Intelligent Behavior in Humans and Machines

Pat Langley
Computational Learning Laboratory
Center for the Study of Language and Information
Stanford University, Stanford, CA 94305

Abstract

In this chapter, I review the role of cognitive psychology in the origins of artificial intelligence and in our continuing pursuit of this objective. I consider some key ideas about representation, performance, and learning that had their inception in computational models of human behavior, and I argue that this approach to developing intelligent systems, although no longer common, has an important place in the field. Not only will research in this paradigm help us understand the nature of human cognition, but findings from psychology can serve as useful heuristics to guide our search for accounts of intelligence. I present some constraints of this sort that future research should incorporate, and I claim that another psychological notion – cognitive architectures – is especially relevant to developing unified theories of the mind. Finally, I suggest ways to encourage renewed interaction between AI and cognitive psychology.

1. Introduction

In its early days, artificial intelligence was closely allied with the study of human cognition, to the benefit of both fields. Many early AI researchers were concerned with using computers to model the nature of people's thinking, while others freely borrowed ideas from psychology in their construction of intelligent artifacts. Over the past 20 years, this link has largely been broken, with very few of the field's researchers showing concern with results from psychology or even taking inspiration from human behavior. I maintain that this trend is an unfortunate one which has hurt our ability to pursue two of AI's original goals: to understand the nature of the human mind and to achieve artifacts that exhibit human-level intelligence.

In the pages that follow, I review some early accomplishments of AI in representation, performance, and learning that benefited from an interest in human behavior. After this, I give examples of the field's current disconnection from cognitive psychology and suggest some reasons for this development. I claim that AI still has many insights to gain from the study of human cognition, and that results in this area can serve as useful constraints on intelligent artifacts. In this context, I argue that research on cognitive architectures, a movement that incorporates many ideas from psychology, offers a promising path toward developing unified theories of intelligence. In closing, I consider steps we can take to remedy the current undesirable situation.

2. Early Links Between Artificial Intelligence and Psychology

At it emerged in the 1950s, artificial intelligence incorporated ideas from a variety of sources and pursued multiple goals, but a central insight was that we might use computers to reproduce the complex forms of cognition observed in humans. Some researchers took human intelligence as an inspiration and source of ideas without attempting to model its details. Other researchers, including Herbert Simon and Allen Newell, generally viewed as two of the field's co-founders, viewed themselves as cognitive psychologists who used AI systems to model the mechanisms that underlie human thought. This view did not dominate the field, but it was acknowledged and respected even by those who did not adopt it, and its influence was clear throughout AI's initial phase. The paradigm was pursued vigorously at the Carnegie Institute of Technology, where Newell and Simon based their work, and it was well represented in collections like *Computers and Thought* (Feigenbaum & Feldman, 1963) and *Semantic Information Processing* (Minsky, 1969).

For example, much of the early work on knowledge representation was carried out by scientists who were interested in the structure and organization of human knowledge. Thus, Feigenbaum (1963) developed discrimination networks as a model of human long-term memory and evaluated his EPAM system in terms of its ability to match established psychological phenomena. Hovland and Hunt (1960) introduced the closely related formalism of decision trees to model human knowledge about concepts. Quillian (1968) proposed semantic networks as a framework for encoding knowledge used in language, whereas Schank and Abelson's (1977) introduction of scripts was motivated by similar goals. Newell's (1973) proposal for production systems as a formalism for knowledge that controls sequential cognitive behavior was tied closely to results from cognitive psychology. Not all work in this area was motivated by psychological concerns (e.g., there was considerable work in logical approaches), but research of this sort was acknowledged as interesting and had a strong impact on the field.

