Building Bridges between AI and Cognitive Psychology

Stephen K. Reed

■ *My goal in this article is to encourage greater integration of the fields of AI and cognitive psychology by reviewing work on shared interests. I begin with examples that link my early research related to AI with my current efforts to organize knowledge in the cognitive sciences. I then describe how cognitive psychologists have contributed to the methods explained in* The Master Algorithm: How the Quest for the Ultimate Learning Machine Will Remake Our World *(Domingos, 2015), including how these methods can be combined. The final section discusses three benefits of building bridges: using computational models in AI as theoretical models in cognitive psychology, solving joint computational problems, and facilitating the interactions between people and machines.*

One might expect that there would be many bridges connecting AI implemented in computers with natural intelligence implemented in people. However, I have been both surprised and disappointed by the lack of cross references between articles on artificial intelligence written by computer scientists and articles on natural intelligence written by cognitive psychologists. I am surprised because articles on AI have informed and inspired my own work as a cognitive psychologist. I begin with examples that link my early research related to AI with my current efforts to organize knowledge in the cognitive sciences.

Copyright © 2019, Association for the Advancement of Artificial Intelligence. All rights reserved. ISSN 0738-4602

I next describe a more extended effort by reviewing how cognitive psychologists have contributed to the methods explained in the book *The Master Algorithm: How the Quest for the Ultimate Learning Machine Will Remake Our World* (Domingos 2015). The objective of *The Master Algorithm* is to inform readers about the different computational methods used in machine learning and to encourage them to reflect on how these methods can be combined to develop algorithms that would be more powerful than any of the individual methods. I conclude by proposing three benefits from greater collaboration between AI and cognitive psychology.

Personal Examples

I discovered as a graduate student majoring in mathematical psychology at UCLA that it was easier to borrow computational methods than to invent my own. The course that had the most influence on my dissertation was an engineering course on mathematical models of pattern recognition. The models were based on exemplars, prototypes, nearest neighbors, and feature probabilities. I wondered which of these would best predict how people would classify patterns, so I ran a series of experiments in which participants classified patterns into two categories consisting of schematic faces (Reed 1972). As indicated in *The Master Algorithm* (Domingos 2015), these categorization methods continue to be refined as methods for machine learning. They also continue to be refined by cognitive psychologists, as I will explain later.

The same year that I published my dissertation, Newell and Simon (1972) published their classic book *Human Problem Solving*. The book offered new insights into studying human problem solving based in part on their initial efforts in AI. I was particularly interested in how the structure of the problem space constrained problem solving, so I joined forces with two AI faculty members, George Ernst and Ran Banerji, at Case Western Reserve to study transfer between the missionary-cannibal problem and a more challenging variation called the jealous husbands problem (Reed, Ernst, and Banerji 1974). We were surprised to find that there was no transfer in reduced solution time between these two variations of the missionary–cannibal problem unless students were given a hint that missionaries corresponded to husbands and wives corresponded to cannibals. The hint aided transfer from the jealous husbands problem to the missionary–cannibal problem but not in the other direction. The insight of Ernst and Banerji that there was a one-to-many mapping of moves from the missionary–cannibal to the jealous husbands problem helped explain this asymmetric transfer.

My initiation into problem solving continued when I began a visiting appointment at Carnegie Mellon University, which gave me the opportunity to work with Herb Simon. I walked into his office in January 1975 armed with data on the effects of a subgoal on solving a variation of the missionary-cannibal problem that required transporting five missionaries and five cannibals across a river on a boat holding three people. The subgoal reduced the average number of moves from 30 for the control group to 20 for the subgoal group. Simon suggested that we develop a stochastic simulation model to predict the average number of moves between each problem state for both of the groups. The resulting strategy-shift model proposed that the subgoal facilitated the shift from an unsuccessful balance strategy to a successful means-end strategy (Simon and Reed 1976).

Two years before I joined Herb Simon on this project, he published "The Structure of Ill Structured Problems" in the journal *Artificial Intelligence* (Simon 1973). Well-structured (puzzle) problems can be represented by a problem space consisting of well-defined initial and goal states that are connected by legal moves. Simon considered the missionary-cannibal problem to be a prototypical example. In contrast, the initial, goal, and intermediate states of ill-structured (design) problems are incompletely specified. Most subsequent reviews of problem solving, including my own (Reed 2016a), had ignored ill-structured problems. I therefore decided to examine Simon's (1973) claim that information-processing principles apply to all problems but apply differently as problems become more ill structured. My article analyzed the similarities and differences among puzzles, insight puzzles, classroom problems, and ill-structured design problems within a theoretical framework consisting of representation construction, schema activation, analogical reasoning, and heuristic search (Reed 2016b). It supported Simon's claim that most ill-structured problems can be decomposed into well-structured subproblems.

Cognitive architectures provide helpful theoretical frameworks for representing tasks, but most, such as Soar (Laird 2012; Laird, Newell, and Rosenbloom 1987), emphasize encoding knowledge as symbols. However, advances in incorporating visuospatial reasoning into cognitive architectures such as biSoar (Chandrasekaran et al. 2011) and Soar/SVS (Lathrop, Wintermute, and Laird 2011) include a visual buffer that supports analog operations. Figure 1 shows the components of Soar/SVS.

I apply Soar/SVS to show which of its major components (visual buffer, spatial scene, predicate extraction, predicate projection, visual generation, procedural memory) are involved in a variety of spatial reasoning tasks (Reed 2019). Although the architecture works well for modeling human visuospatial reasoning, it works less well for modeling human pattern recognition because of the rapid recognition of patterns, in contrast to the slower speeds of cognition and search (Smith and Eckroth 2017). Embedding a neural network model, such as the interactive activation model (McClelland and Rumelhart 1981), within Soar/SVS would create a hybrid model to take advantage of the fast recognition of neural networks and the slower reasoning based on symbolic rules. The next section

