Evaluation of Cognitive Architectures: Views for Assessment of Artificial General Intelligence

Sanket Goutam
Department of Computer Science,
North Carolina State University,
Raleigh, NC

Abstract

The concept of ”Artificial General Intelligence” has emerged as an approach to identify fundamentally distinct property from domain specific capabilities, to create machines with very general cognitive functions that achieve high level of autonomy. However the metrics for accessing the partial progress of such systems, unlike straightforward ways of accessing the achievement of human-level AGI (the Turing test, Robot University Student Test), remain more problematic. In this paper, we review the identified capabilities a cognitive architecture should support, some properties it should exhibit and the evaluation criteria (views) used in prior studies to evaluate the ”intelligent” behavior in cognitive architectures. We highlight the performance of each view in evaluating ”intelligence” and discuss limitations in their approach with a focus on understanding how intelligent behavior is evaluated and the constraints of specific views.

Introduction

The standard approach in the AI discipline (Russell and Norvig 1995) views artificial intelligence largely in terms of the pursuit of discrete capabilities. The term ”narrow AI” coined by Ray Kurzweil (Kurzweil 2010) is used to refer to the creation of systems which demonstrate specific ”intelligent” behaviors in narrow and well-defined contexts. But for such domain specific systems if the environment is changed, if the context or the behavior specification is modified even a little bit, an external reprogramming or reconfiguration effort from humans is required to enable the system to retain its level of intelligence. But the idea of a generally intelligent system, like humans, is quite different in the sense that we possess a broad capability to self-adapt to changes in environment or goals, performing ”transfer learning” (Taylor, Kuhlmann, and Stone 2008) to use the knowledge from a learned task to speed up learning in a novel task. The concept of ”Artificial General Intelligence” (Goertzel 2014) thus has emerged as an antonym to ”narrow AI”, to refer to systems with this sort of broad generalization capability.

Any system that is considered as ”AGI” does not need to possess infinite generality, adaptability, and flexibility; although it can be considered as bridging the gap between current domain-specific targeted solutions and creating general human-like intelligence. General intelligence is defined as the ability to achieve a variety of goals, and carry out different tasks, in heterogeneous contexts and environments. Any system capable of demonstrating general intelligence should be able to handle problems and situations that are more dynamic and varied in nature than those anticipated by its creators. Such systems are expected to be good at generalizing the knowledge gained from a specific task and utilize it to learn to solve different problems in different contexts.

The most promising work for creating machines with general intelligence belongs to the research area of cognitive architectures. A cognitive architecture specifies the underlying infrastructure for an intelligent system, which are essentially a move from specification to implementation. Cognitive Architectures, often designed with some cognitive theory in mind, typically deal with relatively large software systems that have many heterogeneous parts and subcomponents operating together to solve general problems and tasks in more than one domain. Many of these architectures are built to control intelligent agents (Wooldridge and Jennings 1995) that are designed for a specific agent theory. Cognitive Architecture research covers a broad range of topics at all levels: from underlying theoretical assumptions, inspiration, methodology, motivation, structure, requirements, and technology. Research on cognitive architectures is important because it supports a central goal of artificial intelligence and cognitive science: the creation and understanding of synthetic agents that support the same capabilities as humans.

This survey paper reviews the theoretical concepts behind cognitive architectures, starting with the design goals that a cognitive architecture must meet, or the capabilities that a system should be able to demonstrate in order for it to be considered as AGI. This is followed by a summary of various evaluation criteria that is mostly associated with the assessment of cognitive architectures, and then we discuss prior studies on assessment of such architectures and briefly summarize their findings about implementation issues and conclude by reiterating the need to shift to plastic design approaches.

