Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence (AAAI-17)

Progress and Challenges in Research on Cognitive Architectures

Pat Langley
Institute for the Study of Learning and Expertise
2164 Staunton Court, Palo Alto, CA 94306

Abstract

Research on cognitive architectures attempts to develop unified theories of the mind. This paradigm incorporates many ideas from other parts of AI, but it differs enough in its aims and methods that it merits separate treatment. In this paper, we review the notion of cognitive architectures and some recurring themes in their study. Next we examine the substantial progress made by the subfield over the past 40 years, after which we turn to some topics that have received little attention and that pose challenges for the research community.

1 Introduction and Overview

Most research in AI is *analytic*, in that it selects some facet of intelligence and attempts to understand it in detail, typically in isolation from other elements. This is balanced by a smaller movement, *synthetic* in character, that aims to discover how different aspects of intelligence interact. Without efforts of this latter sort, AI may be able to create *idiot savants* that outperform people in narrow arenas, but it cannot create complete intelligent agents that show the same breadth of abilities as seen in humans.

Theoretical physicists seek a grand unified theory that explains all known physical laws within a single consistent framework. The analog in AI is a unified theory of cognition, which Newell (1990) linked to the notion of a *cognitive architecture*. The mapping is imperfect in that most AI researchers focus on creating computational artifacts rather than explaining observations. However, if we want systems that exhibit the full range of human intelligence, they must reproduce all major phenomena associated with the latter.

In the sections that follow, we review the cognitive architecture paradigm and its recurring themes, then discuss the great strides it has made in the decades since its inception. We will argue that, despite this progress, research has focused on some topics to the near exclusion of others, and we examine a number of areas that deserve more attention. One open question is whether replicating these abilities requires changes to existing architectures or simply adding new types of knowledge. Ultimately, this is an empirical question that can only be answered by making the attempt, but we will take some tentative stances on the issue.

2 Previous Work on Cognitive Architectures

The cognitive architecture movement shares many ideas with other branches of AI, but it has sufficiently different emphases that we should clarify them before discussing its status. In this section, we describe its basic aims, some recurring themes, and some well-known examples.

2.1 The Notion of a Cognitive Architecture

A cognitive architecture (Newell 1990) is a theory of intelligent behavior that specifies those facets of cognition hypothesized to remain constant over time and across different domains. This includes memory stores and the representations of elements in those memories, but not their contents, which change as the result of external stimuli and internal processing. In this sense, a cognitive architecture is analogous to a building architecture, which describes its fixed structure (e.g., floors, rooms, and doors), but not its replaceable elements (e.g., tables, chairs, and people).

However, such a framework incorporates more constraints than the 'software architectures' used in mainstream computer science. It makes strong assumptions about the representations and mechanisms that underly cognition, typiclly incorporating ideas from psychology about the nature of the human mind. Most cognitive architectures have distinct modules, but these usually access and alter the same memories and representations. Moreover, they come with a programming language for constructing intelligent agents that adopts a high-level syntax which reflects theoretical assumptions. Production systems (Klahr et al. 1987) were both the first and most common examples of the paradigm; they are not the only variety, but many frameworks labeled as 'cognitive architectures' do not fit the criteria we have specified.

2.2 Recurring Themes

The literature on cognitive architectures reflects a number of common assumptions or recurring themes that give it a distinctive character, although it shares some with other areas of artificial intelligence. These include postulates that:

- *Short-term memories are distinct from long-term stores*. The former, often called *working memories*, include content that changes rapidly over time, such as the agent's beliefs and goals. The latter store stable elements that remain static or change slowly through learning. Short-term

elements are typically specific, while long-term content is often general in form; the former play the role of program variables in traditional languages, whereas the latter often serve as program instructions.

