Advances in Cognitive Systems 1 (2012) 37–45 Submitted 1/2012; published 7/2012

Human-Level Artificial Intelligence Must Be an Extraordinary Science

Nicholas L. Cassimatis CASSIN@RPI.EDU
Department of Cognitive Science, Rensselaer Polytechnic Institute, Troy, NY 12180 USA

Abstract

Aiming to create a cognitive system with human-level intelligence is as different from normal scientific objectives as reaching artificial immortality is to the goals of modern medicine. Most researchers in artificial intelligence, along with the institutions that support them, advocate that AI research should conform to normal scientific standards and methods. I argue that these are often incidental and even antithetical to achieving human-level intelligence and that a different approach is required. In this essay, I propose some principles on which to base such an approach.

1. Why Human-Level Artificial Intelligence?

Developing cognitive systems with human-level intelligence would have tremendous scientific and technological ramifications. However, this objective is significantly different along many dimensions from the objectives of most fields in science and technology. Unless this is recognized, and unless methods and institutions in these fields are changed significantly, we are unlikely achieve the benefits of human-level artificial intelligence in our lifetimes.

For the purposes of this paper, we will say that a system has human-level intelligence if it can accomplish the same cognitive tasks as a human. For example, a system with human-level intelligence would be able to have human language conversations as well as people and be able to identify objects and interpret visual scenes as well as people. Deciding specifically what counts as a cognitive act and what it means to perform it at the same level of a human is not important for this paper, since the field is far from the point at which such fine distinctions will be important.

Understanding and creating human-level intelligence are tremendously important scientific and technological objectives. The key obstacle to achieving these, which I call the *intelligence puzzle*, is to understand how a system composed of unintelligent parts (such as neurons or transistors) can behave as intelligently as people. Without understanding this, we cannot explain how intelligence emerges from natural scientific laws and thus the project of natural science is incomplete. This is also a serious problem because so many important areas of study (e.g., economies, culture, organizational behavior) concern individual humans and how they interact, and thus involve the exercise of human intelligence. In this paper, I will refer to the endeavor to create systems that exhibit human-level intelligence as "human-level artificial intelligence" ("HLAI") and the fields interested in creating intelligence at any level (such as artificial intelligence, machine learning, and computational linguistics) as "artificial intelligence" or "AI".

The most convincing evidence that the intelligence puzzle has been solved would be to build a system with human-level intelligence. The technological implications of this would be at least

as great as the technical. As one example, consider the implications to scientific and technological progress of having millions of minds at least as intelligent as Newton, Darwin, Turing and Edison working on scientific and technological problems. Given that computers are so quick and inexpensive to replicate (compared to high-caliber human minds), this would be a direct benefit of having achieved human-level artificial intelligence.

Where do things stand now relative to these goals? There are many traditional signs of success. AI research has resulted in many extremely lucrative and important applications. There are subfields with standard questions and methods for answering them, frequent research results, respected textbooks, and many professorships. On the other hand, when compared to the definition of HLAI herein and the initial ambition of the field's founders, there remains an enormous gulf between the abilities of current AI technology and the powers of human intelligence.

How do we explain this gap? If we want to see these goals achieved in our lifetime, should we simply rely on normal scientific progress arising from the current ways of the field, or is something else needed? If so, what form would it take? This paper proposes answers to these questions.

2. Human-Level Artificial Intelligence is Not a Normal Science

It is important to fully recognize that the field of HLAI has goals that are very different from the objectives of other fields, in that they are more concrete and much more ambitious. To illustrate, consider a common definition of medicine as the "the science or practice of the diagnosis, treatment, and prevention of disease." This level of ambition and specificity would bring medicine closer to HLAI if it were instead called "artificial immortality" and its goal was to not only detect and cure all diseases, but to also eliminate each of the causes, natural and otherwise, of death. There is a similar contrast with other branches of computer science. For example, software engineering can be defined as "a systematic approach to the analysis, design, implementation and maintenance of software". Its level of ambition and specificity would be more commensurate with HLAI if it were named something like "bug-free software development".

