way to proceed. Let me spin out some of the risks that will bedevil the enterprise.

First, there is what can be called *the Popperian damnation*. The reader at some point spots a particular prediction of the theory, and recalls that in (say) 1965 someone published an article that shows

that the prediction is wrong. Since the aim of scientific endeavor, according to Popper, is to prove that theories are false, once the candidate theory is shown to be false it can be dispensed with. The book can be safely closed and discarded, to await the next candidate unified theory, perhaps in some future William James Lectures.

The second risk can be called the *bugaboos of psychology*. We are acutely sensitive, in psychology, about having theories—as if they are too grand a thing for psychology in its nascent state to have. Indeed, it was so bad at one time that, in order to avoid having to talk about having a theory, we talked about *pretheoretical orientations* (Kendler, 1961). We have perhaps gotten back to where we can talk about theories. But if someone begins to talk about *unified theories*, this sense of modesty can be expected to return. Unified theories are too grand a conception to entertain. Psychology needs to wait until—well, until some indefinite future—before entertaining such a thing.

The next risk is to be found guilty of the *sin of presumption*. Who am I, Allen Newell, to propose a unified theory of cognition? We have never had one of these before (though I have already said that John Anderson really has one). Psychology must wait for its Newton. And we all know that only two people could possibly replace Newton—they are Archimedes and Einstein, both of whom are already dead.

Next is the *necessity of selection*. Developing a unified theory, as will become apparent, is a synthetic enterprise. It is not an enterprise of showing that what psychology has already discovered is wrong. Rather, it is an attempt to synthesize the existing understanding into one unified whole, so we can go forward with the task of fashioning it into a better whole. Alas, selections must still be made, because many of the constructs in psychology are genuinely alternative conceptions, even though there is usually some truth on both sides.³ For instance, if one has productions and semantic nets but not frames or schemas, then that half of the cognitive psychologists who think schemas are grand will be convinced that the enterprise is really not a synthesis at all, but a competitive device to

3. Richman and Simon (1989) have made the point that the same common mechanisms may explain the same phenomenon even when two theories are as different in surface form as symbolic and connectionist theories. This point is related to the discussion above of the random-walk models of decision making.

replace schemas with another construct. There is, of course, a kernel of truth in what they say. Even so, the enterprise loses way.

Another risk is the *half-empty-half-full reaction*. I keep saying, "Look at how much this formulation explains," and the echo keeps coming back, "Look at how much this formulation fails to explain." To one the glass is half full, to another the glass is half empty. No matter how much one tries to show a unified theory and our progress toward it, there will be huge domains in which unification hasn't made significant progress.

The final risk is the *rising tide of connectionism*, which is showing signs of sweeping over all of cognitive science at the moment. The excitement is palpable—we are all hot on the trail of whether neuroscience and the cognitive world can finally be brought together. That is indeed an exciting prospect. But my message relates to symbolic architectures and all the other good things that connectionism sees as the conceptual frame to overthrow. So the final risk is that my timing is terrible. John Anderson had it right when he wrote his book on the architecture of human cognition for publication in 1983. It is too late now.

Believe me, I feel these risks in making the case for unified theories of cognition by putting forward a single candidate and riding that candidate for seven chapters. Iterating through the list of conceivable types of risks at the outset is clearly an attempt on my part to disarm the critics in advance. But having made my magic incantations, it is now time to get on with the business at hand. I do have one last verbal amulet. The late Warren McCulloch, one of the world's great cyberneticists, when attacked sharply at talks, often used to respond with "Don't bite my finger, look where I'm pointing" (McCulloch, 1965, p. iv). Indeed, I will appropriate this epigram as my motto here. I am sure Warren would not mind.

1.4.2. *The Vehicle: Soar*

There are still two more preliminaries. First, there needs to be the vehicle. I will propose a system called *Soar* as a candidate unified theory. *Soar* is an architecture for general intelligence, developed by John Laird, at the University of Michigan, Paul Rosenbloom, at the Information Sciences Institute of the University of Southern California, and myself. *Soar* was first realized as a computer program in 1983 (Laird, Rosenbloom, & Newell, 1986), and there is a current operational version. *Soar*4.4 (Laird, Newell, & Rosen-

bloom, 1987). *Soar* was raised as an AI system, so to speak. It has solved problems and learned on many different tasks, several of which will appear in the following pages. So it is a real system with real properties. As is standard practice in AI, however, a new version (*Soar*5) is on the way. (Indeed, it is being phased in as this book goes to press.) Thus it will sometimes be a little difficult to know what is fully operational and what is not. But in the main, *Soar* may be taken as an operational system.