- *Memories contain modular symbolic structures*. Both short-term and long-term repositories are collections of distinct elements that, usually, are encoded as list structures. These may include numeric information, but each item is a relational structure that also specifies symbolic content. Short-term elements typically share symbols so that, jointly, they can denote large-scale relationships.
- *Relational pattern matching accesses long-term content*. Before an architecture can use the 'program' encoded in its long-term memories, it must find relevant structures. This invariably revolves around matching or unifying relational patterns in these elements against corresponding elements in the short-term stores. Such patterns match when variables in the long-term structure bind consistently with constants in dynamic memories.
- *Cognitive processing occurs in recognize-act cycles*. The core interpreter alternates between matching long-term structures against short-term ones, selecting a subset of the former to apply, and executing their associated actions to alter memory or the environment. Because many long-term elements may be satisfied, a *conflict resolution* mechanism selects among them. This reflects an abiding concern of architectural research with cognitive control.
- *Cognition dynamically composes mental structures*. Sequential application of rules or similar knowledge elements produces working memory elements, which in turn enable matching on the next cycle. This chaining process essentially combines the matched elements into new entities, as in work on logical reasoning. Similarly, structural learning involves the composition of new rules or skills from existing cognitive material.

These assumptions are well suited to the class of phenomena and level of analysis for which most architectures have been designed. Research in this paradigm was influenced strongly by studies of human thinking, with behavior taking seconds to minutes and with basic operations on the order of 50 to 100 milliseconds. In such settings, humans appear to access relevant content from long-term memory in parallel (or at least very rapidly), but sequential bottlenecks keep more than one or a few operations from applying at a time.

The importance of problem-space search has been another recurring theme. Mechanisms to support heuristic search have been built directly into some architectures but not others. Nevertheless, many provide structures and processes that support it in relatively direct ways; Young (1982) has even noted how certain conflict-resolution schemes lead naturally to specific search regimens. At the same time, these frameworks can encode and use expert knowledge, stated as long-term structures, to reduce search or even eliminate it entirely. Finally, as we discuss later, there has also been considerable work on mechanisms for learning such expertise from experience, although most efforts have focused on the acquisition of procedural content.

2.3 Example Architectures

To clarify these ideas, we should briefly review some cognitive architectures that reflect these recurring themes. Here are three examples of such frameworks:

- ACT-R (Anderson 1993) is a mature architectural theory concerned mainly with explaining psychological data. It combines a procedural memory of generalized production rules with a declarative memory of specific facts, with the former accessing the latter through activation-based retrieval and pattern matching. Learning mechanisms include rule compilation and statistical updates, each based on use of knowledge structures.
- Soar (Laird et al. 1987) is another well-developed architecture that is organized around problem-space search. An elaboration stage invokes production rules to draw inferences and evaluate alternatives, with results used by a decision-making stage that selects an operator to apply mentally or in the environment. A chunking mechanism acquires rules based on results from problem solving. Recent versions include other modules that we discuss later.
- ICARUS (Langley, Choi, and Rogers 2009) is a more recent architecture that includes separate modules for conceptual inference, teleoreactive skill execution, means-ends problem solving, and skill acquisition. It differs from predecessors by positing that cognition is grounded in perception and action, conceptual knowledge is distinct from skills, both are organized in a hierarchy, and short-term elements are instances of long-term structures.

These architectures are important examples because they combine mechanisms for inference, routine activity, goal-directed problem solving, and structure learning.

We should also mention frameworks that omit some of these elements but share the core assumptions listed earlier:

- PRODIGY (Veloso et al. 1995) uses search-control rules to guide means-ends analysis and learns new rules from successes and failures. The architecture supports a variety of learning methods, but focuses on planning tasks rather than other forms of high-level cognition.
- CAPS (Thibadeau 1983) encodes knowledge as production rules but applies them in parallel to alter the activations of elements in working memory. This lets the architecture model details of human reading, but it lacks mechanisms for solving problems or learning structures.
- EPIC (Kieras and Meyer 1997) combines parallel application of production rules with resource-limited visual, auditory, and motor mechanisms to model performance times in multi-task settings. Like CAPS, it includes no processes for problem solving or learning.
- CLARION (Sun and Zhang 2004) combines condition-action rules with neural networks to encode 'explicit' and 'implicit' knowledge, as well as complementary mechanisms for learning such content. The architecture includes low-level 'drives' that modulate behavior, but does not support traditional problem-space search.

Langley, Laird, and Rogers (2009) discuss these and other cognitive architectures in greater detail, including issues that arise in their design, construction, and application.

3 Progress in Cognitive Architectures

Research on cognitive architectures has seen substantial progress since their advent in the 1970s. In this section, we review a number of these advances, although they are by no means the only ones. We will not focus on topics primarily of interest to psychologists, such as fitting models' behavior to human reaction times and error rates, as emphasized in work on CAPS, EPIC, and CLARION, or mapping architectural modules onto regions of the brain, a focus of recent ACT-R research (Anderson 2007). Instead, we will examine progress relevant to the construction of intelligent agents, which should hold greater interest for the AI community.

