Advances in Cognitive Systems 1 (2012) 47–58 Submitted 2/2012; published 7/2012

How Minds Will Be Built

Kenneth D. Forbus FORBUS@NORTHWESTERN.EDU
Qualitative Reasoning Group, Northwestern University, 2133 Sheridan Rd, Evanston, IL 60208 USA

Abstract

Artificial intelligence started with the goal of understanding minds by attempting to build them. This essay discussed how our field's research has become unbalanced, and what might be done to change this situation. Specifically, I argue that more of our efforts should be focused on creating integrated cognitive systems. The founding of this journal will hopefully help by promoting such research.

1. Introduction

The scientific goal of artificial intelligence is to understand minds by attempting to build them. Is our field doing this as well as it could? In my view, despite much excellent work and progress, our research portfolio, as it were, is unbalanced. This essay outlines my diagnosis of the problem and how we might rebalance our research efforts to be even more productive. I begin by examining a common metaphor used to explain our current direction, dissecting it to illustrate the problems. Next I argue that we need to spend more of our efforts building integrated cognitive systems. Then as an example I describe some specific bets and hypotheses that my group is exploring. I close with why this journal will hopefully help the scientific community move forward on the goals of artificial intelligence and cognitive science more effectively.

2. The Cathedral Metaphor

Understanding minds well enough to create systems that are as intelligent and capable as people is not, technological singularity fans aside, around the corner. On the other hand, a common view among AI researchers is that any "real" progress is very far off. A historical analogy is often invoked that goes something like this: Building an artificial intelligence is like building a cathedral. The first cathedrals took generations, so most working on them would never see the final outcome. Those working on it took pride in their craft, building bricks and chiseling stones that would be placed into the Great Edifice. So, as AI researchers, we should think of ourselves as humble brick makers, whose job it is to study how to build components (e.g. parsers, planners, learning algorithms) that someday someone, somewhere, will integrate into intelligent systems.

I think this metaphor is inappropriate and counterproductive. First, that is not how humanity learned to build cathedrals. We learned how to build cathedrals by constructing buildings, albeit simpler ones. First huts and shacks, then houses, warehouses, and ever larger and more complex buildings. For any new type of building, there were initially many failures, but as understanding grew, the failures dropped off, or at least were more predictable. Many attempts to build

cathedrals failed because the properties of materials were not sufficiently well understood at the time. As J.E. Gordon (2003) puts it,

> "On the face of it it would seem obvious that the medieval masons knew a great deal about how to build churches and cathedrals, and of course they were often highly successful and superbly good at it. However, if you had had the chance to ask the Master Mason how it was really done and why the thing stood up at all, I think he might have said something like 'The building is kept up by the hand of God – always provided that, when we built it, we duly followed the traditional rules and mysteries of our craft.
>
> Naturally, the buildings we see and admire are those which have survived: in spite of their 'mysteries' and their skill and experience, the medieval masons were by no means always successful. A fair proportion of their more ambitious efforts fell down soon after they were built, or sometimes during construction."

And as we moved on from cathedrals, and learned to build skyscrapers and bridges and the vast kinds of physical infrastructure that underlies our technological civilization, new problems arose that then needed to be studied (e.g., why cracks caused ships to catastrophically fail), some of which we are still trying to figure out today (e.g., how to rescue people from buildings taller than ladders can reach).

This history also illustrates why the cathedral metaphor is counterproductive. No amount of studying bricks in isolation will tell you the problems involved in constructing cathedrals. Only by using components to build integrated cognitive systems can we start to understand the range of problems that are involved in constructing minds. And yet, today, almost all work in artificial intelligence falls into the brick-making mold.