Soar already has the major mechanisms of a cognitive system, as that notion has developed over the last decade in cognitive science. It is largely consonant with the several harbingers above. It has a symbol system and a goal hierarchy. It uses a production system as the foundation of the architecture, thus being a kind of recognize-act system. It uses problem spaces everywhere to do all of its business, that is, to formulate all its tasks. It has a chunking mechanism as part of its architecture, and it learns by chunking on everything it does. In the *Soar* project we are currently busy moving it toward becoming a unified theory of cognition in detail, in ways that this book will describe.⁴

Though in general *Soar* fits into the community of symbolic systems being used in cognitive science, it does have a few highly contentious features. The biggest one, as we shall see, is that its memory seems to be entirely procedural—that is, it is made up entirely of productions. Although it is useful, interesting, and fun to explore such unique assumptions, it may detract a little from using *Soar* simply as an exemplar for a unified theory of cognition, where the emphasis is on being synthetic and noncontentious. It is perhaps not quite vanilla-flavored enough.

1.4.3. *The Role of Artificial Intelligence*

It is worth being explicit about the role of artificial intelligence (AI) in this enterprise, since AI is often seen as contrasting with cognitive science—pure (engineering) AI versus mechanisms of human intelligence—as if these were different subject matters. *Soar* itself

4. In part, this is why *Act** is not the vehicle of choice for this book. To work on a cognitive architecture one must be able to control it and change it. So one must have one's own variant architecture. But also, much has been learned since the specification of *Act** in relatively mature form in 1983, from *Act** itself and from elsewhere. *Soar* has important properties and capabilities that make it a worthwhile exploration in its own right.

already provides a clue. As far as I am concerned, AI is the study of the mechanisms of intelligence and all supporting technologies, extending to perception and motion, namely robotics. AI is itself part of computer science, which is the study of information and its processing in all its variety.

AI provides the theoretical infrastructure for the study of human cognition. It has already provided much of our knowledge of what intelligence is all about and how it is to be realized in information-processing mechanisms. Being an infrastructure, AI fills an entire spectrum of specific roles (Newell, 1970)—from enforcing operational thinking about processes, to a theoretically neutral language of processing, to metaphors about the kinds of processes that might occur in the mind. Various of these specific roles are sometimes taken as the primary way AI relates to psychology. In fact, it plays them all.

AI has one additional critical contribution to make to understanding the mind. What's most important in understanding intelligence is how it functions. The first question to ask about any theory of the nature of intelligent action, whether in humans or machines, is whether it in fact produces intelligent action—effective problem solving, perception, learning, movement, or whatever. Though the issue of functionality might not be central in some areas of psychology, it is critical for perception, cognition, and action. Thus the basic theory of the mechanisms of intelligence is central to human cognition.

I will draw on AI in all these ways throughout the book, without further special comment or acknowledgment.

1.4.4. Preview of the Chapters

This brings us, at last, to a foreshadowing of the chapters themselves. Chapter 2 will introduce the foundations of cognitive systems. What do the terms mean that we all use—knowledge, representation, computation, symbols, architecture, intelligence, search. This will take us back over familiar territory, but it is useful. It gets us started in the right way and prepares the ground. As a minor benefit, it will provide for those readers who feel a need for definitions.

Chapter 3 moves to basic aspects of the human cognitive architecture. We inquire into what general conditions shape the human architecture. Given that the human is constructed out of neural

technology, on the one hand, but is capable of intelligent action, on the other—does this provide constraints on a theory of human cognition? It turns out that it does.

Chapter 4 contains the full detail of the architecture that will be our candidate cognitive architecture, namely Soar. This architecture will, naturally enough, abide by the constraints revealed in Chapter 3. The chapter will focus on the structure of that architecture and how it manages intelligent action. The chapter ends by showing how Soar maps into human cognition.