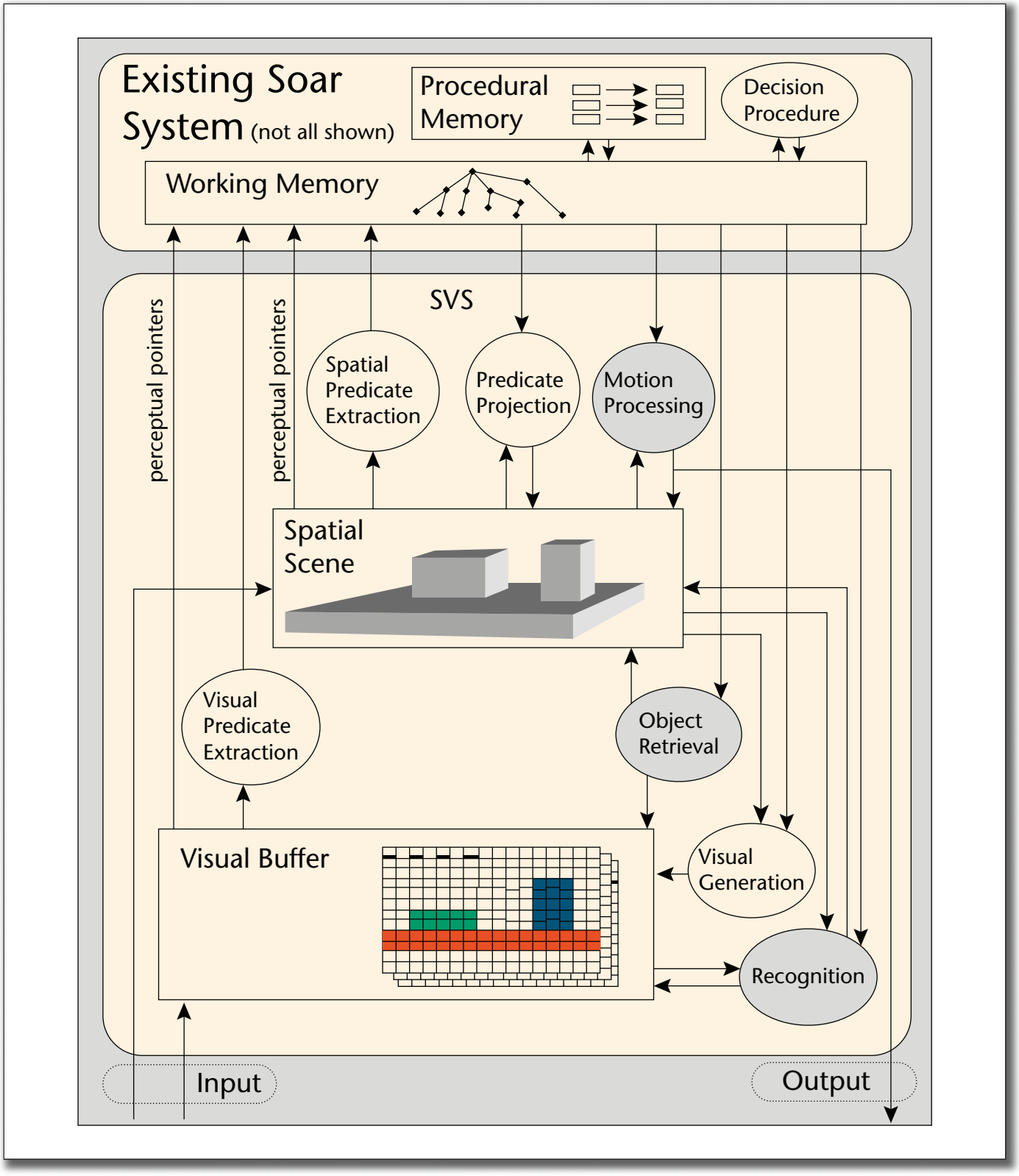

Figure 1. The Soar/SVS Cognitive Architecture.

From Lathrop, Wintermute, and Laird (2011). Reproduced with permission from John Wiley and Sons, ©2010.

discusses a book that seeks to create more powerful learning methods through such combinations.

Analogizers, Bayesians, Connectionists, and Symbolists

In his book *The Master Algorithm: How the Quest for the Ultimate Learning Machine Will Remake Our World*, Pedro Domingos provides many examples demonstrating that machine learning is all around us (Domingos, 2015). Machine learning decides what information to show us when we type a query into a search engine. It filters spam from our e-mails. It makes recommendations when we buy a book from Amazon or select a video from Netflix. It helps pick stocks for our mutual funds. Domingos describes five tribes of machine learning: analogizers, Bayesians, connectionists, evolutionaries, and symbolists. His book gives readers a clear introduction to these methods and their applications. It encourages thinking about how combining the methods can make them more effective.

For analogizers, learning consists of recognizing similarities between situations. One method of measuring similarity is to plot patterns in a multidimensional space and use the distance between them as the measure of similarity. A simple application is to classify a new pattern by selecting the category that contains its nearest neighbors. Variations include comparing a pattern's similarity to all the patterns in a category or to a category prototype that represents the central tendency of the category. Another method is to find a mathematical formula that describes the boundary separating the two categories. Note that Domingos uses the term *analogizers* as a generic term for similarity-based reasoning rather than its more restricted use as reasoning based on a single analogy.

According to Bayesians, learning is a form of uncertain inference that uses Bayes's theorem to incorporate new data into beliefs. A prior probability of a hypothesis becomes a posterior probability after seeing the data. According to the theorem,

$$P(\text{hypothesis} \mid \text{data}) = P(\text{hypothesis}) \times P(\text{data} \mid \text{hypothesis})/P(\text{data})$$

The formula states that the probability of the hypothesis after incorporating the data are equal to the prior probability of the hypothesis times the probability of the data given the hypothesis divided by the probability of the data. Updated probabilities are therefore based on both prior probabilities and evidence.

Connectionists reverse engineer what the brain does by adjusting the strength of connections between neurons. They compare a system's output to the desired one and then change connection weights in layers of neurons to reduce error by using a method called back propagation. Connectionist learning differs from symbolic learning because concepts are distributed across neurons rather than represented by a one-to-one correspondence between concepts and symbols. Another difference is that all connection weights are revised in parallel, whereas symbolic methods are sequential.

Symbolic approaches in the framework of Domingos (2015) are associated with knowledge engineering, in which knowledge is programmed into the computer by experts rather than discovered by learning algorithms. Knowledge for the symbolists occurs by manipulating symbols that replace expressions with other expressions. Manipulating symbols to solve problems typically occurs by learning rules that combine different pieces of preexisting knowledge. Rules can be expressed in logic such as, "If gene *A* is expressed and gene *B* is not, then gene *C* is expressed." An important kind of rule learning is inverse deduction, which identifies missing knowledge needed to make a deduction.

Evolutionaries simulate natural selection to evolve computer programs. A key problem is learning structure, rather than adjusting connection weights.

With the exception of the evolutionaries, members of the five camps can be easily found among psychologic scientists. Although Domingos (2015) emphasized machine learning in *The Master Algorithm*, he also mentioned some contributions by cognitive psychologists. My objective in the next section is the opposite — I will emphasize the contributions of cognitive psychologists.

Cognitive Psychologists' Contributions

It is easy to find analogizers, Bayesians, connectionists, and symbolists among cognitive psychologists. Here are a few examples of their contributions.

Analogizers

The analogizers (in Domingos' terminology) have developed methods to categorize patterns based on their similarity to other patterns. Four generic methods for representing similarity in psychology use geometry, features, alignment, and transformations (Goldstone and Son 2005; Hahn 2014). The geometric approach measures proximity in a multidimensional space, the feature approach examines the number of shared and unique features, the alignment approach creates a mapping between structured elements, and the transformation approach finds the number of transformations required to convert one pattern into another. This section discusses representing patterns in a multidimensional space and assigning weights to features based on their usefulness in distinguishing between categories. The section on the symbolists discusses alignment and transformations.

