A Standard Model of Mind:
AAAI Technical Report FS-17-05

Leveling Up: Strategies to Achieve Integrated Cognitive Architectures

Paul E. Silvey
The MITRE Corporation
psilvey@mitre.org

Abstract

Human-level cognition (most uniquely characterized by our abilities to use language) should be seen as a superset of functional and behavioral capabilities shared by lower life-forms including animals and insects, and this perspective ought to principally guide our strategies for developing integrated cognitive architectures. Just as the study of biological model organisms has led to tremendous advances in our scientific knowledge of genetics and cellular function, the study of embodied cognition in simple agent-environment simulations can yield similar advances in Cognitive Science, Artificial Intelligence, and Robotics. By working first on the foundations of intelligent interaction with one's environment, and by focusing on core functions such as predictive and inductive learning, probabilistic goal-directed behavior compilation, and empathetic reasoning, we can better establish the grounding that the physical symbol system hypothesis assumes (Newell and Simon 1976), yet often without explicit demonstration of a mechanism to derive symbolic relations and semantics from raw sensory data. Logic and language are seen to emerge from our willingness to make discrete simplifying assumptions in a continuous and probabilistic world of experience, and developing a Standard Model of the Mind can help build much-needed bridges between historically non-aligned research communities.

Introduction

A recent effort to map and unify several existing cognitive architectures into a Standard Model of the Mind offers a welcome approach for finding commonality across many disparate but closely related research disciplines, including Artificial Intelligence (AI), Cognitive Science, Neuroscience, and Robotics (Laird, Lebiere, and Rosenbloom in press). However, the candidate models that most directly influenced this work, at least initially, emphasized cognitive characteristics of mental processes (i.e. minds) over more tightly integrated mind-body or agent-environment holistic views. While the proposed standard model does include perceptual and motor components, it clearly encompasses an ambitious human-oriented perspective, in the sense that it seeks to represent, reason, and generally account for conceptual knowledge that can be naturally expressed via linguistic phenomena. This form of knowledge is often described as symbolic relational, and when used in strictly disembodied simulations of intelligence, has given rise to a criticism known as the Symbol Grounding Problem (Harnad 1990). Symbols in the form of natural language words evoke rich semantic interpretations in humans, and many such words are not intended to map directly to tangible, physical objects. Even those that do have object mappings however, challenge the cognitive architect to explain how low-level perceptual and sensory data might give rise to them. Often, it is assumed that sensory data can be progressively abstracted through layers that represent their knowledge sub-symbolically (Steels 2008). The current proposal for a Standard Model of the Mind adopts this view, but doesn't exactly resolve how sub-symbolic knowledge and symbolic relational metadata (which includes frequencies of occurrence, attentional weightings, etc.) might be similar or different. In short, the current proposal leaves room for further research to develop and mature the lowest levels of the cognitive architecture, those closest to animal-behavioral interactions with the world.

Layers of Embodied Cognitive Learning

If perceptual and motor-control processing are seen as being low-level, and knowledge and language as high-level, we can naturally choose to approach human cognition in a top-down, or alternatively, a bottom-up fashion. Most traditional AI research, especially that which emphasizes symbolic knowledge representation, is essentially top-down. Other research however, including connectionist, neural network, and robotics work, prefers a more bottom-up approach (Brooks 1986). Intelligence acquired and demonstrated through the use of language and logical reasoning, vs. simply knowing how to act in real world situations (sometimes referred to as the difference between *book smarts* and

street smarts) has been recognized and discussed within AI (Levesque 2017). Less well observed however, are those clearly cognitive capabilities that, while sub-linguistic in nature, involve self-conscious abstraction and mapping. In the spirit of this, Figure 1 offers a layered model of human learning that emphasizes successively more sophisticated capabilities of an embodied agent, depicted here as a subdivided triangle.