Similarly, studies of human problem solving had a major influence on early AI research. Newell, Shaw, and Simon's (1958) Logic Theorist, arguably the first implemented AI system, which introduced the key ideas of heuristic search and backward chaining, emerged from efforts to model human reasoning on logic tasks.[1] Later work by the same team led to means-ends analysis (Newell, Shaw, & Simon, 1961), an approach to problem solving and planning that was implicated in verbal protocols of humans puzzle solving. Even some more advanced methods of heuristic search that are cast as resulting from pure algorithmic analysis have their roots in psychology. For example, the widely used technique of iterative deepening bears a close relationship to progressive deepening, a search method that De Groot (1965) observed in human chess players.

On a related note, research on knowledge-based decision making and reasoning, which emerged during the 1980s, incorporated many ideas from cognitive psychology. The key method for developing expert systems (Waterman, 1986) involved interviewing human experts to determine the knowledge they used when making decisions. Two related movements – qualitative physics (Kuipers, 1994) and model-based reasoning (Gentner & Stevens, 1983) – incorporated insights into the way humans reasoned about complex physical situations and devices. Yet another theoretical

[1] Simon (1981) explicitly acknowledges an intellectual debt to earlier work by the psychologist Otto Selz.

framework that was active at this time, reasoning by analogy (e.g., Gentner & Forbus, 1991), had even closer ties to results from experiments in psychology. Finally, during the same period, research on natural language understanding borrowed many ideas from structural linguistics, which studied the character of human grammatical knowledge.

Early AI research on learning also had close ties to computational models of human learning. Two of the first systems, Feigenbaum's (1963) EPAM and Hovland and Hunt's (1960) CLS, directly attempted to model results from psychological experiments on memorization and concept acquisition. Later work by Anzai and Simon (1979), which modeled human learning on the Tower of Hanoi, launched the 1980s movement on learning in problem solving, and much of the work in this paradigm was influenced by ideas from psychology (e.g., Minton et al., 1989) or made direct contact with known phenomena (e.g., Jones & VanLehn, 1994). Computational models of syntax acquisition typically had close ties with linguistic theories (e.g., Berwick, 1979) or results from developmental linguistics (e.g., Langley, 1983). In general, early machine learning reflected the diverse nature of human learning, dealing with a broad range of capabilities observed in people even when not attempting to model the details.

The alliance between psychological and non-psychological AI had benefits for both sides. Careful studies of human cognition often suggested avenues for building intelligent systems, whereas programs developed from other perspectives often suggested ways to model people's behavior. There was broad agreement that humans were our only examples of general intelligent systems, and that the primary goal of AI was reproduce this capability with computers. Members of each paradigm knew about each others' results and exchanged ideas to their mutual advantage.

3. The Unbalanced State of Modern Artificial Intelligence

Despite these obvious benefits, the past 20 years have seen an increasing shift in AI research away from concerns with modeling human cognition and a decreasing familiarity with results from psychology. What began as a healthy balance between two perspectives on AI research, sometimes carried on within individual scientists, has gradually become a one-sided community that believes AI and psychology have little to offer each other. Here I examine this trend in some detail, first describing the changes that have occurred in a number of research areas and then considering some reasons for these transformations.

Initial research on knowledge representation, much of which was linked to work in natural language understanding, led to notations such as semantic networks and scripts, which borrowed ideas from psychology and which influenced that field in return. Over time, researchers in this area became more formally oriented and settled on logic as the proper notation to express content. Many became more concerned with guarantees about efficiently processing than with matching the flexibility and power observed in human knowledge structures, leading to representational frameworks that made little contact with psychological theories.

The earliest AI systems for problem solving and planning drew on methods like means-ends analysis, which were implicated directly in human cognition. For mainly formal reasons, these were gradually replaced with algorithms that constructed partial-order plans, which required more memory but retained the means-ends notion of chaining backward from goals. More recently, these have been replaced with “disjunctive” techniques, which reformulate the planning task in terms of constraint satisfaction and which bear little resemblance to problem solving in humans.