Unfortunately, work in other areas of cognitive science is mostly in a similar state. Cognitive simulations, also called computational models, are computer-based accounts of psychological phenomena. A cognitive simulation can show that a particular combination of representations and processes provides an explanation of a phenomenon, by reproducing aspects of people's behavior (answers given, reaction times, error patterns) and by successfully predicting aspects of behavior not previously observed (Cassimatis et al., 2009). Most researchers focus on models that emulate one process or even one step in a process. Such models can indeed be useful. On the other hand, Newell (1973) argued eloquently that playing 20 questions with Nature would never converge, and that we should build larger-scale models. I agree, adding that, when building models of particular processes, we should satisfy the *integration constraint* (Forbus, 2001): a cognitive simulation of a psychological process should be able to serve as a component in simulations of larger-scale cognitive processes. Alas, in my experience, cognitive simulation research rarely satisfies this constraint.

Is the Cathedral metaphor simply an unrealistic caricature of thinking in the field? Unfortunately, it is not. For example, consider the following quote from Russell (1997, p. 68):

> "By analyzing and solving each subcase and producing calculatively rational mechanisms with the required properties, theoreticians can produce the AI equivalent of bricks, beams, and mortar with which AI architects can build the equivalent of cathedrals. Unfortunately, many of the basic components are currently missing. Others are so fragile and non-scalable as to be barely able to support their own weight. This presents many opportunities for research of far-reaching impact."

It is important to note that Russell himself is not advocating working only on the bricks: His own work includes "an attempt to combine all these new bricks to solve an interesting application problem, namely driving a car on a freeway." Similarly, in Koller's (2001) Computers and Thought lecture, she argued that the right approach to AI is to:

- Divide the problem into well-defined pieces
- Make progress on each one
- Build bridges to create a unified whole

The problem with this model is that the individual solutions may be too far apart (as Koller herself points out), and not stable enough, to support bridges. For example, most work in computational linguistics and machine learning, when it looks at semantics or knowledge at all, uses shallow stand-ins, such as WordNet synsets or Wikipedia identifiers. There is currently little evidence that such representations can support the range of reasoning that people exhibit using knowledge gleaned from language. As another example, competitions involving SAT solving have led to optimizations such as re-organizing the working set of constraints so that it remains in CPU's L2 cache as much as possible (Moskewicz et al., 2001). Such optimizations might be valuable for scaling up to large, practical problems, but they seem less likely to be relevant for understanding human reasoning.

Let us imagine a different model. Suppose that the brick-makers are working closely with those trying to construct buildings. Furthermore, suppose that many people are trying to construct buildings, not just a few. Constructing buildings is hard work: It is expensive, time consuming, and requires a wide range of skills and engineering craft. But I believe that spending more of our energies this way would pay off. The feedback cycle would become much faster, becoming months or years, instead of "some day". There is already evidence of the productivity of this approach, as illustrated in the next section.

3. Rebalancing our Research Portfolio

I believe that our field needs to put more effort into building integrated cognitive systems that attempt to capture larger collections of cognitive abilities. We may not be comfortable calling them minds, even very simple minds, but they should be clear steps in that direction. This is crucial for scientific progress, since many issues will only arise at broader scales of operation.

The best examples of this kind of integrated cognitive system are cognitive architectures. Most cognitive models, as noted above, focus in on only one process and use approximations for the rest of the system that provides their inputs and uses their outputs. Cognitive architectures invert this, focusing on how all the pieces might fit together. They explore hypotheses about the nature of intelligence via their assumptions about what kinds of components, and what interactions among them, are central. Components not covered by the theory are approximated, such as sensory-motor systems.

For example, ACT-R (Anderson & Lebiere, 1998) focuses on modeling skill performance and learning, with production rules as its representation for mental procedures, chunks as its representation of declarative knowledge, and a compilation process as its main learning process. In addition to modeling behavioral data on a wide range of tasks, it has been used to make successful predictions about brain activity as measured via fMRI (Anderson, 2007). Another classic cognitive architecture is SOAR (Laird, 2012), which shares the idea of production rules

(although the specifics differ significantly from ACT-R) and adds the idea of universal subgoaling (Laird & Newell, 1983). SOAR, too, has been used to model a wide variety of psychological findings, but it has also been used in a variety of practical applications. For example, SOAR pilots have flown missions in large-scale simulated military exercises side by side with human pilots, dealing with many issues, including radio "chatter" (Laird et al., 1998).

