What Should AGI Learn From AI & CogSci?

Pei Wang[1], Bas R. Steunebrink[2], and Kristinn R. Thórisson[3]

[1] Temple University, Philadelphia PA 19122, USA. pei.wang@temple.edu
[2] The Swiss AI Lab IDSIA, USI & SUPSI
[3] Reykjavik University & Icelandic Institute for Intelligent Machines

Abstract. While the fields of artificial intelligence (AI) and cognitive science (CogSci) both originated from a deep interest in the same phenomenon – intelligence – and both setting themselves high aims in their early days, each has since greatly narrowed its focus, and all but abandoned their core subject for a more limited version of the phenomenon. The many non-obvious causes for this change over the decades are perhaps understandable, but they have significantly reduced the potential of both fields to impact our understanding of the fundamentals of intelligence – in the wild and in the laboratory. This position paper argues that researchers in the field of artificial general intelligence (AGI) should carefully posit their research objectives and methodology to avoid repeating the same mistakes.

1 The Big Picture of Intelligence and Cognition

Roughly speaking, artificial intelligence (AI) and cognitive science (CogSci) come from the same observation and imagination, namely that in a certain sense, the human mind and the electronic computer are – or can become – similar to each other. The similarities (and differences) have been suggested by many people, including Wiener [26], Turing [16], von Neumann [9], McCulloch and Pitt [7], though each from a different perspective.

Initiated in this atmosphere, AI and CogSci can be seen as two sides of the same coin: while the former attempts *to build a mind-like machine* [11], the latter tries *to study the mind as a machine* [1]. Their relation is like that between *engineering* and *science* in general, that is, there is a strong mutual dependence. It is obvious that, to build an intelligent system, one has to have a clear idea about how intelligence works, and most of our knowledge on that topic comes from the study of the human mind. On the other hand, to evaluate the correctness of a theory of cognition, a straightforward way is to model it in an artifact to see if it produces the expected results.

Given this relation, it is natural for AI to get inspiration from CogSci, as well as for CogSci to use AI models. Various theories have been proposed both to explain the phenomena observed in human cognition and to guide the design of machine intelligence (*cf.* [8, 10]).

However, as the difficulties in this research became more and more clear, the mainstream in both fields gradually departed from the original objective to

pursue some more "manageable" and "realistic" goals, and in this process the two fields have been moving away from each other.

In AI, this change is described as "AI adopts the scientific method," since "It is now more common to build on existing theories than to propose brand new ones, to base claims on rigorous theorems or hard experimental evidence rather than on intuition, and to show relevance to real-world applications rather than toy examples" [12], to quote from a well-known textbook. According to some influential and representative opinions (*cf.* [6, 5]), AI should follow the same theory and practice as computer science, thus targeting the problems that are traditionally solvable by the human mind only.

Similarly, in CogSci many researchers have gradually moved away from the objective of studying all types of cognitive system, and instead focused solely on human cognition. Currently the homepage of the Cognitive Science Society explicitly announces that "The Cognitive Science Society, Inc. brings together researchers from many fields who hold a common goal: understanding the nature of the human mind."[4] So CogSci is primarily no longer about minds *in general*. Following this approach, computer systems or AI only serve as tools to model human cognition and intelligence, and the models are desired to be as faithful to the human mind as possible. As a result the conferences and publications of CogSci in recent years have been dominated by cognitive psychology [4] – a field that *exclusively* focuses on human cognition – with the influence and presence of neuroscience continually getting stronger as well.

A consequence of the above trends is that AI and CogSci have both been moving away from a common goal, and retracting back to computer science and cognitive psychology, respectively. Though members of these fields are still producing results of theoretical and practical value, the original motivation of studying "minds" in biological and electronic systems *alike* has been mostly absent in the mainstream of both fields.

As for artificial general intelligence (AGI), it has been very clear from the very beginning that the aim of the field is to return to the original goal of AI, that is, to build general-purpose "minds" that are comparable to human intelligence in general, rather than building special-purpose "tools" [3, 24]. For such systems to be built, it can be argued that we also need to carry out research in CogSci in its original form, that is, to study the cognitive process in computers, rather than merely using computers as tools to study human cognition. Roughly speaking, we need a general science or theory of intelligence and cognition, one which takes human and computer intelligence as special cases [2, 21].

As we have argued in other places [23, 21], a proper treatment of the major concepts is to separate two levels of description. At a general level, concepts like "intelligence," "cognition," "thinking," "mind," and so on, should be treated as *medium-independent*, that is, described and studied from a functional point of view, without assuming any "implementation details." At a more concrete level we can study the realization or implementation of the above concepts in human, computer, animal, as well as in groups or societies, even in extra-terrestrial forms.

[4] Cited from `http://cognitivesciencesociety.org/index.html` in May 12, 2014.