The fact that HLAI has more specific and ambitious goals helps explain why there is among many a sense of failure attached to AI despite its many successes. If the goal of medicine were seen as "artificial immortality", then much of the heretofore impressive advances in medicine would be seen as small, underwhelming steps towards the field's ultimate goal.

HLAI is thus not a normal[1] science. Historically, AI was not conducted as a normal science. The original founders of AI were clearly aiming to reach HLAI (Crevier, 1993) in their lifetimes. Some of the most celebrated "results" of early AI research had few of the trappings of normal science. For example, SHRLDU (Winograd, 1972) had no experimental evaluation, formal proofs, or anything else one tends to see in today's science or engineering research reports.

This all created expectations of progress that, as time passed, did not occur. For this and other[2] reasons, many came to think of AI as a failure. When people tried to explain this perceived

[1] Normal is used here in the sense of "ordinary" although there are also connections to Khun's (1970) notion of normal science.

[2] Three other factors are (1) large-scale research projects founded with grandiose objectives that were not met, (2) the AI venture capital investment bust of the 1980s, and (3) the fact that, once a process is well understood, it appears to be less intelligent. Electronic digital computers are an example of this last point. Turing explicitly (1948) saw his project as enabling machines to have the abilities of human intelligence.

failure, they often pointed to aspects of AI research that made it different from "normal" sciences. For example, people criticized the fact that it was difficult to formally characterize most AI systems or that there were no empirical results that showed that purported advances were actual advances. Having lost faith in the ways of early AI research, and there being no commonly accepted important set of questions and methods within HLAI (Bello, 2008), much of the field succumbed to "science envy" and adopted the trappings of normal science. Since most AI researchers were embedded within universities and computer science departments whose other components were mostly normal in their scientific approach, it was natural to apply these standards to their own research. Before long, Russell and Norvig (2009) could declare (in the most widely used AI textbook of our day): "In terms of methodology, AI has finally come firmly under the scientific method. To be accepted, hypotheses must be subjected to rigorous empirical experiments, and the results must be analyzed statistically for their importance."

However, despite adopting the standards and methods of "real science" for decades, AI is still far from achieving the original HLAI goals. If we want to see HLAI, can we simply rely on the pace of progress afforded by normal scientific[3] practice, or do we need to adopt new methods that are appropriate to the level of ambition inherent to HLAI?

In the remainder of this essay, I will first explain how the current standards in AI and related sciences are not well suited to achieving HLAI and are often regressive distractions. I will then suggest some guidelines for focusing research on the goal of human-level AI so we have a good chance of achieving it within the next few decades.

3. Normal Scientific Methods Fail Human-Level Artificial Intelligence

Modern fields that study intelligence have two problems with respect to achieving HLAI. First, the methods they use to evaluate and incentivize research are not sufficiently focused or strong enough to direct progress towards the goal of human-level intelligence. These methods are proper to normal scientific goals, but, as I explained in the last section, human-level AI is not a normal scientific goal. Second, requiring that researchers adhere to the disciplinary standards of one of these fields substantially retards the progress of researchers, even if they are earnestly dedicated to achieving human-level AI.

3.1 Formal or Empirical Demonstrations of Correctness or Optimality

Many AI researchers believe it is important to prove a theorem or to empirically demonstrate that a particular method is (approximately) optimal according to some standard. An example includes proving that a particular algorithm is guaranteed to find the correct inference given some input according to the axioms of probability. When applied to the HLAI endeavor, this presumes that human-level intelligence is optimal in the senses being studies. However, human intelligence is not correct or optimal in many situations (Simon, 1957).

In fact, it may be the case that incorrectness and non-optimality are essential to human-level intelligence. Any formalization of the problems humans solve that is anywhere close to being

In his day, "computer" was a job description for a human being. The fact that all the vast ramifications of electronic computers are not counted as successes of AI illustrates that a process seems less intelligent when fully understood.

