Artificial Intelligence and Cognitive Science have the Same Problem

Nicholas L. Cassimatis
Department of Cognitive Science, Rensselaer Polytechnic Institute
108 Carnegie, 110 8th St.
Troy, NY 12180
cassin@rpi.edu

Abstract

Cognitive scientists attempting to explain human intelligence share a puzzle with artificial intelligence researchers aiming to create computers that exhibit human-level intelligence: how can a system composed of relatively unintelligent parts (such as neurons or transistors) behave intelligently? I argue that although cognitive science has made significant progress towards many of its goals, that solving the puzzle of intelligence requires special standards and methods in addition to those already employed in cognitive science. To promote such research, I suggest creating a subfield within cognitive science called *intelligence science* and propose some guidelines for research addressing the intelligence puzzle.

The Intelligence Problem

Cognitive scientists attempting to fully understand human cognition share a puzzle with artificial intelligence researchers aiming to create computers that exhibit human-level intelligence: how can a system composed of relatively unintelligent parts (say, neurons or transistors) behave intelligently?

Naming the problem

A few words on terminology will prevent a lot of confusion. I will call the problem of understanding how unintelligent components can combine to generate human-level intelligence the *intelligence problem*; the endeavor to understand how the human brain embodies a solution to this problem *understanding human intelligence*; and the project of making computers with human-level intelligence *human-level artificial intelligence*.

When I say that a system exhibits human-level intelligence, I mean that it can deal with the same set of situations that a human can with the same level of competence. For example, I will say a system is a human-level conversationalist to the extent that it can have the same kinds of conversations as a typical human.

Why the Intelligence Problem is Important

Why is the human-level intelligence problem important to cognitive science? The theoretical interest is that human intelligence poses a problem for a naturalistic worldview insofar as our best theories about he laws governing the behavior of the physical world posit processes that do not include creative problems solving, purposeful behavior and other features of human-level cognition. Therefore, not understanding how the relatively simple and "unintelligent" mechanisms of atoms and molecules combine to create intelligent behavior is a major challenge for a naturalistic world view (upon which much cognitive science is based). Perhaps it is the last major challenge. Surmounting the human-level intelligence problem also has enormous technological benefits which are obvious enough.

The State of the Science

For these reasons, understanding how the human brain embodies a solution to the human-level intelligence problem is an important goal of cognitive science. At least at first glance, we are nowhere near achieving this goal. There are no cognitive models that can, for example, fully understand language or solve problems that are simple for a young child. This paper evaluates the promise of applying existing methods and standards in cognitive science to solve this problem and ultimately proposes establishing a new subfield within cognitive science, which I will call *Intelligence Science*[1], and to outline some guiding principles for that field.

Before discussing how effective the methods and standards of cognitive science are in solving the intelligence problem, it is helpful to list some of the problems or questions intelligence science must answer. The elements of this list – they are not original to this paper – are by no means exhaustive or even the most important. They are merely illustrative examples:

Qualification problem. How does the mind retrieve or infer in so short a time the exceptions to its knowledge? For example, a hill symbol on a map means there is a hill in the corresponding location in the real world except if: the mapmaker was deceptive, the hill was leveled during real estate development after the map was made, the map is of shifting sand dunes. Even the exceptions have exceptions. The sand dunes could be part of a historical site and be carefully preserved or the map could be based on constantly updated satellite images. In these exceptions to the exceptions, a hill symbol does mean there is a hill there now. It is impossible to have foreseen or been taught all these exceptions in advance, yet we recognize them as exceptions almost instantly.

Relevance problem. Of the enormous amount of knowledge people have, how do they manage to retrieve

[1] Obviously, for lack of a better name.

the relevant aspects of it, often in less than a second, to sort from many of the possible interpretations of a verbal utterance or perceived set of events?

Integration problem. How does the mind solve problems that require, say, probabilistic, memory-based and logical inferences when the best current models of each form of inference are based on such different computational methods?

Is it merely a matter of time before cognitive science as it is currently practiced answers questions like these or will it require new methods and standards to achieve the intelligence problem?