ACT-R and SOAR have been worked on for decades. Cognitive architecture research is not a game for the fickle or faint of heart. But the benefits are so valuable that a variety of other architectures have sprung up, including Clarion (Sun, 2001), ICARUS (Langley & Choi, 2006), Polyscheme (Cassimatis, 2006), and our own Companion architecture (Forbus et al., 2009). As Newell (1994) noted, for a long time there will need to be many attempts at building unified theories of cognition. The best way to achieve our goals is to establish a community making different bets, so that we collectively explore all of the promising parts of the space.

There have also been a number of attempts at building intelligent architectures from a purely AI perspective, i.e., unfettered by the constraint of handling particular psychological predictions, and focusing instead on raw ability in one or more areas. In reasoning, examples include PRODIGY (Veloso et al., 1995), Cyc (Lenat & Guha, 1990), and SNePS (Shapiro, 2000). Cyc, after a quarter-century of development, is now being used in a variety of applications and the focus has shifted from hand generation of knowledge to using the knowledge base as a foundation for learning systems (Panton et al., 2006). Several interesting efforts are underway to combine deeper reasoning with perception and robotics, often under the rubrics of cognitive vision and/or cognitive robotics (e.g. Needham et al., 2005).

I believe the field needs many more such efforts. Not everyone needs to work on everything: Many important lessons will be learned from architectures whose view on the world is shaped only by texts, and by robot architectures that reason little but can survive in the physical world. The important thing is that all of them are exploring broader swaths of cognitive processing, instead of only studying a single component or process in isolation.

An extremely interesting example of the importance of integration comes from IBM's Watson effort. This system is based on their Deep QA hypothesis, which they describe in their FAQ as: "…by complementing classic knowledge-based approaches with recent advances in NLP, Information Retrieval, and Machine Learning to interpret and reason over huge volumes of widely accessible naturally encoded knowledge (or "unstructured knowledge") we can build effective and adaptable open-domain QA systems."

Watson's performance on the television game Jeopardy! surpassed by far anything else in the field of question answering. The full scientific story behind Watson will be coming out in the next few years, as the team publishes papers on its findings, but what is already available provides some interesting lessons. First, learning by reading can indeed be used to bootstrap knowledge bases at scale (Fan et al., 2010). Second, pervasive confidence estimation was an essential component of success (Ferrucci et al., 2010). Neither of these lessons could have been learned without actually building Watson. Third, the Watson effort provides evidence of the value of focusing on building integrated systems in improving components. At one point during its development, the team's focus shifted to only evaluating improvements in the system as a whole, instead of focusing on individual components in isolation (Baker, 2011). The components now perform better than they did before the start of the effort (Ferrucci, 2012). This suggests that the acid test of operating in an integrated environment strengthened their development.

To be sure, building integrated systems is not easy. But it is getting easier, especially as the number of architectures under active development grows. More reasonable components are becoming available off the shelf, and more architectures available to which component makers can add their components. Architecture research can, of course, be done badly. Most architectures developed specifically for particular funding programs, for example, tend to vanish when the program is over. At best this leaves behind a residue of improved components, albeit with substantial resources wasted on engineering. The best architecture efforts arise from scientific efforts to explore particular hypotheses, and re-using architectures honed through substantial experience seems wiser.

Other voices in the field have argued for reapportioning our efforts, so as to spend more energy building integrated systems. For example, Peter Stone (2007) eloquently argued against the cathedral metaphor in his 2007 Computers and Thought talk, giving examples of how building task–oriented systems had enriched research on reinforcement learning. That is one approach. In the next section I describe another.

4. My Bets

In science one makes bets about what problems and approaches are the most productive to work on. I believe that there are a number of interesting bets that might pay off about constructing minds. In this section, I review the bets that our group is making.

4.1 Build Minds, Not Brains

Cognitive science is a multidisciplinary effort that combines artificial intelligence, psychology, linguistics, philosophy, neuroscience, and anthropology. Its original inspiration was AI's use of computation as a theoretical model: the idea was that computation could serve as a theoretical language for bringing these groups together to productively understand minds. I believe these other disciplines have a lot to offer us, and I believe we have a lot to offer them in return.