Early work on natural language processing aimed to produce the same deep understanding of sentences and discourse as humans exhibit. However, constructing systems with this ability was time consuming and their behavior was fragile. Over time, they have been largely replaced with statistical methods that have limited connection to results from linguistics or psycholinguistics. More important, they focus on tasks like information retrieval and information extraction that have little overlap with the broad capabilities found in human language processing.

Research in machine learning initially addressed a wide range of performance tasks, including problem solving, reasoning, diagnosis, natural language, and visual interpretation. Many approaches reproduced the incremental nature of human learning, and they combined background knowledge with experience to produce rates of improvement similar to those found in people. However, during the 1990s, work in this area gradually narrowed to focus almost exclusively on supervised induction for classification and reinforcement learning for reactive control. This shift was accompanied by an increased emphasis on statistical methods that require large amounts of data and learn far more slowly than humans.

There are many reasons for these developments, some of them involving technological advances. Faster computer processors and larger memories have made possible new methods that operate in somewhat different ways than people. Current chess-playing systems invariably retain many more alternatives in memory, and look much deeper ahead, than human players. Similarly, many supervised induction methods process much larger data sets, and consider a much larger space of hypotheses, than human learners. This does not mean they might fare even better by incorporating ideas from psychology, but these approaches have led to genuine advances.

Unfortunately, other factors have revolved around historical accidents and sociological trends. Most academic AI researchers have come to reside in computer science departments, many of which grew out of mathematics units and which have a strong formalist bent. Many faculty in such departments view connections to psychology with suspicion, making them reluctant to hire the few scientists who attempt to link the two fields. For such researchers, mathematical tractability becomes a key concern, leading them to restrict their work to problems they can handle analytically, rather than ones that humans tackle heuristically. Graduate students are inculcated in this view, so that new PhDs pass on the bias when they take positions.

The mathematical orientation of many AI researchers is closely associated with an emphasis on approaches that guarantee finding the best solutions to problems. This concern with optimizing stands in stark contrast with early AI research, which incorporated the idea of satisficing (Simon, 1955) from studies of human decision making. The notion of satisficing is linked to a reliance on heuristics, which are not guaranteed to find the best (or even any) solutions, but which usually produce acceptable results with little effort. Formalists who insist on optimality are often willing to restrict their attention to classes of problems for which they can guarantee such solutions, whereas those willing to use heuristic methods are willing to tackle more complex tasks.

Another factor that has encouraged a narrowed scope of research has been the commercial success of AI technology. Methods for diagnosis, supervised learning, scheduling, and planning have all found widespread uses, many of them supported by the content made available on the Web. The financial success and commercial relevance of these techniques have led many academics to focus on the narrowly defined problems like text classification, causing a bias toward near-term applications and an explosion of work on “niche AI” rather than on complete intelligent systems. Component algorithms are also much easier to evaluate experimentally, a lesson that has been reinforced by the many problem repositories and competitions that have become common.

Taken together, these influences have discouraged most AI scientists from making contact with ideas from psychology. Research on computational models of human cognition still exists, but it is almost invariably carried out in psychology departments. This work is seldom referred to as artificial intelligence or published in that field’s typical outlets, even though the systems developed meet all the traditional AI criteria. However, the standards for success in this area tend to be quite different, in that they focus on quantitative fits to reaction times or error rates observed in humans, rather than the ability to carry out a task efficiently or accurately. This leads developers to incorporate parameters and other features that hold little interest to non-psychological researchers, encouraging the mistaken impression that this work has little to offer the broader AI community.

However, a few scientists have instead concentrated on achieving qualitative matches to human behavior. For example, Cassimatis (2004) describes such a theory of language processing, whereas Langley and Rogers (2005) report a theory of human problem solving that extends Newell, Shaw, and Simon’s (1958, 1961) early work. Their goal has been not to develop narrow models that fit the detailed results of psychological experiments, which to them seems premature, but rather to construct models that have the same broad functionality as we find in people. As Cassimatis (2006) has argued, research in this paradigm has more to offer mainstream artificial intelligence, and it comes much closer in spirit to the field’s early efforts at modeling human cognition.