Existing Methods and Standards are not Sufficient to Solve the Intelligence Problem

Historically, AI and cognitive science were driven in part by the goal of understanding and engineering human-level intelligence. There are many reasons goals in cognitive science and, although momentous for several reasons, human-level intelligence is just one of them. Some other goals are to generate models or theories that predict and explain empirical data, to develop formal theories to predict human grammatically judgments and to associate certain kinds of cognitive processes with brain regions. Methods used today in cognitive science are very successful at achieving these goals and show every indication of continuing to do so. In this paper, I argue that these methods are not adequate to the task of understanding human-level intelligence.

Put another way, it is possible to do fantastic research by the current standards and goals of cognitive science and still not make much progress towards understanding human intelligence.

Just to underline the point, the goal of this paper is not to argue that "cognitive science is on the wrong track", but that despite great overall success on many of its goals, progress towards one of its goals, understanding human-level intelligence, requires methodological innovation.

Formal linguistics

The goal of many formal grammarians is to create a formal theory that predicts whether a given set of sentences is judged by people to be grammatical or not. Within this framework, whether elements of the theory correspond to a mechanism humans use to understand language is generally not a major issue. For example, at various times during the development of Chomsky and his students' formal syntax, their grammar generated enormous numbers of syntactic trees and relied on grammatical principles to rule out ungrammatical trees. These researchers never considered it very relevant to criticize their framework by arguing that it was implausible to suppose that humans could generate and sort through this many trees in the second or two it takes them to understand most sentences. That was the province of what they call "performance" (the mechanisms the mind uses) not competence (what the mind, in some sense, knows, independent of how it uses this knowledge). It is possible therefore to do great linguistics without addressing the computational problems (e.g. the relevance problem from the last section) involved in human-level language use.

Neuroscience

The field of neuroscience is so vast that it is difficult to even pretend to discuss it in total. I will confine my remarks to the two most relevant subfields of neuroscience. First, "cognitive neuroscience" is probably the subfield that most closely addresses mechanisms relevant to understanding human intelligence. What often counts as a result in this field is a demonstration that certain regions of the brain are active during certain forms of cognition. A simplistic, but not wholly inaccurate way of describing how this methodology would apply to understanding intelligence would be to say that the field is more concerned with what parts of the brain embody a solution to the intelligence problem, not how they actually solve the problem. It is thus possible to be a highly successful cognitive neuroscientist without making progress towards solving the intelligence problem.

Computational neuroscience is concerned with explaining complex computation in terms of the interaction of less complex parts (i.e., neurons) obviously relevant to this discussion. Much of what I say about cognitive modeling below also applies to computational neuroscience.

Artificial intelligence

An important point of this paper is that cognitive science's attempt to solve the intelligence problem is also an AI project and in later sections I will describe how this has and can still help cognitive science. There are, however, some ways AI practice can distract from that aim, too. Much AI research has been driven in part by at least one of these two goals.

A formal or empirical demonstration that an algorithm is consistent with, approximates or converges on some normative standard. Examples include proving that a Bayes network belief propagation algorithm converges on a probability distribution dictated by probability theory or proving that a theorem prover is sound and complete with respect to a semantics for some logic. Although there are many theoretical and practical reasons for seeking these results (I would like nuclear power plant software to be correct as much as anyone), they do not necessarily constitute progress towards solving the intelligence problem. For example, establishing that a Bayes Network belief propagation algorithm converges relatively quickly towards a normatively correct probability distribution given observed states of the world does not in any way indicate that solving such problems is part of human-level intelligence, nor is there any professional incentive or standard requiring researchers to argue for this. There is in fact extensive evidence that humans are not

normatively correct reasoners. It may even be that some flaws in human reasoning are a tradeoff required of any computational system that solves the problems humans do.

Demonstrating with respect to some metric that an algorithm or system is faster, consumes fewer resources and/or is more accurate than some alternative(s). As with proving theorems, one can derive great professional mileage creating a more accurate part of speech tagger or faster STRIPS planner without needing to demonstrate in any way that their solution is consistent with or contributes to the goal of achieving human-level intelligence.

