Strong CogSci: Guidance from Cognitive Science On the Design of a Test of Artificial Intelligence.

Conference Paper · January 2015

CITATION
1

READS
158

6 authors, including:

Robert H Thomson
United States Military Academy West Point
65 PUBLICATIONS **281** CITATIONS
SEE PROFILE

Christian Lebiere
Carnegie Mellon University
235 PUBLICATIONS **13,577** CITATIONS
SEE PROFILE

Oscar J. Romero
Carnegie Mellon University
47 PUBLICATIONS **226** CITATIONS
SEE PROFILE

Strong CogSci: Guidance from Cognitive Science On the Design of a Test of Artificial Intelligence

Christian Lebiere, Dan Bothell, Don Morrison, Alessandro Oltramari, Michael Martin, Oscar Romero, Robert Thomson, Jerry Vinokurov

Carnegie Mellon University Psychology Department
5000 Forbes Avenue, Pittsburgh PA 15213

Abstract

We propose a test of human-like intelligence inspired by constraints from cognitive science on unified theories of cognition. The most salient characteristics of this test include the capacity for open-ended, generative behavior, the ability to learn and adapt in real time in complex environments, and the ability to interact productively with other human and artificial agents. The proposed test environment is Minecraft, a virtual world reflecting many of the key characteristics of real world embodiments. The core task is to build assemblies of varying complexity and purposes in cooperation with human or artificial entities.

Background

Cognitive Science (CogSci) and Artificial Intelligence (AI) have diverged over the decades from their common beginning focused on understanding the nature of human cognition by reproducing it computationally (Newell & Simon, 1972). CogSci has often focused on laboratory tasks that isolate very specific human capabilities but do not generalize well to broader cognition (Newell, 1973). While AI has also focused on specific tasks, such as playing chess or driving a car, it has tackled them by deriving specialized techniques to maximize performance rather than general, cognitively plausible mechanisms.

The suggestion has been made recently (e.g., Mitchell, 2002; Lebiere & Wray, 2006) that it is time for a rapprochement that would allow each discipline to constrain the other (AI with algorithms, CogSci with behavioral and neural data) to their mutual benefit in exploring a very broad design space of intelligence systems. Therefore, it might behoove us to consider the constraints that have been advanced for theories of human cognition in the design of tests of artificial intelligence. Anderson & Lebiere (2003) reviewed a set of criteria from Newell (1990) to evaluate unified theories of cognition:

1. Capable of flexible, universal behavior
2. Exhibit rational, adaptive behavior
3. Use vast amounts of knowledge
4. Behave robustly given error, uncertainty
5. Acquire capabilities through development
6. Integrate diverse knowledge sources
7. Operate in real time
8. Use natural language
9. Exhibit self-awareness
10. Learn from its environment
11. Arise through evolution
12. Be realizable within the brain

While the latter two criteria may be considered overly specific (although one might consider how the constraints on how the agent arises as informative as its final functionality), the others provide a strong set of desiderata for any entity – natural or artificial – that would claim to be considered generally intelligent. The Turing Test would fail the overall behavior aspect, being limited to disembodied linguistic communication. Other tests such as chess expertise or Robocup fall short of required breadth and knowledge requirements. The Winograd Schema Challenge focuses on linguistic communication and knowledge but does not require any interactive behavior.

In general, the central criteria converge to emphasize the notion of agency: embodied behavior in a real time environment (criteria #1, 7), at different time scales from reactive to deliberative, social and beyond (e.g. Anderson, 2002) (#5, 11, 12), the need for adaptivity and open-ended learning in unpredictable environments (#2, 4, 9, 10), but also the ability to rely on knowledge and instructions to generate structured, goal-directed behavior (#3, 6, 8).

A Cautionary Note

Before proposing a specific instantiation of these criteria, it is worth discussing what does not qualify as desirable attributes of a test, to make sure that we do not look for our keys to intelligence under the lamppost of available techniques. Tests that emphasize the deployment of large amounts of knowledge without requiring learning in interaction with the environment or other entities encourage brittle engineering approaches over the organic development of knowledge. Tests that require answering precise sets of questions, irrespective of their nature, favor brute force techniques that entirely miss the ability to generate constructive, goal-directed behavior in a complex environment. Conversely, tests that require sophisticated perception and mechanical interaction with the environment overly reward the refinement of those techniques at the expense of higher-level cognition and general knowledge. Finally, tests that take place in precise, cut-and-dried environments with well-defined performance functions lack the key requirement of generating effective behavior in ill-defined environments that is a hallmark of human intelligence.