Such a two-level structure acknowledges both the similarities and the differences between human and computers, that is, AI must be similar to human intelligence in *certain* fundamental aspects, though it is not necessary to be human-like in *all* aspects.

2 AGI's Heritage from AI

Given that AGI can be seen as an attempt to return to the original goal of AI, it inherits many ideas and lessons from AI, either in a positive form (as what works) or a negative form (as what does not work). Beside the research objective, AGI also needs to critically evaluate the research paradigm of mainstream AI.

A representative specification of the mainstream AI approach is Marr's "three-level" analysis of problem solving [6]. First, the problem is specified in the *computational* framework, as a function that maps each input into the desired output; then, an *algorithmic* solution is found that uses a fixed and finite operation sequence to process each instance of the problem; finally, the algorithm takes an *implementational* form in a computer system. Though it is Marr who popularized this approach in AI and CogSci, this procedure has been followed in computer science from the very beginning, and its roots can be traced back to how problem solving is defined and carried out in mathematics.

There is no doubt that this methodology of problem solving has played an important role in the successes achieved by computer science and mathematics. However, as we have argued elsewhere [15] it has serious limitations when applied as currently done to AI and CogSci. Notions like "computation," "Turing machine," "algorithm," and so on treat problem solving as a deterministically or probabilistically *repeatable process*, i.e., for a given problem instance, its solving process, result, and resources cost (mostly computer time and space) are all fixed, independent of the past experience of the system and the current environment where the problem appears. In contrast, problem solving processes in the human mind are very different: They are not always accurately repeatable, and the exact same problem instance can be processed differently when it appears in a different time and context.

The above statement should not be used to deny the possibility of AGI, but suggests the need for a paradigm shift. Instead of treating a problem-solving process as following a fixed algorithm, it is more proper to see the process as consisting of many small steps that are dynamically linked together at runtime in a context-sensitive manner [14, 13, 20]. In this way, though each basic step follows a predetermined algorithm, the overall problem-solving process does not, because the selection of the steps at each moment depends on many factors in the external environment and within the system, and these factors are ever-changing, so that their exact combination rarely or never repeats. Even if we were to extend the meaning of "algorithm" to include such processes, we would have to say that all the problem-level "algorithms" are *one-time*, i.e., even the same problem instance is handled by different algorithms when it reoccurs. While it is not impossible that in the future human intelligence may ultimately be found to

implement a small finite set of precise algorithms (we don't necessarily think so, but this is not inconceivable), an intelligent system is still a *system*, and a system is not an algorithm. No matter which terminology we use, a problem-solving process implemented by natural intelligence is no longer accurately repeatable, so it cannot fruitfully be analyzed according to the traditional theories of computability and computational complexity.

Since mainstream AI is so dependent on the computational paradigm, it has not successfully addressed some fundamental features of human intelligence, some obvious ones of which include general adaptivity, creativity, flexibility, and robustness. Even though many AI programs have exceeded humans in terms of speed, capacity, accuracy, complexity, etc., even laypeople still intuitively see a fundamental difference between such systems and human thinking, which explains why whenever a new task is solved in this way, whether a computer winning the world champion at chess or beating the best human in a quiz, people "demote" the task to a status of being one that doesn't in fact require intelligent action.[5] If intelligence is associated with non-algorithmic problem-solving, then AGI should not be considered to sit squarely within computer science, though it is implemented using some of the tools and techniques provided by the latter.

"Intelligence demands adaptation and learning" is not a new idea at all. However, the thinking that intelligence is of a different nature – or requires a different paradigm – than that offered by a pure algorithmic view is not common: Even the study of machine learning has been dominated by the building of "learning algorithms" that treat a learning process as following a repeatable procedure, independent of the history and situation of the system. Such thinking can never lead to the kind of continuous, ever-expanding learning that seems so critical to human cognitive development. Although in recent years some researchers have started to study topics like "online learning," "transfer learning," "one-shot learning," and "lifelong learning," few people have realized that as soon as learning is modeled as a process that does not follow any fixed algorithm, depending instead on the system's past history and current situation, all those features will come together to form a consistent, co-dependent whole [18]. We contend that, given a goal of achieving AGI, it is in fact much better to consider these features of cognitive function as being part of, and forming, a single system, and that separating them in fact makes the challenge of AGI more difficult.

[5] This has been termed "the AI effect" by some, in what seems to us an attempt to explain this uncomfortable effect away. The implicit assumption here seems to be that, in due time, when AI has covered the full range of cognitive skills, people will come to view even natural intelligence as "merely algorithmic." Our view on this is of course that the explanation is in fact quite different, depending instead on the misapplication of the algorithmic view to the study of intelligence.