A geometric measure of the similarity between two patterns is the distance between them in a multidimensional space. Similar patterns have similar coordinate values and are therefore closer to each other. Although a variety of distance measures have been

used in machine learning (Biehl, Hammer, and Villmann 2016), the most frequently used measure in psychology is the Minkowski metric depicted in Equation 1 (Goldstone and Son 2005). The distance between two entities **x** and **y** is the sum of the absolute difference between their coordinate values (x_j, y_j) on each dimension of an *N*-dimensional space. A special case is Euclidean distance in which the exponent is $r = 2$.

$$d(\mathbf{x}, \mathbf{y}) = \left[\sum |x_j - y_j|^r\right]^{1/r} \text{ for } j = 1, N \text{ dimensions.} \quad (1)$$

For Equation 1 to have a symbolic interpretation, the dimensions of the space should be interpretable as recognizable features. This does not always occur or may only imperfectly occur. For instance, a four-dimensional interpretation of animal terms could be imperfectly interpreted as mammalian/nonmammalian, water/land/air, mundane/mythical, and unpleasant/pleasant (Goldstone and Son 2005). Creation of artificial stimuli, such as the schematic faces in figure 2, increases the chance of identifying the dimensions (Reed 1972). The scaling program used in this study aligned the four-dimensional Euclidean solution of the similarity judgments with the physical dimensions of forehead (eye height), eyes (separation), nose (length), and mouth (height).

A limitation of the distance formula in Equation 1 is the assumption that all feature dimensions are equally weighted, and therefore it cannot account for those situations in which some features are better at discriminating between categories. A class-separating transformation (Sebestyen 1962) provides a normative method by weighting features to reduce distances between patterns in the same category and increase distances between patterns in different categories. Equation 2 shows a weighted features distance model that enables higher weights for the more discriminative feature dimensions.

$$d(\mathbf{x}, \mathbf{y}) = \left[\sum w_j^r |x_j - y_j|^r\right]^{1/r} \text{ for } j = 1, N \text{ dimensions.} \quad (2)$$

Applying the class-separating transformation to the faces in figure 2 produced weights of 0.46 for forehead height, 0.24 for eye separation, 0.24 for nose length, and 0.06 for mouth height when normalized to sum to 1. The weighted feature distances improved the prediction of both prototype and exemplar models, indicating that participants placed more emphasis on discriminating features. Their ratings of feature usage also confirmed that they emphasized the more discriminative features in their decisions (Reed 1972).

Another confirmation of the differential weighting of features was found by Nosofsky (1986), whose context model generalized the exemplar theory of categorization developed by Medin and Schaffer (1978). Nosofsky evaluated the model on two observers who categorized stimuli composed of semicircles that varied in four levels of size and four levels of the angle of a radial line. Parameter estimates revealed support for the hypothesis that the classifiers distributed their attention across the size and angle attributes so as to optimize classifications.

Bayesians

The Bayesian approach is illustrated by Anderson's (1991) article on the adaptive nature of categorization, in which similar instances are classified in the same category. Anderson applied his rational (Bayesian) model to a wide range of findings, including category learning. In developing an adaptive theory, Anderson proposed that the first step is to specify what the system is trying to optimize. The second step requires making assumptions about the structure of the environment. The third step makes assumptions about the costs incurred in trying to achieve optimal performance. Anderson applied a Bayesian model to a wide variety of tasks and discovered the model performed as well as the specialized models developed by theorists for each of these tasks.

The categorization tasks studied by Anderson typically consisted of four or five exemplars in each of two categories. A more recent study used a Bayesian analysis to generalize from a single visual exemplar of 20 distinct letter-like forms (Lake, Salakhutdinov, and Tenenbaum 2015). The authors presented Bayesian program learning as an alternative to machine learning methods that require tens or hundreds of examples. They represented the visual characters by probabilistic structural procedures that combine primitives into subparts, subparts into parts, and parts into objects. Their program builds rich concepts from simpler primitives that attempt to both recognize and generate other examples of an object. Although generally successful on both the recognition and generation tasks, Bayesian program learning was nonetheless less successful than people because it lacked explicit knowledge of visual structure such as parallel lines, symmetry, and connections between the ends of strokes. From Anderson's (1991) perspective, this limitation was caused by its less well articulated representation of environmental structure.

The hierarchical Bayesian method has been used to model causal relations. Gopnik and Wellman (2012) applied Bayesian models to relate a higher-level framework theory (input–process–outcome) to specific inputs, processes, and outcomes. Figure 3 contrasts three different representations of this approach to illustrate how children could learn the role of biologic processes in explaining causality. Graph A shows only direct connections between inputs and outputs, such as that getting sleep helps a person run faster and avoid illness. Graph B illustrates how an intervening process (metabolize energy) links specific inputs (causes) to specific outputs (effects). Graph C represents further growth of knowledge by showing how two different processes (metabolize energy and immune defense) link causes to their effects: Getting sleep helps people run faster by metabolizing energy. It

