Advances in Cognitive Systems 1 (2012) 83–92 Submitted 1/2012; published 7/2012

Practical Evaluation of Integrated Cognitive Systems

Randolph M. Jones RJONES@SOARTECH.COM
Robert E. Wray, III WRAY@SOARTECH.COM
Michael van Lent VANLENT@SOARTECH.COM
Soar Technology, 3600 Green Court, Suite 600, Ann Arbor, MI 48105 USA

Abstract

This paper argues that integration is an essential approach to advancing the state of the art in cognitive systems. Integration is the path to aggregating requirements for cognition, as well as the solutions to those requirements. However, evaluation of integrated cognitive systems has proven difficult and expensive, which has hindered scientific advances. We suggest approaches to evaluating integrated cognitive systems by introducing the notion of *practical evaluation*, as well as the imposition of requirements that can only be satisfied with cognitive systems solutions. Practical evaluation involves not only measuring a system's task competence but also the properties of *adaptivity*, *directability*, *understandability*, and *trustworthiness*.

1. Introduction

This journal is dedicated in part to reinvigorating research on systems that "reproduce the entire range of human cognitive capabilities". An essential approach to advancing cognitive systems is to integrate a fixed set of cognitive capabilities and knowledge. Such *integrated cognitive systems* have the potential to implement agent applications that generate behavior that is comparable in breadth and depth to the performance of a human actor within a task domain. For example, an integrated cognitive system for taxi driving would gather destination preferences via a speech dialog, plan a route, and maneuver the vehicle along the route, adapting to the dynamic traffic situation as it is experienced.

In ambition, integrated cognitive systems represent one of the fundamental goals of artificial intelligence. Having such artifacts available for study should result in greater insights about the requirements for human-level cognition. However, a significant obstacle exists in understanding how to evaluate this class of intelligent systems, because there is not yet an agreed set of evaluation criteria, and the criteria that do exist are often not readily quantifiable.

Although our arguments about evaluation are more anecdotal than empirical, they are based on decades of research and development building integrated cognitive systems. We argue first that evaluation of cognitive systems should be *practical*. That is, they should perform precisely those types of tasks that require cognition: tasks that are currently performed well by humans and not by traditional software. We argue next that practical evaluation tasks require an integration of *adaptivity*, *directability*, *explainability*, and *trustworthiness*. In order to perform useful evaluation of systems along these dimensions, tasks and environment must embody a significant set of integrated requirements that fall under these four dimensions.

Finally, we argue that practical evaluation requirements should, as nearly as possible, not be achievable by a system unless it addresses all four dimensions. For any subset of task requirements, there may be many systems and approaches that perform effectively. In contrast, the set of systems that can meet an appropriately large, simultaneous, and integrated set of requirements will be much smaller. It is those systems that we argue are advanced cognitive systems. For this reason, it is not sufficient to show that each component of the system works in isolation from the others or that its components do something interesting in isolation, independent of knowledge and other system components. Cognition *requires* the interdependent operation of mechanisms and knowledge. Therefore, useful evaluation must ensure the imposition of simultaneous requirements on positive outcomes.

The remainder of this paper elaborates our arguments for the importance of such integrated approaches to cognition, as well as for practical evaluations that center on system adaptivity, directability, explainability, and trustworthiness. The paper additionally describes some potential approaches to evaluation motivated by this insight that useful evaluations must impose requirements that make integrated cognition *necessary* rather than just sufficient.

2. Integrated Approaches to Cognition

The components of cognitive systems depend intimately on each other to produce behavior. An advanced cognitive system must *integrate* components rather than merely *combine* them. By this, we mean that there is significant interaction among components, which impacts their design. Such systems must also effectively integrate knowledge with their cognitive capabilities.

This view contrasts with the dominant trend in artificial intelligence and cognitive science, which is to identify and investigate individual cognitive constructs and functional capabilities. Research on specific components is important to advancing cognitive systems, because individual capabilities must be computationally instantiated if we hope to create advanced systems. However, cognition is known to involve a complex interplay among a variety of mechanisms, each operating on mutually shared (and sometimes competing) knowledge representations (Damasio, 2010). Focusing on any component in isolation from the others is limiting, because the interoperation of multiple components satisfies requirements that are essential to cognition.