The next three chapters will range over different domains of cognition, showing how Soar provides a unified treatment. Chapter 5 will start with immediate behavior—tasks that go from presentation to completion in just a second or two. Chapter 6 covers learning, memory, and skill acquisition—behavior that cuts across the time dimension. Chapter 7 moves up to tasks that take long enough so the system can attain modest degrees of rationality, namely tasks of the order of minutes and more.

Finally, Chapter 8 explores the frontier that separates what Soar can currently treat, in at least modestly concrete terms, and what is still pure speculation. A unified theory of cognition must indeed be able to address the full range of cognition, even if the actual investigation is still to be done. It is important at least to ask what approach would be indicated for such areas as language and development, and what the relation could be to the bordering system layers, neurophysiology below and social systems above. Then, having made these few probes of the frontier, we bring the case to a close.

2.10. Summary

This chapter has been a tour of the key terms that lie at the foundation of cognitive science. Iterating through these concepts seems a reasonable way to lay the foundations for the rest of the book. Figures 2-21 and 2-22 provide a summary that can serve as a ready reference when the concepts reappear in later chapters. Figure 2-21

24. The names come from the constants in the logistic equation used in quantitative population models.

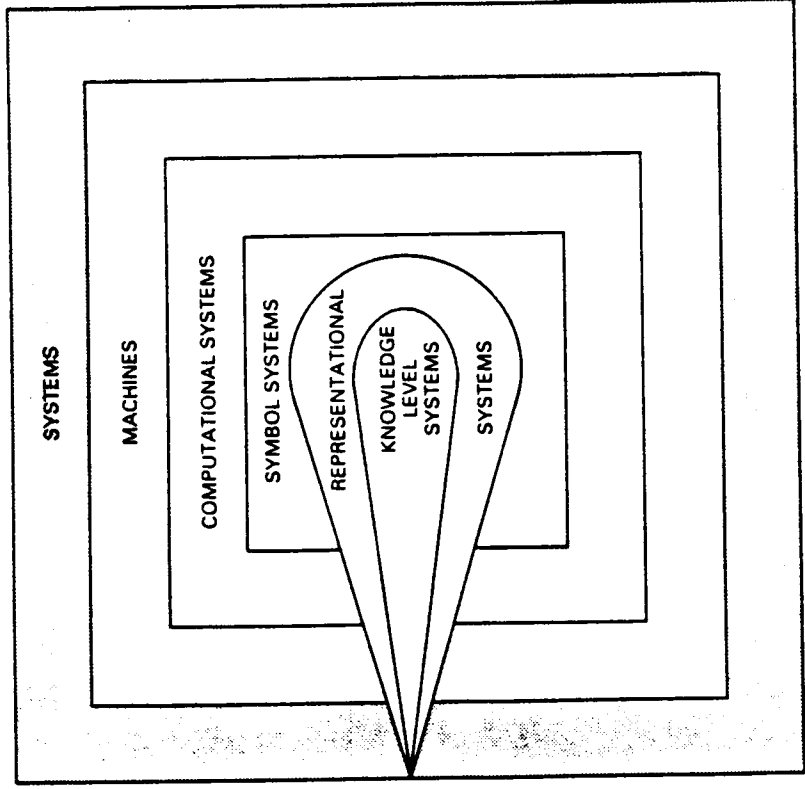


Figure 2-22. The different types of systems.

lists all the key terms and phrases teamed with short statements giving their essential character.

During the course of the chapter several systems were introduced. Figure 2-22 shows how they relate to one another. Focusing on the internal structure of the systems, it shows four nested classes of systems (the rectangles), each successively more specialized. The most general is just the class of all systems. Systems can be specialized to machines, which are systems that have a medium, so that inputs and outputs can be defined for them such that every machine effects a function from its inputs to its outputs (actually from the history of its inputs to its outputs). Computational systems are machines that are capable of producing multiple functions; they work by using some of their inputs to specify the functions that hold between other inputs and outputs. Symbol systems are computational systems with symbols, that is, with devices for distal access. The diagram doesn't reflect adequately that the diversity of computational systems is much greater than that of symbol systems, the diversity of machines much greater than that of computers, and so