In summary, over the past few decades, AI has become increasingly focused on narrowly defined problems that have immediate practical applications or that are amenable to formal analysis. Concerns with processing speed, predictive accuracy, and optimality have drawn attention away from the flexible forms of intelligent behavior at which humans excel. Research on computational modeling does occur in psychology, but it emphasizes quantitative fits to experimental results and also exhibits a narrowness in scope. The common paradigms differ markedly from those pursued at the dawn of artificial intelligence.

4. Promised Benefits of Renewed Interchange

Clearly, the approach to developing intelligent systems adopted by early AI researchers is no longer a very common one, but I maintain that it still has an important place in the field. The reasons for pursuing research in this paradigm are the same as they were 50 years ago, at the outset of our discipline. However, since most AI scientists appear to have forgotten them, I repeat them here at the risk of stating the obvious to those with longer memories. I also discuss a paradigm that holds special promise for bridging the rift.

First, research along these lines will help us understand the nature of human cognition. This is a worthwhile goal in its own right, not only because it has implications for education and other applications, but because intelligent human behavior is an important set of phenomena that demand scientific explanation. Understanding the human mind in all its complexity remains an open and challenging problem that deserves more attention, and computational models of cognition offer the best way to tackle it.

Second, findings from psychology can serve as useful heuristics to guide our development of intelligent artifacts. They tell us ways that people represent their knowledge about the world, the processes that they employ to retrieve and use that knowledge, and the mechanisms by which they acquire it from experience. AI researchers must make decisions about these issues when designing intelligent systems, and psychological results about representation, performance, and learning provide reasonable candidates to consider. Humans remain our only example of general intelligent systems and, at the very least, insights about how they operate should receive serious consideration in the design of intelligent artifacts. Future AI research would benefit from increased reliance on such design heuristics.

Third, observations of human capabilities can serve as an important source of challenging tasks for AI research. For example, we know that humans to understand language at a much deeper level than current systems, and attempting to reproduce this ability, even by means quite different than those people use, is a worthwhile endeavor. Humans can also generate and execute sophisticated plans that trade off many competing factors and that achieve multiple goals. They have the ability to learn complex knowledge structures in an incremental manner from few experiences, and they exhibit creative acts like composing musical pieces or devising scientific theories. Most current AI research sets its sights too low by focusing on simpler tasks like classification and reactive control, many of which can hardly be said to involve intelligence. Psychological studies reveal the impressive abilities of human cognition, and thus serve to challenge the field and pose new problems that require extensions to existing technology.

In addition, I claim that another psychological notion – *cognitive architectures* (Newell, 1990; Langley, Laird, & Rogers, 2006) – is especially relevant to developing unified theories of the mind. A cognitive architecture specifies aspects of an intelligent system that are stable over time, much as in a building's architecture. These include the memories that store perceptions, beliefs, and knowledge, the representation of elements that are contained in these memories, the performance mechanisms that use them, and the learning processes that build on them. Such a framework typically comes with a programming language and software environment that supports the efficient construction of knowledge-based systems.

Most research on cognitive architectures shares a number of important theoretical assumptions. These include claims that:

- short-term memories are distinct from long-term memories, in that the former contain dynamic information and the latter store more stable content;
- both short-term and long-term memories contain symbolic list structures that can be composed dynamically during performance and learning;

- the architecture accesses elements in long-term memory by matching their patterns against elements in short-term memory;
- cognition operates in cycles that retrieve relevant long-term structures, then use selected elements to carry out mental or physical actions; and
- learning is incremental, being tightly interleaved with performance, and involves the monotonic addition of new symbolic structures to long-term memory.

Most of these ideas have their origins in theories of human memory, problem solving, reasoning, and skill acquisition. They are widespread in research on cognitive architectures, but they remain relatively rare in other branches of artificial intelligence to their detriment. The benefits of this paradigm include a variety of constraints from psychology about how to approach building intelligent systems, an emphasis on demonstrating the generality of architectures across a variety of domains, and a focus on systems-level research that moves beyond component algorithms toward unified theories of intelligence.