Integrated approaches take as a foundational assumption that the interplay between knowledge and mechanisms is central to achieving broad, robust computational intelligence. Much research in integrated cognition focuses on *cognitive architectures* (Langley, Laird, & Rogers, 2009). A particular cognitive architecture might contain one or more short-term and long-term memories, mechanisms for relating knowledge and making choices, processes for perception, attention, and action, methods for learning new long-term knowledge, and other components.

The importance of an integrated approach is clarified by considering how the components of a cognitive architecture impose requirements on each other. For example, representation of short-term memory items is closely related to methods for storage and retrieval. In turn, decision-making techniques must be sensitive to details of memory retrieval. Attention mechanisms must work with the components that use attention, such as perceptual systems and memories. In general, in an integrated cognitive architecture, one cannot make changes to one component without propagating new requirements to other components.

Our opinions about evaluation and advanced cognitive systems derive from our development of a wide variety of capability-rich intelligent systems for applied domains that were built within an integrated cognitive architecture. We and our colleagues have developed intelligent systems for applications such as fixed-wing air combat, rotary-wing air operations, indirect fire, culturally-aware interrogation models, and tactical controllers, among others (Jones et al., 1999; Laird, 2012; Stensrud, Taylor, & Crossman, 2006; Stensrud, et al., 2008; Taylor et al., 2007).

We have built these systems within the Soar architecture (Laird, 2012; Newell, 1990), and they exploit the advantages of prior work to ensure the integrated operation of working memory, long-term memory, preference-based deliberation, least-commitment reasoning, and the other components of Soar's design, representations, and mechanisms. However, the goal of our applied systems was not merely to evaluate or use Soar as a platform for intelligent systems. Rather, it was to build useful and practical intelligent systems. Many of these systems encode large knowledge stores, and Soar made it easier to build these systems than it would have been if we started with some other programming language or reasoning engine. However, we had to perform additional task analysis and knowledge engineering to meet the requirements of the applications.

Examples of these requirements included managing multiple independent goals simultaneously, interleaving and interrupting tasks dynamically and quickly, taking advantage of serendipitous opportunities to achieve goals, mixing serial and parallel reasoning activities appropriately and effectively, focusing attention to avoid perceptual overload, explaining decisions, flexibly accepting tasking, and performing the tasks in verifiable and trustworthy ways. These requirements were not always *directly* addressed or solved by the Soar architecture, but solutions to them were significantly constrained and informed by Soar's interdependent components and knowledge.

3. Evaluating Integrated Cognitive Systems

Integration imposes coherence on the wide variety of processes that contribute to cognition. However, a notable difficulty with pursuing research into cognitive architectures concerns evaluation. For an integrated system, evaluation must take place at a higher level of abstraction than for evaluation of the individual components, and it is particularly difficult to specify formal evaluation criteria at these higher levels, as it is for complex systems in general.

As an example, Laird et al. (2009) have suggested a number of measures for evaluating complex cognitive systems. These include "concrete" measures of performance and scalability, which are relatively well defined and quantifiable (although even these can take many forms in a complex system). But most of the measures are "abstract", like generality, expressivity, robustness, instructability, taskability, and explainability. These are certainly characteristics for advanced cognitive systems to strive for, but it remains ill-defined how to measure them. Wray and Lebiere (2007) describe some of the challenges to creating domain-independent metrics for such abstract system characteristics.

In spite of the difficulties in evaluating integrated cognitive architectures, if we wish to advance the state of cognitive systems research, we must look increasingly to integrated frameworks, rather than focusing on the individual components of intelligence. Our general approach to such evaluation also focuses on integration. If, as we argue, integration is essential to cognition, then we should be able to evaluate cognitive systems by measuring their performance on tasks that *require* the set of integrated cognitive capabilities and knowledge. If the tasks do not require such integration, then they do not require the depth and breadth of human reasoning capability, and they are therefore the wrong tasks to be using for evaluation.