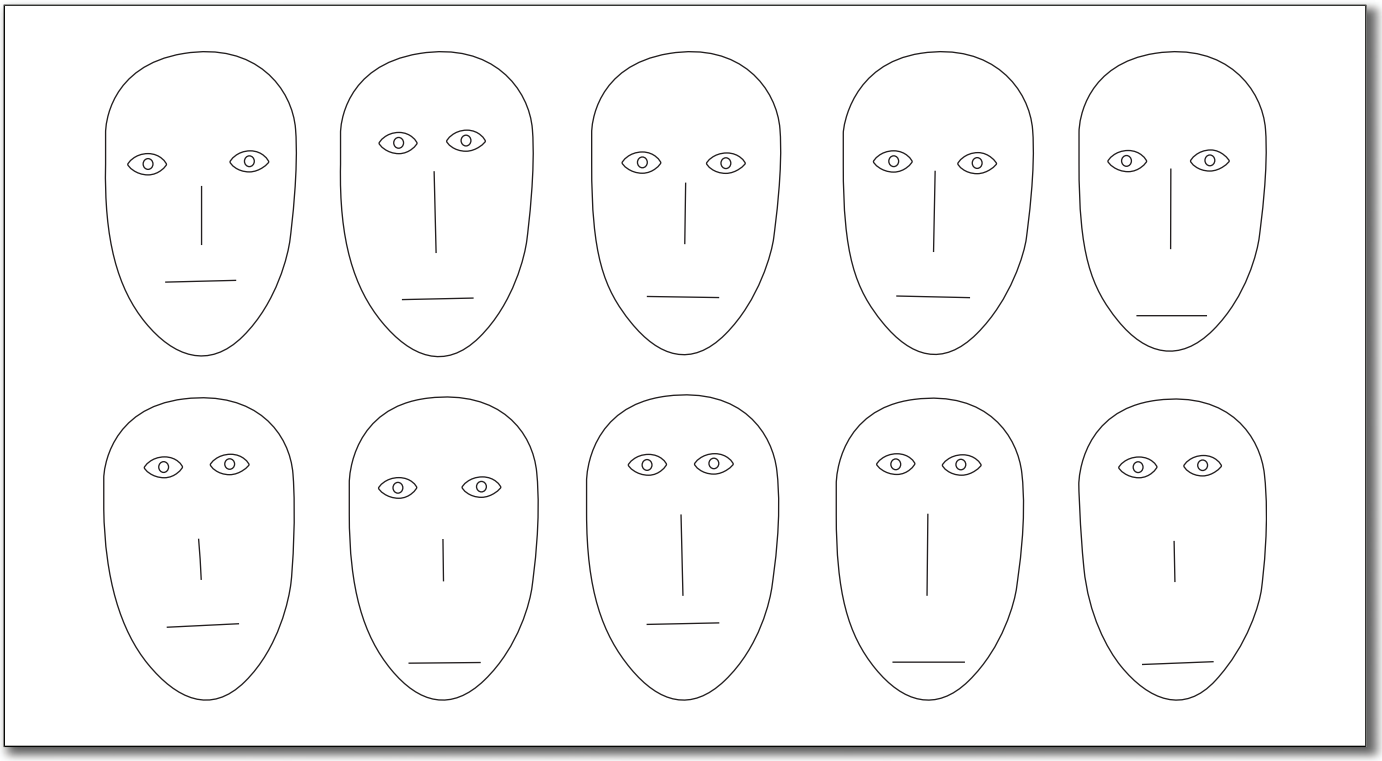

Figure 2. Two Categories of Schematic Faces.

From Reed and Friedman (1973). Reproduced with permission from Springer Nature, ©1973 the Psychonomic Society.

also helps people avoid illness through improving their immune defense.

The Bayesian perspective is a very general approach to child development that requires specific hypotheses to test developmental theories (Gopnik and Bonawitz 2014). Its principle advantage is that it allows theories to be formulated in precise and transparent ways.

Connectionists

The connectionists have also addressed the problem of modeling similarity but from a different perspective than the rational analysis used by Anderson (1990) and the Bayesians. Rogers and McClelland (2014) discuss this distinction within the framework of Marr's (1982) levels of analysis. The rational level corresponds to Marr's computational level, which focuses on an analysis of the problem, including mathematical methods to solve it. In contrast, the connectionist approach considers how the brain — neurons and their connections — constrain the nature of the solutions as formulated by Rumelhart, Hinton, and McClelland (1986).

The TRACE model of auditory word recognition (McClelland and Elman 1986) is a typical example. The model contains three layers to represent the temporal dynamics of word recognition by taking as input (1) auditory features such as voiced and acute (2) that activate phonemes (3) that activate a word. Words are recognized incrementally by increasing the activation level of the correct phoneme and word units. Activation occurs in parallel across these units and includes top-down processing that enables activation at the word level to influence activation at the phoneme level. The top-down activation provides the same constraints on the recognition of phonemes as the interactive activation model provides on the recognition of letters.

Another connectionist model—a simple recurrent network—learns the semantic and syntactic properties of words by attempting to predict the next word in a sentence (Elman 2004). The network uses each new word in a sentence (the input) to predict the next word in the sentence (the output). Learning occurs by comparing the prediction with the actual occurrence. A key step in the recurrent network is that the hidden-unit weights depend on the context unit from the previous word, which also depended on the context unit of the previous word, so a history of previous information is preserved.

The similarity among the words used in the study resulted in easily interpretable clusters based on their hidden-unit activation patterns (Elman 2004). The initial split partitioned the words into verbs and nouns. The verbs are clustered into transitive verbs and intransitive verbs. Transitive verbs, such as like and chase, take a direct object, whereas intransitive verbs, such as think and sleep, do not. The nouns are clustered into animates and inanimates, which are

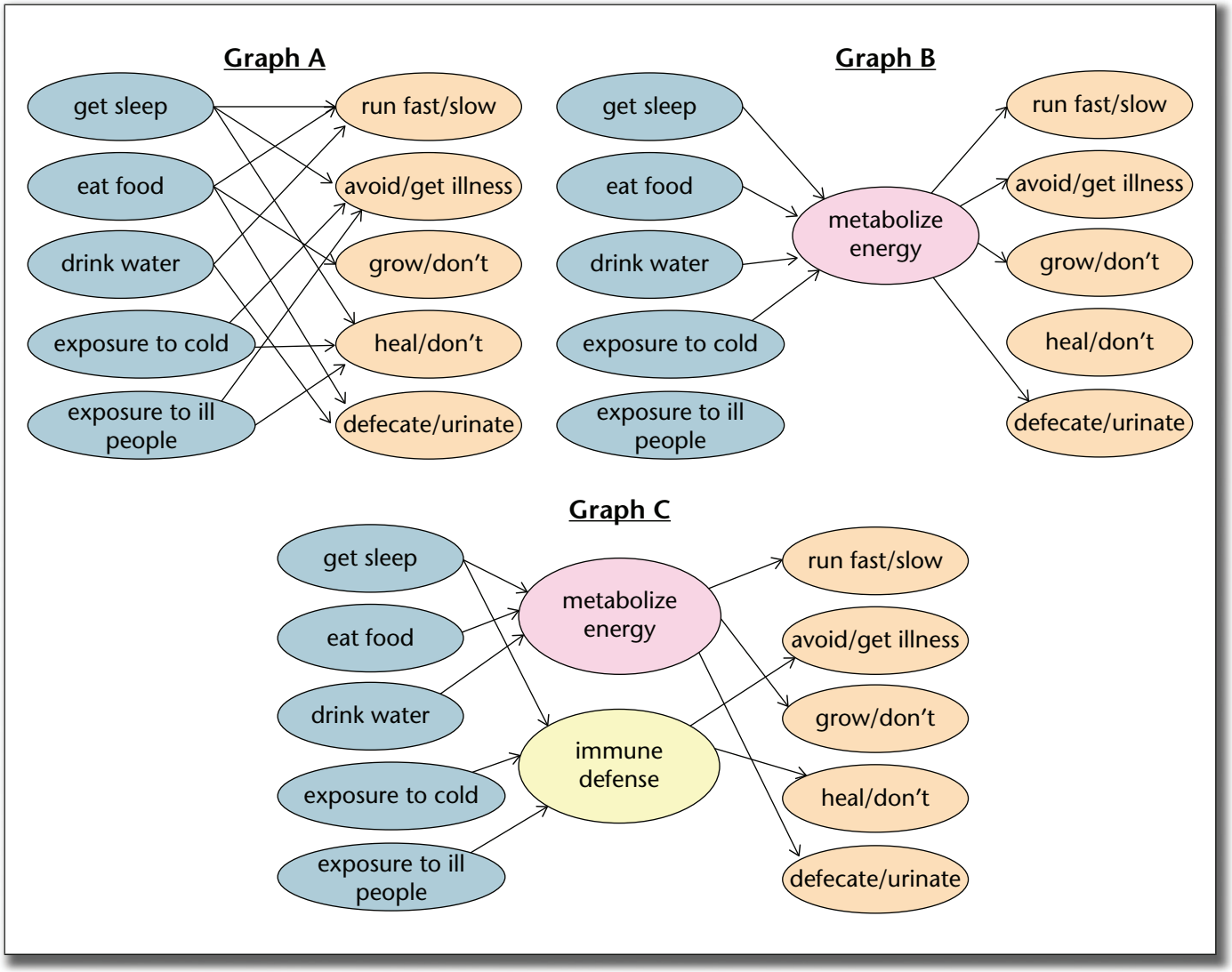

Figure 3. Hierarchical Bayesian Model of Causality.

