Artificial Intelligence 59 (1993) 285–294
Elsevier

ARTINT 1010

Book Review

Allen Newell, *Unified Theories of Cognition* *

Daniel C. Dennett

Center for Cognitive Studies, Tufts University, Medford, MA 02155-7068, USA

Received August 1991
Revised September 1992

The time for unification in cognitive science has arrived, but who should lead the charge? The immunologist-turned-neuroscientist Gerald Edelman [6] thinks that neuroscientists should lead—or more precisely that he should (he seems to have a low opinion of everyone else in cognitive science). Someone might think that I had made a symmetrically opposite claim in *Consciousness Explained* [4]: philosophers (or more precisely, those that agree with me!) are in the best position to see how to tie all the loose ends together. But in fact I acknowledged that unifying efforts such as mine are proto- theories, explorations that are too metaphorical and impressionistic to serve as the model for a unified theory. Perhaps Newell had me in mind when he wrote in his introduction (p.16) that a unified theory "can't be just a pastiche, in which disparate formulations are strung together with some sort of conceptual bailing wire", but in any case the shoe more or less fits, with some pinching. Such a "pastiche" theory can be a good staging ground, however, and a place to stand while considering the strengths and weaknesses of better built theories. So I agree with him.

> It is not just philosophers' theories that need to be made honest by modeling at this level; neuroscientists' theories are in the same boat. For instance, Gerald Edelman's (1989) elaborate theory of

Correspondence to: D.C. Dennett, Center for Cognitive Studies, Tufts University, Medford, MA 02155-7068, USA. E-mail: ddennett@pearl.tufts.edu.

* (Harvard University Press, Cambridge, MA, 1990); xvi+549 pages.

> "re-entrant" circuits in the brain makes many claims about how such re-entrants can accomplish the discriminations, build the memory structures, coordinate the sequential steps of problem solving, and in general execute the activities of a human mind, but in spite of a wealth of neuroanatomical detail, and enthusiastic and often plausible assertions from Edelman, we won't know what his re-entrants can do—we won't know that re-entrants are the *right* way to conceive of the functional neuroanatomy—until they are fashioned into a whole cognitive architecture at the grain-level of Act* or Soar and put through their paces. [4, p. 268]

So I begin with a ringing affirmation of the central claim of Newell's book. Let's hear it for models like Soar. Exploring whole cognitive systems at roughly that grain-level is the main highway to unification. I agree, moreover, with the reasons he offers for his proposal. But in my book I also alluded to two reservations I have with Newell's program without spelling out or defending either of them. This is obviously the time to make good on those promissory notes or recant them. "My own hunch", I said, "is that, for various reasons that need not concern us here, the underlying medium of production systems is *still* too idealized and oversimplified in its constraints" [4, p. 267]. And a little further along I expressed my discomfort with Newell's support for the traditional division between working memory and long-term memory, and the accompanying notion of *distal access* via symbols, since it encourages a vision of "movable symbols" being transported here and there in the nervous system—an image that slides almost irresistibly into the incoherent image of the Cartesian Theater, the place in the brain where "it all comes together" for consciousness.

Preparing for this review, I re-read *Unified Theories of Cognition*, and read several old and recent Newell papers, and I'm no longer confident that my reservations weren't based on misunderstandings. The examples that Newell gives of *apparently* alternative visions that can readily enough be accommodated within Soar—semantic and episodic memory within the single, unified LTM, Koler's proceduralism, Johnson-Laird's mental models, for instance—make me wonder. It's not that Soar can be all things to all people (that would make it vacuous), but that it is easy to lose sight of the fact that Soar's level is a *low* or foundational architectural level, upon which quasi- architectural or firmware levels can be established, at which to render the features and distinctions that at first Soar seems to deny. But let's put my reservations on the table and see what we make of them.