This leads us to the question of how we can characterize in detail appropriate tasks for evaluation of advanced cognitive systems. The mechanisms of an integrated cognitive system provide critical value only when the task domain is complex enough that it becomes intractable to perform the task without using a range of cognitive capabilities and knowledge. Our experience in creating intelligent agents for a wide variety of real-world applications has suggested a particular set of application properties that, when taken together, require an integrated cognitive approach. We discuss these further in Section 5.

Application tasks that do not impose an appropriate and combined set of requirements can usually be solved with software systems that we would not consider to be "cognitive". However, cognitive systems that can achieve such tasks only do so if they appropriately integrate a range of cognitive capabilities and knowledge, because that is what the application requirements dictate. To reiterate, if an application does not *require* cognition or rich knowledge, then we could find some non-cognitive solution. The evaluation of integrated cognitive systems must be *practical* and not just theoretical. The practical goal of advanced cognitive systems is to capture the essential advantages of human intelligence that currently make humans the best choice for many tasks.

4. Challenges for Evaluation

We have thus far advocated an ambitious approach to developing advanced cognitive systems, and we have acknowledged that evaluation challenges increase with these ambitions. In this section, we discuss some of those challenges, keeping in mind that evaluation is essential to scientific progress, even if it is difficult or expensive. Effective analytic and empirical methods exist for evaluating algorithmic components of cognitive systems. However, it is substantially more difficult to perform systematic evaluations of systems that integrate numerous capabilities and knowledge.

The biggest problem for evaluation of advanced cognitive systems is requirements definition. Evaluation should be directed toward providing evidence that the system meets some standard, but the problem lies in defining that standard. The grand goal of cognitive systems is to build systems that reproduce the entire range of human cognitive capabilities, but this goal is not an evaluable standard or requirement. If we use human performance as the standard to achieve, we still have to define what "entire range" means, which capabilities count as "cognitive", and how we handle individual differences in human behavior and capability.

From a scientific perspective, we can strive to use human data on cognitive performance as the standard for evaluating our cognitive theories. But we cannot yet fix all the independent variables to match the conditions of the human performance for which data are available. This is particularly true if we are evaluating systems on tasks that require knowledge. It is difficult to determine which knowledge a human subject had before performing a task, although there are methods for approximating prior knowledge of this sort (e.g., Ericsson & Simon, 1993; Jones & VanLehn, 1992).

As we have argued, the evaluation of advanced cognitive systems should also include an applied, practical perspective, rather than just a scientific perspective. When building and evaluating applied cognitive systems, the customer for the system often has in mind some degree or quality of capability that the system should provide, and this can drive the question of whether the implemented system meets the goals and requirements. Unfortunately, when it comes to cognitive systems, customer requirements are seldom any more precise than the scientific goal to reproduce the entire range of human capabilities. Often the requirements are of forms such as to "demonstrate human expert-level capability" on a task, or to "perform this task correctly, with an ability to handle unanticipated situations", or to "perform all functions and components of the specified mission in a realistic fashion." These types of requirements are to be expected, to some extent, because they reflect a desire for the system to exhibit human-like cognitive properties. But they do little to drive specific modeling choices or to measure success.

Especially for applied tasks, we must be precise about defining what it means to exhibit human levels of intelligence. These requirements can certainly be subjective, and they can also be task independent. An example requirement might be to react to all significant events within human reaction times. Another might be to exhibit "natural" interactions with humans, where "natural" means that the humans' subjective sense is that they do not have to accommodate the cognitive system's idiosyncrasies. Requirements definition can often occur simultaneously with task and knowledge analysis during the development of an applied cognitive system. The more complex the application, the more requirements there are on the knowledge and capabilities necessary to perform the broad suite of tasks.

Another dominant factor in the development and evaluation of applied cognitive systems is practicality. In applied (and even non-cognitive) systems, the important question is often not "Does the system provide capability X?" but rather "How much would it cost for the system to provide capability X?" Applied evaluation focuses on capabilities, practicality, and cost, whereas scientific evaluation focuses on theory, coherence, understanding, and predictive value.