[3] Whether HLAI is a scientific and/or technological problem is not a question addressed in this paper. To simplify the writing, "science" or "scientific" is intended to encompass "science and technology" and "scientific and technological."

comprehensive implies that the human-level inference problem is at the very least NP-complete. If so, and if humans are not super-Turing computers, then it is impossible to create computational systems that make human-level inferences in human-scale times. Human departures from optimality may in fact be a manifestation of tradeoffs in human cognitive mechanisms that enable human-level intelligence. To focus research on optimality could thus direct effort away from discovering these mechanisms.

3.2 Incremental Increases in Speed

One of the greatest challenges of HLAI is the amount of time it takes to make the necessary computations. Suppose for a moment that all problems could be framed as belief propagation in probabilistic graphical models or as logic theorem proving. It thus seems reasonable to favorably evaluate research that makes computational systems faster. The problem with this approach is that, in modern AI practice, systems that make relatively incremental improvements are rewarded. Thus, in many cases, an algorithm that solves a problem twenty percent faster than the previously fastest algorithm is seen as a good result. However, as just mentioned, the human-level intelligence problem is very computationally complex. Inference over human-scale graphical models or logical systems requires computational resources far greater than the size of the known physical universe to deal with them. Literal lifetimes of incremental improvements on the order just described will not be enough to reach HLAI. Thus, the kinds of computational speedups valued in almost every other kind of computer science are nearly pointless in HLAI.

3.3 Empirical Psychology

Presumably, it would be much easier to create cognitive systems with human-level intelligence if the mechanisms of human intelligence were already known. The current fields that study human intelligence (e.g., linguistics, neuroscience and cognitive psychology), however, all have their own disciplinary standards that are not aligned with human-level intelligence.

Consider first empirical cognitive psychology. In this field, one generally proposes theories, or even precise computational or mathematical models, and tests them against human data. Examples of such data include reaction times or error rates obtained in behavioral experiments. However, there is nothing in the field that forces attention on the problems of human-level intelligence. One can, for example, study the relative effects of recency and frequency on the accuracy of memory recall regardless of whether this question has any real import to solving the intelligence puzzle. While the answers to some questions are considered more important than others, solving the intelligence puzzle is at best one of many factors, and not the most important, in deciding which questions are important. Relying on such standards to lead to a resolution of the intelligence puzzle would be like the Wright Brothers having relied on ornithological studies of the color of birds and the frequency of their droppings to achieve flight.

Not only does the drive to empirically confirm precise predictions offer weak focus on the intelligence puzzle, it can distract from its solution. It may be fundamentally impossible to precisely predict behavior that results from higher-order cognition. The impossibility of such precise predictions is common to many sciences and does not mean they are failures. For example, the inability to predict temperature and precipitation precisely beyond a few days or to predict the precise sequence and timing of newly evolving species does not imply that meteorology and evolutionary science have failed. It can be impossible to predict the behavior of even deterministic systems when their behavior is "sensitive to initial conditions" and those conditions are

impossible to know. For example, slight changes in initial atmospheric conditions, most famously caused by a butterfly flapping its wings (Lorenz, 1963), can lead to significantly different weather behavior after even modest intervals.

The inability to precisely determine the initial conditions of many complex systems (e.g., the weather or large economies) is one reason their behavior can be so difficult to predict. The initial conditions that are relevant to much human intelligence are numerous and they vary considerably among people, which suggests that it may be infeasible to generate precise predictions about much complex cognitive behavior. For example, to predict whether a person will find an analogy (Duncker, 1926; Holyoak & Thagard, 1997) between infantry attacking a fortress and approaches to killing tumors, one must have details about his military knowledge, what films he has seen, his knowledge of history, and many other factors. Unfortunately, these all vary greatly even within relatively homogenous subject pools (e.g., students in the same university).

The response to this within experimental psychology, as in other sciences, is to not attempt to make predictions about individual events, but to predict averages over those events. In psychology, this involves holding the relevant factors constant while varying other factors across experimental trials. For instance, when a person reacts to a visual display, many factors are at play, including how much caffeine he has consumed, the quality of his eyesight, and the lighting conditions. Experimenters keep relevant visual factors constant across trials and subjects in hopes that the irrelevant factors will cancel each other in the mean.