On the first charge, that Soar (and production systems in general) are still too idealized and oversimplified, Newell might simply agree, noting that we must begin with oversimplifications and use our experience with them to uncover the complications that matter. Is Soar *the* way to organize cognitive

science, or is it "just" a valiant attempt to impose order (via a decomposition) on an incredibly heterogeneous and hard-to-analyze tangle? There's a whole messy world of individualized and unrepeatable mental phenomena out there, and the right question to ask is not: "Does Soar idealize away from these?"—because the answer is obvious: "Yes, so what?" The right question is: "Can the *important* complications be reintroduced gracefully as elaborations of Soar?" And the answer to that question depends on figuring out which complications are really important and why. Experience has taught me that nothing short of considerable mucking about with an actual implementation of Soar, something I still have not done, would really tell me what I should think about it, so I won't issue any verdicts here at all, just questions.

First, to put it crudely, what about pleasure and pain? I'm not just thinking of high-urgency interrupts (which are easy enough to add, presumably), but a more subtle and encompassing focusing role. Newell recognizes the problem of focusing, and even points out—correctly, in my view—that the fact that this can be a problem for Soar is a positive mark of verisimilitude. "Thus the issue for the standard computer is how to be interrupted, whereas the issue for Soar and Act* (and presumably for human cognition) is how to keep focused" (Newell, Rosenbloom and Laird [10]). But the Soar we are shown in the book is presented as hyperfunctional.

> Soar's mechanisms are dictated by the functions required of a general cognitive agent. We have not posited detailed technological limitations to Soar mechanisms. There is nothing inappropriate or wrong with such constraints. They may well exist, and if they do, they must show up in any valid theory. (p. 354)

Doesn't this extreme functionalism lead to a seriously distorted foundational architecture? Newell provides an alphabetized list (Fig. 8.1, p. 434) of some mental phenomena Soar has not yet tackled, and among these are daydreaming, emotion and affect, imagery, and play. Soar is all business. Soar is either working or sound asleep, always learning-by-chunking, always solving problems, never idling. There are no profligate expenditures on dubious digressions, no along-for-the-ride productions cluttering up the problem spaces, and Soar is never too tired and cranky to take on yet another impasse. Or so it seems. Perhaps if we put just the right new menagerie of operators on stage, or the right items of supplementary knowledge in memory, a sprinkling of sub-optimal goals, etc., a lazy, mathophobic, lust-obsessed Soar could stand forth for all to see. That is what I mean about how easy it is to misplace the level of Soar; perhaps all this brisk, efficient problem solving should be viewed as the biological (rather than psychological) activities of elements too small to be visible to the naked eye of the folk-psychological observer.

But if so, then there is a large element of misdirection in Newell's advertising about his functionalism. "How very functional your teeth are, Grandma!" said Red Riding Hood. "The better to model dysfunctionality when the time comes, my dear!" replied the wolf. Moreover, even when Soar deals with "intendedly rational behavior" of the sort we engage in when we are good experimental subjects—comfortable, well-paid, and highly motivated—I am skeptical about the realism of the model. Newell acknowledges that it leaves out the "feelings and considerations" that "float around the periphery" (p. 369), but isn't there also lots of *non*-peripheral waste motion in human cognition? (There certainly seems to me to be a lot of it when I think hard—but maybe Newell's own mental life is as brisk and no-nonsense as his book!)

Besides, the hyperfunctionality is *biologically* implausible (as I argue in my book). Newell grants that Soar *did* not arise through evolution (Fig. 8.1), but I am suggesting that perhaps it *could* not. The Spock-like rationality of Soar is a very fundamental feature of the architecture; there is no room *at the architectural level* for some thoughts to be harder to think *because they hurt*, to put it crudely. But isn't that a fact just as secure as any discovered in the psychological laboratory? Shouldn't it be a primary constraint? Ever since Hume got associationism under way with his quasi-mechanical metaphors of combination and attraction between ideas, we have had the task of describing the dynamics of thought: what makes the next thought follow in the heels of the current thought? Newell has provided us, in Soar, with a wonderfully deep and articulated answer—the best ever—but it is an answer that leaves out what I would have thought was a massive factor in the dynamics of thought: pain and pleasure. Solving some problems is a joy; solving others is a bore and a headache, and there are still others that you would go mad trying to solve, so painful would it be to contemplate the problem space. Now it *may just be* that these facts are emergent properties at a higher level, to be discerned in special instances of Soar chugging along imperturbably, but that seems rather unlikely to me. Alternatively, it may be that the *Sturm und Drang* of affect can be piped in as a later low-level embellishment without substantially modifying the basic architecture, but that seems just as unlikely.