This form was chosen to emphasize the foundational aspect that the lower levels are believed to provide. Views such as this are also adopted by researchers in the nascent field of Sociocognitive Science, where interactivity and grounding are seen as key elements of social context and learning, and embodied cognition helps define the central notion of *agency* (Neumann and Cowley 2013).

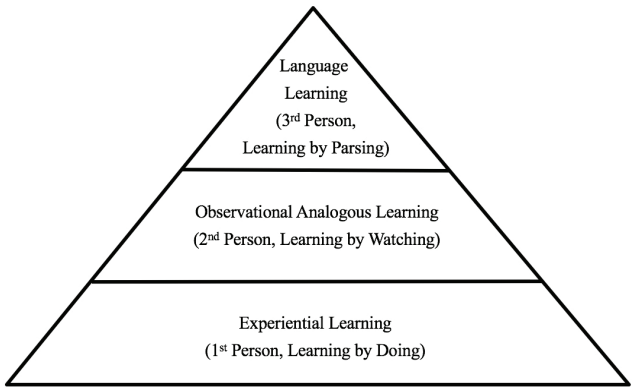

Figure 1. Levels of Learning for Embodied Cognition.

The two upper layers, taken together, reflect social awareness and communication skills, as they are used to teach others and to learn from them. Similarly, the two lower layers reflect more primitive cognitive skills, motivated by the observation that animals and even insects, lacking the language and verbal skills which clearly distinguish humans as higher life-forms, are nevertheless able to learn using both methods. This has been surprisingly well demonstrated in bees, who can learn by mapping a second person perspective onto themselves using what we might describe as empathetic or analogous reasoning (Mirwan and Kevan 2013).

The lowest layer of experiential learning is possible in a world devoid of other agents, where patterned regularities such as determinism and laws of physics can be discovered via exploration and trial-and-error, but also through careful interactive experimentation and hypothesis testing. Biological minds rarely if ever exist in such simplistic worlds, but our artificial agents can be studied under these conditions, and robots sent to explore space and desolate planetary worlds can experience them. Achieving an eventual detailed understanding of how this foundation enables and facilitates the upper layers might serve to unlock many remaining mysteries of the mind.

This layered decomposition was motivated by an analysis of an agent's information sources for learning, and by extending embodiment into social and contemplative realms. However, it may also serve to frame the discussion regarding heterogeneous and pluralistic representations for experience and knowledge. The lowest layer, needing the closest binding to raw sensory data, is most suitable for using what have been called sub-symbolic representations. These could include the learned weights of a neural network, but also might be captured by feature trajectories or other temporal sequence memory structures (as described more later). At the highest layer, we have the large and historic body of work that uses symbolic relational models, including declarative languages based on assertions and ontologies, or graphical semantic networks. Somewhere between these rather separated schools of thought might lie the knowledge representation approaches that use diagrammatic or analogical models, case-based reasoning, or ones that attempt to map qualitative features into multi-dimensional conceptual spaces (Lieto, Chella, and Frixione 2017).

Time is of the Essence

In symbolic AI models, lone symbols intended to represent objects are abstracted away from their specific and experiential temporal contexts (often by a purely manual engineering process), allowing their conceptual forms to be powerfully and more easily inserted or removed from imagined scenarios. The psychologist's distinction between episodic and semantic memory attempts to delineate at least some of this temporal factoring, and we can use this to argue that episodic memory is an important component in the experiential first-person learning of Figure 1 (Gershman and Daw 2017), whereas semantic memory belongs somewhere higher up. In addition, many important semantic relationships are in fact temporal in nature, and these kinds of generalizations are used in symbolic AI to help construct and leverage what is known as Procedural Memory.