Marr's (1982) articulation of levels of explanation provides a valuable methodological lens. He described three different levels of cognitive models:

- *Information-level models* focus on understanding what needs to be computed and why. These models are often formulated as the constraints that a system ought to satisfy (e.g., Bayesian models).
- *Process-level models* focus on understanding how to compute things, including both representations and the algorithms that operate over them (e.g., ACT-R models of solving equations).
- *Implementation-level models* focus on how processes are implemented within particular substrates (e.g., biological systems).

Cognitive science seeks models at all three levels, and different methods are better for tackling problems at different levels. For example, Bayesian models in cognitive science typically focus on the information level, most symbolic models focus on a combination of the information and process levels, and most connectionist models focus on the process and implementation levels.

While others make different bets, my group focuses on the information and process levels, and tends to avoid research at the implementation level. One reason we avoid this level is that

computational modeling of neural systems relies on a scientific understanding of the biology of neural systems, an area that is currently in flux. For example, there is now considerable evidence that glial cells, whose total volume in the human brain is equal to that of neurons, play an important role in how synapses function (Eroglu & Barres, 2010). No current connectionist modeling system that I am aware of includes glial cells, so these models are at best seriously incomplete, and at worst completely wrong. Computational modeling at this level is an important and essential activity, but is best done by those trying to unravel how neural systems work per se.

An analogy between artificial intelligence and "artificial flight" (Ford & Hayes, 1998) is illuminating. The Wright brothers did not start by trying to understand feathers. Instead, they sought the principles underlying flight, what we now call aerodynamics. As aerodynamics and materials became better understood, how feathers worked was ultimately uncovered. Even so, only now are people are starting to build successful flying machines whose wings are, at the implementation level, the same as biological systems. I believe the same will be true with understanding minds. That is, we will achieve human-level artificial intelligences first, and this will help us understand how brains work, rather than the other way around. Others are making different bets, of course.

4.2 Sources of Evidence

One of the strengths of cognitive science research is that many types of evidence can be brought to bear to achieve insights. For example, laboratory studies use instruments as simple as surveys and interviews or as complex as eye trackers, EEG, and neural imaging (e.g. fMRI and MEG). Field studies, in classrooms and across cultures, yield other kinds of data.

The popularity of neuroscience today tempts many AI researchers, and cognitive scientists more broadly, into over-interpreting imaging results. There is evidence that people are more likely to believe a study when there is neuroscience information involved, even when it is irrelevant, and that such information can mask logical flaws in the study, especially for non-experts (Weisberg et al., 2008). There are many excellent scientific uses of neural imaging (e.g., Bowden et al., 2005; Chang et al., 2011; Kuhl & Rivera-Gaxiola, 2008). Alas, not all practitioners are so careful. For example, a survey examining methods used in social neuroscience studies concluded that as many as half have serious methodological flaws (Vul et al., 2009).[1] So while imaging studies are already a valuable source of data, we must be careful consumers of their results, as with any experimental method.

Generally the phenomenon under study determines the appropriate kind of evidence. Newell (1990) proposed a decomposition of cognitive phenomena based on time scale:

- *Biological band*: 10^{-4} to 10^{-2} seconds: organelles to neural circuits
- *Cognitive band*: 10^{-1} to 10^{1} seconds: deliberate acts, basic problem-solving operations
- *Rational band*: 10^{2} to 10^{4} seconds (i.e. minutes to hours)
- *Social band*: 10^{5} to 10^{7} seconds (i.e. days to months)

While incomplete in some ways (e.g., development takes far longer and involves biology as well as social aspects), the basic point is that different tools are appropriate for studying phenomenon

[1] Indeed, Bennett et al. (2009) showed that, using the same method as in some prior imaging studies, fMRI data could be used to argue that a dead Atlantic salmon can determine what emotion a person in a photograph is experiencing.

at different time scales. Neural models tend to focus on the biological and cognitive bands, while traditional cognitive architectures focus on the cognitive and rational bands. (Indeed, the divergence between work on ACT-R and SOAR can be seen as the former pushing into the biological band and the latter pushing into the social band.) Our Companions architecture is focused on the rational and social bands, because of our interest in higher-order cognition, learning, and conceptual change. In testing models of conceptual change, for example, interview data (Friedman et al., 2011), laboratory data (Friedman & Forbus, 2011) and classroom data (Friedman & Forbus, 2010) can all be useful.