3.1 Hybrid Representations and Processing

Early production-system frameworks like PSG (Newell and McDermott 1975) and OPS2 (Forgy and McDermott 1978) were almost entirely symbolic, incorporating only a few numbers like the ordering of rules and recency of working memory elements. This was consistent with the general emphasis on symbolic processing in both AI and cognitive psychology at the time, which was still engaged in distinguishing itself from the earlier, number-oriented traditions of the associationist and behaviorist paradigms that preceded them.

However, early versions of ACT (Anderson 1982) introduced strengths on productions and activations on elements in working memory, with architectures like CAPS (Thibadeau 1983) and PRISM (Langley 1983) following suit. These served as modulators on symbol structures during conflict resolution and helped focus cognitive attention. Somewhat later, ACT-R (Anderson 1993) reinterpreted numeric annotations in probabilistic and decision-theoretic terms, and early versions of ICARUS (Choi et al. 2004) associated values with cognitive skills. Even Soar (Laird 2012) now attaches quantitative scores to production rules that it uses in decision making and control. Many modern architectures are hybrid in character rather than purely symbolic.

3.2 Learning Procedural Knowledge

Cognitive architectures had their roots in accounts of problem solving and heuristic search (Newell and Simon 1972), and the first production systems relied on a fixed knowledge base to perform sequential tasks. However, they also proved quite useful for modeling successive stages in children's cognitive development, which in turn led to research on *adaptive* production systems that learned by adding new condition-action rules, or by modifying existing structures, based on their experience (Klahr et al. 1987).

Early versions of ACT (Anderson 1982) introduced processes for rule generalizaton, discrimination, proceduralization, and composition, with frameworks like PRISM (Langley 1983) adopting similar ideas. Soar and PRODIGY developed analytic methods for acquiring search-control from success and failure during problem solving, and many others pursued related approaches. ICARUS contributed mechanisms for learning hierarchical skills from problem solving, while CLARION relied on quite different techniques for statistical learning. Incorporation of procedural learning has been one of the clear success stories of the architectural paradigm, and it has influenced AI's other subfields.

3.3 Large-Scale Structures

Most cognitive architectures encode long-term knowledge, at least the procedural variety, as condition-action rules. Such production systems have many attractive features, including modularity that supports automated composition, flexibility of use, and ease of acquisition. The success of frameworks like Soar, ACT-R, and EPIC, as well as their use in constructing many expert systems, suggest that the benefits of this formalism are real. Nevertheless, other frameworks for intelligent systems, such as frames (Minsky 1975) and scripts (Schank and Abelson 1977), instead propose larger-scale structures that encode more content per element. For example, each 'method' in the SHOP2 formalism (Nau et al. 2003) for hierarchical task network maps onto a number of separate rules in ACT-R and Soar.

Some architectures have incorporated large-scale structures into their framework and syntax. Veloso et al. (1995) reported an extension to PRODIGY that stores justified solutions to problems and uses them analogically to guide search on new tasks. Similary, Langley et al.'s (2009) ICARUS encodes its hierarchical skills in terms of subgoals they should achieve, much as hierarchical task networks decompose complex tasks into subtasks. Both appear to retain the modularity and flexibility seen in production systems, which suggests that adopting larger knowledge elements is a viable option. Still, this approach remains uncommon in the paradigm and deserves more attention from researchers.

3.4 Embodied Agency

Like early computational models of problem solving and language, initial cognitive architectures focused on mental capacities and were effectively disembodied. There was no denial that human intelligence arises in a physical body that operates in an external environment, but both AI and cognitive psychology were concerned mainly with internal phenomena, and researchers were inclined to abstract away from sensorimotor issues. Both fields had made great strides by ignoring such matters, but it remained clear that, eventually, a unified theory of cognition must address them.