While some AI researchers may believe that semantic level symbolic knowledge can be directly crafted into dis-embodied agents to bestow them with human-level intelligence, or that symbol grounding is unnecessary even for achieving Artificial General Intelligence (AGI), it appears shortsighted to de-emphasize the role lower-level experience in the world has on our ability to learn, think, and intelligently act. Central to this viewpoint is the recognition that time plays a crucial role. Sensory data comes to an agent continuously, or taking a discrete modeling perspective, as a series of data frames. Orderings matter, and patterns in time are

more fundamental than patterns in space (which require motion and perspective to establish). Causality, even if near-instantaneous, depends on temporal sequence. Furthermore, causality often is evident only with respect to some temporal lag, which will not be discernable unless one's cognitive architecture explicitly provides a means to look for it in its variable form. Generally speaking, in any environment where processes exist independent of the agent (obviously including those occupied by other agents), time itself can be said to introduce pressure to think and decide quickly, in the form of deadlines or what engineers might call real-time requirements (Cohen et. al. 1989).

To its credit, the Standard Model is built around the notion of a cognitive cycle, where a deliberative act is chosen approximately every 50ms, as informed by studies of human behavior. Time is clearly evident in the model in the form of this cyclic process, but also inherently in both the Episodic and Procedural Memory components. Similarly, learning to become reactive after practiced deliberation (e.g. the notion of Chunking in Soar), captures another importance of time in a cognitive architecture. Whether innate or learned, thinking fast is a hallmark of intelligence, and is complementary to more reasoned contemplation and goal-driven planning (Kahneman 2011). This is illustrated in Figure 2.

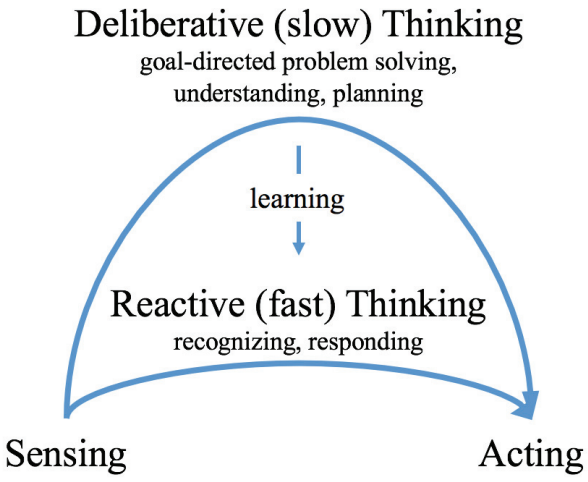

Figure 2. Temporal Cognition.

While present in various forms throughout the Standard Model, temporal aspects of cognition are somewhat hidden, to accommodate representational and functional memory decompositions. In contrast, consider the view of intelligent processing that sees the core function of the brain's neocortex as continuously executing a common algorithm to perform temporal sequence prediction and learning (Hawkins and Blakeslee 2007), or Reinforcement Learning models which use recursive expansions of time in their formalisms.

Symbols and Statistical Patterns

Symbolic AI has long recognized that the symbols being manipulated by computer programs are equivalent to arbitrary bit strings, and that it is the relationships between symbols as well as how they are processed which are necessary to give any meanings to them. Black-box models of intelligence, including the Turing Test and Searle's Chinese Room, attribute understanding to an agent or system's ability to "do the right thing" in response to particular stimuli. This is true whether the stimulus is linguistic or sensory-motor in nature. So, we might say that all semantics are behavioral (and in the eye of the beholder), and that understanding unfolds and manifests itself over time through experience. This perspective is important, because it helps keep us from expecting all natural language expressions to have a true and correct meaning independent of the background of the listener. While behavioral semantics may be a somewhat unifying concept to work from, the level of abstraction of the stimulus still serves to separate researchers.

Consider how those working from a linguistic perspective of cognition more naturally seem to adopt symbolic representations, specifically ones that align with logic and formal systems of reasoning, including rules of grammatical syntax and discrete mechanisms for thought, planning, and problem-solving. Approaching semantic grounding from a declarative or knowledge-based perspective has produced some linguistic theories that seek to make connections to sensory-motor phenomena. However, these approaches often remain abstracted above raw sensory data, as they seek to maintain various temporal relational factorings (e.g. Lian et. al. 2017).