4.3 Kinds of Minds

Our group is interested in both building models of minds and building useful cognitive systems. Our working hypothesis, shared by many others, is that creating cognitive systems that are also accurate cognitive models will result in smarter systems. However, at this point it is very much an empirical question: there could be entirely different architectures and methods that lead to equal or better flexibility, learning, and performance than anything biology has produced. Evolution, after all, does not optimize, and the particular solutions used in biological organisms may not be the most efficacious for software organisms. Even those of us who take clues from other fields in cognitive science are generally open to the possibility of deliberately sacrificing aspects of psychological plausibility when thinking about applications. A cognitive system that is superhuman in terms of, for example, working memory limitations or accuracy of memory retrieval might be less good as a cognitive model, but better as a partner working jointly with people. Building systems that complement our strengths and weaknesses could lead to new kinds of cognitive prostheses (Ford et al., 1997).

Progress in comparative cognition – the comparison of human cognition with that of other animals – provides a source of optimism. In light of this literature, phrases like "the mind" seem rather quaint: There are many kinds of minds across the animal kingdom. This suggests that, as we continue to attempt to build minds, there will be interesting intermediate points along the way to full human capabilities. Human-level AI is a goal, but often phenomena are best understood by contrasts. As we get better at building software organisms, comparisons between them and biological intelligences will very likely become a productive source of insights, just as comparative cognition studies are today.

4.4 Analogy, Logic, and Statistics: The Three Pillars of Intelligence

It seems clear that human-level intelligence relies on extremely expressive knowledge representations. (If you doubt this, try to model what it took for you to read and understand this sentence and the previous one.) The study of logic began as an attempt to formalize human reasoning, and it morphed over the centuries into an unparalleled tool for studying and expressing formal arguments. What kind of logic(s) are needed to capture the range of human reasoning is, of course, very much an open question. The statistical revolution in AI brought probabilistic methods for inference and learning that provide ways to handle uncertain information. As the IBM Watson effort illustrates, such information is crucial in building complex systems: knowing what sources of evidence matter in particular contexts, what methods might be best, how likely is an answer produced by a particular method given its inputs. There are a number of attempts to

bring together these two pillars of human intelligence, to bring back our construction metaphor (e.g., Domingos et al., 2006; Milch et al., 2007; Rosenbloom 2010).

I believe these two approaches are incomplete by themselves. There is a substantial body of psychological evidence that analogy is central to human cognition (Gentner, 2003). By analogy, I mean a process of alignment over structured, relational representations, as defined by structure-mapping theory (Gentner, 1983). Such processes also provide a good model for similarity (Gentner & Markman, 1995). Why might analogy be so central? There are functional properties of analogical processing that make it particularly appropriate for intelligent systems. First, since it relies on structured, relational representations, it can handle the expressiveness needed for human cognition. Unlike, say, feature vectors or multidimensional spaces, structure-mapping can (and has) handled causal models, proofs, explanations, plans, stories, and other rich, complex descriptions (e.g., Ouyang & Forbus 2006; Dehghani et al., 2008). Second, it allows examples to be immediately reused, via within-domain analogies. This by itself is sufficient to achieve near transfer over a variety of conditions (Klenk & Forbus, 2009). It can also support deductive and abductive reasoning, by importing an entire proof or argument, without extensive chaining. Third, it supports incremental generalization, which enables relational abstractions to be learned at a faster rate than connectionist or statistical models. So far, our models learn within the same number of examples as required for human learning (e.g. Kuehne et al., 2000, Lockwood et al., 2008). Moreover, analogy solves one of the core problems implicit in probabilistic models, namely determining which aspects of complex stimuli go together (Halstead & Forbus, 2005). As analogical generalizations are built up, the frequencies for each statement occurring in it are derived, thus grounding priors in experience.

