Cognitive Architectures

Paul Thagard
University of Waterloo
pthagard@uwaterloo.ca

Thagard, P. (2012). Cognitive architectures. In K. Frankish & W. Ramsay (Eds.), *The Cambridge handbook of cognitive science* (pp. 50-70). Cambridge: Cambridge University Press.

A cognitive architecture is a general proposal about the representations and processes that produce intelligent thought. Cognitive architectures have primarily been used to explain important aspects of human thinking such as problem solving, memory, and learning. But they can also be used as blueprints for designing computers and robots that possess some of the cognitive abilities of humans. The most influential cognitive architectures that have been developed are either rule-based, using if-then rules and procedures that operate on them to explain thinking, or connectionist, using artificial neural networks. This chapter will describe the central structures and processes of these two kind of architectures, and review how well they succeed as general theories of mental processing. I argue that advances in neuroscience hold the promise for producing a general cognitive theory that encompasses the advantages of both rule-based and connectionist architectures.

What is an explanation in cognitive science? In keeping with much recent philosophical research on explanation, I maintain that scientific explanations are typically descriptions of mechanisms that produce the phenomena to be explained (Bechtel and Abrahamsen, 2005; Machamer, Darden and Craver, 2000). A mechanism is a system of related parts whose interactions produce regular changes. For example, to explain how a bicycle works, we describe how its parts such as the pedals, chain, and wheels are connected to each other and how they interact to produce the movement of the bike.

December 12, 2012

Similarly, explanation in physics, chemistry, and biology identifies relevant parts such as atoms, molecules, and cells and describes how they interact to produce observed changes in things and organisms. Explanations in cognitive science are typically mechanistic in that they describe how different kinds of thinking occur as the result of mental representations (parts) operated on by computational procedures (interactions) that change mental states.

A cognitive architecture is a proposal about the kinds of mental representation and computational procedure that constitute a mechanism for explaining a broad range of kinds of thinking. A complete unified general theory of cognition would provide mechanisms for explaining the workings of perception, attention, memory, problem solving, reasoning, learning, decision making, motor control, language, emotion, and consciousness. Let us now review the history of cognitive architectures.

Brief History of Cognitive Architectures

The term "cognitive architecture" developed from the idea of a computer architecture, which originated with a description of the first widely used computer, the IBM 360 (Amdahl, Blaaw, and Brooks, 1964). A computer architecture is the conceptual structure and functional behavior of a system as seen by a programmer, not the computer's physical implementation. John Anderson's 1983 book, *The Architecture of Cognition* was the main text that introduced the term "cognitive architecture", defined (p. ix) as a "the basic principles of operations of a cognitive system". That book describes the ACT architecture, which is a synthesis of Anderson's earlier ideas about propositional memory with previous ideas about rule-based processing. The idea of a cognitive architecture was already implicit in the rule-based information processing theories of

Newell and Simon (1972). Allan Newell further popularized the idea in his 1990 book, *Unified Theories of Cognition,* which described his work with John Laird and Paul Rosenbloom on a particular rule-based architecture, SOAR (Rosenbloom, Laird, and Newell, 1993). Rule-based systems were originally used by Newell and Simon to explain problem solving, but later work has applied them to account for a much broader range of psychological phenomena, including memory and learning. The rule-based approach continues to thrive in ongoing research by proponents of ACT, SOAR, and related cognitive architectures; for more references, see the discussion below of psychological applications of rule-based systems.

Rule-based systems are not the only way to think about cognition. In the 1970s, researchers such as Minsky (1975) and Schank and Abelson (1977) proposed a different way of understanding cognition as involving the matching of current situations against concept-like structures variously called frames, schemas, scripts, and prototypes. On this view, the fundamental kind of mental representation is a schema that specifies what holds for a typical situation, thing, or process. Proponents of schemas have used them to explain such phenomena as perception, memory, and explanation. For example, you understand what happens when you go out to eat by applying your restaurant schema, which specifies the typical characteristics of restaurants. However, schema-based systems have not survived as general theories of cognition, although they have been included in hybrid systems that use both rules and schemas such as PI, which models aspects of scientific reasoning such as discovery and explanation (Thagard, 1988).

Another supplement to the rule-based approach involves analogical reasoning, in which problems are solved not by the application of general rules but by the matching of

a stored mental representation of a previous case against a description of the problem to be solved. For example, you might understand a new restaurant by comparing it to a highly similar restaurant that you have previously experienced, rather than by using a general schema or rule. Although analogical reasoning has been much discussed in psychology (Holyoak and Thagard, 1995), and in artificial intelligence under the term "case-based" reasoning (Kolodner, 1993), it is implausible to base a whole cognitive architecture on just schema-based or case-based reasoning.

The major alternative to rule-based cognitive architectures emerged in the 1980s. Neural network models of thinking had been around since the 1950s, but they only began to have a major impact on theorizing about the mind with the development of the PDP (parallel distributed processing) approach (Rumelhart and McClelland, 1986). This approach is also called *connectionism*, because it views knowledge as being encoded, not in rules, but via the connections between simple neuron-like processors. More details will be provided below about how such processors work and how connectionist architectures differ from rule-based architectures. Connectionism has been applied to a broad range of psychological phenomena ranging from concept learning to high-level reasoning. Like rule-based cognitive architectures, connectionist ones are a thriving intellectual industry, as seen for example in the applications to categorization and language found in Rogers and McClelland (2004) and Smolensky and LeGendre (2006). We can conduct a more systematic comparison of rule-based and connectionist approaches to explaining cognition by reviewing what they say about representations and procedures.

Representations

Since its origins in the mid-1950s, cognitive science has employed a fundamental hypothesis, that thinking is produced by computational procedures operating on mental representations. However, there has been much controversy about *what* kind of representations and *what* kind of procedures are best suited to explain the many varieties of human thinking. I will not attempt to review all the different versions of rule-based and connectionist architectures that have been proposed. Instead, I will provide an introduction to the representations and procedures used by rule-based and connectionist systems by showing how they can deal with a familiar area of human thinking: personality and human relations.