From Gopnik and Wellman (2012). Reproduced with permission from the American Psychologic Association, ©2012.

further partitioned based on their semantic features into categories such as breakable objects and edible objects. It is important to note that the investigators assigned these category labels based on their interpretation of the hierarchical cluster analysis. The labels were not given to the learning network. Its activation patterns in the hidden layer depended on the position of the words in the sentences and their relation to other words in the sentences.

Symbolists

Despite these successes Forbus, Liang, and Rabkina (2017) argued that the connectionist approach has limitations as a model of human reasoning. One limitation is that this approach requires massive amounts of data to learn—far more than required by people. A second limitation is that all of the data must be available at the beginning, which does not capture the incremental nature of human learning that adds new information. A third limitation is that it is not always apparent what is being learned in distributed representations. These limitations indicate that symbolic representations should play a central role in efforts to explain human cognition, particularly those showing structural alignments (Forbus, Liang, and Rabkina 2017).

A method related to structural alignment is a transformational account that measures the similarity between two objects by the amount of effort required to transform one representation into another (Hahn 2014). The two methods are related because the transformational distance is influenced by the alignment of the components. For instance, only the addition of the letter *s* is required to transform the word lack into the word slack. However, aligning the first letters of the two words—*s* with *l*, *l* with *a*, and so on—creates differences between all the letters. This

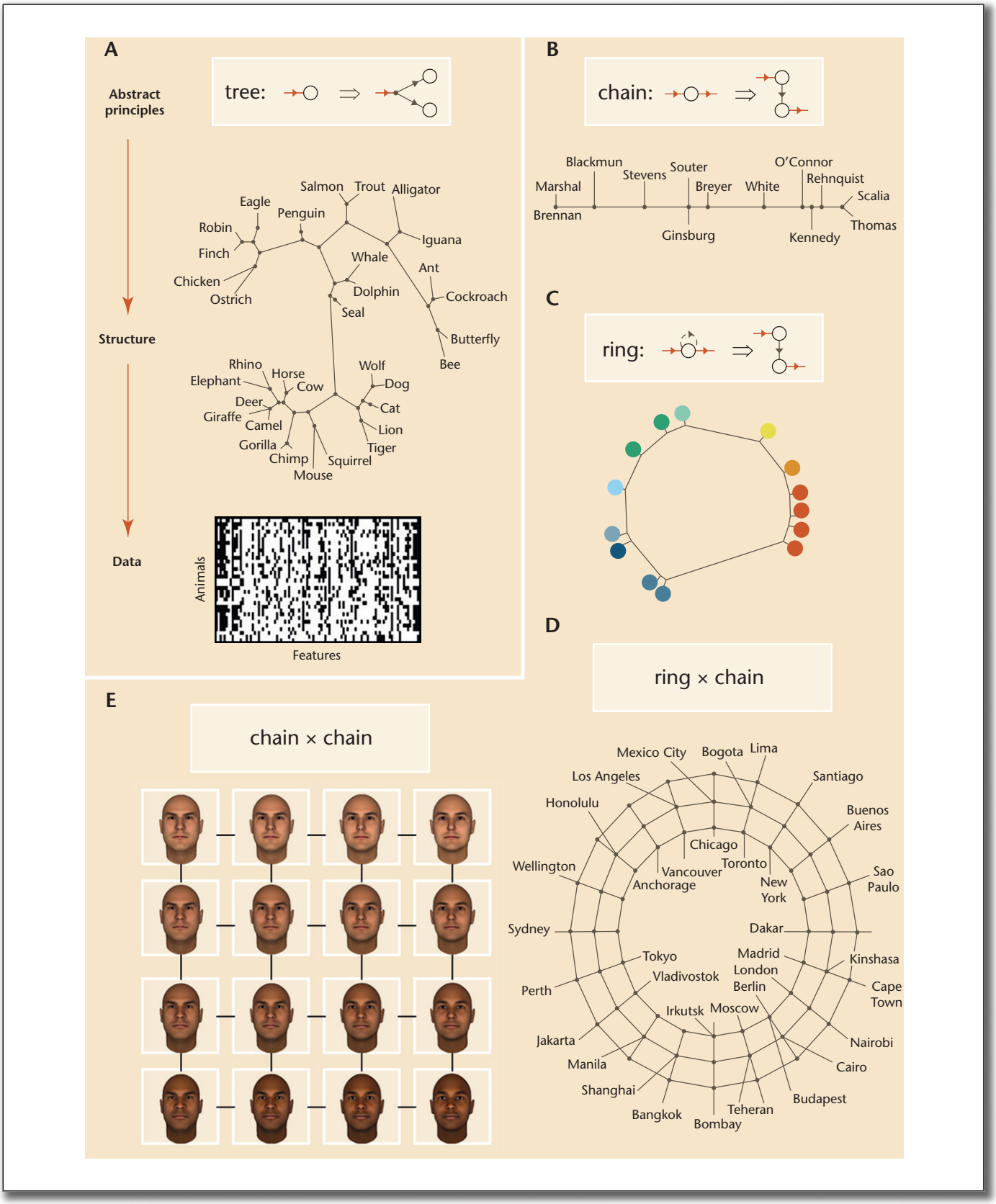

Figure 4. Use of Hierarchical Bayesian Models to Find Structure in Data.

From Tenenbaum et al. (2011). Reproduced with permission from the American Association for the Advancement of Science, ©2011.

comparison creates an unreasonably high measure of transformational effort, but it illustrates the interaction between transformational effort and alignment (Hahn 2014).

Hahn (2014) concluded her review of various similarity measures by asking whether differences between methods involve different types of representations or whether the methods can be unified within a common framework. Domingos' (2015) goal in writing *The Master Algorithm* was to encourage readers to reflect on how combining the analogizer, Bayesian, connectionist, and symbolic methods would create more powerful techniques of machine learning. The next section discusses how psychologists have combined these methods to model human performance.

Combining Methods

One of the most impressive integrations of the analogizer and Bayesian approaches evaluates different graph representations of similarity (Kemp and Tenenbaum 2008). Figure 4 shows how the abstract principles of graph structure—trees, chains, and rings—can capture the similarity of many different types of stimuli (Tenenbaum et al. 2011). The data are similarities based on biologic features (panel A), Supreme Court votes (B), judgments of pure color wavelengths (C), the actual distances between world cities (D), and Euclidean distances between faces (E).

As illustrated in figure 4, the best-fitting structures are a tree for animals (A), a chain for voting records (B), a ring for colors (C), a ring × chain combination for cities (D), and a chain × chain combination for faces (E). Bayesian models can also model developmental changes. For instance, when only five animal features were used in the model, the animal species clustered into five categories but did not form a tree structure (Kemp and Tenenbaum 2008). The initial categories consisted of birds, insects, two classes of mammals, and water animals such as a dolphin, a salmon, and an alligator. As the number of features increased from 5 to 20, there was a qualitative shift between these partitions and a tree structure. The tree became increasing complex with the further addition of features that corresponded more closely to the adult classifications shown in figure 4A.