Behaving system

A system characterized by multiple response functions on its environment

Knowledge

The medium of a knowledge-level system

What can be used to select actions to attain goals

- Abstracted from representation and processing

The knowledge level is just another system level

Representation

A system that obeys representation laws

Composing transformations

A general approach to representing

Uses a nonspecific medium capable of combinatorial variety

Composes transformations to satisfy representation laws

Required if enough variety is required of the representation

Computational system

A machine that obtains multiple functions on some inputs and outputs by using other inputs to specify functions

Symbol

A mechanism for obtaining distal access

Required if enough variety is required of the system

Architecture

The fixed structure that realizes a symbol-level system

Intelligence

The degree of approximation to a knowledge-level system

Search

Problem search: Search in the problem space to solve a task

Knowledge search: Search in memory to guide problem search

Problem space

A set of states and a set of operators from state to state

- Operators may have conditions of applicability

A task is to find a desired state, applying operators freely

The states are generated by the agent, they do not pre-exist

Deliberation vs. preparation trade-off

Deliberation: Activity after task is set

Preparation: Activity prior to setting of task

Trade-off: The relative mix of two types of activity

Figure 2-21. Foundational terms for cognitive science.

on, but the figure is sufficient to reflect our interest in systems that can support mind-like capabilities.

The two comet-shaped regions indicate the systems that support adaptation with respect to the external world. The larger class are the representational systems, which are systems whose internal structure obeys representation laws. Such laws support adaptation.²⁵ Within the class of representational systems are the knowledge-level systems, which are so structured that their behavior can be derived just from the content of the representations about the world and their goals. There are many representational systems that cannot be described at the knowledge-level.

Representational systems must be realized in some structure, so the two comet-shaped regions intersect the four types of system structures. Representational systems with substantial scopes for domains and goals require symbol systems for their realization, but as scope diminishes then structures of lesser capability can be used as well. The areas of intersection with these more general system structures become smaller and smaller, indicating their more specialized character. There are of course many structures of each type that do not satisfy representation laws, hence are not about the external world, and hence do not support active adaptation.

25. I did not discuss systems that support adaptation but are not representational. They would be another outer layer of the adaptational class of systems. The obvious candidates would be open-loop adaptations, such as those constructed by evolution and human designers. Examples would be a bird's wing or a wheel.

3.12. Summary

The objective of the book is to be convincing that cognitive science is ready for unified theories of human cognition. So far, I have been laying the groundwork. I have introduced the major concepts needed to describe cognitive systems generally, and assayed the main outlines of the human cognitive architecture. My strategy is to present at the outset anything that can be discerned generally about the human cognitive architecture so there would be just that much less that would have to be posited in a proposed theoretical architecture.

I have proposed three ingredients as given, from which to extract these general features of the architecture: (1) the human architecture is constructed out of neural technology; (2) cognitive behavior is clearly evident at about one second; and (3) cognitive behavior requires a symbol system. Because cognitive behavior shows up rapidly in terms of how fast neural technology operates—the real-time constraint on cognition—they provide enough constraint to obtain some useful results.

First of all, different cognitive worlds attend different time scales of actions—the biological band, which is still in the realm of physical-chemical law; the cognitive band, which is the realm of representational law and computational mechanisms; the intendedly rational band, which is the realm of knowledge and goal-driven reason; and above that, with increasing fuzziness, the social, historical, and evolutionary bands. Our focus, however, was not on such a grand scale but on what could be said about the mechanisms and structure of the cognitive band, in which there are four levels. The lowest level is that of memory access, which occurs at ~ 10 ms, the lowest level of neural circuits. The next level is that of elementary deliberations, which occurs at ~ 100 ms. It is the level of automatic behavior. The next level is that of simple operations, which occurs at ~ 1 sec and already provides minimal cognitive behavior. The fourth level is that of composed operations, which occurs at ~ 10 sec. This is longer than the minimal time for cognitive behavior and shows a continual shift to a recognition-based system.