Minecraft: An Open-Ended Environment

We propose to use the popular computer game Minecraft for creating meaningful challenges. Minecraft is a sandbox game that has no fixed objectives or goals. It provides an agent with a simplified 3D world and a well-defined set of actions that can be performed. That world consists of inanimate items like rocks, wood, and water that the player can collect, move, and combine into new items. It may also contain other agents that can be game-controlled creatures or human agents. This open-endedness has made it very popular both as a game with 100 million registered users (Reilly, 2014), and as a teaching tool through the MinecraftEdu project (http://minecraftedu.com).

The use of a virtual world requires that the agents be embodied and interact with the world, but alleviates the need for having to initially solve complex perceptual and motor processing challenges. The ability to operate in the world without communication or other agents also allows to initially create tasks that do not require language comprehension or generation. Those can be introduced as additional agents are added and social interactions are required. The nature of the world also provides many opportunities for the agent to demonstrate learning. Learning can occur from the environment (learning properties of items, where to find them, how to navigate around the world), about other creatures (how they react and what to expect from them), and can involve learning from other agents, such as being taught how to construct some item or travel to a particular location.

Task: Minecraft Assembly Assistant

The most fundamental task in Minecraft is constructing complex objects out of basic materials. Construction of artifacts is a hallmark of human intelligence that brings into play many cognitive capabilities, from complex object identification, to problem-solving to generate both a design to accomplish a given function and a plan to execute it, to motor actions that can manipulate the environment to produce the desired effect. The task can be individual or involve cooperation with one or more agents. In particular, the agent could be instructed at various levels of details in how to accomplish the task by a human or simulated agent.

In order to perform this task, the agent will be forced to manipulate the environment in which it finds itself and learn the various affordances of the objects within it. In addition to learning the discrete affordances of objects, the agent will also have to learn the continuous physics of the Minecraft world, which are simple enough to be tractable but sufficiently complex to offer a substantial challenge. Learning can involve self-directed exploration and experimentation as well as receiving instructions from other agents, which requires capabilities for social interaction as well as incremental knowledge development and integration of knowledge from various sources. Introducing stochasticity or external events in the environment can increase the need for robustness and adaptivity. Creativity can be required to find new solutions to original problems. Evaluation metrics include time and accuracy in building assemblies of various complexity, level and amount of instruction required to complete the task and success in generalizing to new problem domains.

Conclusion

We argue that the development of meaningful tests of artificial intelligence can be guided by adopting the same criteria used in evaluating theories of human cognition. Adopting those constraints would not require the use of evaluation metrics that constrain artificial intelligence to exactly reproduce human intelligence in all its details. However, it would provide safeguards against common pitfalls that have affected past tests, such as the incentives to develop specialized techniques designed to perform well in narrow domains but that consistently fail to generalize to broader contexts, or the tendency to focus on particular functionality that is not central to true intelligence. Our hope is that the adoption of those constraints will guide the development of tests that result in true incremental progress toward mechanisms, representations and architectures that provide a general account of human and artificial intelligence across all domains of human activity.

References

Anderson, J. R. (2002). Spanning seven orders of magnitude: A challenge for cognitive modeling. *Cognitive Science,* 26, 85-112

Anderson, J. R. & Lebiere, C. (2003). The Newell test for a theory of cognition. *Behavioral & Brain Sciences* 26, 587-637.

Lebiere, C. & Wray, R. (2006). Between a Rock and a Hard Place: Cognitive Science Principles Meet AI-Hard Problems. *AAAI Spring Symposium Technical Report SS-06-02*. Menlo Park, CA: AAAI Press

Mitchell, T. (2002). AAAI Presidential Address. Retrieved from http://aitopics.org/link/tom-mitchells-aaai-presidential-address-august-2002.

Newell, A. (1973). You can't play 20 questions with nature and win: Projective comments on the papers of this symposium. In W. G. Chase (Ed.), *Visual information processing* (pp. 283-308). New York: Academic Press.

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

Newell, A., and Simon, H. A. (1972) *Human problem solving* Englewood Cliffs, NJ: Prentice-Hall

Reilly, Luke (2014). Original Minecraft Reaches 100 Million Registered Users. *IGN*. Ziff Davis. Retrieved from http://www.ign.com/articles/2014/02/26/original-minecraft-reaches-100-million-registered-users.