As is the case for causal relations in figure 3, the model is a hierarchical Bayesian model because there are two components in determining the best fit (Holyoak 2008). One component is finding whether the best structure is a tree, a chain, a ring, a chain × chain, or a ring × chain. This component corresponds to P(hypothesis|data), where the hypothesis is one of the structures in figure 4, and the data are the similarities between pairs of items. The other component is finding the location of examples within the structure. In figure 4B, the task requires not only determining that a chain is sufficient for comparing the votes of Supreme Court justices but also determining the location of the justices on the chain. The voting patterns of Justices Marshall and Brennan were very similar to each other and very different from the voting patterns of Justices Scalia and Thomas.

An integration of the analogizer and symbolist perspectives occurs in the hybrid architectures discussed by Love (2015). The hybrid architectures contain both similarity measures and rules such as, "If it has feathers and wings, then it is a bird." A rule can be considered a special case of similarity when only a single or a small subset of an object's attributes are relevant (Pothos 2005). Hybrid architectures such as SUSTAIN (Love, Medin, and Gureckis 2004) incorporate selective attention mechanisms to enable similarity models to follow rules by focusing on a critical dimension. For instance, in classifying the make of cars, SUSTAIN learns that shape is a better predictor than color. It can account for human behavior in learning problems that require a simple rule and exceptions to those rules by creating a small set of clusters that encode items that follow rules and other clusters for those items that are exceptions.

COVIS is a neurologically inspired theory of category learning that proposes both a rule-based system that can quickly learn rule-based categories and a procedural learning system that more slowly learns a wide variety of other category structures (Ashby and Valentin 2017). One of their tasks differed along two dimensions, orientation and narrowness of bars. A rule-based task involves categorizing on a single dimension, such as narrow bars belonging in one category and wide bars in the other category. An information integration task involves using both dimensions. COVIS proposes that the two tasks require different memory and neuroscience structures. For instance, rule-based tasks depend more on working memory and other declarative memory systems, whereas prototype distortion tasks depend more on perceptual representations in the visual cortex (Ashby and Valentin 2017).

A Bayesian and symbolist integration is a guiding principle of the architecture ACT-R. As stated in the preface to its formulation, "One should begin with a rational analysis to figure out what the system should be doing and then worry about how the prescriptions of a rational analysis are achieved in the mechanism of the architecture" (Anderson 1990, xi). Although the Bayesian formulation drives the rational analysis, symbolic production systems drive the implementation (Anderson 1983). The production rules have the form "If [condition] Then [action]" that specifies the necessary conditions for executing an action. Choice among competing production rules is based on their utilities—estimates of the rule's probability of success and cost in leading to the goal.

Connectionism has also been important in the development of the ACT models, which have a long history of connecting a subsymbolic activation-based memory to a symbolic system of production rules

(Anderson and Lebiere 2003). The subsymbolic level tunes the rules to the statistical structure of the environment. This development resulted in the ACT-R version of the models in which the R refers to rational analysis (Anderson 1990).

The relationship between the connectionists and the symbolists is evident in another hybrid architecture that combines a lower-level connectionist network with higher-level symbolic processing. CLARION is an integrative cognitive architecture that consists of a top-level explicit representation and a bottom-level implicit representation (Sun and Zhang 2006). Explicit knowledge is represented by easily interpretable symbols that have clear conceptual meaning. Implicit knowledge is represented by a subsymbolic distributed representation within a back-propagation network. In contrast to an explicit memory that encodes rules as all or none, implicit memory supports a graded accumulation of knowledge.

Heile and Sun (2010) developed an explicit-implicit interaction theory based on CLARION to analyze the four stages of problem solving proposed in Wallas's (1926) influential book *The Art of Thought.* Preparation is the initial search for a solution, incubation is a period of inactivity following an impasse, illumination (or insight) is a sudden discovery of a possible solution, and verification is a determination of whether the discovered solution is valid. The implementation of CLARION assumes that the initial preparation phase is predominately rule based as people respond to verbal instructions, form a representation of the problem, and establish goals. In contrast, the second (incubation) stage is predominately implicit processing in which people may not consciously think about the problem. The third stage, insight, occurs when the activation level crosses a threshold that makes the output available for verbal report. The final verification stage, like the initial stage, requires explicit processing to evaluate the potential of the discovered solution.

These hybrid architectures demonstrate how analogizer, Bayesian, connectionist, and symbolist methods can work together. However, not all combinations may work. A Bayesian and connectionist integration presents the biggest challenge for combining pairs of methods because the two approaches have often been viewed as competitors (Griffiths et al. 2010, McClelland et al. 2010). As stated by Rogers and McClelland (2014, 1061), "While some may see probabilistic models as replacing or subsuming PDP [parallel distributed processing] models, another perspective is that probabilistic framework and PDP approach have overlapping but diverging aspirations. We expect both approaches to continue to evolve and to challenge each other, and possibly to benefit from a degree of competition between them."

But competition may lead to cooperation for the benefit of both. In their article *Building Machines That Learn and Think Like People,* Lake et al. (2017) argue for an integration of the deep learning connectionist methods with the building blocks (attention, working memory, stacks, queues) of traditional cognitive and computer science. They view the former approach as exceling in statistical pattern recognition and the latter approach as exceling in building models to understand the world. These models—whether learned, built in, or enriched—are core ingredients of human intelligence.

Building Machines That Learn and Think Like People is a must read for anyone interested in computational intelligence. The article provides a comprehensive survey of current thought and is supplemented by many expert commentaries. One of these from the connectionist camp argues for the flexibility of connectionist learning by avoiding an initial commitment to domain-specific knowledge structures (Hansen et al. 2017). The authors propose that joining forces is the best approach for understanding how human cognitive abilities arise in richly structured learning environments.

Benefits of Building Bridges

There are likely many benefits in building bridges between AI and cognitive psychology, but I want to emphasize three. The first is that computational programs in AI can serve as potential theoretical models in cognitive psychology. I mentioned at the beginning of this article how an engineering course on mathematical methods of pattern recognition served as the basis for my modeling human categorization (Reed 1972). An early collaborative effort between a cognitive psychologist (Alan Collins) and a computer scientist (Ross Quillian) resulted in the hierarchical network model for representing semantic organization in human memory (Collins and Quillian 1969). But it was the efforts of Newell and Simon to apply computational methods in computer science (Newell, Shaw, and Simon 1958) to human problem solving (Newell and Simon 1972) that introduced many new ideas into cognitive psychology.

A second benefit is that AI and cognitive psychology share common methods. One example is the work summarized here on the methods developed by the analogizers, Bayesians, connectionists, and symbolists. They also share common challenges such as the reasoning from imperfect knowledge. My collaboration with computer scientist Adam Pease considered how both people and machines must process ambiguous, conditional, contradictory, fragmented, inert, misclassified, and uncertain information (Reed and Pease 2017).