In thinking about all the people you know, you employ a familiar set of concepts, describing them as kind or cruel, intelligent or dumb, considerate or self-centered, polite or crude, outgoing or antisocial, confident or fearful, adventurous or cautious, conscientious or irresponsible, agreeable or difficult, and so on. Rule-based and connectionist approaches offer very different pictures of the nature of these concepts. From a rule-based perspective, your knowledge about other people consists of a set of rules, that can be stated as if-then structures. For example, here are some rules that might capture part of your knowledge about kindness, letting P stand for any person.

If P is kind, then P helps other people.

If P is kind, then P cares about other people.

If P is kind, then P is not cruel.

If P cares about other people and helps other people, then P is kind.

If P has the goal of being kind, then P should think about the feelings of others.

If P is cruel, then avoid P.

As an exercise you should try to write down rules for a few other social concepts such as *outgoing* and *polite*. Unless you find it terribly difficult to construct such rules, you should find it plausible that the representations in your mind of social concepts consist of rules.

Connectionist cognitive architectures propose a very different kind of mental representation. As a first approximation, we can think of a concept as a node in a network that is roughly analogous to networks of neurons in the brain. Figure 1 shows a very simple network that has a few nodes for the concepts *kind, cruel,* and *mean*. But these concepts are not related by if-then rules that employ word-like symbols, but instead by simple connections that can be either positive or negative, just as neurons in the brain are connected by synapses that enable one neuron to either excite or inhibit another. The network in figure 1 uses a kind of representation called *localist*, which means that each concept is represented by a single neuron-like node.

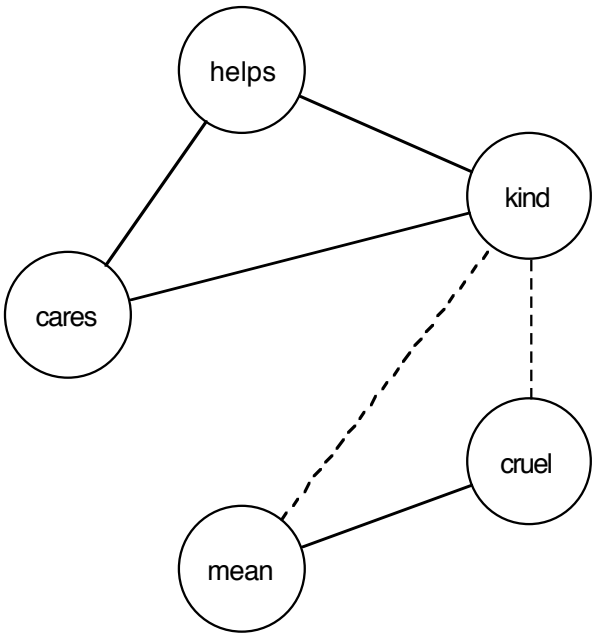

Figure 1. Localist network showing some of the connections between social concepts. The solid lines indicate excitatory links and the dotted lines indicate inhibitory links. Links in this network are symmetric, that is, they run in both directions.

Much more radically, connectionism can represent concepts by *distributed* representations that use many nodes for each concept. Figure 2 shows a typical three-layer network that consists of an input layer of simple features and an output layer of concepts, with an intervening layer called *hidden* because it is neither input nor output. As in the localist network in figure 1, the nodes are connected by links that are positive or negative depending on how the network is trained. Whereas if-then rules and localist connections are typically specified in advance, connections in a distributed representation are usually learned by experience. I will say more about how such networks are trained in the section below about procedures. The key point to note now is that a concept such as *cruel* is not the single node in the output layer, nor any simple rule connecting the input and output layers. Rather, it is a whole pattern of connections involving the input, output, and hidden layers; the nodes in the hidden layer do not need to acquire any specific interpretation. Neural networks in the brain are much more complicated than the simple three-layer network in figure 2, but they share the property that representation of concepts is distributed across many neurons.