On the other hand, those working close to sensory-motor interfaces with the world are prone to think in terms of quantitative values and statistical patterns, vs. symbols. Combining high-bandwidth sensing with complex multi-agent environments produces an enormous pattern space of temporal sequence data to which relatively few symbols must somehow be grounded. While the symbols may be relatively short bit strings, the experiential patterns from which they derive their useful meanings may be arbitrarily complex in structure and length. A sensory data-driven view of cognition sees statistical and probabilistic representations, including the real-valued numeric weights of a neural network, as a way to abstract and summarize in the face of this combinatorically vast underlying variability. Unfortunately, these bottom up approaches have their own difficulty connecting naturally to linguistic and symbolic reasoning.

Interestingly, a middle-out strategy has proven hard to develop, yet there could be a lot to learn from features of in-

telligence that are sufficiently abstract yet still pre-linguistic. Associating names with patterns of experience (or other forms of indexing sensory data memories) appears to be an important aspect of symbol grounding, but language and thought clearly support more than this. The 2nd person perspective of the middle layer of Figure 1 is useful for highlighting a significant feature of minds, namely what is known as *metacognition,* including various kinds of *self-consciousness*. While humans are easily capable of reasoning about the state of knowledge of others, sometimes recursively to several levels and combined with hypothetical and counterfactual reasoning, basic forms of metacognition have been demonstrated in animals as well. That animals can know when they remember, or have a sense of social fairness, is quite remarkable (Foote and Crystal 2007) (Brosnan and De Waal 2003). It seems reasonable that many human cognitive linguistic abilities, including making assertions about assertions (reification), might be built in some way on these sub-linguistic foundations.

Finding a Goldilocks approach to all this does hold promise. In retrospect, it may be recognized as simply an historical artifact that truth-theoretic formalisms such as deductive inference played such an important part in early AI research, owing to the significant influence of mathematics and philosophical schools of thought (such as Logical Positivism). Over time, as it has become more obvious that human inference is only plausible at best and that even the best-intentioned rationality is inherently resource-bounded, methods of inexact reasoning have gained prominence (Pearl 2014). Furthermore, by studying that which is theoretically learnable to only a probabilistically and approximately correct degree, we are reshaping our very concept of the meaning of the word knowledge (Valiant 2013). But to see semantics emerge from behavior, and behavior emerge from experience-based learning, we find it useful to try to first understand the most primitive of our cognitive abilities.

In addition to robotics research, there are increasing efforts to extend and generalize the Reinforcement Learning paradigm as a form of bottom-up cognitive modeling, which can be studied using multi-agent and agent-environment simulations. Cast in the form of games, these investigative frameworks are able to add increasingly complex real-world challenges for learning and intelligent behavior, including partial observability, stochastics, time pressure, and non-ergodic process complexities (Aslanides, Leike, and Hutter 2017). One aspect that is particularly important in this work is seeking to understand how agents might deal with trade-offs in multi-goal situations, which is ubiquitously present in its simplest form as the well-recognized dilemma of exploration vs. exploitation. Motivations that shift with context, and goals that are often at odds with each other, make this particularly challenging, yet inspiration from brain research, combining parallel evaluation with sparse coding ideas, may assist in demonstrating simple architectural principles that can account for intelligent dynamic behavior.

This basic research agenda offers many opportunities to study and understand embodied cognitive processes without the complexities of building and testing real-world robots, and with the considerable advantage of being able to start with benign environments and simple agent capabilities, but to then progressively introduce realism over time. This is reminiscent of the affordance of using Model Organisms in Biology and Genetics research, where simple forms of life are extensively studied in laboratory settings, and from which tremendous advances in our knowledge of cellular function and the fundamental genetic aspects of all life on earth have emerged. Or it might better be compared to behavioral psychology studies done *in-silico*, where our creatures can be endowed with natural and even unnatural abilities, and then tested under highly controlled environmental conditions.