4.5 Building Social Organisms

Our current goal for the Companion cognitive architecture is to create software social organisms. Why organisms? One amazing property of minds is their stability. Today's AI systems tend to tread a thin line between catatonia and attention-deficit disorder. They generally cannot survive extended bouts of learning, falling prey to either crippling losses of accuracy (Carlson et al., 2010) or to filling up their memories with useless material. Thinking of software as an organism brings such issues to the forefront. For example, I suspect that the substrate we share with other mammals - e.g., emotions and mechanisms for sensing cognitive state - are part of the solution to the stability problem. Importantly, smarter organisms also tend to be social organisms (Tomasello, 1999). Perhaps this should not be surprising, since much of our knowledge is learned via cultural transmission (Vygotsky, 1962). This suggests that the organisms we build should be social ones. Sociality may offer a robust solution to the accuracy degradation problem, it might accelerate the bootstrapping of intelligent systems, and it could make them more effective collaborators. Hence it seems very important to explore.

5. Conclusions

I have argued that that more of the field's efforts should be spent on attempting to build minds. These will be very simple minds at first, to be sure, just as the first buildings were shacks and lean-tos. Research on particular areas and problems must continue, but with an increased awareness of where their solutions might fit in some broader cognitive system. A thriving relationship between the brick builders and the building crews will benefit both.

The founding of this journal is part of an ongoing attempt to rebalance the field. It is intended to serve as a complement to existing venues, focusing on the issues that arise when building cognitive systems, i.e. minds, even if simple minds, or at least capabilities that are closely tied to what will be needed to understand minds.

We have an unprecedented historical opportunity to make significant progress towards the goal of artificial intelligence – we have made useful progress in many sub-areas, and the resources available (both intellectual and computational) have radically improved (Forbus, 2010). Let us take advantage of it.

Acknowledgements

This essay has benefited from discussions with Dedre Gentner, John Laird, Paul Rosenbloom, Paul Bello, Pat Langley, David Ferrucci, Tom Hinrichs, Andrew Lovett, and Johan de Kleer. The errors remain mine. This essay was written while a Fellow at the Hanse Wissenschaftskolleg, with additional support from the Alexander von Humboldt Foundation.

References

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

Anderson, J. R. (2007). Using brain imaging to guide the development of a cognitive architecture. In W. D. Gray (Ed.), *Integrated models of cognitive systems* (pp. 49–62). New York: Oxford University Press.

Baker, S. (2011). *Final jeopardy: Man vs. machine and the quest to know everything*. Boston: Houghton Mifflin Harcourt.

Carlson, A. Betteridge, J., Kisiel, B., Settles, B., Hruschka, E. R., & Mitchell, T. M. (2010). Toward an architecture for never-ending language learning. *Proceedings of the Twenty-Fourth AAAI Conference on Artificial Intelligence*. Atlanta, GA: AAAI Press.

Cassimatis, N. (2006). A cognitive substrate for human-level intelligence. *AI Magazine, 27*, 45–56.

Cassimatis, N., Bello, P., & Langley, P. (2008). Ability, breadth, and parsimony in computational models of higher-order cognition. *Cognitive Science, 32*, 1304–1322.

Chang, K. K., Mitchell, T., & Just, M. A. (2011). Quantitative modeling of the neural representation of objects: How semantic feature norms can account for fMRI activation. *NeuroImage*, *56*, 716–727.

Dehghani, M., Tomai, E., Forbus, K., & Klenk, M. (2008). An integrated reasoning approach to moral decision-making. *Proceedings of the Twenty-Third AAAI Conference on Artificial Intelligence.* Chicago: AAAI Press.