Cognitive Architecture Design

An architecture includes those aspects of a cognitive agent that are constant over time and across different application domains, including:

- short term and long term memories to store agent beliefs, goals, and knowledge
- representation of elements present inside these memories as some data structure
- functional processes that operate on such defined structures, including mechanisms to not only utilize them but also learning mechanisms to alter them

However, for an agent capable of demonstrating general intelligence, these memories are bound to change over time and thereby knowledge and beliefs cannot be encoded as a part of the agents' architecture - different knowledge bases and beliefs should be interpreted by the same underlying cognitive architecture. Thus architectural research, as opposed to research on expert systems aims for breadth of coverage across a diverse set of tasks and domains and offers accounts of intelligent behavior at the systems level, rather than at the level of component methods designed for domain specific tasks.

In this section we briefly summarize the theory on cognitive architecture designs and challenges (Langley, Laird, and Rogers 2009) and cover various capabilities that a cognitive architecture should demonstrate.

Capabilities

Intelligent systems are designed to engage in activities that, as a whole, constitute to its functional capabilities. Although activities such as recognition and decision making require a well-defined architecture on their own, in this section we briefly cover different functionalities without specifying the underlying mechanisms used to implement them.

Recognition and categorization An intelligent agent must be able to recognize situations or events from its surroundings and should be able to interpret it as instances of known or recognizable patterns. This is closely related to categorization as the agent should also be able to assign objects, situations, and events from its environment to known concepts or categories.

Although recognition is often considered a primitive process that underlies many higher level functions,and categorization as one such higher level function, they are more closely related and cognitive architectures in order to support them must provide a way to represent patterns and situations in memory. A recognition process must also be included that lets a system identify when a particular situation matches a stored pattern and the degree to which it matches, thereby allowing the agent to adapt its learning behavior to changes in the environment.

Decision making and choice A critical part of any intelligent system is not just being able to recognize changes in its surrounding but also to make decisions and select among alternatives. Decisions are very closely associated with the underlying recognition of a situation or pattern, and most cognitive architectures combine the two processes as a unified recognize-act cycle underlying all cognitive behavior. In order to support this sort of one-step decision making in the systems level, the architecture must provide ways to represent alternative choices or actions, that encompass both internal cognitive functions and external ones.

But making decisions in a dynamic environment purely based on pattern matching is not entirely an "intelligent" behavior. An ideal cognitive model should also be able to improve upon its decisions through learning (meta-learning (Vilalta and Drissi 2002)). This results in improvement in the decision making which will be reflected in the overall behavior of the agent in new environments.

Perception and situation assessment Systems will often be equipped with a variety of sensors from which it must sense, perceive, and interpret information about its environment. A cognitive architecture thus should be able confront the issue of attention, that is it should be able to decide how to allocate and direct its perpetual resources to detect relevant information in complex environments. It should also be able to track rapid changes in a dynamic environment and should have a perpetual knowledge of what sensors to use, when and where to focus them, and should be able to overcome the various interferences that are plausible.

Prediction and monitoring Perpetually existing intelligent systems should be in a position to predict future situations and events accurately which requires the knowledge of how the actions of an agent will affect the outcome or the environment the agent is in. An ideal architecture should also include the ability to learn predictive models from experience and to refine them over time.

Planning Cognitive architectures allow intelligent systems to achieve their goals in new situations by enabling them to generate plans and to solve problems. Planning is only possible when the agent has an environmental model that predicts the effects of its actions. Thus a cognitive architecture must have mechanisms to represent plans as ordered set of actions, with their expected effects, and how each effect enable later actions. It should also allow an agent to construct a plan from components available in memory, which in turn provides problem solving capabilities to the agent through multi-step construction of a problem solution.

Reasoning and belief maintenance In contrast to planning where we achieve objectives by taking actions, reasoning draws mental conclusions from other beliefs or assumptions that the agent already holds. So in order to support reasoning, a cognitive architecture must first represent relationships among beliefs, using first order logic, production rules, and neural networks or Bayesian networks. Belief maintenance and their relations is essentially important in dynamic environments where the situations may change in unexpected ways and the agent must track the changes to determine whether it should continue its former beliefs.