David Joslin has pointed out to me that the business-like efficiency we see in the book is largely due to the fact that the various implementations of Soar that we are shown are all special-purpose, truncated versions, with tailor-made sets of operators and productions. In a fully general-purpose Soar, with a vastly enlarged set of productions, we would probably see more hapless wandering than we would want, and have to cast about for ways to focus Soar's energies. And it is here, plausibly, that an affective dimension might be just what is needed, and it has been suggested by various people (Sloman and Croucher [13], de Sousa [5]) that it cannot be packaged

within the contents of further knowledge, but must make a contribution orthogonal to the contribution of knowledge.

That was what I had in mind in my first reservation, and as one can see, I'm not sure how sharply it cuts. As I said in my book, we've come a long way from the original von Neumann architecture, and the path taken so far can be extrapolated to still brainier and more biological architectures. The way to find out how much idealization we can afford is not to engage in philosophical debates.

My second reservation, about symbols and distal access, opens some different cans of worms. First, there is a communication problem I want to warn other philosophers about, because it has bedeviled me up to the time of revising the draft of this review. I think I now understand Newell's line on symbols and semantics, and will try to explain it. (If I still don't get it, no harm done—other readers will set me straight.) When he introduces symbols he seems almost to go out of his way to commit what we philosophers call use–mention errors. He gives examples of symbol tokens in Fig. 2-9 (p. 73). He begins with words in sentences (and that's fine), but goes on to *numbers* in equations. We philosophers would say that the symbols were *numerals*—names for numbers. Numbers aren't symbols. He goes on: atoms in formulas. No. Atom-symbols in formulas; formulas are composed of symbols, not atoms; molecules are composed of atoms. Then objects in pictures. No. Object-depictions in pictures. I am sure Newell knows exactly what philosophers mean by a use–mention error, so what is his message supposed to be? "For the purposes of AI it doesn't matter"? Or "We AI-types never get confused about such an obvious distinction, so we can go on speaking loosely"? I don't believe it. There is a sort of willful *semantic descent* (the opposite of Quine's semantic ascent, in which we decide to talk about talk about things) that flavors many AI discussions. It arises, I think, largely because in computer science the expressions up for semantic evaluation do in fact refer very often to things inside the computer—to subroutines that can be called, to memory addresses, to data structures, etc. Moreover, because of the centrality of the domain of arithmetic in computers, the topic of "discussion" is often numbers, and arithmetical expressions for them. So it is easy to lose sight of the fact that when you ask the computer to "evaluate" an expression, and it outputs "3", it isn't *giving* you a number; it's *telling* you a number. But that's all right, since all we ever want from numbers is to have them identified—you can't eat 'em or ride 'em. (Compare "Gimme all your money!" "OK. $42.60, including the change in my pocket.") Can it be that this confusion of symbols and numbers is also abetted by a misappreciation of the fact that, for instance, the binary ASCII code for the *numeral* "9" is not the binary expression of the number 9?

Whatever its causes—or even its justifications—this way of speaking cre-

ates the impression that, for people in AI, semantics is something entirely internal to the system. This impression is presumably what led Jerry Fodor into such paroxysms in "Tom Swift and his Procedural Grandmother" [7]. It is too bad he didn't know how to put his misgivings constructively. I tried once:

> We get the idea [from Newell [9]] that a symbol designates if it gives access to a certain object or if it can affect a certain object. And this almost looks all right as long as what we're talking about is internal states .... But of course the real problem is that that isn't what reference is all about. If that were what reference was all about, then what would we say about what you might call my Julie Christie problem? I have a very good physically instantiated symbol for Julie Christie. I know it refers to her, I know it really designates her, but it doesn't seem to have either of the conditions that Professor Newell describes, alas. [2, p. 53] (See also Smith [14].)