Experimental psychology

Cognitive psychologists generally develop theories about how some cognitive process operates and run experiments to confirm these theories. There is nothing specifically in this methodology that focuses the field on solving the intelligence problem. The field's standards mainly regard the accuracy and precision of theories, not the level of intelligence they help explain. A set of experiments discovering and explaining a surprising new phenomenon in (mammalian-level) place memory in humans will typically receive more plaudits than another humdrum experiment in high-level human reasoning. To the extent that the goal of the field is solely to find accurate theories of cognitive processes, this makes sense. But it also illustrates the lack of an impetuous towards understanding human-level intelligence. In addition to this point, many of Newell's (Newell, 1973) themes apply to the project of understanding human-level intelligence with experimental psychology alone and will not be repeated here.

A subfield of cognitive psychology, cognitive modeling, does, at its best, avoid many of the mistakes Newell cautions against and I believe understanding human cognition is ultimately a cognitive modeling problem. I will therefore address cognitive modeling extensively in the rest of this paper.

Cognitive Modeling and the Intelligence Problem: The Model Fit Imperative

Cognitive modeling is indispensable to the project of understanding human-level intelligence. Ultimately, you cannot say for sure that you have understood how the human brain embodies a solution to the intelligence problem unless you have 1. a computational model that behaves as intelligently as a human and 2. some way of knowing that the mechanisms of that model, or at least its behavior, reflect what is going on in humans. Creating computer models to behave like humans and showing that the model's mechanisms at some level correspond to mechanism underlying human cognition is a big part of what most cognitive modelers aim to do today. Understanding how the human brain embodies a solution to the intelligence problem is thus in part a cognitive modeling problem.

This section describes why I think some of the practices and standards of the cognitive modeling community, while being well-suited for understanding many aspects of cognition, are not sufficient to, and sometimes even impede progress towards, understanding human-level intelligence.

The main approach to modeling today is to create a model of human cognition in a task that fits existing data regarding their behavior in that task and, ideally, predicts behavior in other versions of the task or other tasks altogether. When a single model with a few parameters predicts behavior in many variations of a task or in many different tasks, that is good evidence that the mechanisms posited by the model correspond, at least approximately, to actual mechanisms of human cognition. I will call the drive to do this kind of work the *model fit imperative*.

What this approach does not guarantee is that the mechanisms uncovered are important to understanding human-level intelligence. Nor does it do impel researchers to find important problems or mechanisms that have not yet been addressed, but which are key to understanding human-level intelligence.

An analogy with understanding and synthesizing flight will illustrate these points[1]. Let us call the project of understanding birds *aviary science*; the project of creating computational models of birds *aviary modeling* and the project of making machines that fly *artificial flight*. We call the problem of how a system that is composed of parts that individually succumb to gravity can combine to defy gravity the *flight problem*; and we call the project of understanding how birds embody a solution to this problem *understanding bird flight*.

You can clearly do great aviary science, i.e., work that advances the understanding of birds, without addressing the flight problem. You can create predictive models of bird mating patterns that can tell you something about how birds are constructed, but they will tell you nothing about how birds manage to fly. You can create models that predict the flapping rate of a bird's wings and how that varies with the bird's velocity, its mass, etc. While this work studies something related to bird flight, it does not give you any idea of how birds actually manage to fly. Thus, just because aviary science and aviary modeling are good at understanding many aspects of birds, it does not mean they are anywhere near understanding bird flight. If the only standard of their field is to develop predictive models of bird behavior, they can operate with great success without ever understanding how birds solve the flight problem and manage to fly.

I suggest that the model fit imperative in cognitive modeling alone is about as likely to lead to an understanding of human intelligence as it would be likely to drive aviary science towards understanding how birds fly. It is possible to collect data about human cognition, build fine models that fit the data and accurately predict new

[1] I have been told that David Marr has also made an analogy between cognitive science and aeronautics, but I have been unable to find the reference.

observations – it is possible to do all this without actually helping to understand human intelligence. Two examples of what I consider the best cognitive modeling I know of illustrate this point. (Lewis & Vasishth, 2005) have developed a great model of some mechanisms involved in sentence understanding, but this and a dozen more fine pieces of cognitive modeling could be done and we would still not have a much better idea of how people actually mange to solve all of the inferential problems in having a conversation, how they sort from among all the various interpretations of a sentence, how they manage to fill in information not literally appearing in a sentence to understand the speaker's intent. Likewise, Anderson's (Anderson, 2005) work modeling brain activity during algebraic problem solving is a big advance in confirming that specific mechanisms in ACT-R models of cognition actually reflect real, identifiable, brain mechanisms. But, as Anderson himself claimed[1], these models only shed light on behavior where there is a preordained set of steps to take, not where people actually have to intelligently figure out a solution to the problem on their own.