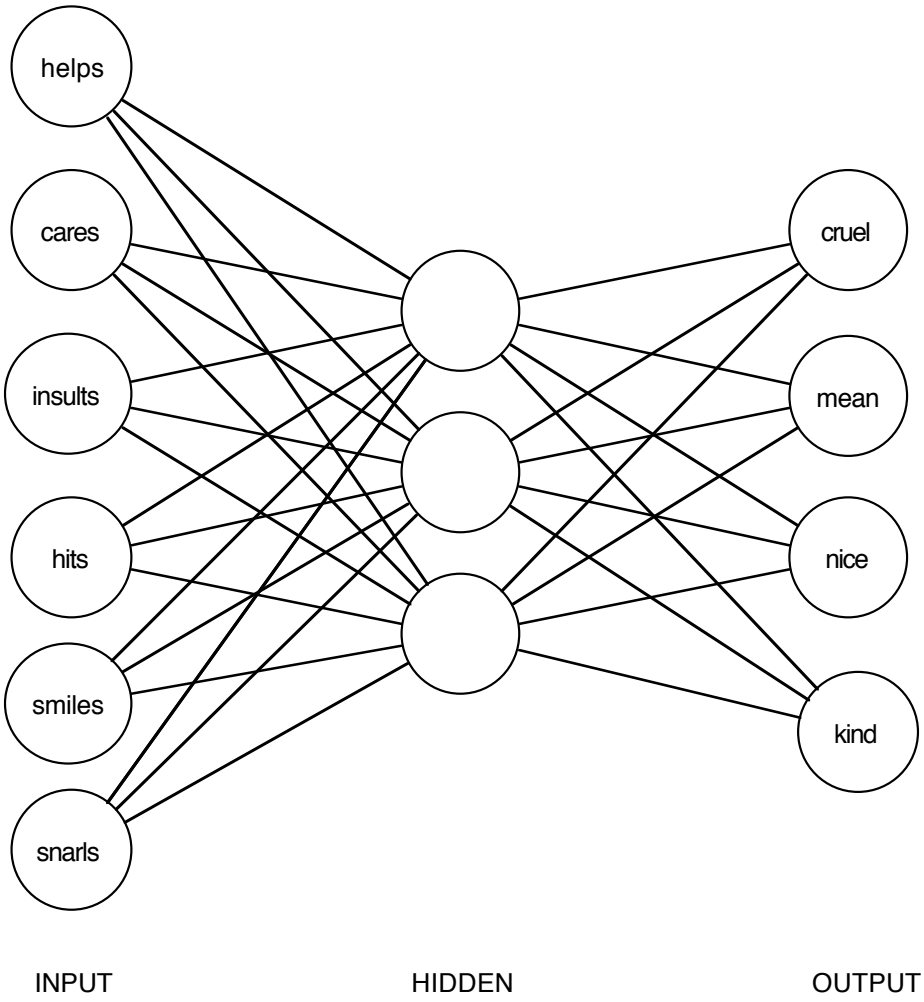

Figure 2. Distributed representation of social concepts. The links are not symmetric, but feed activation forward from left to right. Weights on the links are learned by training.

To summarize, social and other concepts in a rule-based cognitive architecture primarily consist of sets of if-then rules, but in a connectionist architecture concepts are patterns of connections between nodes in a network, including hidden nodes that by themselves do not have any specific interpretation. Rather, they serve by virtue of their links to input and output layers to furnish a statistical connection between inputs and outputs that is often hard to characterize in words and is rarely replaceable by general if-

then rules. To appreciate fully the difference between rule-based and connectionist representations, it is crucial to notice how they support different kinds of procedures for reasoning, problem solving, and learning.

Rule-based Procedures

Just as you cannot make a cake without doing things to the ingredients, you cannot think without mental procedures that operate on your representations. For rule-based systems, the simplest kind of procedure is the obvious one where you match the IF part against something you know and then fill in the THEN part. For example, you might make the following inference:

If P cares about other people and helps other people, then P is kind.

Sandra cares about other people and helps other people.

Therefore, Sandra is kind.

In a computational model of a rule-based system, this sort of inference is made by having a list of facts of current interest, such as that Sandra cares about other people, in addition to a large set of rules that encapsulate information about social concepts. Here is the main forward procedure performed by a cognitive architecture based on rules:

1. Match what is currently known (the facts) against a database of rules.
2. If the facts match the IF parts of a rule, then infer the THEN part.
3. Repeat.

The repetition is crucial, because a rule-based system usually needs to make a whole series of inferences to come to an interesting conclusion. For example, having inferred that Sandra is kind, we could then use the rule *if P is kind then P is not cruel* to infer that

Sandra is not cruel. Thus if-then rules can be chained together to produce complex inferences.

Often it is useful to chain rules backwards instead of forward in order to answer questions or solve problems. Suppose, for example, your aim is to answer the question whether Sandra is cruel and you want to find rules that can answer it. You can then work backwards using the following procedure:

1. Match want you want to know (the goals) against a database of rules.
2. If the goal matches the THEN part of a rule, then add the IF part to the set of goals.
3. Repeat.

This procedure may enable you to chain backward from the goals you want to accomplish to find aspects of the current situation that would identify the information you need to then chain forward to provide an answer to your question or a solution to your goal. For example, generating the goal to determine if Sandra is cruel may lead you to retrieve rules such as *If P insults people then P is cruel* that can then spur you to ask whether Sandra insults people.

Thus a rule-based system accomplishes reasoning and problem solving by forward or backward chaining using a sequence of rules. To make such reasoning psychologically effective, other procedures are needed for retrieving rules from memory, resolving conflicts between competing rules, and learning new rules. First, consider retrieval from memory. My description of the procedures for forward and backward chaining assumed that there is an accessible list of relevant rules, but an educated adult has accumulated many thousands of rules constituting the thousands of concepts that have been acquired. (It has been estimated that the typical vocabulary of an adult is

more than 100,000 words, so there must be least this many concepts and rules.) It would be too slow and awkward to match thousands of rules one by one against the rule-based system's list of known facts or goals to be solved. Hence there needs to be a procedure to ensure that the matching is only done against a set of rules somehow selected to be potentially relevant. Anderson's (1983) ACT architecture uses spreading activation among the constituents of rules, facts, and goals as a way to select from memory a set of rules that appear relevant for matching. For example, if the concepts *cruel* and *insult* are associated in your memory because of your previous experiences, then activating one of them can lead to the activation of the other, making available a new set of relevant rules.

Second, additional procedures are needed to determine what rules to apply in cases where they provide conflicting answers. Suppose you want to determine whether Solomon is outgoing, and you have the following rules in your memory base:

If P likes to go to parties, then P is outgoing.

If P likes to read a lot of books, then P is not outgoing.

If you know that Solomon likes to go to parties *and* to read lots of books, your rules suggest that you should infer that Solomon is both outgoing and not outgoing. To resolve this conflict, which is even more acute when the THEN part of the rules suggests incompatible actions such as both talking to someone and walking away, there needs to be a procedure to select which rules apply best to the problem situation. Procedures that have been used in various cognitive architectures include using rules that are most specific to the current situation and using rules that have been highly successful in past problem solving episodes.

The third sort of procedure that is important for rule-based cognitive architectures involves learning new rules and new strategies for solving problems more effectively. How did you acquire rules like *If P is kind, then P helps homeless people?* This rule is not part of the central meaning of the concept *kind,* so it is unlikely that you were simply told it as part of learning what kindness is. Instead, you may have learned it from experience, seeing a collection of examples of people who are both kind and help homeless people, producing a new rule by generalization. Another way of acquiring a rule is by stringing together other rules you already have, perhaps reasoning as follows:

If P is kind, then P cares about people.

If P cares about people, then P helps homeless people.

So: If P is kind, then P helps homeless people.

Here a new rule is acquired by combining two or more other rules. In sum, rule-based architecture can have various procedures for learning new rules, including being given the rule, generalizing from experience, and compiling new rules from previous rules.