Newell's answer:

> The criticisms seemed to me to be a little odd because to say that one has access to something does not mean that one has access to *all* of that thing; having some information about Julie Christie certainly doesn't give one complete access to Julie Christie. That is what polite society is all about .... The first stage is that there are symbols *which lead to internal structures.* I don't think this is obscure, and it is important in understanding where the aboutness comes from ... the data structures *contain knowledge about things in the outside world.* So you then build up further symbols which access things that you can think of as knowledge about something—knowledge about Julie Christie for instance. If you want to ask why a certain symbol says something about Julie Christie, you have to ask why the symbolic expression that contains the symbol says something about Julie Christie. And the answer may be ... because of processes that put it together which themselves have knowledge about Julie Christie .... Ultimately it may turn out to depend upon history, it may depend on some point in the history of the system when it came in contact with something in the world which provided it with that knowledge. ([2, p. 171], emphasis mine)

What we have here, I finally realize, is simply a two-stage (or *n*-stage) functional role semantics: *in the end* the semantics of symbols is anchored to the world via the knowledge that can be attributed to the whole system *at the knowledge level* in virtue of its capacity, exercised or not, for perspicuous

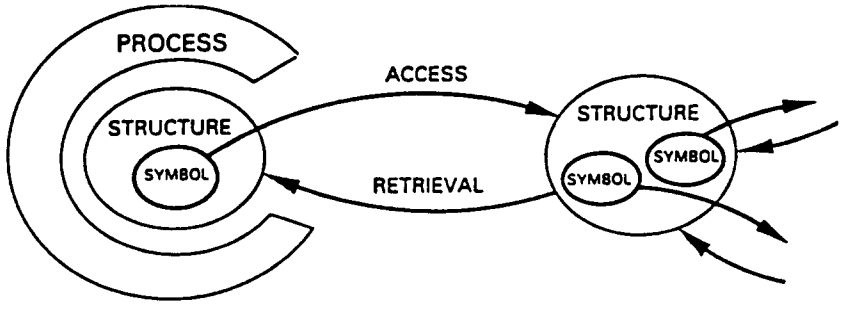

Fig. 1. Symbols provide distal access. (Originally, Fig. 2.10 (p. 75) in *Unified Theories of Cognition.*)

behavior vis-a-vis the items in the world its knowledge is about. And that's my view, too. What makes a data structure about Julie Christie is that it's the part of the system the presence of which explains my capacity to pick her out of a crowd, answer questions about her in quiz shows, etc., etc. That's all there is to it. But it is certainly misleading to say that the symbol gives one *any* "access" (partial access, in polite society!) to the object itself. (It turns out that Julie Christie and I have a mutual friend, who sent her an offprint of [2]. And what do you know, she ... sent me a Christmas card. "Getting closer", I thought. "Maybe Newell's right after all! You just have to be patient. Porsche, Porsche, Porsche.")

Newell's diagram in Fig. 1 makes it all clear (in retrospect) as long as you realize that it is not just that he concentrates (in his book and in his earlier writing) on the semantic link-arrows in the middle of the diagram—the access links tying symbols to their distal knowledge-stores—but that he simply *assumes* there is a solution to any problems that might arise about the interpretation of the arrows on the right-hand side of the diagram: those arrows, as I understand him now, lead one *either* to more data structures or eventually to something in the external world—but he is close to silent about this final, anchoring step. This is fine by me, but then I'm one of the few philosophers who thinks Twin-Earth and narrow content are artifactual philosophical conundrums of no importance to cognitive science [1,3]. Make no mistake, though: serious or not, Newell sweeps them under the rug right here. [1]

[1] In a more recent paper, he goes a bit further in defense of this interpretation: "The agent's knowledge is embodied in the knowledge of the four problem space components. However, this latter knowledge is about the problem space, states and operators; hence it cannot of itself be the knowledge of the agent, which is about the goal, actions and environment. It becomes the agent's knowledge by means of the relationships just described. That is, states are about the external world because of KL perception; operators are about the external world because of KL actions; the desired states are about the goal of the KL agent because of formulate-task; and the means-ends knowledge of select-operator is about performing tasks in the environment because it links environment-referring operators on environment-referring states to descriptions of environment-referring desired states." (Newell et al. [11, p. 23])