The point of these examples is not that they are failures. These projects are great successes. They actually achieved the goals of the researchers involved and the cognitive modeling community. That they did so without greatly advancing the project of understanding human intelligence is the point. The model fit imperative is geared towards understanding cognition, but not specifically towards making sure that human-level intelligence is part of the cognition we understand. To put the matter more concretely, there is nothing about the model fit imperative that forces, say, someone making a cognitive model of memory to figure out how their model explains how humans solve the qualification and relevance problems. When one's goal is to confirm that a model of a cognitive process actually reflects how the mind implements that process, the model fit imperative can be very useful. When one has the additional goal of explaining human-level intelligence, then some additional standard is necessary to show that this model is powerful enough to explain human-level performance.

Further, I suggest that the model fit imperative can actually impeded progress towards understanding human intelligence. Extending the analogy with the fight problem will help illustrate this point. Let us say the Wright Brothers decided for whatever reason to subject themselves to the standards of our hypothetical aviary modeling community. Their initial plane at Kitty Hawk was not based on detailed data on bird flight and made no predictions about it. Not only could their plane not predict bird wing flapping frequencies, its wings did not flap at all. Thus, while perhaps a technological marvel, their plane was not much of an achievement by the aviary modeling community's model fit imperative. If they and the rest of that community had instead decided to measure bird wing flapping rates and create a plane whose wings flapped, they may have gone through a multi-decade diversion into understanding all the factors that contribute to wing flapping rates (not to mention the engineering challenge of making plane whose wings flaps) before they got back to the nub of the problem, to discover the aerodynamic principles and control structures that can enable flight and thereby solve the flight problem. The Wright Flyer demonstrated that these principles were enough to generate flight. Without it, we would not be confident that what we know about bird flight is enough to fully explain how they fly. Thus, by adhering to the model fit imperative, aviary science would have taken a lot longer to solve the flight problem in birds.

I suggest that, just as it would in aviary science, the model fit imperative can retard progress towards understanding how the human brain embodies a solution to the intelligence problem. There are several reasons for this, which an example will illustrate. Imagine that someone has created a system that was able to have productive conversations about, say, managing one's schedule. The system incorporates new information and answer questions as good as a human assistant can. When it is uncertain about a statement or question it can engage in a dialog to correct the situation. Such a system would be a tremendous advance in solving the intelligence problem. The researchers who designed it would have had to find a way, which has so far eluded cognitive science and AI researchers, to integrate multiple forms of information (acoustic, syntactic, semantic, social, etc.) within milliseconds to sort through the many ambiguous and incomplete utterance people make. Of the millions of pieces of knowledge about this task, about the conversants and about whatever the conversants could refer to, the system must find just the right knowledge, again, within a fraction of a second. No AI researchers have to this point been able to solve these problems. Cognitive scientists have not determined how people solve these problems in actual conversation. Thus, this work is very likely to contain some new, very powerful ideas that would help AI and cognitive science greatly.

Would we seriously tell these researchers that their work is not progress towards understanding the mind because their system's reaction times or error rates (for example) do not quite match up with those of people in such conversations? If so, and these researchers for some reason wanted our approval, what would it have meant for their research? Would they have for each component of their model run experiments to collect data about that component and calibrate the component to that data? What if their system had dozens of components, would they have had to spend years running these studies? If so, how would they have had the confidence that the set of components they were studying was important to human-level conversation and that they were not leaving out components whose importance they did not initially anticipate? Thus, the data fit model of research would either have forced these researchers to go down a long experimental path that they had little confidence would

[1] In a talk at RPI.

address the right issues or they would have had to postpone announcing, getting credit for and disseminating to the community the ideas underlying their system.

For all these reasons, I conclude that he model fit imperative in cognitive modeling does not adequately drive the field towards achieving an understanding of human intelligence and that it can even potentially impede progress towards that goal.