In intelligent human reasoning, problem solving, and language understanding, however, the factors that vary across subjects (such as their knowledge and experience) are generally relevant to how they perform a task, and they are often impossible to control (at least ethically). In the analogy example, by averaging over subjects, we would therefore average not only over irrelevant factors, but over the factors most relevant to the problem at hand. To confine investigations to factors we can only reproduce in a laboratory severely restricts the investigator's ability to study the intelligence puzzle.

3.4 Neuroscience

Neuroscience is a vast field, but for the most part it conforms to the standards of normal science at least as much as experimental psychology. As a consequence, it has many of the same limitations. For example, one can study with great professional success the operation of a neurotransmitter without ever considering its relevance to resolving the intelligence puzzle. As with experimental psychology, the drive for carefully controlled experiments also precludes research on many aspects of human-level intelligence. The field may thus generate all manner of other medical and technological benefits, but nothing in its constitution directs its efforts towards human-level AI.

3.5 Formal Linguistics

Human language is perhaps the cognitive faculty that is most obviously different from those of animals and thus it is natural to study it for clues about intelligence. However, within the field of linguistics the central research problem is to create a formal system that generates a set of sentences that includes all those that people deem to be grammatical and none of those they consider to be ungrammatical. Yet the process by which these systems generate sentences is often too complex to embody in an actual computer. Although such a factor is crucially relevant to HLAI, it is not a concern of almost all formal linguists. Further, humans have the ability to

understand many ungrammatical sentences and their ability to do so separates them from current computers. Thus, it is possible to do research that is well regarded within the field, but incidental to the goals of HLAI.

3.6 The Best and the Brightest: Climbing Mountains to Reach the Moon

As we have seen, scientific disciplines studying cognition have standards and methods that are often incidental and even counter to understanding and synthesizing human-level intelligence. This explains why it is difficult to aid progress by simply directing more resources to the field, or by assembling a team of the best-regarded researchers. This approach succeeded in the Manhattan Project and Apollo program because there was already a general understanding of how to build an atomic bomb and a long-range rocket. The issues involved there mostly pertained to integrating component technologies and making incremental refinements. In AI, there are no widely-known principles for integrating research ideas that will achieve human-level AI. Without such a guiding vision, the "top" researchers in the fields of intelligence will, left to their own normal scientific devices, make at best incremental advances. They will climb the mountains of their individual fields and get a little closer to the moon, but they will never reach it.

4. Some Steps Towards Human-Level Artificial Intelligence

Since HLAI is not a normal science, it will not succeed by adopting the normal scientific AI standards of the broader AI research community. The most important reform is to recognize this fact and begin to ask for each element of research how it promotes progress towards HLAI. This needs to happen in all areas, from intellectual concerns such as the development of algorithms and system architectures to intuitional issues such as funding and evaluating research. In this section, I offer some guidelines and proposals that have resulted from such reflections. They are not a comprehensive proposal for the field's conduct, but earnestly adopting them will lead to significant progress and give us a good chance of seeing HLAI in our lifetimes.

Many of these ideas are based in part on the belief that identifying a "cognitive substrate" is an important component of understanding and replicating human-level intelligence. The key precept is that a relatively small set of mechanisms for reasoning[4] over a relatively small set of concepts underlies all of human cognition. I mention one argument for a substrate here, and refer readers to Cassimaits (2006) for additional evolutionary, psychological, neuroscientific, linguistic and computational arguments in its favor. Humans evolved in an environment that did not include most of the objects and relations that hold in our present-day environment. Parliaments, interest rates, speed boats, the spread option offense, digital computers, electronic social networks, and much more simply did not exist then. Thus, the mechanisms people use to reason about these things must have originally evolved to reason about other concepts. Since navigating the social and physical world were extremely important to humans when they evolved, we hypothesize that these concepts were social and physical. Thus, if we can solve the problems of HLAI for dealing with these basic physical and social concepts, the rest of HLAI will "merely" involve mapping other domains on to these physical and social domains. This line of reasoning has several implications for how to perform and evaluate HLAI research.