Certainly, some expensive software systems are still worth building, but the value proposition for advanced cognitive systems is often unclear. They can be difficult to build, and it is not always obvious which level of cognitive capability is needed for a particular application. However, this is the reason we continue to explore new cognitive mechanisms, new approaches to knowledge representation and knowledge acquisition, and new ways to integrate knowledge and capabilities. Thus, the advantages associated with advanced cognitive systems are not simply advantages related to cognitive capability. They are also practical advantages, leading to more cost-effective levels of autonomy and automation.

5. Proposed Evaluation Approaches

As we have argued above, a key to evaluating advanced cognitive systems is to be clear about requirements. The common criterion of "reproducing the entire range of human cognitive abilities" is not well-defined or measurable. So for any work in this area, we are left with a serious question of *how* to measure the degree to which a particular system reproduces some range of human cognitive abilities. In considering evaluation of complex systems, we must accept that the primary forms of evaluation must be empirical. This demands that we specify requirements, define independent variables for evaluation, and ensure that we are collecting data on dependent variables that are actually appropriate to the requirements. This section outlines some approaches to these issues for integrated cognitive systems.

5.1 Focus on Practical Evaluation

Practical evaluation involves looking at advanced cognitive systems as a means to an end rather than as ends in themselves. That is, we focus evaluation on the question of why we want cognitive systems from an application perspective. The high-level answer is that we expect cognitive systems will perform functions that currently require (or are beyond) the capabilities of humans. This leads us to examine which properties these cognitive systems would need to possess before we could generally expect human-level performance from them. What we have discovered in our application development is that these properties cannot merely perform a complex task well. They must additional address four important classes of requirements that derive from applied tasks:

- The task requires the system to be *adaptive*. Adaptivity requires recognition and categorization of many task-relevant situations, special cases, and exceptions. In addition, adaptivity requires the system to generate decisions, actions, and expectations, and evaluate alternative hypotheses that are highly situation dependent and that change fluidly as the dynamics of a situation unfold.
- The task requires the system to be *directable*, in that it flexibly and competently accepts, interprets, and implements tasks, information, advice, and direction received from external agents. We use the term "directable" as a more general concept than Laird et al.'s (2009) "taskable", because it includes more than just "tasking", which holds for even simple software systems. We intend "directable" to connote a more collaborative relationship that incorporates shared situational understanding.
- The task requires the system to be *understandable* to a human collaborator, with an ability to explain its interpretations, decisions, and actions. We use the term "understandable" to suggest a level of proactivity, interaction, and collaboration beyond the related terms "explainable", "comprehensible", and "rational".
- The task requires the system to be *trustworthy*, because there is little advantage to an intelligent system that is not allowed to use its intelligence autonomously or semi-autonomously. Advanced cognitive systems must be competent enough that their human team-mates *let* them use their competence.

These properties are somewhat different from the traditional view that advanced cognitive systems should be "autonomous". Humans and cognitive systems should be autonomous to the degree that they are adaptive and trustworthy, but trustworthiness comes in a full context of competence and interaction with others. Thus, for any particular cognitive system, a practical evaluation will focus on whether it can actually perform the task as ably as a human. That is, the question of autonomy is replaced by whether the system carries out the task in adaptive, directable, understandable, and trustworthy ways.

5.2 Use Real(istic) Environments

Jones and Laird (1997) argue that task complexity and realism impose significant and useful requirements on the design and evaluation of a cognitive system that performs a range of tasks. For any individual task, there is a space of "correct" solutions, but it is difficult or impossible to build a system that "works" but is not "cognitive", especially as the breadth, number, complexity, and realism of task requirements increase. This follows the spirit of evaluation for theories in the hard sciences. The more the theory is observed to match reality, the more confidence we have in it. If a theory fails to match reality, then we consider refining it. We have less confidence in evaluations that do not *require* the use of integrated capabilities and knowledge. By increasing the complexity and realism of the evaluation environment, we increase the degree to which cognitive capabilities are required, especially where we have a methodology for identifying specific cognitive requirements from task descriptions.