Thus rule-based systems can employ many powerful procedures for problem solving and learning: forward and backward chaining, retrieval by spreading activation, conflict resolution, and generation of new rules.

Connectionist Procedures

Connectionist cognitive architectures have very different methods for reasoning and learning. In rule-based systems, problem solving consists primarily of using rules to *search* a space of possible actions. In contrast, the connectionist perspective conceives of problem solving as *parallel constraint satisfaction*. Suppose your problem is to categorize someone as either kind or cruel, perhaps as part of a hiring decision.

Instead of using rule-based reasoning, you might apply the kind of network shown in figure 1. The excitatory links in the network represent positive constraints, factors that tend to go together, such as being kind and helping others. The inhibitory links represent negative constraints, factors that tend not to go together, such as being kind and being cruel. The inference problem here is to figure out the best way to satisfy the most constraints, which is done in parallel by spreading activation through the network. Activation is a property of each node in the network, roughly analogous to the firing rate of a neuron (how many times it fires per second compared to how fast it could fire). Activation of a node represents the acceptability of the representation to which the node corresponds. Just as the brain operates by parallel activity of multiple neurons, constraint satisfaction in a neural network should be a parallel process that takes into account all relevant constraints simultaneously.

Here is an outline of the procedure used to solve a constraint satisfaction problem in connectionist fashion:

1. Express the problem as a set of nodes connected by excitatory and inhibitory links.
2. Establish the givens of the problem as inputs to some of the nodes.
3. Spread activation among the nodes based on their excitatory and inhibitory inputs, until the network settles, i. e. all nodes have reached stable activation.
4. Read off the networks solution to the problem as represented by the nodes that have highest activation.

For example, the network shown in figure 1, with inputs from the evidence that a person helps others and cares about them, will settle with the node for *kind* having high activation and the node for *cruel* having low activation. The next section lists many

other kinds of problems that can be solved by parallel constraint satisfaction, from decision making to vision to language comprehension.

In the connectionist procedure I just sketched for solving parallel constraint satisfaction problems, the links between the nodes are given, but how might they be learned? Moreover, how do the nodes in networks with distributed representations like those in figure 2 acquire meaning? The most common connectionist procedure used to learn weights is called *backpropagation*, because it propagates errors back from output nodes to adjust all the weights in the network. Here is a simple description of how backpropagation works:

1. Assign weights randomly to all the connections in the network.
2. Provide inputs to the input units, feed activation forward through the network, and see whether the outputs produced are correct.
3. If the outputs are wrong, then change the weights that produced them, including weights between the input and hidden layer and between the hidden and output layer.
4, Repeat with many input examples until the network has acquired the desired behavior.

This procedure is a kind of supervised learning, in that it requires telling the network whether it is getting the right answer. There are also learning procedures for artificial neural networks that do not require a supervisor. The simplest is one proposed by Hebb (1949) that has been found to operate in real neural networks: if two neurons are connected and they fire at the same time, then increase the strength of the connection between them; whatever fires together, wires together. More complicated procedures for

unsupervised learning using an internal model of the task to be completed have also been developed.

To sum up, connectionist networks make inferences and solve problems by parallel constraint satisfaction, and they learn to improve their performance by procedures that adjust the weights on the links between nodes. I will now review some of the many psychological applications that have been found for rule-based and connectionist cognitive architectures.

Psychological Applications

Both rule-based and connectionist architectures embody powerful theories about the representations and procedures that explain human thinking. Which cognitive architecture, which theory of thinking, is the best? There have been many great battles in the history of science between competing theories, for example heliocentric Copernican astronomy vs. Ptolemy's geometric theory, the wave theory of light vs. particle theories, and Darwin's theory of evolution vs. creationism. These battles are adjudicated by evaluating how well the competing theories explain all the relevant evidence.

Both rule-based and connectionist architectures have had many impressive applications to psychological phenomena. Table 1 shows that rule-based architectures have had explanatory successes in many psychological domains, especially problem solving and language. Table 2 shows that connectionism has also done very well in generating explanations. Which kind of cognitive architecture is the best explanation of the full range of psychological phenomena? Neither tables 1 and 2 nor the very large additional literature espousing these two approaches establishes a winner. I see no

immediate prospect of one of the two kinds of cognitive architecture superseding the other by showing itself capable of explaining everything that the other one does in addition to what it currently explains. Moreover, there are some aspects of thinking such as consciousness that have largely been neglected by *both* approaches.

The current battle between rule-based and connectionist architecture is analogous to a previous episode in the history of science, the controversy between wave and particle theories of light. From the seventeenth through the nineteenth centuries, there was an oscillation between the wave theory, advocated by scientists such as Huygens and Young, and the particle theory, advocated by Gassendi and Newton. The battle was only settled in the twentieth century by the advent of quantum theories of light, according to which light consists of photons that exhibit properties of *both* particles and waves. Similarly, I think that the most reasonable conclusion from the current impasse of rule-based and connectionist architectures is that the mind is both a rule-based and a connectionist system, and that problem solving can sometimes be search through a space of rules and sometimes parallel constraint satisfaction.

Domains	*Applications*	*References*
Problem solving	Domains such as logic and chess	Newell and Simon (1972), Newell (1990)
	Human-computer interaction	Kieras and Meyer (1997)
	Perceptual-motor system	Anderson et al. (2004)
Learning	Arithmetic procedures	Anderson (1983)
	Scientific discovery	Langley et al. (1987), Thagard (1988)
	Skill acquisition	Newell (1990)
	Tutoring	Anderson (1993)
	Induction	Holland et al., (1986)
Language	Acquisition	Anderson (1983), Pinker (1989)
	Regular and irregular verbs	Pinker (1999)
Reasoning	Syllogisms	Newell (1990)
	Statistical heuristics	Nisbett (1993)
Memory	List memory	Anderson et al. (1998)
Explanation	Hypothesis generation	Thagard (1988)

Emotion	Cognitive appraisal	Scherer (1993)

Table 1. Selection of psychological phenomena that can be explained by processing of rules.