[4] I use the term "reasoning" here in an extremely general sense to encompass everything called "inference" and "problem solving". This includes "language understanding" since that can be viewed as a reasoning problem (Hobbs et al., 1993).

4.1 Quasi-Formal Natural Language Semantics

In order to understand the structure of the vast array of concepts a cognitive substrate must address, it would be useful to have worked out precise representation of this structure. While human language may not be able to express all relevant aspects of human intelligence, it can capture a great deal. Thus, by working to achieve a precise formalization of the meaning of human language, we would be performing a survey expedition for HLAI research.

The role of formalization would be different in this endeavor than in conventional formal linguistics. Questions of soundness, completeness, and expressive power would largely be incidental. The purpose of formalization would be to make sure researchers understand precisely what they are discussing and to make it easier for computational investigations to proceed. The objective of the formalism would be to capture the distinctions among elements of human knowledge representation and to express the dependencies among them.

4.2 Reductions

If a substrate exists, understanding it would focus research on the mechanism for reasoning about the concepts in the substrate. One step in this direction would be to demonstrate that the reasoning problems of one or more domains can be reduced to the reasoning problems in one other domain. Such reductions can serve as evidence for the substrate hypothesis. For example, if the task of grammatical processing can be reduced to the problem of reasoning in a superficially very different domain such as physical reasoning (Cassimatis, 2004), then it seems much more plausible that the vast diversity of domains can be reduced to a substrate. Another consequence of such reductions is that they generate ideas for good microcosms in which to perform research.

4.3 Microcosms

A single research project that lasts one or two years is most likely not going to create an HLAI. Thus, the short-term goals of such research must be more modest. This is one factor that has made it easy for AI to regress into a normal science that chooses many subgoals that are, at best, incidental to HLAI. Formal proofs and experimental results are comparatively easy to generate by the bunch every year. Thus, we need to choose goals that enable shorter-term progress while taking us towards the real goal. The notion of a microcosm can help deal with this problem (Cassimatis & Bignoli, 2011).

A microcosm is a simulated world for carrying out HLAI research. The goal is to achieve HLAI, or at least to make significant advances towards it, within that world. For example, a small baby crib world with a few objects in it could be a potential microcosm. A good microcosm would be one in which solving the same problems a human can is beyond the reach of current technology, but not so far as to be impossible to make progress on. Further, to ensure that the results of this work have as much import as possible, the microcosm would instantiate problems involved in human intelligence in a broad range of domains.

The kinds of reductions proposed in the last subsection can guide towards evaluating how the selection of microcosms for this purpose. For example, if there are reductions for a wide array of domains to physical reasoning and very few to, say, wine tasting, then, barring undiscovered wine-tasting reductions, a physical reasoning microcosm would be a better choice.

4.4 Building Individual Systems that Integrate Multiple Abilities

For microcosms to drive research, it is important that they be used correctly. I will illustrate some of the pitfalls using the "blocks worlds" microcosm as an example. Although this was the catalyst for some very important research advances, it has also has been misused by both its advocates and critics.

This domain was the focus of a substantial body of "Good Old Fashioned AI" (GOFAI) research. The shortcomings of that approach relative to initial expectations have often been blamed on the blocks world being an unrealistic and oversimplified microcosm of the real world. However, blocks worlds are not too easy. There are many problems today within the blocks world that humans can easily solve and that computers still cannot. For example, most blocks worlds systems are evaluated in situations which include no uncertainty about the identity (e.g., is the block that disappeared behind an occluder the same as the block that just appeared from behind it), no change over time, and extremely simple spatial arrangement (blocks are only on top of other blocks or on a table, not situated with respect to any realistic coordinate system). With these assumptions relaxed, blocks world reasoning is well beyond the reach of current AI systems. Blocks worlds are therefore not "too easy".