Research on cognitive architectures varies widely in the degree to which it attempts to match psychological data. ACT-R (Anderson & Lebiere, 1998) and EPIC (Kieras & Meyer, 1997) aim for quantitative fits to reaction time and error data, whereas Prodigy (Minton et al., 1989) incorporates selected mechanisms like means-ends analysis but otherwise makes little contact with human behavior. Architectures like Soar (Laird, Newell, & Rosenbloom, 1987; Newell, 1990) and Icarus (Langley & Choi, in press; Langley & Rogers, 2005) take a middle position, drawing on many psychological ideas but also emphasizing their strength as flexible AI systems. What they hold in common is an acknowledgement of their debt to theoretical concepts from cognitive psychology and a concern with the same intellectual abilities as humans.

5. Healing the Intellectual Rift

If we assume, for the sake of argument, that artificial intelligence would indeed benefit from renewed links to cognitive psychology, then there remains the question of how we can best achieve this objective. As usual for complex problems, the answer is that we must take many steps, each of which will take us some distance toward our goal.

Clearly, one response must involve education. Most AI courses ignore connections to cognitive psychology, and few graduate students read papers that are not available on the Web, which means they are not even aware of early paradigms. We need to provide a broader education in AI that incorporates ideas from cognitive psychology, linguistics, and logic, which are far more important to the field's original agenda than ones from mainstream computer science. One example comes from a course on reasoning and learning in cognitive systems (http://cll.stanford.edu/reason-learn/), which I have offered through the Computer Science Department at Stanford University (despite some internal opponents who claimed it did not belong there), but we need many more.

We should also encourage more research within the cognitive modeling tradition. The past few years have seen encouraging developments in funding agencies like NSF and DARPA, with the latter supporting a number of large-scale programs with an emphasis on "cognitive systems" (Brachman

& Lemnios, 2002). These projects have already galvanized many researchers to work on intelligent systems with the same broad capabilities as found in humans, and ideas from cognitive psychology have found their way into a number of these efforts. In terms of publication venues, the annual AAAI conference has added a distinct track for integrated systems that includes cognitive models as one theme, and a recent AAAI Spring Symposium (Lebiere & Wray, 2006) dealt explicitly with relations between AI and cognitive science. We need more changes along these lines, but recent events have moved us in the right direction.

In summary, the original vision of AI was to understand the principles that support intelligent behavior and to use them to construct computational systems with the same breadth of abilities as humans. Many early systems doubled as models of human cognition, while others made effective use of ideas from psychology. The last two decades have seen far less research in this tradition, and work by psychologists along these lines is seldom acknowledged by the AI community. Nevertheless, observations of intelligent behavior in humans can provide an important source of mechanisms and tasks to drive research. Without them, AI seems likely to become a set of narrow, specialized subfields that have little to tell us about the nature of the mind. Over the next 50 years, AI must reestablish connections with cognitive psychology if it hopes to achieve its initial goal of achieving human-level intelligence.

Acknowledgements

This material is based on research sponsored by DARPA under agreement number FA8750-05-2-0283. The U. S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright notation thereon. Discussions with Herbert Simon, Allen Newell, John Anderson, David Nicholas, John Laird, Randy Jones, and many others over the years have contributed to the ideas presented in this chapter.

References

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

Anzai, Y., & Simon, H. A. (1979). The theory of learning by doing. *Psychological Review*, *86*, 124–140.

Berwick, R. (1979). Learning structural descriptions of grammar rules from examples. *Proceedings of the Sixth International Conference on Artificial Intelligence* (pp. 56–58). Tokyo: Morgan Kaufmann.

Brachman, R., & Lemnios, Z. (2002). DARPA's new cognitive systems vision. *Computing Research News*, *14*, 1.