Execution and action Cognition also allows supports and drives our activity in the environment. Thus being able to store, represent and execute motor skills that enable such activity is also a requirement for a cognitive architecture. Not only limited to execution, an ideal cognitive model should also be capable of learning about skills and execution policies from experience.

Interaction and communication The most effective form of obtaining knowledge in humans is through communication, so naturally being able to communicate acquired knowledge between systems becomes a major requirement in any cognitive architecture. It should support mechanisms to support transformation of knowledge into a communicable form so that other agents are able to acquire syntactic and semantic knowledge for use from the communicated message.

Learning Learning usually involves generalization beyond specific beliefs and events. Many architectures treat learning as an automatic process that is not subject to inspection or conscious control, but also use meta-reasoning to support learning in a more deliberate manner. Since the data required for agents learning comes from various sources, it should be able to support processing on different memory structures to improve the agents' capabilities.

Views of cognitive architecture: choosing an approach

Each science is differentiated from the others not merely by the set of phenomena it claims as its object of study, but also by the approach it takes (or the paradigm (Kuhn 2012)). Similar to the view of man as a symbolic processor for the study of the phenomena of human intelligence (Lenat 1977), the study of cognitive architecture requires various different approaches (views) in order to evaluate the various capabilities that it demonstrates.

But the evaluation of cognitive systems poses a greater challenge than evaluation of component knowledge structures and methods, primarily because of the architectural research occurring at the systems level. Additionally a cognitive architecture offers a unified theory of cognition (Newell 1994) with different modules tightly bound together in order to support synergistic effects. Since evaluating the synergy in cognitive systems cannot be tested empirically, the evaluation criteria covered below helps in understanding the functionality of the architecture.

Generality Since we consider cognitive architectures as the building block for artificial general intelligence, it would be really weird if we do not include generality of the architecture as a key dimension along which to evaluate a candidate framework. And the most straightforward way of evaluating generality in an architecture is to construct an intelligent agent using the same architecture and evaluating its performance under diverse environmental conditions. The intelligence of the said architecture is then determined by the number of environments in which it is able demonstrate intelligent behavior. Thus broader the range of these environments, the greater is its generality.

Rationality Humans make rational decisions in our day to day lives, and in a way intelligence can be defined when the pursuit of certain behavior is accompanied by specific logical reasoning. Rationality in agents can be identified from the relationship between its goals, knowledge, and its actions. (Newell 1982) states "If an agent has knowledge that one of its actions will lead to one of its goals, then the agent will select that action". However given the choice of multiple actions that can be taken to achieve the same goal, an agent should ideally choose the most optimal action. Although due to the limited cognitive resources available to an agent, the concept of bounded rationality (Simon 1957) is put forward that states that the agent will behave nearly optimal to its goals though limited by its resources.

Efficiency and scalability The sole idea of research in artificial general intelligence is to build agents that can be used in practice to solve problems at least as efficiently as humans, implying that the agents must be able to perform tasks under a time and space constraints. The design of architectures that include a recognize-act cycle, as discussed in the previous section, allows for the architecture to support real-time systems.

However architectures must be in a position to handle different difficulties of task and situations, and thus need to be scalable across different scenarios. We evaluate the scalability of a system by testing it against tasks of varying difficulty, environmental uncertainty and other more complicating factors. The scalability factor of the architecture is then determined by how unaffected its performance is by these factors.

Reactivity Intelligent agents are expected to demonstrate intelligent behavior even in extreme and unpredictable environments, and the underlying cognitive architecture should support their operation. Thus the ability to react to unpredictable changes in the environment is also a criteria for the intelligence in cognitive architectures. The more rapidly an architecture responds or the greater its chances of responding to unpredictable changes, the greater is the reactivity of the system.

Improvability One of the key aspects of human intelligence is our ability to improve ourselves over time through learning, experiences, and self-evaluation. Improvability in cognitive architectures is evaluated based on the agents' ability to perform a particular task that it could not perform earlier after acquiring some knowledge. An agent is also expected to learn from its experiences, so improvability in that case could be measured in terms of its ability to perform new tasks.