What concerns him is rather the interesting question of Plato's aviary: how does an intelligent agent with more knowledge than it can "contemplate" all at once get the right birds to come when it calls? (Dennett [4, p. 222–225]). And how do you do this without relying on a dubious *transportation* metaphor, which would require shipping symbol-tokens here and there in the system? I'm not sure I understand his answer entirely, but the crucial elements are given on p. 355:

> Functionally, working memory must be a short-term memory. It is used to hold the coded knowledge that is to be processed for the current task. It is necessary to replace that knowledge when the task changes. That replacement can be achieved in many ways, by moving the data [bad idea!—DCD], by moving the processes [better!—DCD], or by changing the access path [best!—DCD] .... Working memory for cognition has no continued functional existence outside these limits, however, since elements that are no longer linked to the goal stack become unavailable. Furthermore, problem spaces themselves have no existence independent of the impasses they are created to resolve.

I find these ideas some of the most difficult to understand in cognitive science, for they require setting aside, for once, what we might call the concrete crutch: the lazy picture of places (with boxes around them) and things moving to and fro. *That* vision, for all its utility at the symbol level, is a dangerous companion when we turn to the question of mapping computational processes onto brain processes. When Newell says "Search leads to the view that an intelligent agent is always operating within a *problem space.*" (p. 98) we should recognize that this is really being presented as an *a priori* constraint on how we shall interpret intelligent agents. Show me an intelligent agent, and whatever it does, I'll show you a way of interpreting it as setting up a problem space. Since the key term "distal" is defined relative to *that* space—that logical space—we should be cautious of interpreting it too concretely (cf. Fodor and Pylyshyn [8]).

So my second reservation is blunted as well. Two strikes, or maybe foul balls. There is one more issue I want to take a swing at as long as I'm up at bat. Newell's silence on the issue of natural language as a symbolic medium of cognition in human beings is uncanny. We know that Soar can (in principle) learn from taking *advice* (e.g., p. 312), and Newell sketches out the way Soar would or might handle language acquisition and comprehension (pp. 440–449; see especially his discussion of redundant encoding, p. 453), but I cannot figure out from these brief passages what Newell thinks happens to the overall shape of the competence of a cognitive system when it acquires a natural language, and I think his reticence on this score hides major issues. Early on he gives an eloquent survey of what he calls the "efflorescence of

adaptation" by the human (and only the human) species (pp. 114–115), but does this paean to productive versatility proclaim that the symbols of an *internalized natural language* are necessary, or is it rather that one needs a pre-linguistic language of thought—in which case we may wonder why the human language of thought gives us such an edge over the other species, if it does not get most of its power from the external language we learn to speak. For instance, Newell's discussion of annotated models (pp. 393ff) is a fine perspective on the mental models debates, but I am left wondering: can a non-human intelligent agent—a dog or dolphin or ape, for instance—avail itself of an annotated model, or is that level of cognitive sophistication reserved for language-users? This is just one instance of a sort of empirical question that is left curiously burked by Newell's reticence.

This gap is all the more frustrating since in other regards I find Newell's treatment in Chapters 1 and 2 of the standard debating topics in the philosophy of cognitive science a refreshing challenge. These chapters are simply required reading henceforth for any philosophers of cognitive science.[2] Newell doesn't waste time surveying the wreckage; he gets down to business. He says, in effect: "Sit down and listen; I'll show you how to think about these topics." He simply *makes moves* in all the games we play, and largely leaves it to us to object or play along. This should be a model for all non-philosopher scientists who aspire (correctly!) to philosophical probity. Don't try to play the philosophers' games. Just make your moves, clearly and explicitly, and see if you can get away with them.