Combining low-level sensory data and temporal sequence learning, we have been studying models of memory constructed as variable-depth probabilistic suffix trees, where nodes in the tree can be seen as categorical state space values that emerge from patterns over select or focus bits in each data frame. Each node-to-root path represents a hypothetical Markov state, allowing the agent to eventually determine whether it is safe to assume the past and future are conditionally independent given that state (or whether the notion of *the present* might be better defined in that context using a longer path or higher Markov order). As such, each node maintains a frequency-of-occurrence distribution of experienced successor states, and these can be initialized and updated in ways that support Bayesian reasoning as well as biases designed to adapt to statistical non-stationarities through selective forgetting. The occurrence of particular patterns in such trees can be thought of as opaque yet symbolic, and the whole tree as a Variable Order Markov Model (Ron, Singer, and Tishby 1994) (Volf and Willems 1995) (Bühlmann and Wyner 1999). It can be updated continuously and in parallel with its primary use for prediction as a decision process model, which enables both reactive recognition as well as deliberative time-sensitive planning, using for example Monte Carlo planning (Chung, Buro, and Schaeffer 2005) as a kind of Anytime Algorithm (Dean and Boddy 1988). This architecture has been chosen to most naturally accommodate the key temporal aspects of embodied cognition, and to leverage the success of weak-methods for data-driven reinforcement learning in the era of low-cost storage and massive computing parallelism.

Branching tree representations of time, whether into the past for memories (suffix trees) or into the future for expectations and plans, are natural and efficient ways of grounding sequential dependencies. They can be customized via choices in determining the nodal states, for example by using attention-focused raw data or derived features. They can adopt a fixed frame of reference from the perspective of a sensor, or they can be built to take a target perspective if coupled with sensory-motor object tracking (think of a cartoon of a running character who stays fixed in the middle of the frame as the background moves behind them). Paths through these trees represent behavioral trajectories, which become aggregate mental objects that can themselves be recognized and frequency counted, classified and clustered, and used to form higher-level patterns.

The data-intensive nature of these architectures leaves us with the formidable problem of induction and generalization. Currently, we see three possible approaches to abstracting patterns from the suffix trees to produce hierarchical classes of discernable states. The first is a method that looks for syntactic similarity in suffix tree paths, based on the heuristic that some intermediate node states might not matter, as when coincidental noise intervenes in a causal pattern with temporal lag (Schmill and Cohen 1995). The second method is one that strictly considers the similarity of the successor state distributions (Shalizi 2001), which could be combined with the first method to test the strength of proposed pattern clusters. This is sometimes thought of as a kind of *semantic* similarity, since states with different syntactic forms that give rise to similar outcome distributions can be thought of as having similar behavioral meaning. Finally, there is the method that tries to group states by their temporal proximity (George 2008), which would view the successor states in the probabilistic suffix tree as weighted edges in a DeBruijn Graph, and would cluster nodes that have strong temporal correlations. It appears to be an open research problem to evaluate these alone and in combinations, to apply them to multiple levels of abstraction, or to determine key characteristics of problem environments to which they most appropriately might be suited.

Spectrum of Reasoning and Representation

Having identified induction as a key to leveling up in our models of the mind, it might be worthwhile to review other forms of natural reasoning that have been studied, and to consider how they leverage and depend on each other. Very early in the history of scientific thought, the syllogisms of the ancient Greeks captured the essence of inference via *deduction*. Much later, in the mid 1700's, Hume and his contemporaries debated the importance and nature of *induction*. And it was more than another 100 years before what is now commonly called *abduction* was properly identified and characterized by C.S. Pierce (Minnameier 2010).