Does all this mean that cognitive science is somehow exceptional, that in every other part of science, the notion of creating a model, fitting it to known data and accurately predicting new observations does not apply to understanding human-level intelligence?

Not at all. There are different levels of detail and granularity in data. Most cognitive modeling involves tasks were there is more than one possible computer program known that can perform in that task. For example, the problem of solving algebraic equations can be achieved by many kinds of computer programs (e.g., Mathematica and production systems). The task in that community is to see which program the brain uses and to select a program that exhibits the same reaction times and error rates as humans is a good way to go about this. However, in the case of human-level intelligence, *there are no known programs that exhibit human-level intelligence*. Thus, before we can get to the level of detail of traditional cognitive modeling, that is, before we can worry about fitting data at the reaction time and error rate level of detail, we need to explain and predict the most fundamental datum: people are intelligent. Once we have a model that explains this, we can fit the next level of detail and know that the mechanisms whose existence we are confirming are powerful enough to explain human intelligence.

Creating a models that predict that people are intelligent means writing computer programs that behave intelligently. This is also a goal of artificial intelligence. Understanding human intelligence is therefore a kind of AI problem.

Artificial Intelligence and Cognitive Science Can Help Each Other on the Intelligence Problem

I have so far argued that existing standards and practices in the cognitive sciences do not adequately drive the field towards understanding human intelligence. The main problems are that (1) each field's standards make it possible to reward work that is not highly relevant to understanding human intelligence; (2) there is nothing in these standards to encourage researchers to discover each field's gaps in its explanation of human intelligence and (3) that these standards can actually make it difficult for significant advances towards understanding human-intelligence to gain support and recognition. This section suggests some guidelines for cognitive science research into human intelligence.

Understanding human-intelligence should be its own subfield. Research towards understanding human intelligence needs to be its own subfield, *intelligence science*, within cognitive science. It needs its own scientific standards and funding mechanisms. This is not to say that the other cognitive sciences are not important for understanding human intelligence; they are in fact indispensable. However, it will always be easier to prove theorems, fit reaction time data, refine formal grammars or measure brain activity if solving the intelligence problem is not a major concern. Researchers in an environment where those are the principle standards will always be at a disadvantage professionally if they are also trying to solve the intelligence problem. Unless there is a field that specifically demands and rewards research that makes progress towards understanding how the brain solves the intelligence problem, it will normlaly be, at least from a professional point of view, more prudent to tackle another problem. Just as it is impossible to seriously propose a comprehensive grammatically theory without addressing verb use, we need a field where it is impossible to propose a comprehensive theory of cognition or cognitive architecture without at least addressing the qualification, relevance, integration and other problems of human-level intelligence.

Model the right data. I argued earlier that the most important datum for intelligence scientists to model is that humans are intelligent. With respect to the human-level intelligence problem, for example, to worry about whether, say, language learning follows a power or logarithmic law before actually discovering how the learning is even possible is akin to trying to model bird flap frequency before understanding how wings contribute to flight.

The goal of building a model that behaves intelligently, instead of merely modeling mechanisms such as memory and attention implicated in intelligent cognition, assures that the field addresses the hard problems involved in solving the intelligence problem. It is hard to avoid a hard problem or ignore an important mechanisms if, say, it is critical to human-level physical cognition and building a system that makes the same physical inferences that humans can is key to being published or getting a grant renewed.

A significant part of motivating and evaluating a research project in intelligence science should be its relevance for (making progress towards) answering problems such as the qualification, relevance and integration problems.

Take AI Seriously. Since there are zero candidate cognitive models that exhibit human-level intelligence, researchers in intelligence science are in the same position as AI researchers aiming for human-level AI: they are both in need of and searching for computational mechanisms that exhibit a human-level of intelligence. Further, the history of AI confirms its relevance to cognitive science. Before AI many philosophers and psychologists did not trust themselves or their colleagues to posit internal mental representations without implicitly smuggling in some form of mysticism or homunculus. On an technical level, search, neural networks, Bayesian networks, production rules, etc.

were all in part ideas developed by AI researchers but which play an important role in cognitive modeling today.