Creating complex and realistic environments may seem infeasibly expensive, but our efforts with TacAir-Soar suggest that this can be a practical approach to evaluation (Jones & Laird, 1997; Jones et al., 1999). When we can build a realistic enough task environment and impose realistic requirements in areas such as reaction times, interaction, and quality of performance, we are essentially requiring the same activities and level of performance that we would require of a human expert performing the task *in situ*. These increasing requirements bring us ever closer to ensuring that the system under evaluation cannot be "cheating". Note that this type of evaluation is in a similar spirit to the Turing Test (without actually being a complete Turing Test). The point is to ensure that requirements are complex enough that we can say with confidence they would only be achievable by a system that we would be willing to call cognitive.

It is also important to emphasize that in this approach we assume we are evaluating a single system that meets *all* of the requirements. We are not interested in requirements that are achievable by separable components of the cognitive system. Rather, the integrated system must bring all of its capabilities to bear on the range of evaluation tasks, because this is what we expect of systems (or humans) that we are willing to call intelligent.

Prior work demonstrates that using realistic environments can work well for evaluating non-learning systems. However, the approach offers even greater benefit when learning capabilities acquire knowledge automatically. As we have argued, a key property of integrated cognitive systems is that there are no separable pieces. Every mechanism is sensitive to the operation of other mechanisms, and this is particularly true of the relationship between learning and other cognitive mechanisms. Much of the work in machine learning has taken a component-oriented approach, measuring how well individual algorithms work under varying conditions. However, that research has not looked in depth into integrating such algorithms into a larger cognitive system (Langley, 1997). If we build advanced cognitive systems that include learning capabilities sufficient for effectively acquiring substantial knowledge, then we can evaluate them using task environments that *require* learning and performance to take place simultaneously. Such an evaluation approach would provide ample evidence that a particular integrated cognitive system is scientifically valuable, simply based on the fact that it operates successfully in the task environment at all.

5.3 Use Human Assessment Techniques

A complementary approach is to consider how we would evaluate whether a particular *human* "reproduces the entire range of human cognitive abilities". We assume that most humans meet this standard by definition, but we still have tests and methods for evaluating human abilities, skills, "innate intelligence", experience, and knowledge. If we intend our systems to meet a human standard, then it makes sense to evaluate them (at least partially) in the ways that we evaluate humans (Bringsjord & Schimanski, 2003). We might even consider this approach to trump all others, because we consider it to be a sufficient approach to evaluating humans themselves.

However, an obvious limitation is that much of human behavior in the execution of complex tasks is not readily reducible to quantitative measures. There are different ways to intercept an enemy aircraft, drive to a hotel across town, make a pie, or solve a physics problem. One of the primary difficulties in using human behavior as the standard is that it may require human judgment to evaluate that performance. However, even in these cases, computational tools can be used to reduce bias and identify envelopes of acceptable behavior within a large space of possible behaviors (Wallace & Laird, 2003). As an aside, when evaluating humans, we cannot generally observe or measure directly the internal workings of human task performance, as we can with cognitive systems. Thus, in the long run, we should have the advantage of being able to evaluate cognitive systems even more thoroughly than we can evaluate humans.

5.4 Identify Independent Variables

It is difficult to run systematic experiments on complex tasks because there are so many independent variables. Additionally, when trying to match human data, many of the independent variables are unobservable. Perhaps the most troublesome independent variable concerns the initial knowledge state of the agent from whom we have collected data. We must face the question of which knowledge a human or cognitive system possesses before the evaluation task is performed. For learning applications, we also need to assess which knowledge is acquired during the course of the task.

As with human assessment, we should look to the field of education. One goal of education is to identify at least a portion of an individual's knowledge state and then alter that state by increasing or improving it. Intelligent tutoring systems (Woolf, 2008) take on this task in an applied way. They use the results of human performance on various tasks to tease out which "chunks" of knowledge an individual must have, might be missing, or might be incorrect. To run careful scientific experiments on cognitive systems by matching them to human data, we should use similar techniques to ensure that we are appropriately identifying the initial knowledge state.