Domains	*Applications*	*References*
Vision	Stereoscopic vision Figure interpretation Visual expectation	Marr and Poggio (1976) Feldman (1981) Bressler (2004)
Language	Letter perception Discourse comprehension Irony Grammar Semantic cognition	McClelland and Rumelhart (1981) Kintsch (1998) Shelley (2001) Smolensky and Legendre (2006) Rogers and McClelland (2004)
Concepts	Schema application Impression formation	Rumelhart, et al. (1986) Kunda and Thagard (1996)
Analogy	Mapping and retrieval	Holyoak and Thagard (1989, 1995)
Explanation	Theory evaluation Social explanations	Thagard (1992, 2000) Read and Marcus-Newhall (1993)
Social behavior	Cognitive dissonance Personality Social perception Attitude change	Shultz and Lepper (1996) Shoda and Mischel (1998) Read and Miller (1998) Spellman, Ullman, and Holyoak (1993)
Decision	Plan selection Preference construction	Thagard and Millgram (1995) Simon, Krawcyyck, and Holyoak (2004)
Emotion	Appraisal and inference	Nerb and Spada (2001), Thagard (2000, 2006)

Table 2. Selection of psychological phenomena that can be explained by parallel constraint satisfaction.

Neural Architecture

How can the brain be both a rule-based and a connectionist system? It might seem that connectionism has a head start in taking into account knowledge about the brain, given that its parallel processing seems to employ a kind of brain-style computation. But there are many respects in which connectionist cognitive architectures

have not accurately captured how the brain works. First, at the level of individual neurons, connectionist models usually describe neural activity in terms of activation, understood as the rate of firing. But there are both neurological and computational reasons to think that it matters that neurons show particular patterns of spiking (Maass and Bishop, 1999; Rieke et al., 1997). Imagine a neuron whose firing rate is 50 times per second. Such a rate is consistent with many very different patterns of firing, for example (FIRE REST FIRE REST …) versus (FIRE FIRE REST REST …). Biologically realistic neural networks encode information using spiking patterns, not just rates of firing. A population of neurons can become tuned to a set of stimuli such as faces by acquiring synaptic connections that generate different spiking patterns.

Second, neural networks are not simply electrical systems, sending charges from one neuron to another; they are also chemical systems employing dozens of neurotransmitters and other molecules to carry out signaling in complex ways. Important neurotransmitters include glutamate for excitatory connections, GABA for inhibitory connections, and dopamine for circuits that evaluate the reward potential of stimuli. A single synaptic connection can involve multiple neurotransmitters and other chemicals operating at different time scales (Leonard, 1997).

Third, the brain should not be thought of as one big neural network, but as organized into areas that have identifiable functions. For example, the occipital area at the back of your head is the main visual processing center. The prefrontal cortex, the part of your brain roughly behind your eyes, is important for high-level reasoning and language. More specifically, the ventromedial (bottom-middle) prefrontal cortex facilitates decision making by providing connections between high-level reasoning in the

dorsolateral (top-sides) prefrontal cortex and emotional reactions in the amygdala, which lies below the cortex. Hence traditional connectionist models are typically not biologically realistic either at the level of individual neurons or at the level of brain organization.

There is, however, a wealth of current research aimed at producing more biologically realistic models of cognitive processes. Whether these models should currently be called "cognitive architectures" is not clear, because they have mostly been applied to low-level kinds of cognition such as perception and memory, rather than to high-level kinds of inference such as problem solving. But these models have the potential to develop into broader accounts of human thinking that I hope will supersede the current apparent conflict between rule-based and connectionist approaches. Table 3 points to the work of five researchers in theoretical computational neuroscience who are pursuing promising directions.

Researcher	*Applications*	*Sample publications*
Jonathan Cohen, Princeton University	Decision making, attention, categorization	Miller & Cohen (2001)
Chris Eliasmith, University of Waterloo	Perception, memory, motor control	Eliasmith and Anderson (2003),
Stephen Grossberg, Boston University	Perception, attention, learning	Carpenter and Grossberg (2003),
Randy O'Reilly, University of Colorado	Learning, memory, attention	O'Reilly, R. C., & Munakata, Y. (2000).
Terry Sejnowski, University of California-San Diego	Learning, memory, motor control	Quartz and Sejnowski (2002),

Table 3. Some prominent work in the emerging field of theoretical neuroscience, which develops biologically realistic computational models of cognition.

Research in theoretical neuroscience along the lines of table 3 is highly technical, and I will not attempt to summarize the similarities and differences among the various researchers. Instead, I will return to my previous example and indicate how concepts such as *kind* and *cruel* might be represented in a more biologically realistic fashion than is possible in rule-based and connectionist cognitive architectures. Eliasmith (2003) provides a more specific argument about the advantages of theoretical neuroscience for going beyond the limitations of rule-based and connectionist approaches.

Concepts in human brains are represented in a distributed fashion across multiple neurons, just as in the parallel distributed processing version of connectionism. Whereas connectionist models distribute a concept such as *kind* across a small number of closely attached units, a more biologically realistic model would have thousands or millions of spiking neurons distributed across multiple brain areas. Using spiking neurons has the computational advantage of making it possible to model the dynamic properties of neural networks such as temporal coordination of different neural populations. Moreover, in some models (e.g. ones by Cohen and O'Reilly) the role of particular neurotransmitters such as dopamine can be emphasized. Dopamine is associated with positive emotional reactions, so it is likely involved in the fact that the concept of kindness is for most people a positive one. When you think of someone as kind, you usually have a positive feeling toward them, whereas applying the concept *cruel* prompts negative emotions for most people. Thus theoretical neuroscience is developing models that take into account the spiking and chemical properties of neurons.