A third benefit is that the growing impact of AI on our lives requires understanding how computers and people can best work together. An impressive current collaborative effort based on IBM's WatsonPaths expands on the Watson question answering system (Ferrucci et al. 2010) that became famous on the television show *Jeopardy*. A new project with the

Cleveland Clinic Lerner College of Medicine of Case Western Reserve University presents a patient summary and asks for the most likely diagnosis or most appropriate treatment (Lally et al. 2017).

Kitano (2016) discussed how AI has been driven by the success of previous grand challenges, such as IBM's chess program Deep Blue defeating Kasparov, IBM's Watson winning on *Jeopardy*, and humanoid robots eventually beating humans in RoboCup. Although these victories surpassed human efforts, Kitano (2016) recommended a new collaborative grand challenge to develop an AI system that can assist in a scientific discovery that is worthy of a Nobel Prize in the biomedical sciences.

In their article on a standard model of the mind, Laird, Lebiere, and Rosenbloom (2017) proposed that a fundamental hypothesis in AI is that minds are cognitive systems that can be implemented by either natural brains or general-purpose computers. Their long-term objective is to develop a standard model of a human-like mind that can serve as a common computational framework across artificial intelligence, cognitive science, neuroscience, and robotics. The development of such a model would establish many bridges.

Acknowledgements

I dedicate this article to the memory of Jeff Elman (1948–2018), who made many contributions in advancing cognitive science as a researcher and an administrator. Jeff had read and approved my summary of his work in this article. I also thank anonymous reviewers for their very helpful assistance.

References

Anderson, J. R. 1983. *The Architecture of Cognition*. Cambridge, MA: Harvard University Press.

Anderson, J. R. 1990. *The Adaptive Character of Thought*. Hillsdale, NJ: Erlbaum.

Anderson, J. R. 1991. The Adaptive Nature of Human Categorization. *Psychological Review* 98(3): 409–29. doi.org/10.1037/0033-295X.98.3.409

Anderson, J. R., and Lebiere, C. 2003. The Newell Test for a Theory of Cognition. *Behavioral and Brain Sciences* 26(5): 587–640. doi.org/10.1017/S0140525X0300013X

Ashby, F. G., and Valentin, V. 2017. Multiple Systems of Perceptual Category Learning: Theory and Cognitive Tests. In *Handbook of Categorization in Cognitive Science*, edited by H. Cohen and C. Lefebvre, 157–88. Amsterdam: Elsevier. doi.org/10.1016/B978-0-08-101107-2.00007-5

Biehl, M.; Hammer, B.; and Villmann, T. 2016. Prototype-Based Models in Machine Learning. *Wiley Interdisciplinary Reviews: Cognitive Science* 7(2): 92–111.

Chandrasekaran, B.; Banerjee, B.; Kurup, U.; and Lele, O. 2011. Augmenting Cognitive Architectures to Support Diagrammatic Imagination. *Topics in Cognitive Science* 3(4): 760–77. doi.org/10.1111/j.1756-8765.2011.01156.x

Collins, A. M., and Quillian, M. R. 1969. Retrieval Time from Semantic Memory. *Journal of Verbal Learning and Verbal Behavior* 8(2): 240–7. doi.org/10.1016/S0022-5371(69)80069-1

Domingos, P. 2015. *The Master Algorithm: How the Quest for the Ultimate Learning Machine Will Remake Our World*. New York: Basic Books.

Elman, J. L. 2004. An Alternative View of the Mental Lexicon. *Trends in Cognitive Sciences* 8(7): 301–6. doi.org/10.1016/j.tics.2004.05.003

Ferrucci, D. A.; Brown, E.; Chu-Carroll, J.; Fan, J.; Gondek, D.; Kalyanpur, A.; Lally, A.; Murdock, J. W.; Nyberg, E.; Prager, J.; Schlaefer, N.; and Welty, C. 2010. Building Watson: An Overview of the DeepQA Project. *AI Magazine* 31(3): 59–79. doi.org/10.1609/aimag.v31i3.2303

Forbus, K. D.; Liang, C.; and Rabkina, I. 2017. Representation and Computation in Cognitive Models. *Topics in Cognitive Science* 9(3): 694–718. doi.org/10.1111/tops.12277

Goldstone, R. L., and Son, J. Y. 2005. Similarity. In *The Cambridge Handbook of Thinking and Reasoning*, edited by K. J. Holyoak and G. Morrison, 155–76. New York: Cambridge University Press.

Gopnik, A., and Bonawitz, E. 2014. Bayesian Models of Child Development. *WIREs Cognitive Science* 6(2): 75–86.

Gopnik, A., and Wellman, H. M. 2012. Reconstructing Constructivism: Causal Models, Bayesian Learning Mechanisms, and the Theory Theory. *Psychological Bulletin* 138(6): 1085–108.

Griffiths, T. L.; Chater, N.; and Kemp, C. 2010. Probabilistic Models of Cognition: Exploring Representations and Inductive Biases. *Trends in Cognitive Sciences* 14(8): 357–64. doi.org/10.1016/j.tics.2010.05.004

Hahn, U. 2014. Similarity. *Wiley Interdisciplinary Reviews: Cognitive Science* 5(3): 271–80.

Hansen, S. S.; Lampinen, K.; Suri, G.; and McClelland, J. L. 2017. Building on Prior Knowledge Without Building It In. *Behavioral and Brain Sciences* 40: e268. doi.org/10.1017/S0140525X17000176

Hélie, S., and Sun, R. 2010. Incubation, Insight, and Creative Problem Solving: A Unified Theory and a Connectionist Model. *Psychological Review* 117(3): 994–1024. doi.org/10.1037/a0019532

Holyoak, K. J. 2008. Induction as Model Selection. *Proceedings of the National Academy of Sciences of the United States of America* 105(31): 10637–8. doi.org/10.1073/pnas.0805910105

Kemp, C., and Tenenbaum, J. B. 2008. The Discovery of Structural Form. *Proceedings of the National Academy of Sciences of the United States of America* 105(31): 10687–92. doi.org/10.1073/pnas.0802631105

Kitano, H. 2016. Artificial Intelligence to Win the Nobel Prize and Beyond: Creating the Engine for Scientific Discovery. *AI Magazine* 37(1): 39–49. doi.org/10.1609/aimag.v37i1.2642

Laird, J. E. 2012. *The Soar Cognitive Architecture*. Cambridge: MIT Press. doi.org/10.7551/mitpress/7688.001.0001

Laird, J. E.; Lebiere, C.; and Rosenbloom, P. S. 2017. A Standard Model of the Mind: Toward a Common Computational Framework Across Artificial Intelligence, Cognitive Science, Neuroscience, and Robotics. *AI Magazine* 38(4): 13–26. doi.org/10.1609/aimag.v38i4.2744