Laird et al.'s (1991) Robo-Soar was an early example of providing a cognitive architecture with external sensors and effectors to let it control a robot. More recently, Trafton et al. (2013) reported a version of ACT-R with similar capabilities. However, both endowed their frameworks with new modules that, arguably, were not part of their core theories. The same holds for Soar and ACT-R systems that controlled the bodies of synthetic characters in virtual environments, which operated in similar but simulated settings. In contrast, ICARUS has controlled virtual agents (Choi et al. 2007) using an approach that grounds all structures in perceptions and actions, although it still lacks a theory of peripheral perceptual and motor processing. Each effort successfully extended the notion of cognitive architecture beyond purely mental processing, although more work remains to be done.

A related line of research has used cognitive architectures to handle situations that involve interaction with other agents. Examples include TacAir-Soar (Jones et al. 1999), which modeled expert fighter pilots in simulated air battles,

and ACT-R/E (Trafton et al. 2013), which interacted with humans to carry out joint tasks. Both systems incorporated not only an ability to communicate with others, but also to construct at least limited models of their beliefs and goals. This raises issues somewhat different from sensing and control, but interaction with other agents is an important class of behaviors that goes beyond purely internal processing.

3.5 Declarative and Episodic Memories

Initial cognitive architectures encoded all long-term knowledge as production rules. This reflected an emphasis on procedural content that supports activity over time, rather than on declarative information about static facts. Some early work attempted to represent the latter as condition-action rules, but the results were awkward. However, even the first versions of ACT (Anderson 1982) included a separate declarative respository in addition to production memory. Elements had the same form as those in working memory, which was viewed as the active part of the declarative store. ACT complemented this notion with a spreading activation mechanism that retrieved elements from the latter, and Soar has adopted similar ideas in recent years.

Related research has addressed the problem of episodic memory, which records an agent's experiences over time. This has included adaptations of ACT-R's declarative repository for this purpose, but other frameworks have also tackled the issue. Veloso et al.'s (1995) extension of PRODIGY stored and used episodic traces, as did Jones and Langley's (2006) EUREKA architecture, although both were limited to records of successes and failures used during analogical problem solving. More recently, Soar (Laird 2012) has expanded to include a separate episodic memory with its own retrieval mechanisms. Not all cognitive architectures provide support for this capability, but the topic is important enough that it seems likely to happen with time.

4 Open Research Issues

Despite the steady progress seen in the cognitive architecture community, a number of important aspects of human intelligence have not received the attention they deserve. In this section, we describe some of these capabilities, along with questions researchers should tackle when addressing them. We will not assume that each one must be suppported at the architecture level; they may be handled adequately by a combination of existing mechanisms and knowledge. However, taking them seriously seems likely, at the very least, to push current frameworks in new directions.

4.1 Understanding and Interpretation

Traditional cognitive architectures have adopted an *action* metaphor. The terminology used to describe production-system frameworks like ACT-R, Soar, and EPIC reflects this idea: rules comprise a condition side and an action side. This emphasis is natural given their history, in that production systems grew out of theories of problem solving merged with behaviorist notions of stimulus-response pairs. One of the earliest papers in the paradigm, by Newell (1973), presented them as accounts of cognitive control. Naturally, this has made them well suited for modeling both novice problem solving and expert behavior. They can even handle sentence processing, as evidenced by shift-reduce parsers, which are a variety of production systems.

However, the problem of understanding sequences of connected events, whether they arrive through language, vision, or some other medium, has received little attention from cognitive-architecture researchers. Consider a typical example from story understanding: *John wanted a raise. He told his boss that he knew where she went when she told her husband she was working late*. Interpreting this story requires reasoning not only about domain content, such as decision-making authority, but also John's inferences about his boss's beliefs and goals. There has certainly been work on this topic (e.g., Schank and Abelson 1977), but little of it has been associated with cognitive architectures.

The problem is not representational. Traditional architectures can encode content like that involved in our story. However, they would have difficulty reasoning over the partial information it provides, as they assume that all a rule's conditions must match before it applies. In contast, connected understanding appears to be *abductive* and relies on introduction of plausible assumptions. This has clear implications for architectural design, suggesting the need for some form of partial matching. One option is to incorporate analogical reasoning, like that in Forbus' (2016) Companions framework, which typically operates over large-scale structures and which is inherently abductive in character.

4.2 Dynamic Memory

As noted earlier, learning has been a central concern for cognitive architectures almost since their inception. An early argument for production systems was that the modular encoding of content should ease acquisition of knowledge, and repeated successes have supported this idea. However, the bulk of research on this topic focused on learning routine procedural skills or heuristics for problem-space search. Such knowledge is essential for intelligent agents that pursue goals over time, but, again, it is primarily about actions, whether they are physical or mental.