3 AGI's Heritage from CogSci

From an AGI perspective, if we blame AI for being too far from the reality of human intelligence and cognition, then we should also blame CogSci for being too close to the details of human intelligence and cognition.

As already mentioned, mainstream CogSci has turned away from the goal of establishing theories of cognition in general, ones that cover humans, computers, and other cognitive systems alike, and focused primarily on describing and explaning *human* cognition. Consequently, its results, though abundant and valuable for other purposes, become less relevant to AGI for the following reasons:

Lack of generalization. CogSci does not *generalize* its findings from human cognition to a more abstract level, to render the conclusions relevant to non-human or non-biological systems. Instead, the cognitive models usually attempt to be as faithful to the details of human cognition as possible, without the separation between the aspects that are universal in all cognitive systems and the aspects that are specific in human cognition. Such a separation is, however, crucial for AGI, since it is neither desired nor possible for a computer system to duplicate certain human-specific properties.

Lack of justification and predictive power. CogSci is increasingly becoming like cognitive psychology: A natural science whose theories are mostly *descriptive* by nature. That is, when studying a cognitive function, the expected explanation depends on the identification of the responsible psychological and biological mechanisms and processes, while what AGI needs are *functional* explanations that see cognitive mechanisms as contextualized processes, selected by evolution to serve certain functions for the individual or species. We do not want to reproduce a process in computer merely because it is observed in human cognition, unless it can be justified as *normative*. Also, some features of human cognition may simply be historical peculiarities of no relevance, interest, or usefulness when studying intelligence in general.

Generally speaking, all approaches of AGI are more or less inspired by the best known form of intelligence – human intelligence, though they are based on descriptions of the human mind/brain complex that differ in level, scope, and granularity. The real problem is not *whether* to be "human-inspired" (which is different from "biologically inspired"), but where the *similarity* lies [19]. Since it is neither necessary nor possible for an AGI to be identical to the human mind/brain in all aspects (for instance biochemistry), pure descriptive theories of the human mind are not of much help to AGI. What AGI needs from CogSci is not merely "This is what the human does ...", but "This is what all cognitive/intelligent systems should do, because ...".

The above issues do not only delay the progress of AGI, but also have implications for CogSci itself. As we have discussed in [22], improperly applied normative models in CogSci are fairly common. Models like mathematical logic and probability theory only specify rational inferences in highly idealized situations, where the situation is fundamentally different from the normal environ-

ment in which human cognition happens. Therefore, some systematic deviations from these models observed in human cognition should not be simply judged as fallacy or bias [25, 17]. Instead, these observations call for the building of new normative models whose fundamental assumptions are closer to the reality of cognitive systems, both human and computer.

4 Implications for AGI

For AGI to reach its goal successfully over the coming decades – the same as the original goal of AI and CogSci but which they have since in major respects abandoned – it is important to clarify the relations among three central notions: *Human Intelligence* (HI), *Computer Intelligence* (CI), and *(General) Intelligence* (GI) [23]. Here GI should be described in a medium-independent way, with HI and CI as its special realizations. In this context, "intelligence" can be replaced by "cognition," "mind," or "thinking," without too much difference. After all, we still uphold the intuitive idea that intelligence is information processing, and thus many or most mental processes should be reproducable in computer systems.

Since the human mind/brain complex and a computer system are clearly very different both in their internal substance and structure and in how they interact with the environment, their commonality (the above GI) must be captured at an abstract level. A theory about GI should be descriptive for HI (i.e., it explains the observation from the human mind) and normative for CI (i.e., it guides the construction of computer systems).

With respect to this demand, mainstream AI has not been successful in identifying the principles of GI that should be realized in computers. Instead of exploring new ways to design computer systems, AI has focused on specific problem-solving capabilities, and mostly retreated back to computer science, adopting its methodology and theoretical stance wholly [15]. On the other hand, mainstream CogSci has not been successful in abstracting the principles of GI from the specifics of HI. Instead of experimenting with models of GI on computers, CogSci has seen the computer merely as a platform for modeling HI, and mostly retreated back to cognitive psychology (or even to neuropsychology). It makes for an interesting excercise to look further back in time, to the prime years of cybernetics [26]. Cybernetics was in many ways the overarching field from which CogSci and AI descended. Why AI and CogSci became separated from cybernetics is an interesting question, as is the question whether this separation was beneficial or detrimental to the pursuit of AGI (we lean towards the latter). Surprisingly, perhaps due to its emphasis on *systems*, many of the ideas from the "last days" of cybernetics produced by its second-generation scholars seem more relevant to AGI today than recent results in AI and CogSci circles.

Though AGI can surely learn a lot from AI and CogSci (and cybernetics), we should carefully analyze the heritages we received from these two fields, and try our best to prevent ourselves from repeating the past mistakes.