However, in accordance with the normal scientific standards mentioned in Section 2, research in blocks worlds today is, to the near exclusion of other approaches, focused on formalizations and demonstrations of optimality and on incremental increases in performance. As argued in that section, these standards are beside the point. Thus, the blocks world does not illustrate the futility of microcosms, but it does illustrate the need for different standards for evaluating research progress.

In general, the success of a system in a microcosm should be evaluated according to which situations in that microcosm it would deal with as well as a human would. It is important that these questions be asked of a single integrated system rather than a collection of systems. It is generally much easier to tailor a system to deal with a narrow class of problems. However, these optimizations often come at the cost of failure on other aspects of a problem. For example, the classification abilities of a typical statistical neural network come at the cost of the reasoning abilities of a typical search algorithm and *vice versa*. One of the main challenges of HLAI research thus is to somehow overcome these tradeoffs and create a single system that can manifest all the aspects of intelligence. It is thus important that work in microcosms be evaluated, not according to the success of a class of systems on a set of problems, but instead according to the performance of a single system that integrates the ability to deal with each of these problems.

5. Conclusions

To reiterate, the goal of HLAI is qualitatively different from the goals of other fields in science and technology. The methods of normal sciences have been increasingly prevalent within AI for some time. I have argued that they are in fact a distraction and often even regressive to progress towards HLAI. The notion of a cognitive substrate motivates a different set of standards and methods for research towards HLAI. These include "quasi-formal" research into linguistic semantics, reductions between reasoning problems in differing domains, the development of microcosms for evaluating research, and constructing systems that integrate multiple abilities. Both the goals and the benefits of human-level artificial intelligence are extraordinary. To reach them our research methods must also be extraordinary.

Acknowledgements

Pat Langley and Diana Cooper read and commented on earlier drafts of this essay. The research reported in this paper was supported by Grants N00014-09-1-1029 and N00014-09-1-0094 from the Office of Naval Research, as well as Grant FA9550-07-1-0072 from the Air Force Office of Scientific Research.

References

Bello, P. (2008). Cognitive development: Informing the design of architectures for naturally intelligent systems. *Proceedings of the AAAI 2008 Workshop on Naturally Inspired Artificial Intelligence*. Chicago, IL.

Cassimatis, N. L. (2004). Grammatical processing using the mechanisms of physical inference. *Proceedings of the Twentieth-Sixth Annual Conference of the Cognitive Science Society* (pp. 192–197). Chicago, IL.

Cassimatis, N. L. (2006). A cognitive substrate for human-level intelligence. *AI Magazine, 27*, 45–56.

Cassimatis, N. L., & Bignoli, P. (2011). Testing common sense reasoning abilities. *Journal of Theoretical and Experimental Artificial Intelligence*, *23*, 279–298.

Crevier, D. (1993). *AI: The tumultuous history of the search for artificial intelligence*. New York: Basic Books.

Duncker, K. (1926). A qualitative (experimental and theoretical) study of productive thinking (solving of comprehensible problems). *Journal of Genetic Psychology, 33*, 642–708.

Hobbs, J. R., Stickel M., Appelt, D., & Martin, P. (1993). Interpretation as abduction. *Artificial Intelligence*, *63*, 69–142.

Holyoak, K. J., & Thagard, P. (1997). The analogical mind. *American Psychologist, 52*, 35–44.

Kuhn, T. S. (1970). *The structure of scientific revolutions*. Chicago: University of Chicago Press.

Lorenz, E. N. (1963). Deterministic nonperiodic flow. *Journal of Atmospheric Sciences*, *20*, 130–141.

Russel, S., J. & Norvg, P. (2009). *Artificial intelligence: A modern approach (Third edition)*. New York: Prentice Hall.

Simon, H. A. (1957). *Models of man: Social and rational*. New York: Wiley

Turing, A. M. (1946). *Proposed electronic calculator*. National Physical Laboratory, Teddington, UK.

Winograd, T. (1972). Understanding natural language. *Cognitive Psychology*, *3*, 1–91.