This chapter is highly speculative. It attempts to establish strong structural statements from highly generalized considerations. We don't understand the full variety of mechanisms that contribute to

all these system levels. In fact, almost every aspect of the analysis has substantial uncertainty associated with it. That is no reason for not engaging in such speculation. The potential gains are very great, namely to get some general guidelines for the cognitive architecture regardless of any specific proposal. Moreover, taken with the right attitude, such speculation is almost totally benign. An argument has been laid out. It links together several general aspects of the human cognitive system. If some facts contravening the conclusion can be found, that is interesting. They can be taken back through the argument to find where the argument was wrong. That will be instructive, because the argument can be taken as the obvious line of reasoning. Thus, the form of the argument is important and needs to be shaped up. For instance, one can attempt sensitivity analyses, such as Figure 3-10, to make it easy to find the soft spots.

In the following chapters such concerns are put aside. Any candidate architecture for human cognition must exhibit these characteristics. The architecture must be a further specification of the outline sketched here.

230 ■

4.9. Summary

Iterating through the list of Figure 4-24 instills confidence that Soar is a good candidate for a unified theory of cognition. That is, of course, deliberate on my part. One wants as complete a picture as

17. But see Maes and Nardi (1988) for a variety of reflections in computational systems.

possible, however, so I list here some of the problems and issues that Soar faces.

Soar, like any computer system, is a system always in a state of becoming something else. Many aspects are not yet operational, or only operational in some limited way. I describe here a current version, Soar 4.5, but then immediately claim that certain improvements are coming in Soar5, the next major release. These include some important aspects, otherwise we would not be modifying or enhancing Soar to include them. Of course, some of these changes will not make it even then, so it will have been premature to speak of them. It is a genuine Tristram Shandy situation. In the two years since I delivered the William James Lectures, Soar has changed significantly. In particular, Soar5 has gone from the design stage to implementation and is now being introduced. To update the book to take into account Soar5 would require extensive rewriting, during which Soar would change some more. Thus, what is represented to be Soar is really a moving average through a two-year window.

It is easy enough to enumerate specific examples of important aspects of Soar in this in-between state. The main new aspects of Soar5 are components of the perceptual and motor systems and the adoption of the principle that Soar5 is able to remember only a single state within any problem space, as opposed to Soar4's current capability for remembering all of the states that have been visited. Beyond Soar5 are modifications to reduce the power of the matching capability of productions. The current one, inherited from Ops5, seems in some ways clearly too powerful for what can be done in a single recognition. A more minor issue concerns the requirement for some productions to keep Soar from descending into infinite impasse pits. Somehow, one feels that the default behavior ought to be defined entirely in the architecture, but this has not occurred yet.

Certain things have not been demonstrated with Soar, although nothing in our current understanding indicates where the difficulties will arise. One long-term goal of artificial intelligence has been to construct *nonbrittle* systems—systems that do not abruptly fail when they move out beyond some predefined task scope. Certainly humans show some degree of nonbrittleness, although we may overestimate it (transport a steelworker to the floor of the stock exchange or a psychoanalyst to the old manual Vax-configuration task, and see what happens). Soar has enough attributes of general

immediately fatal. Big investments demand that abandonment be a deliberate business, done only when the preponderance of evidence shows it must be done.

(6) Unified theories of cognition must be made easy to use and easy to make inferences from. This becomes especially important for theories of very wide scope and for theories that are, in important ways, embedded in simulations. We have learned in computer science the immense effort required to produce a successful programming language or a new operating system or a new statistical package. It requires much more than the lone investigator with the good idea and the initial implementation—as many lone programmers have found out to their sorrow. It requires attention to the user interface, to user documentation, to training workshops, to system support. These tasks are especially important for building up a community around a unified theory and have become a significant component of the Soar community effort.

(7) Finally, I repeat the exhortation to acquire domains of application, which can support and invigorate a theory intellectually. Building application domains, by the way, creates a community with a large investment in ease of use, and hence with a willingness to expend the effort to create the tools to make it happen.

■

So we end where we began, but with understanding.

Psychology has arrived at the possibility of unified theories of cognition—theories that gain their power by positing a single system of mechanisms that operate together to produce the full range of human cognition.

I do not say they are here. But they are within reach and we should strive to attain them.