Acknowledgment

Some of the opinions in this paper were triggered by the authors' previous discussions with Joscha Bach. Comments from Hongbin Wang on an early version of the paper were also helpful. This was in part supported by an FP7 grant from the European Union (HUMANOBS: Humanoids That Learn Socio-Communicative Skills by Observation, contract no. FP7-STReP-231453 – www.humanobs.org), as well as Nascence (FP7-ICT-317662), SNF grant #200020-138219, and by a Centres of Excellence project grant (www.iiim.is) from the Icelandic Council for Science and Technology Policy.

References

1. Boden, M.A.: Mind As Machine: A History of Cognitive Science. Oxford University Press, Oxford (2006)
2. Cassimatis, N.L.: Artificial intelligence and cognitive science have the same problem. In: Papers from the AAAI Spring Symposium on Between a Rock and a Hard Place: Cognitive Science Principles Meet AI-Hard Problems. pp. 27–32 (2006)
3. Goertzel, B., Pennachin, C. (eds.): Artificial General Intelligence. Springer, New York (2007)
4. Goldstone, R.: Returning to a new home. Cognitive Science 29, 1–4 (2005)
5. Hayes, P., Ford, K.: Turing Test considered harmful. In: Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence. pp. 972–977 (1995)
6. Marr, D.: Vision: A Computational Investigation into the Human Representation and Processing of Visual Information. W. H. Freeman & Co., San Francisco (1982)
7. McCulloch, W.S., Pitts, W.H.: A logical calculus of ideas immanent in neural activity. Bulletin of Mathematical Biophysics 5, 115–133 (1943)
8. Minsky, M.: The Society of Mind. Simon and Schuster, New York (1985)
9. von Neumann, J.: The Computer and the Brain. Yale University Press, New Haven, CT (1958)
10. Newell, A.: Unified Theories of Cognition. Harvard University Press, Cambridge, Massachusetts (1990)
11. Nilsson, N.J.: The Quest for Artificial Intelligence: A History of Ideas and Achievements. Cambridge University Press, Cambridge (2009)
12. Russell, S., Norvig, P.: Artificial Intelligence: A Modern Approach. Prentice Hall, Upper Saddle River, New Jersey, 3rd edn. (2010)
13. Thórisson, K.R., Nivel, E.: Achieving artificial general intelligence through peewee granularity. In: The Proceedings of the Second Conference on Artificial General Intelligence. pp. 222–223 (2009)
14. Thórisson, K.R.: A New Constructivist AI: From Manual Construction to Self-Constructive Systems. In: Wang, P., Goertzel, B. (eds.) Theoretical Foundations of Artificial General Intelligence. Atlantis Thinking Machines, 4:145–171 (2012)
15. Thórisson, K.R.: Reductio ad Absurdum: On Oversimplification in Computer Science and its Pernicious Effect on Artificial Intelligence Research. In: Abdel-Fattah, A.H.M., Kühnberger K.-U. (eds.) Proc. of the workshop Formalizing Mechanisms for Artificial General Intelligence & Cognition (Formal MAGIC), pp. 31–35 (2013)
16. Turing, A.M.: Computing machinery and intelligence. Mind LIX, 433–460 (1950)
17. Tversky, A., Kahneman, D.: Judgment under uncertainty: heuristics and biases. Science 185, 1124–1131 (1974)

18. Wang, P.: Rigid Flexibility: The Logic of Intelligence. Springer, Dordrecht (2006)
19. Wang, P.: What do you mean by 'AI'. In: Proceedings of the First Conference on Artificial General Intelligence. pp. 362–373 (2008)
20. Wang, P.: Case-by-case problem solving. In: Proceedings of the Second Conference on Artificial General Intelligence. pp. 180–185 (2009)
21. Wang, P.: A General Theory of Intelligence (2010), an on-line book under development. URL: *http://sites.google.com/site/narswang/EBook*
22. Wang, P.: The assumptions on knowledge and resources in models of rationality. International Journal of Machine Consciousness 3(1), 193–218 (2011)
23. Wang, P.: Theories of artificial intelligence – Meta-theoretical considerations. In: Wang, P., Goertzel, B. (eds.) Theoretical Foundations of Artificial General Intelligence, pp. 305–323. Atlantis Press, Paris (2012)
24. Wang, P., Goertzel, B.: Introduction: Aspects of artificial general intelligence. In: Goertzel, B., Wang, P. (eds.) Advance of Artificial General Intelligence, pp. 1–16. IOS Press, Amsterdam (2007)
25. Wason, P.C., Johnson-Laird, P.N.: Psychology of Reasoning: Structure and Content. Harvard University Press, Cambridge, Massachusetts (1972)
26. Wiener, N.: Cybernetics, or control and communication in the animal and the machine. John Wiley & Sons, Inc., New York (1948)