Deduction is usually associated with formal, exact, and infallible reasoning, where universal and necessary generalizations meet incontrovertible facts, to produce indisputable conclusions. Abduction, as a form of only plausible reasoning, also uses generalizations, but posits likely explanations for after-the-fact observations given them. Induction, as suggested in the previous section, is the holy grail, in the sense that both of the other types of inference depend on it. While some universally quantified natural language statements are formal and definitional (mathematically axiomatic), human linguistic and common-sense reasoning regularly adopt the use of deduction based on induced rules, and therefore in practice deduction is only as plausible as the foundation it is built upon.

These points illustrate that natural uses of logic and language are a form of digital thinking derived from simplifying assumptions, that is, the convenient and powerful methods we use to make our analog and continuous world easier to understand, and to facilitate communication and social cooperation. Animals reason about their environments without visible forms of symbolic language, but in spite of this they successfully use behavioral interactions and observations of others to communicate. The Physical Symbol System Hypothesis can still hold over temporal patterns of sensory data and their abstract generalizations, which at their lowest levels at least, appear to be emergent symbols whose meaning comes from their contextual occurrences and use.

Being able to recognize, generate, and reason with knowledge in linguistic forms seems to lie at the pinnacle of human cognitive capabilities. But jumping to representational conclusions modeled after linguistic forms may be preventing us from letting the trees more naturally define the forest. The gradual but powerful rise of data-driven and statistical weak-methods should give us pause, to consider and favor performance and behavior over explainable transparency. To wit, many abductive explanations of our reasoning are arguably telling stories that make us comfortable without strong justification for their particular selection (e.g. "the stock market was down today on concerns over oil prices"). Perhaps the precision of logic and language is an illusion, useful in its own right, but more derivative than foundational.

Integrating Across Technical Perspectives

Over the course of its history, Artificial Intelligence has kept a mostly consistent and widely accepted goal-oriented view, namely to understand intelligent behavior well enough to

emulate it with machines. However, as a professional discipline, the field has been splintered repeatedly in its technological approaches. Functional decomposition of human abilities has led some researchers to study vision, others language, others problem-solving or learning. Even within a subfield such as Machine Learning, there are numerous technical schools of thought and algorithmic orientations that cry for unification (Domingos 2015). Fortunately, AI has also played a part in spinning-off many useful engineering technologies, including Object-Oriented Programming, Operations Research, and Information Retrieval (search engines), so the divide and conquer approach has had its benefits. Yet a grand unification theory of cognition beckons us.

One major divide that has unfortunately largely persisted, is that between inexact and exact representation and reasoning. The Standard Model's discussion of metadata appears to try to accommodate both, but grounded knowledge in the form of sensory patterns and their generalizations should be recognized as symbolic in their own right. Using low-level data structures that capture temporal relationships in their very structure (as opposed to only using temporal predicates, the way we represent relations like ownership) seems justifiable as well, given their fundamental nature.

This brings us to our final call for integrative cognitive modeling, to build bridges between the oft-separated technical perspectives of probability and logic, or connectionist and symbolic AI. To the extent that the Standard Model helps to achieve this, it will have done Science a huge favor.

References

Aslanides, J., Leike, J., and Hutter, M. (2017). Universal Reinforcement Learning Algorithms: Survey and Experiments. arXiv preprint arXiv:1705.10557.

Brooks, R. (1986). A robust layered control system for a mobile robot. IEEE journal on robotics and automation, 2(1), 14-23.

Brosnan, S. F., and De Waal, F. B. (2003). Monkeys reject unequal pay. Nature, 425(6955), 297-299.

Bühlmann, P., and Wyner, A. J. (1999). Variable length Markov chains. The Annals of Statistics, 27(2), 480-513.

Chung, M., Buro, M., and Schaeffer, J. (2005). Monte Carlo planning in RTS games. In CIG.

Cohen, P. R., Greenberg, M. L., Hart, D. M., and Howe, A. E. (1989). Trial by fire: Understanding the design requirements for agents in complex environments. AI magazine, 10(3), 32.