Another crucial form of knowledge concerns *categories* that the agent encounters and relations among them, and acquisition of such long-term content – precisely the type needed for complex understanding – has rarely been examined in the cognitive architecture paradigm. Schank (1982) referred to both the structures and mechanisms for learning them as *dynamic memory*. One can argue that the framework he proposed, and those descending from it, were themselves cognitive architectures, although they were seldom cast in these terms and they lacked standard features, such as a well-specified interpreter or a programming language.

Traditional architectures have had difficulty in this area because their long-term knowledge is oriented around action. Dynamic memory depends on the ability to create and organize new conceptual symbols that have associated descriptions; in contrast, action-oriented architectures like production systems typically combine existing symbols to create new conditional responses. Langley et al. (1991) presented an architectural design based on a variety of dynamic

memory, but it was never fully implemented. More recently, Li et al. (2012) have reported a refinement of ICARUS that extends its conceptual memory by defining new terms, but it seems clear we need more work on this important topic.

4.3 Creativity

One distinctive feature of human cognition is the ability to solve novel problems in unexpected and surprising ways, something often referred to as *creativity*. As already noted, the initial development of cognitive architectures was influenced strongly by results on human problem solving, and most frameworks have supported this process. Moreover, Weisberg (1993) has argued that scientific discovery, artistic composition, and other invention can be explained in terms of problem-space search guided by heuristics. Despite the intriguing nature of creativity, there has been remarkably little work on it within the cognitive architecture paradigm.

A few exceptions have used traditional architectures for heuristic search in arenas associated with creative endeavors such as scientific discovery (e.g., Langley et al. 1987). However, there has long been evidence for architecture-level operations in creative inquiry, especially ones related to storage and recall. Jones and Langley's (2005) EUREKA architecture combined problem-space search with retrieval through spreading activation to explain insight effects, but the community has not built on their results. More recently, Helie and Sun (2010) have adapted CLARION to model insight effects using another activation-based mechanism that supports soft constraint satisfaction. Both results suggest that cognitive architectures may require new elements to provide a complete explanation of creative thought.

One promising topic concerns humans' ability to *reformulate* problems in new terms. Insight puzzles like the mutilated checkerboard illustrate this idea most clearly, but changes in formulation have also led to important breakthroughs in science. These are often linked to shifts in representations and ontologies that make problems easier to solve or that suggest new theories to explain phenomena. Some aspects of reformulation, such as abstraction that ignores aspects of states or operators, fit well within existing architectures, but other forms, such as conceptual reorganization, offer more challenges to current theories.

4.4 Emotions and Metacognition

One bias that architecture research has shared with most of AI is an emphasis on intellectual abilities such as planning, reasoning, and language processing. As noted earlier, the movement grew originally from models of human problem solving, which focused on puzzles, game playing, and mathematical domains. There is no question that such abstract processing is a critical way in which humans differ other mammals, like dogs and cats, at least in the degree to which we exhibit this capability.

However, people also experience *emotions* when working on a difficult puzzle or playing a challenging opponent in chess. A common assumption, exacerbated by the popular media, is that emotions are irrational holdovers from an earlier stage of evolution. In contrast, Simon (1967) argued that they play an important role in controlling cognitive attention, and more recent empirical analyses support this view. This has obvious implications for the design of cognitive architectures, which require some form of conflict resolution to select among mental actions. Recent years have seen clear progress on computational models of emotion, some of it building directly on existing architectural frameworks (Marsella et al. 2010). But there have been few efforts to incorporate them directly into such architectures and to explain how they modulate other cognitive activities.[1]

This omission may be related to another important area that the paradigm has long overlooked – *metacognition* or 'thinking about thinking' (Cox 2007). Soar has some metacognitive aspects, but it does not include specific structures and processes – such as recording traces of mental operations for later inspection – at the architecture level, and research that does address them has not adopted many of the core assumptions described earlier. Note that many common emotions, such as *relief* and *disappointment*, arise when an agent experiences certain combinations of goals, expectations, and beliefs with respect to an object or event. In other words, the process of emotion elicitation appears to inspect traces of regular cognition, which suggests it is metacognitive in nature. This idea has interesting archtitectural implications that deserve further exploration.