Another laborious but proven approach to identifying knowledge state is *protocol analysis* (Ericsson & Simon, 1993). When building and evaluating the Cascade cognitive system, Jones and Vanlehn (1992) used protocol analysis to identify which units of physics knowledge were present, missing, or incorrect in each human subject. They also identified specific learning events and the knowledge acquired during those events. This was possible by ensuring that the protocols did not merely record subject task performance, but also recorded their actions (such as referring to examples in a book) and their verbalized reasoning processes. By analyzing more than just performance data, protocol analysis can do a good job of specifying independent variables associated with pre-existing knowledge.

5.5 Evaluate Specific Qualitative Capabilities

Researchers in the area of cognitive architectures have proposed methods for subjective evaluation of qualitative capabilities (e.g., Newell, 1990; Anderson & Lebiere, 2003; Laird et al., 2009). The idea is to define, at least at an abstract level, what it would mean for a system to reproduce some set of cognitive capabilities. Laird et al. (2009) identify "abstract" measures of generality, expressivity, robustness, instructability, taskability, and explainability. Some of these overlap with the proposed measures for practical evaluation, described above.

To evaluate a particular cognitive system, there remains the daunting task of refining each of these abstract metrics into concrete, measurable variables. As one example, Kaminka (2002) describes an empirical measure for the robustness of team behavior in robot soccer and shows how systematic variation in team characteristics leads to changes in this variable. Developing such measures for a cognitive application is difficult, so it is important to think about the kinds of measures we should identify. We advocate combining this qualitative-capability approach with the use of realistic environments. The presence or absence of abstract cognitive capabilities can serve as a sanity check on the reality and complexity of a set of evaluation tasks.

5.6 Reuse and Aggregate

Our final point about evaluating integrated cognitive systems is to ensure that one aggregates results that reuse the same capabilities and knowledge across multiple, preferably widely divergent, tasks. If this can be done carefully, then requirements can cut across tasks and evaluations, and we can truly measure the breadth and depth of the cognitive system. This is a primary approach advocated by researchers who develop cognitive architectures. It suggests that the architecture should remain fixed across models and experiments, and this reusability demonstrates its strength and value, much as a strong scientific theory can be reused across different experiments.

However, this approach has not always been applied carefully in practice. Cognitive architectures change and evolve over time, which is not a bad thing from a scientific perspective, but it weakens the approach of aggregating evaluation across experiments. There are not usually enough resources to rerun all previous studies every time a change occurs. Thus, the experimental results produced by ACT-R (Anderson & Lebiere, 1998) and Soar in the 1980s may not be the same as they would be with the current versions of those architectures. This is not a reason to abandon this approach to evaluation, but it should be taken into account.

An additional problem with this approach is that, historically, it has not studied reuse of knowledge along with reuse of architectures. If each new experiment with a particular cognitive architecture relies on the development of a new model with its own knowledge base, it is fair to question how much of the result is due to the knowledge and how much is due to the architecture. For this reason, we advocate reusing both the architecture *and* the knowledge base.

6. Summary and Conclusions

We have argued that the development of advanced cognitive systems requires a focus on integration. Traditionally, integration approaches have emphasized cognitive architecture but not knowledge content. However, the interoperation of the architecture with a non-trivial knowledge base is essential if we want to build systems that "reproduce the entire range of human capabilities." Evaluation of such complex systems is difficult, but there are there are responses to this challenge, and we have described some that we consider promising. The issue is not that integrated cognitive systems cannot be evaluated, but that the process is complicated, time consuming, and expensive. These factors often inhibit evaluation and subsequently delay scientific advances.

However, finding appropriate methods for evaluation is essential to progress in cognitive systems. We advocate a future focus on *practical evaluation* that looks particularly at the *adaptivity, directability, understandability,* and *trustworthiness* of applied systems. Additionally, the field should aggregate requirements and lessons learned, which in turn should lead to an eventual convergence of theories and solutions. We can see evidence of this in the evolution of cognitive architectures. For example, Soar and ACT-R began their development with different emphases, strengths, and weaknesses. But as each has been applied to an increasing scope and complexity of tasks, many aspects of their designs have converged. This suggests that there are forceful requirements on integrated cognitive systems that can successfully replicate all of the capabilities in which our research community is interested. This is further evidence that advancing cognitive systems requires an integrated approach that accumulates requirements from broad sets of cognitive tasks. Component-level approaches do not impose enough constraints on the solution space to ensure that one solution is really better than another.