Autonomy We expect intelligent agents to be able to function on their own for extended periods of time and the architectures supporting them should allow them to create their own tasks and goals. They should also exhibit robustness in novel environments and should not fail when they encounter unexpected situations. Autonomy in intelligent agents is measured by presenting them with high-level tasks that require autonomous decision making and then evaluating their performance in that environment.

Choosing a viewpoint

As we have defined in the previous section, the different views for evaluating "intelligence" helps us in quantifying each capability that the cognitive system should demonstrate. While it would be highly beneficial to have a benchmark or common test problem for cognitive architectures to facilitate more general comparison, no appropriate benchmark of reasonable maturity is known to exist (Thórisson and Helgasson 2012). So in order to measure intelligent behavior, we need to choose some viewpoint among the ones discussed above and evaluate a given cognitive architecture using that. We now present a summary of two previously used viewpoints for evaluation of cognitive architectures and discuss the weakness in their approach.

Intelligence as Efficient searching of an a priori space The very first intelligent systems ever built by humans were based on the understanding that most of the behavior that we regard as "intelligent" involves some sort of discovery process. These models were built with the view of Man as information processor (Newell, Shaw, and Simon 1957; Newell and Simon 2007). Significant efforts were made to model a wide variety of cognitive activities (recognizing, problem solving, inventing) as a search in which the performer is guided by 'heuristics rules'. Thus different cognitive models were developed using Heuristic Rule Guided Search and the ability to zero in on something from a vast search space as "intelligent" behavior.

Probably one of the earliest AI programs ever written was the Logic Theorist (Newell, Shaw, and Simon 1957), which was repeatedly given symbolic logic theorems for it to find a formal proof for each. The search for proof in LT was done in a completely exhaustive manner with the help of a few heuristics to constrain its search space. The learnings about rule guided search from LT was used to build another system, GPS (General Problem Solver), with the aim to embed few general heuristics in a domain-independent form and thus allowing it to solve any problem once it is specified in GPS formalism. However these few general heuristics could not possibly be as powerful as the general human-like intelligence that we desire.

But the importance of using both general heuristics and task-specific heuristics for guidance was soon realized and a scientific invention tool (AM) was developed (Lenat 1977). AM performed discovery of new mathematics concepts and relationships between them. It viewed open-ended math research as a search and explored in a space of partially-developed concepts while guided by a few hundred heuristic rules. AM was able to demonstrate that open-ended scientific theory formation, including defining and exploring of new concepts and relationships) could be mechanized, and modeled as a heuristic rule guided search. However the major flaw in considering a rule guided search as a verification of intelligence brings forward the view of generality discussed earlier, is it really a general intelligence if we have to manually provide numerous heuristics again if the environment of the agent is changed.

In essence, AM and other knowledge based expert programs like MYCIN (Shortliffe 1974), MOLGEN (Feitelson and Stefik 1977), PROSPECTOR (Duda et al. 1978), perform complex problem solving by utilizing the power of additivity in rule-guided search, where many small pieces of local knowledge combine to produce sophisticated global effects, with a consequent ease of introducing new pieces of knowledge. However the major flaw in this is that rule guided search could theoretically be synonymous to discovery process but that does not provide enough bias for defining intelligence, let alone general intelligence.

Autonomous behavior of intelligent agents Autonomy, as discussed earlier, is also a key approach to evaluate any system that is considered generally intelligent and quite recently (Thórisson and Helgasson 2012) this concept has been used as an organizing principle for the comparison of cognitive systems. The viewpoint considered is that of an exploration robot that can be deployed, without special preparation, into virtually any environment, and move between them without serious problems. The environments that the robot encounters may vary significantly in dynamics and complexity. The goal of the robot is exploration, in a time constrained but otherwise open-ended world, which translates into learning about the environment through observation and action.