Dean, T. L., and Boddy, M. S. (1988). An Analysis of Time-Dependent Planning. In AAAI (Vol. 88, pp. 49-54).

Domingos, P. (2015). The master algorithm: How the quest for the ultimate learning machine will remake our world. Basic Books.

Foote, A. L., and Crystal, J. D. (2007). Metacognition in the rat. Current Biology, 17(6), 551-555.

Gershman, S. J., and Daw, N. D. (2017). Reinforcement learning and episodic memory in humans and animals: An integrative framework. Annual review of psychology, 68, 101-128.

George, D. (2008). How the brain might work: A hierarchical and temporal model for learning and recognition. Ph.D. diss, Stanford University, Stanford, CA.

Harnad, S. (1990). The symbol grounding problem. Physica D: Nonlinear Phenomena, 42(1-3), 335-346.

Hawkins, J., and Blakeslee, S. (2007). On intelligence: How a new understanding of the brain will lead to the creation of truly intelligent machines. Macmillan.

Kahneman, D. (2011). Thinking, fast and slow. Macmillan.

Laird, J. E., Lebiere, C. and Rosenbloom, P. S. (In press). A Standard Model of the Mind: Toward a Common Computational Framework across Artificial Intelligence, Cognitive Science, Neuroscience, and Robotics. AI Magazine.

Levesque, H. J. (2017). Common Sense, the Turing Test, and the Quest for Real AI: Reflections on Natural and Artificial Intelligence. MIT Press.

Lian, R., Goertzel, B., Vepstas, L., Hanson, D., and Zhou, C. (2017). Symbol Grounding via Chaining of Morphisms. arXiv preprint arXiv:1703.04368.

Lieto, A., Chella, A., and Frixione, M. (2017). Conceptual Spaces for Cognitive Architectures: A lingua franca for different levels of representation. Biologically Inspired Cognitive Architectures, 19, 1-9.

Minnameier, G. (2010). The logicality of abduction, deduction, and induction. In Ideas in action: Proceedings of the applying Peirce conference (pp. 239-251). Helsinki: Nordic Pragmatism Network.

Mirwan, H. B., and Kevan, P. G. (2013). Social learning in bumblebees (Bombus impatiens): worker bumblebees learn to manipulate and forage at artificial flowers by observation and communication within the colony. Psyche: A Journal of Entomology, 2013.

Neumann, M. and Cowley, S. (2013). Human Agency and the Resources of Reason. In Cognition Beyond the Brain (pp. 13-30). Springer.

Newell, A., & Simon, H. A. (1976). Computer science as empirical inquiry: Symbols and search. Communications of the ACM, 19(3), 113-126.

Pearl, J. (2014). Probabilistic reasoning in intelligent systems: networks of plausible inference. Morgan Kaufmann.

Ron, D., Singer, Y., and Tishby, N. (1994). Learning probabilistic automata with variable memory length. In Proceedings of the seventh annual conference on Computational learning theory (pp. 35-46). ACM.

Schmill, M. D., and Cohen, P. R. (1995). Learning Predictive Generalizations for Multiple Streams: An Incremental Algorithm. Department of Computer Science Technical Report 95-36, University of Massachusetts, Amherst.

Shalizi, C. R. (2001). Causal architecture, complexity and self-organization in the time series and cellular automata. Ph.D. diss., University of Wisconsin, Madison, WI.

Steels, L. (2008). The symbol grounding problem has been solved. So what's next. Symbols and embodiment: Debates on meaning and cognition, 223-244.

Valiant, L. (2013). Probably Approximately Correct: Nature's Algorithms for Learning and Prospering in a Complex World. Basic Books (AZ).

Volf, P. A., and Willems, F. M. (1995). A study of the context tree maximizing method. In Proc. 16th Benelux Symp. Inf. Theory, Nieuwerkerk Ijsel, Netherlands (pp. 3-9).