As we increasingly incorporate learning mechanisms into cognitive systems, they will become even more capable of acquiring large and *effective* knowledge bases. As this occurs, we will move even more quickly to the types of advanced systems the research community desires to understand and that application developers seek to exploit for practical, real-world needs.

Acknowledgments

We are grateful to Bob Marinier and Pat Langley for many helpful comments that improved the content and presentation of this paper.

References

Anderson, J., & Lebiere, C. (1998). *The atomic components of thought*. Mahwah, NJ: Lawrence Erlbaum.

Anderson, J. R., & Lebiere, C. L. (2003). The Newell test for a theory of cognition. *Behavioral and Brain Science, 26*, 587–637.

Bringsjord, S., & Schimanski, B. (2003). What is artificial ntelligence? Psychometric AI as an answer. *Proceedings of the Eighteenth International Joint Conference on Artificial Intelligence* (pp. 887–893). Acapulco: Morgan Kaufmann.

Damasio, A. (2010). *Self comes to mind: Constructing the conscious brain.* New York: Pantheon.

Ericsson, K. A., & Simon, H. A. (1993). *Protocol analysis: Verbal reports as data*. Cambridge, MA: MIT Press.

Jones, R. M., & Laird, J. E. (1997). Constraints on the design of a high-level model of cognition. *Proceedings of the Nineteenth Annual Conference of the Cognitive Science Society* (pp. 358–363). Stanford, CA: Psychology Press.

Jones, R. M., Laird, J. E., Nielsen, P. E., Coulter, K. J., Kenny, P., & Koss, F. V. (1999). Automated intelligent pilots for combat flight simulation. *AI Magazine, 20*, 27–41.

Jones, R. M., & VanLehn, K. (1992). A fine-grained model of skill acquisition: Fitting Cascade to individual subjects. *Proceedings of the Fourteenth Annual Conference of the Cognitive Science Society* (pp. 873–878). Hillsdale, NJ: Lawrence Erlbaum.

Kaminka, G. (2002). Towards robust teams with many agents. *Proceedings of the First Joint Conference On Autonomous Agents and Multi-Agent Systems*. Bologna: ACM.

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

Laird, J. E., Wray, R. E., Marinier, R. P., & Langley, P. (2009). Claims and challenges in evaluating human-level intelligent systems. *Proceedings of the Second Conference on Artificial General Intelligence.* Arlington, VA: Atlantis Press.

Langley, P. (1997). Machine learning for intelligent systems. *Proceedings of the Fourteenth National Conference on Artificial Intelligence* (pp. 763–769). Providence, RI: AAAI Press.

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

Newell, A. 1990. *Unified theories of cognition.* Cambridge, MA: Harvard University Press.

Stensrud, B., Taylor, G., & Crossman, J., (2006). IF-Soar: A virtual, speech-enabled agent for indirect fire training. *Proceedings of the 25th Army Science Conference*. Orlando, FL.

VanLehn, K., Jones, R. M., & Chi, M. T. H. (1992). A model of the self-explanation effect. *Journal of the Learning Sciences, 2*, 1–59.

Wallace, S., & Laird, J. E. (2003). Behavior bounding: Toward effective comparisons of agents and humans. *Proceedings of the Eighteenth International Joint Conference on Artificial Intelligence*. Acapulco: Morgan Kaufmann.

Woolf, B. P. (2008). *Building intelligent interactive tutors: Student-centered strategies for revolutionizing e-learning.* San Francisco: Morgan Kaufman.

Wray, R. E., & Lebiere, C. (2007). Metrics for cognitive architecture evaluation. *Proceedings of the AAAI-07 Workshop on Evaluating Architectures for Intelligence.* Vancouver: AAAI Press.