Chess-playing programs are often used as examples of how AI can succeed with brute-force methods that do not illuminate human intelligence. Note, however, that chess programs are very narrow in their functionality. They only play chess. Humans can play many forms of games and can learn to play these rather quickly. Humans can draw on skills in playing one game to play another. If the next goal after making computer programs chess masters was not to make them grandmasters, but to make them learn, play new games and transfer their knowledge to other games, brute force methods would not have been sufficient and researchers would have had to develop new ideas, many of which would probably bear on human-level intelligence.

Have a success. Many AI researchers have retreated from trying to achieve human-level AI. The lesson many have taken from this is that one should work on more tractable problems or more practical applications. This attitude is tantamount to surrendering the goal of solving the human intelligence problem in our lifetimes. The field needs a success to show that real progress is capable soon. One obstacle to such a success is that the bar, especially in AI, has been raised so high that anything short of an outright demonstration of full human-level AI is considered by many to be hype. For a merely very important advance towards human-level intelligence that has no immediate application, there is no good way to undeniably confirm that importance. We thus need metrics that push the state of the art but are at the same time realistic.

Develop realistic metrics. Developing realistic methods for measuring a system's intelligence would make it possible to confirm that the ideas underlying it are an important part of solving the intelligence problem. Such metrics would also incrase confidence in the prospects of intelligence science enabling quicker demonstrations of proress. My work on a model of physical cognition has illustrated the value of such metrics. I have so far tested this model by presenting it with sequences of partially occluded physical events that I have partly borrowed from the developmental psychology literature and have partly crafted myself. My strategy has been to continually find new classes of scenarios that require different forms of reasoning (e.g., probabilistic, logical, defeasible, etc.) and update my model so that it could reason about each class of scenarios. Using superficially simple physical reasoning problems in this way has had several properties that illustrate the value of the right metric.

Difficulty. Challenge problems should be difficult enough so that a solution to them requires a significant advance in the level of intelligence it is possible to model. Human-level intelligence in the physical cognition domain requires advances towards understanding the frame problem, defeasible reasoning and how to integrate perpetual and cognitive models based on very different algorithms and data structures.

Ease. While being difficult enough to require a real advance, challenge problem should be as simple as possible so that real progress is made while avoiding extraneous issues and tasks. One benefit of the physical cognition domain over, for example, Middle East politics is the smaller amount of required for a system to have before it can actually demonstrate intelligent reasoning.

Incremental. It should be possible to demonstrate advances towards the goal short of actually achieving it. For example, it is possible to show progress in the physical cognition domain without actually providing a complete solution by showing that an addition to the model enables and explains reasoning in a significantly wider, but still not complete, set of scenarios.

General. The extent to which a challenge problem involves issues that underlie cognition in many domains makes progress towards solving that problem more important. For example, I have shown (Cassimatis, 2004) how syntactic parsing can be mapped onto a physical reasoning problem. Thus, progress towards understanding physical cognition amounts to progress in two domains.

Conclusions

I have argued that cognitive scientists attempting to understand human intelligence can be impeded by the standards of the cognitive sciences, that understanding human intelligence will require its own subfield, intelligence science, and that much of the work in this subfield will assume many of the characteristics of good human-level AI research. I have outlined some principles for guiding intelligence science that I suggest would support and motivate words towards solving the intelligence problem and understanding how the human brain embodies a solution to the intelligence problem.

In only half a century we have made great progress towards understanding intelligence within fields that, with occasional exceptions, have not been specifically and wholly directed towards solving the intelligence problem. We have yet to see the progress that can happen when large numbers of individuals and institutions make this their overriding goal.

References

Anderson, J. R. (2005). Human symbol manipulation within an integrated cognitive architecture. *Cognitive Science*, 313-341.

Cassimatis, N. L. (2004). *Grammatical Processing Using the Mechanisms of Physical Inferences.* Paper presented at the Twentieth-Sixth Annual Conference of the Cognitive Science Society.

Lewis, R. L., & Vasishth, S. (2005). An activation-based model of sentence processing as skilled memory retrieval. *Cognitive Science, 29*, 375-419.

Newell, A. (1973). You can't play *20 questions* with nature and win. In W. G. Chase (Ed.), *Visual Information Processing*: Academic Press.