Domingos, P., Kok, S., Poon, H., Richardson, M., & Singla, P. (2006). Unifying logical and statistical AI. *Proceedings of the Twenty-First National Conference on Artificial Intelligence* Boston: AAAI Press.

Eroglu, C., & Barres, B. (2010). Regulation of synaptic connectivity by glia. *Nature*, *468*, 223–231.

Fan, J., Ferrucci, D., Gondek, D., & Kalyanpur, A. (2010). PRISMATIC: Inducing knowledge from a large scale lexicalized relation resource. *NAACL Workshop on Formalisms and Methodology for Learning by Reading.* Los Angeles, CA.

Ferrucci, D. (2012). Introduction to "This is Watson". *IBM Journal of Research and Development*, *54*, 1–15.

Ferrucci, D., Brown, E., Chu-Carroll, J., Fan, J., Gondek, D., Kalyanpur, A., Lally, J., Murdock, W., Nyberg, E., Prager, J., Schlaefer, N., & Welty, C. (2010). Building Watson: An overview of the Deep QA project. *AI Magazine*, *31*, 59–79.

Forbus, K. (2010). AI and cognitive science: The past and next 30 years. *Topics in Cognitive Science*, *2*, 345–356

Forbus, K., Klenk, M., & Hinrichs, T. , (2009). Companion cognitive systems: Design goals and lessons learned so far. *IEEE Intelligent Systems*, *24*, 36–46.

Ford, K. and Hayes, P. (1998). On computational wings: Rethinking the goals of artificial intelligence. *Scientific American Presents*, *9*, 78–83.

Ford, K., Glymour, C., & Hayes, P. (1997). Cognitive prostheses. *AI Magazine*, *18*, 104.

Friedman, S. E., & Forbus, K. (2010). An integrated systems approach to explanation-based conceptual change. *Proceedings of the Twenty-Fourth AAAI Conference on Artificial Intelligence*. Atlanta, GA: AAAI Press.

Friedman, S., & Forbus, K. (2011). Repairing incorrect knowledge with model formulation and metareasoning. *Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence.* Barcelona, Spain.

Friedman, S. E., Forbus, K. D., & Sherin, B. (2011). Constructing and revising commonsense science explanations: A metareasoning approach. *Proceedings of the AAAI Fall Symposium on Advances in Cognitive Systems*. Arlington, VA: AAAI Press.

Gentner, D. (1983). Structure-mapping: A theoretical framework for analogy. *Cognitive Science*, *7*, 155–170.

Gentner, D. (2003). Why we're so smart. In D. Gentner & S. Goldin-Meadow (Eds.), *Language in mind: Advances in the study of language and thought* (pp. 195–235). Cambridge, MA: MIT Press.

Gentner, D., & Markman, A. B. (1995). Similarity is like analogy: Structural alignment in comparison. In C. Cacciari (Ed.), *Similarity in language, thought and perception* (pp.111–147). Brussels: BREPOLS.

Gordon, J. (2003). *Structures: Or why things don't fall down*. New York: Da Capo Press.

Halstead, D., & Forbus, K. (2005). Transforming between propositions and features: Bridging the gap. *Proceedings of the Twentieth AAAI Conference on Artificial Intelligence*. Pittsburgh, PA: AAAI Press.

Klenk, M., & Forbus, K. (2009). Analogical model formulation for AP physics problems. *Artificial Intelligence*, *173*, 1615–1638.

Koller, D. (2001). Representation, reasoning, and learning. Computers and Thought talk from IJCAI 2001. http://robotics.stanford.edu/~koller/CnT-web.htm

Kuehne, S., Gentner, D., & Forbus, K. (2000). Modeling infant learning via symbolic structural alignment. *Proceedings of the Twenty-Second Annual Meeting of the Cognitive Science Society*. Philadelphia, PA.

Kuhl, P. K., & Rivera-Gaxiola, M. (2008). Neural substrates of language acquisition. *Annual Review of Neuroscience, 31*, 511–534.

Laird, J. & Newell, A. (1983). A universal weak method: Summary of results. *Proceedings of Eighth International Joint Conference on Artificial Intelligence* (pp. 771–773). Karlsruhe.