In addition, theoretical neuroscience can describe the contributions to the representation of a concept from different brain areas. The semantic characteristics of

kind and *cruel* that are captured by approximate rules describing the behavior of people are probably represented in the prefrontal cortex, which plays a large role in reasoning. But other brain areas are likely involved too, for example the primary visual cortex which would be activated if you created a mental image of a person being kind or cruel, perhaps by kicking a homeless person. Some concepts, e. g. *automobile,* are closely tied to specific modalities such as vision (Barsalou et al., 2003). Moreover, the emotional component of concepts such as *kind* and *cruel* suggests the involvement of brain areas that are known to be active in positive emotions (e.g. the nucleus accumbens, which is tied to various pleasurable activities) and negative emotions (e.g. the insula which has been found to be active in both physical and social pain). Thagard and Aubie (2008) show how satisfaction of both cognitive and emotional constraints can be performed in a neurally plausible manner. In sum, from the perspective of theoretical neuroscience, a concept is a pattern of spiking and chemical behaviors in a large population of neurons distributed across multiple brain areas.

Rule-based models have also been moving in the direction of greater neurological plausibility. John Anderson and his colleagues have used brain scanning experiments to relate the ACT system to specific brain regions such as the prefrontal cortex, used for memory and matching of rules against facts, and the basal ganglia, used for the implementation of production rules (Anderson et al., 2004). Other brain areas they postulate to be involved in the matching and firing of rules include the striatum for selection of rules and parts of the prefrontal cortex for memory buffers. Thus rule-based cognitive architectures are becoming neural architectures, just as connectionist approaches are giving way to computational neuroscience.

Earlier, I mentioned the great synthesis accomplished by the quantum theory of light, according to which light consists of photons, which have properties of both particles and waves. A similar synthesis has yet to occur in cognitive science, as no one has figured out fully how to blend the ideas emerging from theoretical neuroscience about the behavior of spiking chemical neurons in multiple brain areas with the more high-level behavior of neurally grounded rule-based systems. Among the exciting new ideas are mathematical ways of showing how artificial neurons can implement features of rules such as their complex symbolic structure (e.g. Smolensky and Legendre, 2006). My hope is that a grand synthesis will be accomplished by identifying how neural mechanisms that are well suited for low-level operations such as perception can also serve to support high-level kinds of symbolic inferences. Such a synthesis will show that the competition that raged in the 1980s and 1990s between rule-based and connectionist cognitive architectures was merely a prelude to deep reconciliation by virtue of a unified theory of neural mechanisms.

Accomplishment of this synthesis would not eliminate the usefulness of rule-based and connectionist cognitive models, although it would undercut their claims to be universal cognitive architectures. Cognitive theories are claims about the mental representations and computational procedures that produce different kinds of thinking. Computational models implemented as running programs provide simplified approximations of such representations and procedures. Scientific models are like maps, in that different ones can be useful for different purposes. Just as you use different scales of maps depending on whether your aim is to travel across the city or to travel across the country, so different kinds of model are useful for explaining different aspects

of human thinking. A full model of the brain, encompassing all of its billions of neurons, trillions of synapses, and hundreds of chemicals would be as useless as a map of a country that was the same size and detail as the country itself. A cognitive or neural theory does not aim to describe everything about thought and the brain, but rather to describe the mechanisms that underlie the fundamental causal processes most relevant to explaining those aspects of thinking we find most interesting. Simplifications of the sort provided by rule-based and connectionist models will remain useful for explaining particular phenomena at comprehensible levels of detail. Current rule-based and connectionist models successfully capture many aspects of thinking, particularly sequential problem solving and parallel constraint satisfaction. Hence it will continue to be methodologically legitimate to employ them, even if it becomes established that the ultimate cognitive architecture is provided by theoretical neuroscience.

If principles of neuroscience make possible the unification of rule-based and connectionist explanations under a common framework, then they should also serve to bring into a single theoretical fold other aspects of cognition that have been discussed using different theoretical ideas. For example, it would be theoretically exciting to integrate ideas about probabilistic inference into a general framework that also applies to rule-based and connectionist processing. Reasoning involving mental models, which are rich psychological representations of real or imagined situations, should also be incorporated. Then cognitive science would have the sort of unifying theory that relativity and quantum theories provide to physics and that evolution and genetics provide to biology. Such a grand unification may, however, requires decades or even centuries.

Artificial Intelligence

If it turns out that the deepest cognitive architecture is furnished by theoretical neuroscience, what are the implications for artificial intelligence? When cognitive science began in the mid-1950s and got named and recognized as an interdisciplinary field in the mid-1970s, there was a common perception that psychology and artificial intelligence were natural allies. A unified cognitive architecture based on rules or other sorts of representation would provide a way simultaneously of understanding how human minds work and how computers and robots can be made to work in comparable ways. The reconceptualization of cognitive architectures as neural architectures raises the possibility that what kind of hardware an intelligent system is running on matters much more than the pioneers of cognitive science realized. Compared to computers, whose chips can perform operations more than a billion times per second, a neuron looks hopelessly slow, typically firing only around a hundred times per second. But we have billions of neurons, with far more biological and chemical complexity than research on simple neural networks has recognized. There are thousands of different kinds of neurons adapted for different purposes, and each neuron has thousands of chemical connections to other neurons that allow many kinds of chemical modulation as well as transmission of electrical impulses. The best way to get a computer to do things that are intelligent may be to develop software more suited to the extraordinary speed and lack of evolutionary history of its central processing unit. Then there will be a bifurcation of cognitive architectures into ones best suited for operating with the messy biological hardware of the brain and those best suited for operating with digital

processing. Langley (2006) provides a thorough discussion of the role of cognitive architectures in artificial intelligence.

Another possibility besides bifurcation is that there will be a set of statistical principles that describe how both brains and intelligent machines operate in the world. Perhaps there is a convergence between the recent trend in neuroscience to describe what brains do as a kind of Bayesian statistical inference (Doya et al., 2007) and the major trend in artificial intelligence and robotics to approach problems statistically using Bayesian inference mechanisms (Thrun, Burgard, and Fox, 2005). Bayesian inference is a way of evaluating a hypothesis about what is going on in an environment by mathematically taking into account the prior probability of the hypothesis, the probability of the evidence given the hypothesis, and the probability of the evidence. Perhaps then, at some level, both the human brain and digital computers can be viewed as engines for statistical inference. It remains to be seen whether that level will be the most fruitful for understanding human and artificial intelligence.