4.5 Personality and Goal Reasoning

Another topic that has received substantial attention in psychology is *personality*. This is a broad area that we cannot review in detail here, but there is general agreement about high-level features of human personalities, specifically that they vary across people to produce distinct behavioral styles, they remain reasonably stable over time, they influence behavior globally across many situations, and they affect both coarse-gained and fine-grained behavior. Natural language includes many words to describe them, such as *persistent*, *confident*, and *dogmatic*. Personality differences are part of our everyday experience and they deserve a computational account, yet the cognitive architecture community has made few attempts to address them.

Some psychological theories attempt to explain these phenomena in behaviorist terms, treating personality as a collection of stimulus-response pairs; others instead posit a set of fixed personality 'traits' that influence behavior. The latter have been adopted in AI work on synthetic characters, which typically encode personality as a point in N-dimensional space. As usually presented, neither is consistent with the cognitive architecture paradigm, but variations offer potential. For example, one can imagine architectural parameters – such as persistence in pursuing goals and readiness to revise beliefs – that map onto everyday personality terms. Similarly, one might include knowledge about the conditions under which to pursue different goals – say helping those in need or acquiring wealth – and their priorities.

The second scheme is far from behaviorist in character,

[1] Marinier and Laird's (2007) work in Soar is a rare exception. We do not include dimensional theories of emotion used in some synthetic characters, which are essentially noncognitive accounts.

but it offers an account in terms of conditional responses that could, in princple, change over time. Rizzo et al. (1999) report an initial extension of PRODIGY that explores this idea, but, to our knowledge, no one has built on their promising work. This approach also relates to recent work on goal reasoning (Aha, Cox, and Muñoz-Avila 2013), which studies the origin and management of agents' goals over time. In this view, personalities are simply collections of goal-generating rules and the associated priorities. If so, then we can view personalities, like emotions, as playing metacognitive roles in the architecture. This hypothesis may or may not be fruitful, but it suggests one way to incorporate an important class of phenomena into unified theories of the mind.

5 Some Peripheral Topics

Our agenda for research on cognitive architectures has omitted a number of topics that may concern some readers. We should explain the reasons for bypassing them, especially given the attention they have received recently in the AI and cognitive science communities.

One such area concerns sensorimotor processing, which an embodied agent needs to interact with the external environment. We have noted that the central processes in EPIC and ICARUS accept perceptual inputs and produce effector outputs, but they do not model these peripheral mechanisms. We have also reviewed research on Soar and ACT-R that incorporates modules for sensor interpretation and motor control, but these are not part of their core theories. The reason is that cognitive architectures are concerned with the nature of *intelligence*, and sensorimotor processing is not central to this phenomenon. Rats, pigeons, and cockroaches exhibit sophisticated perceptual and effector abilities, but they are not intelligent. Cognitive psychology is distinct from perceptual psychology and kinesiolgy for good reasons.

We have also downplayed discussion of statistical learning that occurs gradually over time. Most versions of the ACT architecture have included a mechanism for production strengthening, PRODIGY collected statistics to determine which control rules to retain, and both CLARION and recent versions of Soar incorporate varieties of reinforcement learning. However, these are mainly background processes that serve to evaluate cognitive structures which are created from very few experiences, so that improvements in statistical techniques are unlikely extend the coverage and ability of architectures. Humans share such learning with rats, pigeons, and insects, which suggests only a minor role in intelligence. Moreover, they distract one from the central insight of AI, which is that computers are not mere number crunchers, but rather *general symbol processors*.

Finally, we have not examined connections to neuroscience, despite interest in 'biologically inspired' architectures[2] (Stocco, Lebiere, and Samsonovich 2010). Some researchers identify the mind with the brain and assume we cannot understand the former without the latter. But theories of intelligence can be independent of the hardware or wetware on which they operate, just as the same computer program can run on entirely different architectures and operating systems. Neuroscience has made great strides in recent years, but most results have focused on perception and action. They have little to say about how to represent beliefs, goals, or knowledge, use such structures for reasoning, problem solving, and language processing, or acquire this content at human learning rates. Functional studies of human thinking have revealed deep insights about intelligence, but neuroscience has not, because the mind and the brain involve different levels of scientific description.