Laird, J. E., Jones, R. M., & Nielsen, P. E. (1998). Lessons learned from TacAir-Soar in STOW-97. *Proceedings of the Seventh Conference on Computer Generated Forces and Behavioral Representation*. Orlando, FL

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

Lenat, D., & Guha, R. 1990. *Building large knowledge-based systems*. Boston: Addison-Wesley.

Langley, P., & Choi, D. (2006). A unified cognitive architecture for physical agents. *Proceedings of the Twenty-First AAAI Conference on Artificial Intelligence*. Pittsburgh, PA: AAAI Press.

Lockwood, K., Lovett, A., & Forbus, K. (2008). Automatic classification of containment and support spatial relations in English and Dutch. *Proceedings of the International Conference on Spatial Cognition VI: Learning, Reasoning, and Talking about Space* (pp. 283–294). Freiburg, Germany

Marr, D. (1982). *Vision.* New York: W. H. Freeman & Co.

Moskewicz, M., Madigan, C., Zhao, Y., Zhang, L., & Malik, S. (2001). Chaff: Engineering an efficient SAT Solver. *Proceedings of the 39th Design Automation Conference*. Las Vegas.

Needham, C. J., Santos, P. E., Magee, D. R., Devin, V., Hogg, D. C., & Cohn, A. G. (2005). Protocols from perceptual observations. *Artificial Intelligence*, *167*, 103–136.

Milch, B., Marthi, B., Russell, S., Sontag, D., Ong, D. L., & Kolobov, A. (2007). BLOG: Probabilistic models with unknown objects. In L. Getoor & B. Taskar (Eds.), *Introduction to statistical relational learning.* Cambridge, MA: MIT Press.

Newell, A. (1973). You can't play 20 questions with nature and win: Projective comments on the papers of this symposium. In W. G. Chase (Ed.), *Visual information processing*. New York: Academic Press.

Newell, A. (1994). *Unified theories of cognition*. Harvard University Press.

Ouyang, T., & Forbus, K. (2006). Strategy variations in analogical problem solving. *Proceedings of the Twenty-First AAAI Conference on Artificial Intelligence*. Boston: AAAI Press.

Panton, K., Matuszek, C., Lenat, D., Schneider, D., Witbrock, M., Siegel, N., & Shepard, B. (2006). Common sense reasoning – from Cyc to intelligent assistant. In Y. Cai & J. Abascal (Eds.) *Ambient intelligence in everyday life* (pp. 1–31). Springer.

Rosenbloom, P. S. (2010). An architectural approach to statistical relational AI. *Proceedings of the AAAI-10 Workshop on Statistical Relational AI*. Atlanta, GA: AAAI Press.

Russell, S. (1997). Rationality and intelligence. *Artificial Intelligence*, *94*, 57–77.

Shapiro, S. (2000). SNePS: A logic for natural language understanding and commonsense reasoning. In L. Iwanska & S. Shapiro (Eds): *Natural language processing and knowledge representation: Language for knowledge and knowledge for language.* Menlo Park, CA: AAAI Press.

Stone, P. (2007). Learning and multiagent reasoning for autonomous agents. *Proceedings of the Twentieth International Joint Conference on Artificial Intelligence* (pp. 13–30). Hyderabad.

Sun, R. (2001). *Duality of the mind: A bottom-up approach toward cognition*. New York: Psychology Press.

Tomasello, M. (1999). *The cultural origins of human cognition*. Cambridge, MA: Harvard University Press.

Vul, E., Harris, C., Winkielman, P., & Pashler, H. (2009). Puzzlingly high correlations in fMRI studies of emotion, personality, and social cognition. *Perspectives on Psychological Science*, *4*, 274–290.

Weisberg, D., Keil, F., Goodstein, J., Rawson, E., & Gray, J. (2008). The seductive allure of neuroscience explanations. *Journal of Cognitive Neuroscience*, *20*, 470–477.

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

Vygotsky, L. (1962). *Thought and language*. Cambridge, MA: MIT Press.