Now given that the robots' processing capacity is limited and the environment(s) information is rich, the robot must be able to demonstrate attention capability in order to select which sensory data to process and how deeply. I should also have some expectations for upcoming events, to steer it focus of attention, thereby requiring the capability of prediction. It should also be able to couple reasoning with prediction to avoid trial and error approaches in situations of irreversibility. It also requires introspective capabilities that allow it to evaluate and reason about itself in order to improve its internal, and thus external, operation. So in totality four main themes are considered vital to the systems' operation: Real-time, resource management, learning, and meta-learning.

The architectures covered in the study using autonomous behavior of systems range from architectures designed at the robotics end like Ymir (Thórisson 1999) to more traditional cognitively focused architectures like ACT-R (Anderson 1996; Anderson, Matessa, and Lebiere 1997), Soar (Laird 2008), NARS (Wang 1995), and also the Ikon Flux architecture (Nivel 2007).

Based on their exhaustive comparison on the parameters discussed earlier, they identify a common tendency among most of the architectures as to ignore realtime operation and resource management aspects even though these capabilities are quite essential for any higher-level intelligence. Time is often not considered as absolute in most systems, rather the tasks are scheduled with time as a relative function. Cognitive architectures should also be able to implement resource management in much more efficient way so as to allow the system to prioritize processes by itself without having a constructionist approach (Thórisson 2009) of human defined heuristics. Additionally, a major importance is given to the requirement of an agent to support a plastic infrastructure, one in which it should be able to reconfigure its own learning structure using meta-learning.

The study also highlights the importance of designing architectures from the get-go with a more complete set of cognitive functions and operational capabilities and not just having a number of different cognitive processes interacting with each other. The authors stress upon the move from constructionist approach of building AI systems to constructivist approach with design goals of building architectures that constitute of smaller components, and thus making modifications easier.

The architecture comparisons reveal that very few existing cognitive architectures are based on viable methodologies that would help reach human level autonomy. They also argue that present research should search for methodologies that are able to handle systems of substantial size and complexity. Since meta-learning and improvability is also a part of these architectures, the effort of a system being able to reconfigure itself should be made simpler.

Limitations

Although the study of cognitive architectures from the viewpoint of autonomy does provide valuable insights into the importance of a number of the capabilities of an intelligent system, the evaluation stands constricted to our definition of autonomy. Autonomous agent in a broad sense, implies that it should have its own body and should be able to employ its own body to sustain long-time tight interactions with the external environment to pursue its own goals (Chella and Manzotti 2009). Embodiment here refers not to a robot being able to move using actuators but the kind of development and causal processes engaged between an agent, its body, and the external environment. This raises a concern in the definition of autonomous agents that whether a non-embodied agent would every really be autonomous. This begs the question of redefining the autonomous behavior in humans, what aspects of the cognitive behavior and the pragmatic architecture in humans is considered as autonomy.

Additionally, similar constraints also exist on the design and formulations of the capabilities for any cognitive architecture. The design of most architectures only focus on the generalization of solutions to problems or executions of actions, but categorization, understanding and recognition of events in the environment should be considered a crucial aspect in designing such architectures.

The extensive applications of cognitive architectures in deployment of systems focusing on problem solving has led to researchers move away from implementing episodic memory into these architectures. However the inclusion of episodic memory is an essential part for any system that attempts to use its learned knowledge in diverse environments.

From an engineering perspective, and even from meta-learning perspective, as identified by (Thórisson and Helgasson 2012), architectures would become much more interesting if the granularity of the subsystems is increased to such a level so as to allow both the agent and the human developers to be able to modify the architecture easily based on requirements of the environment.

Conclusion

Using the various views for the evaluation of cognitive architectures in order to assess the artificial general intelligence exhibited by an agent does seem a very accurate way of measuring partial progress of the system, but a design specification needs to be formulated for building the architecture complete with formal definitions of each capability of the system. This would prevent the ambiguity in the views that each system is bound to, and would allow developers to build truly general cognitive models by leveraging granular abstractions of the function modules. This design specification would even allow evaluation using more common test problem like the one proposed in (Johnston 2010).