Laird, J. E.; Newell, A.; and Rosenbloom, P. S. 1987. SOAR: An Architecture for General Intelligence. *Artificial Intelligence* 33(1): 1–64. doi.org/10.1016/0004-3702(87)90050-6

Lake, B. M.; Salakhutdinov, R.; and Tenenbaum, J. B. 2015. Human-Level Concept Learning Through Probabilistic Program Induction. *Science* 350(6266): 1332–8. doi.org/10.1126/science.aab3050

Lake, B. M.; Ullman, T. D.; Tenenbaum, J. B.; and Gershman, S. J. 2017. Building Machines That Learn and Think Like People. *Behavioral and Brain Sciences* 40: e253. doi.org/10.1017/S0140525X16001837

Lally, A.; Bagchi, S.; Barborak, M. S.; Buchanan, D. W.; Chu-Carroll, J.; Ferrucci, D. A.; Glass, M. R.; and Prager, J. M. 2017. WatsonPaths: Scenario-Based Question Answering and Inference Over Unstructured Information. *AI Magazine* 38(2): 59–76. doi.org/10.1609/aimag.v38i2.2715

Lathrop, S. D.; Wintermute, S.; and Laird, J. E. 2011. Exploring the Functional Advantages of Spatial and Visual Cognition from an Architectural Perspective. *Topics in Cognitive Science* 3(4): 796–818. doi.org/10.1111/j.1756-8765.2010.01130.x

Love, B. C. 2015. Concepts, Meaning, and Conceptual Realtionships. In *Oxford Handbook of Cognitive Science*, edited by S. Chipman. New York: Oxford University Press.

Love, B. C.; Medin, D. L.; and Gureckis, T. M. 2004. SUSTAIN: A Network Model of Category Learning. *Psychological Review* 111(2): 309–32. doi.org/10.1037/0033-295X.111.2.309

Marr, D. C. 1982. *Vision: A Computational Investigation into the Human Representational System and Processing of Visual Information*. San Francisco: W. H. Freeman.

McClelland, J. L.; Botvinick, M. M.; Noelle, D. C.; Plaut, D. C.; Rogers, T. T.; Seidenberg, M. S.; and Smith, L. B. 2010. Letting Structure Emerge: Connectionist and Dynamical Systems Approaches to Cognition. *Trends in Cognitive Sciences* 14(8): 348–56. doi.org/10.1016/j.tics.2010.06.002

McClelland, J. L., and Elman, J. L. 1986. The TRACE Model of Speech Perception. *Cognitive Psychology* 18(1): 1–86. doi.org/10.1016/0010-0285(86)90015-0

McClelland, J. L., and Rumelhart, D. E. 1981. An Interactive-Activation Model of Context Effects in Letter Perception: 1. An Account of Basic Findings. *Psychological Review* 88(5): 375–407. doi.org/10.1037/0033-295X.88.5.375

Medin, D. L., and Schaffer, M. M. 1978. Context Theory of Classification Learning. *Psychological Review* 85(3): 207–38. doi.org/10.1037/0033-295X.85.3.207

Newell, A., and Simon, H. A. 1972. *Human Problem Solving*. Englewood Cliffs, NJ: Prentice-Hall.

Newell, A.; Shaw, J. C.; and Simon, H. A. 1958. Elements of a Theory of Human Problem Solving. *Psychological Review* 65(3): 151–66.

Nosofsky, R. M. 1986. Attention, Similarity, and the Identification-Categorization Relationship. *Journal of Experimental Psychology* 115(1): 39–57. doi.org/10.1037/0096-3445.115.1.39

Pothos, E. M. 2005. The Rules Versus Similarity Distinction. *Behavioral and Brain Sciences* 28(1): 1–14. doi.org/10.1017/S0140525X05000014

Reed, S. K. 1972. Pattern Recognition and Categorization. *Cognitive Psychology* 3(3): 382–407. doi.org/10.1016/0010-0285(72)90014-X

Reed, S. K. 2016a. Problem Solving. In *Oxford Handbook of Cognitive Science*, edited by S. Chipman, 231–48. New York: Oxford University Press.

Reed, S. K. 2016b. The Structure of Ill-Structured (and Well-Structured) Problems Revisited. *Educational Psychology Review* 28(4): 691–716. doi.org/10.1007/s10648-015-9343-1

Reed, S. K. 2019. Modeling Visuospatial Reasoning. *Spatial Cognition & Computation* 19: 1–45.

Reed, S. K.; Ernst, G. W.; and Banerji, R. 1974. The Role of Analogy in Transfer Between Similar Problem States. *Cognitive Psychology* 6(3): 436–50. doi.org/10.1016/0010-0285(74)90020-6

Reed, S. K., and Friedman, M. P. 1973. Perceptual Versus Conceptual Categorization. *Memory & Cognition* 1(2): 157–63. doi.org/10.3758/BF03198087

Reed, S. K., and Pease, A. 2017. Reasoning from Imperfect Knowledge. *Cognitive Systems Research* 41: 56–72. doi.org/10.1016/j.cogsys.2016.09.006

Rogers, T. T., and McClelland, J. L. 2014. Parallel Distributed Processing at 25: Further Explorations in the Microstructure of Cognition. *Cognitive Science* 38(6): 1024–77. doi.org/10.1111/cogs.12148

Rumelhart, D. E.; Hinton, G. E.; and McClelland, J. L. 1986. A General Framework for Parallel Distrubuted Processing. In *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*, edited by D. E. Rumelhart and J. L. McClelland. Cambridge, MA: Bradford Books.

Sebestyen, G. S. 1962. *Decision-Making Processes in Pattern Recognition*. New York: Macmillan.

Simon, H. A. 1973. The Structure of Ill-Structured Problems. *Artificial Intelligence* 4(3-4): 181–201. doi.org/10.1016/0004-3702(73)90011-8

Simon, H. A., and Reed, S. K. 1976. Modeling Strategy Shifts in a Problem-Solving Task. *Cognitive Psychology* 8(1): 86–97. doi.org/10.1016/0010-0285(76)90005-0

Smith, R. G., and Eckroth, J. 2017. Building AI Applications: Yesterday, Today, and Tomorrow. *AI Magazine* 38(1): 6–22. doi.org/10.1609/aimag.v38i1.2709

Sun, R., and Zhang, X. 2006. Accounting for a Variety of Reasoning Data within a Cognitive Architecture. *Journal of Experimental & Theoretical Artificial Intelligence* 18(2): 169–91. doi.org/10.1080/09528130600557713

Tenenbaum, J. B.; Kemp, C.; Griffiths, T. L.; and Goodman, N. D. 2011. How to Grow a Mind: Statistics, Structure, and Abstraction. *Science* 331(6022): 1279–1285. doi.org/10.1126/science.1192788

Wallas, G. 1926. *The Art of Thought*. New York: Harcourt, Brace.

Stephen K. Reed is an emeritus professor of psychology at San Diego State University and a visiting scholar in the Department of Psychology at the University of California, San Diego. His academic interests focus on organizing knowledge in the cognitive sciences. He is currently writing a book on organizing knowledge to describe recent advances in AI and cognitive psychology to a general audience.