[2]We should note that many 'cognitive architectures' reported in this literature do not satisfy the definition presented earlier.

6 Concluding Remarks

In this paper, we reviewed the notion of a cognitive architecture and some common themes in research on the topic. We found that the subfield shares key ideas with other branches of AI, such as the use of symbolic structures and relational pattern matching, but that it also has distinctive features, such as a central concern with unified accounts of mental capacities. We also recounted areas in which the paradigm has made impressive progress since it was launched four decades ago. These included the development of hybrid representations that combine symbolic and numeric content, mechanisms for learning procedural and control knowledge, incorporation of large-scale knowledge structures, construction of embodied and interactive agents, and support for both declarative and episodic memories.

However, we also examined other important topics that have received little attention from the community. They included accounts for abductive understanding, dynamic memories that acquire new conceptual structures, creative aspects of problem solving, emotional processing, and agent personality, along with the plausibly related topics of metacognition and goal reasoning. One way to address these phenomena is to introduce high-level knowledge structures into an existing architecture, but it seems likely that at least some of them will require revisions to established theories. We believe that serious efforts at responding to these challenges will drive the cognitive architecture paradigm in new directions that extend its coverage and bring it closer to comprehensive theories of the mind. This in turn will lead to more general and effective methods for constructing systems that exhibit human levels of intelligence.

Acknowledgements

This research was supported in part by Grant N00014-15-1-2517 from the Office of Naval Research, which is not responsible for its contents. We thank John Anderson, Paul Bello, Dongkyu Choi, Michael Cox, Randolph Jones, John Laird, Allen Newell, Stellan Ohlsson, and Seth Rogers for useful discussions that aided our architectural analysis.

References

Aha, D. A.; Cox, M. T.; and Muñoz-Avila, H. eds. 2013. *Goal reasoning: Papers from the ACS workshop*. Baltimore, MD.

Anderson, J. R. 1982. Acquisition of cognitive skill. *Psychological Review* 89: 369–406.

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

Anderson, J. R. 2007. *How can the human mind occur in the physical universe?* Oxford, UK: Oxford University Press.

Choi, D.; Kaufman, M.; Langley, P.; Nejati, N.; and Shapiro, D. 2004. An architecture for persistent reactive behavior. *Proceedings of the Third International Joint Conference on Autonomous Agents and Multi Agent Systems*, 988–995. New York: ACM Press.

Choi, D.; Könik, T.; Nejati, N.; Park, C.; and Langley, P. 2007. A believable agent for first-person shooter games. *Proceedings of the Third Annual Artificial Intelligence and Interactive Digital Entertainment Conference*, 71–73. Stanford, CA: AAAI Press.

Cox, M. T. 2007. Perpetual self-aware cognitive agents. *AI Magazine* 28: 32–45.

Forbus, K. 2016. Software social organisms: Implications for measuring AI progress. *AI Magazine* 37: 85–90.

Forgy, C. L.; and McDermott, J. 1978. *The OPS2 reference manual*. Technical Report, Department of Computer Science, Carnegie-Mellon University, Pittsburgh, PA.

Helie, S.; and Sun, R. 2010. Incubation, insight, and creative problem solving: A unified theory and a connectionist model. *Psychological Review* 117: 994-1024.

Jones, R. M.; Laird, J. E.; Nielsen P. E.; Coulter, K.; Kenny, P.; and Koss, F. 1999. Automated intelligent pilots for combat flight simulation. *AI Magazine* 20: 27–42.

Jones, R. M.; and Langley, P. 2005. A constrained architecture for learning and problem solving. *Computational Intelligence* 21: 480–502.

Kieras, D.; and 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.

Klahr, D.; Langley, P.; and Neches, R. eds. 1987. *Production system models of learning and development*. Cambridge, MA: MIT Press.

Laird, J. E. 2012. *The Soar cognitive architecture*. Cambridge, MA: MIT Press.

Laird, J. E.; Newell, A.; and Rosenbloom, P. S. 1987. Soar: An architecture for general intelligence. *Artificial Intelligence* 33: 1–64.

Laird, J. E.; Yager, E. S.; Hucka, M.; and Tuck, C. M. 1991. Robo-Soar: An integration of external interaction, planning, and learning using Soar. *Robotics and Autonomous Systems* 8: 113–129.