Conclusion

This chapter has reviewed the two main current approaches to cognitive architecture: rule-based systems and connectionism. Both kinds of architecture assume the central hypothesis of cognitive science that thinking consists of the application of computational procedures to mental representations, but they propose very different kinds of representations and procedures. Rule-based systems apply procedures such as forward chaining to if-then representations with word-like symbols, whereas connectionist systems apply procedures such as parallel activation adjustment to representations comprised of neuron-like units with excitatory and inhibitory connections

between them. Both rule-based and connectionist architectures have had many successes in explaining important psychological phenomena concerning problem solving, learning, language use, and other kinds of thinking. Given their large and only partially overlapping range of explanatory applications, it seems unlikely that either of the two approaches to cognitive architecture will come to dominate cognitive science. I suggested an alternative scenario, consistent with current developments in both rule-based systems and connectionist modeling, that will see a reconciliation of the two approaches by means of theoretical neuroscience. Unified understanding of how the brain can perform both serial problem solving using rules and parallel constraint satisfaction using distributed representations will be a major triumph of cognitive science.

Further Reading

Thagard (2005) gives an accessible introduction to approaches to mental representation and computation. Boden (2006) provides a review of the history of different approaches to cognitive science. For rule-based systems, Newell (1990) is a good introduction. For connectionism, see Bechtel and Abrahamsen (2002). Dayan and Abbott (2001) provide an introduction to theoretical neuroscience.

References

Amdahl, G. M., Blaauw, G. A., & Brooks, L. R. (1964). Architecture of the IBM System/360. *IBM Journal of Research and Development, 8*(2(April)), 87-101.

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

Anderson, J. R. (1993). *Rules of the mind*. Hillsdale, NJ: Erlbaum.

Anderson, J. R., Bothell, D., Byrne, M. D., Douglas, S., Lebiere, C., & Qin, U. (2004). An integrated theory of the mind. *Psychological Review, 111*, 1030-1060.

Anderson, J. R., Bothell, D., Lebiere, C., & Matessa, M. (1998). An integrated theory of list memory. *Journal of Memory and Language, 38*, 341-380.

Barsalou, L. W., Simmons, W. K., Barbey, A. K., & Wilson, C. D. (2003). Grounding conceptual knowledge in modality-specific systems. *Trends in Cognitive Sciences, 7*, 84-91.

Bechtel, W., & Abrahamsen, A. A. (2002). *Connectionism and the mind: Parallel processing, dynamics, and evolution in networks* (2nd ed.). Oxford: Basil Blackwell.

Bechtel, W., & Abrahamsen, A. A. (2005). Explanation: A mechanistic alternative. *Studies in History and Philosophy of Biology and Biomedical Sciences, 36*, 421-441.

Boden, M. (2006). *Mind as machine: A history of cognitive science*. Oxford: Oxford University Press.

Bressler, S. L. (2004). Inferential constraint sets in the organization of visual expectation. *Neuroinformaics, 2*, 227-237.

Carpenter, G. A., & Grossberg, S. (2003). Adaptive resonance theory. In M. A. Arbiib (Ed.), *Handbook of brain theory and neural networks*. Cambridge, MA: MIT Press.

Dayan, P., & Abbott, L. F. (2001). *Theoretical neuroscience: Computational and mathematical modeling of neural systems*. Cambridge, MA: MIT Press.

Doya, K., Ishii, S., Pouget, A., & Rao, A. (Eds.). (2007). *Bayesian brain*. Cambridge, MA: MIT Press.

Eliasmith, C. (2003). Moving beyond metaphors: Understanding the mind for what it is. *Journal of Philosophy, 100*, 493-520.

Eliasmith, C., & Anderson, C. H. (2003). *Neural engineering: Computation, representation and dynamics in neurobiological systems*. Cambridge, MA: MIT Press.

Feldman, J. A. (1981). A connectionist model of visual memory. In G. E. Hinton & J. A. Anderson (Eds.), *Parallel models of associative memory* (pp. 49-81). Hillsdale, NJ: Erlbaum.

Hebb, D. O. (1949). *The organization of behavior*. New York: Wiley.

Holland, J. H., Holyoak, K. J., Nisbett, R. E., & Thagard, P. R. (1986). *Induction: Processes of inference, learning, and discovery*. Cambridge, MA: MIT Press/Bradford Books.

Holyoak, K. J., & Thagard, P. (1989). Analogical mapping by constraint satisfaction. *Cognitive Science, 13*, 295-355.

Holyoak, K. J., & Thagard, P. (1995). *Mental leaps: Analogy in creative thought*. Cambridge, MA: MIT Press/Bradford Books.

Kieras, D. E., & Meyer, D. E. (1997). An overview of the EPIC architecture for cognition and performance with application to human-computer interaction. *Human-Computer Interaction, 12*, 391-438.

Kintsch, W. (1998). *Comprehension: A paradigm for cognition*. Cambridge: Cambridge University Press.

Kolodner, J. (1993). *Case-based reasoning*. San Mateo, CA: Morgan Kaufmann.

Kunda, Z., & Thagard, P. (1996). Forming impressions from stereotypes, traits, and behaviors: A parallel-constraint-satisfaction theory. *Psychological Review, 103*, 284-308.