With this survey on the theory, design, and evaluation of cognitive architectures we have covered the design goals behind these architectures, the capabilities that they need to address, and the various approaches to evaluate the intelligent behavior in these architectures. We also covered the shortcomings of the reviewed literature and presented further research goals that needs to be addressed in the development of architectures for Artificial General Intelligence (AGI) systems.

References

Anderson, J. R.; Matessa, M.; and Lebiere, C. 1997. Act-r: A theory of higher level cognition and its relation to visual attention. *Human-Computer Interaction* 12(4):439–462.

Anderson, J. R. 1996. Act: A simple theory of complex cognition. *American Psychologist* 51(4):355.

Chella, A., and Manzotti, R. 2009. Machine consciousness: a manifesto for robotics. *International Journal of Machine Consciousness* 1(01):33–51.

Duda, R. O.; Hart, P. E.; Nilsson, N. J.; and Sutherland, G. L. 1978. Semantic network representations in rule-based inference systems. In *Pattern-directed inference systems*. Elsevier. 203–221.

Feitelson, J., and Stefik, M. 1977. A case study of the reasoning in a genetics experiment. *Heuristic Programming Project Memo HPP-77-18 (working paper)(May 1977)*.

Goertzel, B. 2014. Artificial general intelligence: Concept, state of the art, and future prospects. *J. Artificial General Intelligence* 5:1–48.

Johnston, B. T. S. 2010. The toy box problem (and a preliminary solution).

Kuhn, T. S. 2012. *The structure of scientific revolutions*. University of Chicago press.

Kurzweil, R. 2010. *The singularity is near*. Gerald Duckworth & Co.

Laird, J. E. 2008. Extending the soar cognitive architecture. *Frontiers in Artificial Intelligence and Applications* 171:224.

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

Lenat, D. B. 1977. The ubiquity of discovery. *Artif. Intell.* 9:257–285.

Newell, A., and Simon, H. A. 2007. *Computer science as empirical inquiry: Symbols and search*. ACM.

Newell, A.; Shaw, J. C.; and Simon, H. A. 1957. Empirical explorations of the logic theory machine: a case study in heuristic. In *Papers presented at the February 26-28, 1957, western joint computer conference: Techniques for reliability*, 218–230. ACM.

Newell, A. 1982. The knowledge level. *Artificial intelligence* 18(1):87–127.

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

Nivel, E. 2007. Ikon flux 2.0.

Russell, S. J., and Norvig, P. 1995. Artificial intelligence - a modern approach: the intelligent agent book. In *Prentice Hall series in artificial intelligence*.

Shortliffe, E. H. 1974. A rule-based computer program for advising physicians regarding antimicrobial therapy selection. In *Proceedings of the 1974 annual ACM conference-Volume 2*, 739–739. ACM.

Simon, H. A. 1957. Models of man; social and rational.

Taylor, M. E.; Kuhlmann, G.; and Stone, P. 2008. Transfer learning and intelligence: an argument and approach. In *AGI*.

Thórisson, K. R., and Helgasson, H. 2012. Cognitive architectures and autonomy: A comparative review. *J. Artificial General Intelligence* 3:1–30.

Thórisson, K. R. 1999. Mind model for multimodal communicative creatures and humanoids. *Applied Artificial Intelligence* 13:449–486.

Thórisson, K. R. 2009. From constructionist to constructivist ai. In *AAAI Fall Symposium: Biologically Inspired Cognitive Architectures*.

Vilalta, R., and Drissi, Y. 2002. A perspective view and survey of meta-learning. *Artificial Intelligence Review* 18:77–95.

Wang, P. 1995. *Non-axiomatic reasoning system: Exploring the essence of intelligence*. Citeseer.

Wooldridge, M., and Jennings, N. R. 1995. Intelligent agents: theory and practice. *Knowledge Eng. Review* 10:115–152.