I very largely agree with his moves, and it will be a pity if philosophers who disagree with him don't rise to the bait. They may not, alas. At times Newell underestimates how ingrown his jargon is. I have pushed portions of his text on some very smart philosophers and neuroscientists, and they are often completely at sea. (These issues are awfully hard to communicate about, and I am well aware that the alternative expository tactics I have tried in my own writing run their own risks of massive misconstrual.)

It might seem odd, finally, for me not to comment at all on Newell's deliberate postponement of consideration of consciousness, which gets just a brief apology on p. 434. Is this not unconscionable? Not at all. Newell's

[2]Philosophers will find important material throughout the book, not just in the foundational chapters at the beginning. For instance, the discussion of the discovery of the data-chunking problem in Soar and its handling (pp. 326–345) can be interpreted as a sort of inverse version of Meno's paradox of inquiry. The problem is not how can I search for something if I don't already know what it is, but how can I set myself up so that when I confront a real Meno-problem, there will be a way I can solve it? (Alternatively, if Soar couldn't solve the data-chunking problem, Meno's claim would not be paradoxical when applied to Soar, but simply true.) I think the memory-management search control strategies that are adopted can be read as part of an explicit answer—much more explicit than any philosopher's answer—to Meno's challenge.

project is highly compatible with mine in *Consciousness Explained* [4]. For instance, I endorse without reservation his list of multiple constraints on mind in Fig. 1-7 (p. 19). How can he achieve this divorce of consciousness? Just look! The enabling insight, for Newell and for me, is that handsome is as handsome does; you don't need any *extra witnesses* in order to explain cognition. Newell modestly denies that he has yet touched on consciousness; I disagree. He's made a big dent.

References

[1] D.C. Dennett, Beyond belief, in: A. Woodfield, ed., *Thought and Object* (Clarendon Press, Oxford, 1982).
[2] D.C. Dennett, Is there an autonomous "Knowledge Level"? in: Z.W. Pylyshyn and W. Demopoulos, eds., *Meaning and Cognitive Structure* (Ablex, Norwood, NJ, 1986) 51–54.
[3] D.C. Dennett, *The Intentional Stance* (MIT Press/Bradford Books, Cambridge, MA, 1987).
[4] D.C. Dennett, *Consciousness Explained* (Little Brown, Boston, 1991).
[5] R. de Sousa, *The Rationality of Emotion* (MIT Press, Cambridge, MA, 1987).
[6] G.M. Edelman, *The Remembered Present: A Biological Theory of Consciousness* (Basic Books, New York, 1989).
[7] J.A. Fodor, Tom Swift and his Procedural Grandmother, *Cognition* **6** (1978) 229–247.
[8] J.A. Fodor and Z.W. Pylyshyn, Connectionism and cognitive architecture: a critical analysis, *Cognition* **28** (1988) 3–71; also in: S. Pinker and J. Mehler, eds., *Connectionism and Symbol Systems* (MIT Press, Cambridge, MA, 1988) 3–71.
[9] A. Newell, The symbol level and the knowledge level, in: Z.W. Pylyshyn and W. Demopoulos, eds., *Meaning and Cognitive Structure* (Ablex, Norwood, NJ, 1986) 169–193.
[10] A. Newell, P.S. Rosenbloom and J.E. Laird, Symbolic architectures for cognition, in: M. Posner, ed., *Foundations of Cognitive Science* (MIT Press, Cambridge, MA, 1989).
[11] A. Newell, G. Yost, J.E. Laird, P.S. Rosenbloom and E. Altmann, Formulating the problem-space computational model, in: R. Rashid, ed., *Carnegie Mellon Computer Science: A 25-Year Commemorative* (ACM Press/Addison-Wesley, Reading, MA, 1992).
[12] Z.W. Pylyshyn and W. Demopoulos, eds., *Meaning and Cognitive Structure* (Ablex, Norwood, NJ, 1986).
[13] A. Sloman and M. Croucher, Why robots will have emotions, in: *Proceedings IJCAI-81*, Vancouver, BC (1981).
[14] B.C. Smith, The link from symbol to knowledge, in: Z.W. Pylyshyn and W. Demopoulos, eds., *Meaning and Cognitive Structure* (Ablex, Norwood, NJ, 1986) 40–50.