Langley, P. 1983. Exploring the space of cognitive architectures. *Behavior Research Methods and Instrumentation* 15: 289–299.

Langley, P.; Simon, H. A.; Bradshaw, G. L.; and Żytkow, J. M. 1987. *Scientific discovery: Computational explorations of the creative processes*. Cambridge: MIT Press.

Langley, P.; Choi, D.; and Rogers, S. 2009. Acquisition of hierarchical reactive skills in a unified cognitive architecture. *Cognitive Systems Research* 10: 316–332.

Langley, P.; Laird, J. E.; and Rogers, S. 2009. Cognitive architectures: Research issues and challenges. *Cognitive Systems Research* 10: 141–160.

Langley, P.; McKusick, K. B.; Allen, J. A.; Iba, W. F.; and Thompson, K. 1991. A design for the ICARUS architecture. *SIGART Bulletin* 2: 104–109.

Li, N.; Stracuzzi, D. J.; and Langley, P. 2012. Improving acquisition of teleoreactive logic programs through representation extension. *Advances in Cognitive Systems* 1: 109–126.

Marinier, R. P.; and Laird, J. E. 2007. Computational modeling of mood and feeling from emotion. *Proceedings of Twenty-Ninth Meeting of the Cognitive Science Society*, 461–466. Nashville, TN.

Marsella, S.; Gratch, J.; and Petta, P. 2010. Computational models of emotion. In K. Scherer, T. Banziger, and E. Roesch eds., *A blueprint for affective computing: A sourcebook and manual*. Oxford University Press.

Minsky, M. 1975. A framework for representing knowledge. In P. Winston Ed.), *The psychology of computer vision*. New York: McGraw Hill.

Nau, D.; Au, T.; Hghami, O.; Kuter, U.; Murdock, J. Wu, D.; and Yaman, F. 2003. SHOP2: An HTN planning system. *Journal of Artificial Intelligence Research* 20: 379–404.

Newell, A. 1973. Production systems: Models of control structures. In W. G. Chase Ed.), *Visual information processing.* New York: Academic Press.

Newell, A.; and McDermott, J. 1975. *PSG manual*. Technical Report, Department of Computer Science, Carnegie-Mellon University, Pittsburgh, PA.

Newell, A.; and Simon, H. A. 1972. *Human problem solving*. Englewood Cliffs, NJ: Prentice-Hall.

Rizzo, P.; Veloso, M.; Miceli, M.; and Cesta, A. 1999. Goal-based personalities and social behaviors in believable agents. *Applied Artificial Intelligence* 13: 239–272.

Schank, R. 1982. *Dynamic memory: A theory of learning in computers and people*. New York: Cambridge University Press.

Schank, R, and Abelson, R. P. 1977. *Scripts, plans, goals, and understanding*. Hillsdale, NJ: Lawrence Erlbaum.

Simon, H. A. 1967. Motivational and emotional controls of cognition. *Psychological Review* 74: 29–39.

Stocco, A.; Lebiere, C.; and Samsonovich, A. 2010. The B-I-C-A of biologically inspired cognitive architectures. *International Journal of Machine Consciousness* 2: 171–192.

Sun, R.; and Zhang, X. 2004. Top-down versus bottom-up learning in cognitive skill acquisition. *Cognitive Systems Research* 5: 63–89,

Thibadeau, R. 1983. CAPS: A language for modeling highly skilled knowledge-intensive behavior. *Behavior Research Methods and Instrumentation* 15: 300–304.

Trafton, J. G.; Hiatt, L. M.; Harrison, A. M.; Tamborello, F.; Khemlani, S. S.; and Schultz, A. C. 2013. ACT-R/E: An embodied cognitive architecture for human robot interaction. *Journal of Human-Robot Interaction* 2: 30–55.

Veloso, M.; Carbonell, J.; Perez, A.; Borrajo, D.; Fink, E.; and Blythe, J. 1995. Integrating planning and learning: The PRODIGY architecture. *Journal of Experimental and Theoretical Artificial Intelligence* 7: 81–120.

Young, R. M. 1982. Architecture-directed processing. *Proceedings of the Fourth Annual Conference of the Cognitive Science Society*, 164–166. Ann Arbor, MI: Lawrence Erlbaum.

Weisberg, R. W. 1993. *Creativity: Beyond the myth of genius*. New York: W. H. Freeman.