Cassimatis, N. L. (2004). Grammatical processing using the mechanisms of physical inferences. *Proceedings of the Twentieth-Sixth Annual Conference of the Cognitive Science Society*. Chicago.

Cassimatis, N. L. (2006). Cognitive science and artificial intelligence have the same problem. In C. Lebiere & R. Wray (Eds.). (2006). *Between a rock and a hard place: Cognitive science principles meet AI-hard problems*. Technical Report SS-06-02, AAAI, Menlo Park, CA.

de Groot, A. D. (1965). *Thought and choice in chess*. The Hague: Mouton.

Feigenbaum, E. A. (1963). The simulation of verbal learning behavior. In E. A. Feigenbaum & J. Feldman (Eds.), *Computers and thought*. New York: McGraw-Hill.

Feigenbaum, E. A., & Feldman, J. (Eds.) (1963). *Computers and thought*. New York: McGraw-Hill.

Gentner, D., & Forbus, K. (1991). MAC/FAC: A model of similarity-based retrieval. *Proceedings of the Thirteenth Annual Conference of the Cognitive Science Society* (pp. 504–509). Chicago: Lawrence Erlbaum.

Gentner, D., & Stevens, A. L. (1983). *Mental models*. Hillsdale, NJ: Lawrence Erlbaum.

Hovland, C. I. & Hunt, E. A. (1960). The computer simulation of concept attainment. *Behavioral Science*, *5*, 265–267.

Jones, R. M., & VanLehn, K. (1994). Acquisition of children's addition strategies: A model of impasse-free, knowledge-level learning. *Machine Learning*, *16*, 11–36.

Kieras, D., & 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.

Kuipers, B. (1994). *Qualitative reasoning: Modeling and simulation with incomplete knowledge*. Cambridge, MA: MIT Press.

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

Langley, P. (1982). Language acquisition through error recovery. *Cognition and Brain Theory*, *5*, 211–255.

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

Langley, P., Laird, J. E., & Rogers, S. (2006). *Cognitive architectures: Research issues and challenges*. Technical Report, Computational Learning Laboratory, CSLI, Stanford University, CA.

Langley, P., & Rogers, S. (2005). An extended theory of human problem solving. *Proceedings of the Twenty-seventh Annual Meeting of the Cognitive Science Society*. Stresa, Italy.

Lebiere, C., & Wray, R. (Eds.). (2006). *Between a rock and a hard place: Cognitive science principles meet AI-hard problems*. Technical Report SS-06-02, AAAI, Menlo Park, CA.

Minsky, M. (Ed.) (1969). *Semantic information processing*. Cambridge, MA: MIT Press.

Minton, S., Carbonell, J. G., Knoblock, C. A., Kuokka, D., Etzioni, O., & Gil, Y. (1989). Explanation-based learning: A problem solving perspective. *Artificial Intelligence*, *40*, 63–118.

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

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

Newell, A., Shaw, J. C., & Simon, H. A. (1958). Elements of a theory of human problem solving. *Psychological Review*, *65*, 151–166.

Newell, A., & Simon, H. A. (1961). GPS, A program that simulates human thought. In H. Billing (Ed.), *Lernende automaten*. Munich: Oldenbourg KG. Reprinted in E. A. Feigenbaum & J. Feldman (Eds.), *Computers and thought*. New York: McGraw-Hill, 1963.

Quillian, M. R. (1968). Semantic memory. In M. Minsky (Ed.), *Semantic information processing*. Cambridge, MA: MIT Press.

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

Simon, H. A. (1955). A behavioral model of rational choice. *Quarterly Journal of Economics*, *69*, 99–118.

Simon, H. A. (1981). Otto Selz and information-processing psychology. In N. H. Frijda & A. de Groot (Eds.), *Otto Selz: His contribution to psychology*. The Hague: Mouton.

Waterman, D. A. (1986). *A guide to expert systems*. Reading, MA: Addison-Wesley.