Langley, P. (2006). Cognitive architectures and general intelligent systems. *AI Magazine, 27*(2), 33-44.
Langley, P., Simon, H., Bradshaw, G., & Zytkow, J. (1987). *Scientific discovery*. Cambridge, MA: MIT Press/Bradford Books.
Leonard, B. E. (1997). *Fundamentals of pharmacology* (2nd ed.). Chichester: John Wiley.
Maass, W., & Bishop, C. M. (Eds.). (1999). *Pulsed neural networks*. Cambridge, MA: MIT Press.
Machamer, P., Darden, L., & Craver, C. F. (2000). Thinking about mechanisms. *Philosophy of Science, 67*, 1-25.
Marr, D., & Poggio, T. (1976). Cooperative computation of stereo disparity. *Science, 194*, 283-287.
McClelland, J. L., & Rumelhart, D. E. (1981). An interactive activation model of context effects in letter perception: Part 1: An account of basic findings. *Psychological Review, 88*, 375-407.
Miller, E. K., & Cohen, J. D. (2001). An integrative theory of prefrontal cortex function. *Annual Review of Neuroscience, 24*, 167-202.
Minsky, M. (1975). A framework for representing knowledge. In P. H. Winston (Ed.), *The psychology of computer vision* (pp. 211-277). New York: McGraw-Hill.
Nerb, J., & Spada, H. (2001). Evaluation of environmental problems: A coherence model of cognition and emotion. *Cognition and Emotion, 15*, 521-551.
Newell, A. (1990). *Unified theories of cognition*. Cambridge, MA: Harvard University Press.
Newell, A., & Simon, H. A. (1972). *Human problem solving*. Englewood Cliffs, NJ: Prentice-Hall.
Nisbett, R. E. (Ed.). (1993). *Rules for reasoning*. Hillsdale, NJ: Erlbaum.
O'Reilly, R. C., & Munakata, Y. (2000). *Computational explorations in cognitive neuroscience*. Cambridge, MA: MIT Press.
Pinker, S. (1989). *Learnability and cognition: The acquisition of argument structure*. Cambridge, MA: MIT Press.
Pinker, S. (1999). *Words and rules: The ingredients of language*. New York: HarperCollins.
Quartz, S. R., & Sejnowski, T. J. (2002). *Liars, lovers, and heroes: What the new brain science reveals about how we become who we are*. New York: Morrow, William.
Read, S. J., & Marcus-Newhall, A. (1993). Explanatory coherence in social explanations: A parallel distributed processing account. *Journal of Personality and Social Psychology, 65*, 429-447.
Read, S. J., & Miller, L. C. (1998). On the dynamic construction of meaning: An interactive activation and competition model of social perception. In S. J. Read & L. C. Miller (Eds.), *Connectionist models of social reasoning and behavior* (Vol. 27-68). Mahwah, NJ: Erlbaum.
Rieke, F., Warland, D., de Ruyter van Steveninick, R. R., & Bialek, W. (1997). *Spikes: Exploring the neural code*. Cambridge, MA: MIT Press.
Rogers, T. T., & MClelland, J. L. (2004). *Semantic cognition: A parallel distributed processing approach*. Cambridge, MA: MIT Press.

Rosenbloom, P., Laird, J., & Newell, A. (1993). *The SOAR papers: Research on integrated intelligence*. Cambridge, MA: MIT Press.
Rumelhart, D. E., & McClelland, J. L. (Eds.). (1986). *Parallel distributed processing: Explorations in the microstructure of cognition*. Cambridge MA: MIT Press/Bradford Books.
Rumelhart, D. E., Smolensky, P., Hinton, G. E., & McClelland, J. L. (1986). Schemata and sequential thought processes in PDP models. In J. L. McClelland & D. E. Rumelhart (Eds.), *Parallel distributed processing: Explorations in the microstructure of cognition* (Vol. 2, pp. 7-57). Cambridge MA: MIT Press/Bradford Books.
Schank, R. C., & Abelson, R. P. (1977). *Scripts, plans, goals, and understanding: An inquiry into human knowledge structures*. Hillsdale, NJ: Lawrence Erlbaum.
Scherer, K. R. (1993). Studying the emotion-antecedent appraisal process: An expert system approach. *Cognition and Emotion, 7*, 325-355.
Shelley, C. (2001). The bicoherence theory of situational irony. *Cognitive Science, 25*, 775-818.
Shoda, Y., & Mischel, W. (1998). Personality as a stable cognitive-affective activation network: Characteristic patterns of behavior variation emerge from a stable personality structure. In S. J. Read & L. C. Miller (Eds.), *Connectionist models of social reasoning and behavior* (pp. 175-208). Mahwah, NJ: Erlbaum.
Shultz, T. R., & Lepper, M. R. (1996). Cognitive dissonance reduction as constraint satisfaction. *Psychological Review, 103*, 219-240.
Simon, D., Krawczyk, D. C., & Holyoak, K., J. (2004). Construction of preferences by constraint satisfaction. *Psychological Science, 15*, 331-336.
Smolensky, P., & Legendre, G. (2006). *The harmonic mind*. Cambridge, MA: MIT Press.
Spellman, B. A., Ullman, J. B., & Holyoak, K. J. (1993). A coherence model of cognitive consistency: Dynamics of attitude change during the Persian War. *Journal of Social Issues, 49*, 147-165.
Thagard, P. (1988). *Computational philosophy of science*. Cambridge, MA: MIT Press/Bradford Books.
Thagard, P. (1992). *Conceptual revolutions*. Princeton: Princeton University Press.
Thagard, P. (2000). *Coherence in thought and action*. Cambridge, MA: MIT Press.
Thagard, P. (2005). *Mind: Introduction to cognitive science* (2nd ed.). Cambridge, MA: MIT Press.
Thagard, P. (2006). *Hot thought: Mechanisms and applications of emotional cognition*. Cambridge, MA: MIT Press.
Thagard, P., & Aubie, B. (2008). Emotional consciousness: A neural model of how cognitive appraisal and somatic perception interact to produce qualitative experience. *Consciousness and Cognition, 17*, 811-834.
Thagard, P., & Millgram, E. (1995). Inference to the best plan: A coherence theory of decision. In A. Ram & D. B. Leake (Eds.), *Goal-driven learning:* (pp. 439-454). Cambridge, MA: MIT Press.
Thrun, S., Burgard, W., & Fox, D. (2005). *Probabilistic robotics*. Cambridge, MA: MIT Press.