THEORETICAL FOUNDATIONS OF ARTIFICIAL GENERAL INTELLIGENCE

PEI WANG, BEN GOERTZEL (EDS.)

ATLANTIS THINKING MACHINES
SERIES EDITOR | K.-U. KÜHNBERGER

ATLANTIS
PRESS

ATLANTIS THINKING MACHINES

VOLUME 4

SERIES EDITOR: KAI-UWE KÜHNBERGER

Atlantis Thinking Machines

(ISSN: 1877-3273)

Aims and scope of the series

This series publishes books resulting from theoretical research on and reproductions of general Artificial Intelligence (AI). The book series focuses on the establishment of new theories and paradigms in AI. At the same time, the series aims at exploring multiple scientific angles and methodologies, including results from research in cognitive science, neuroscience, theoretical and experimental AI, biology and from innovative interdisciplinary methodologies.

For more information on this series and our other book series, please visit our website at:

www.atlantis-press.com/publications/books

ATLANTIS PRESS

AMSTERDAM – PARIS – BEIJING

Theoretical Foundations of Artificial General Intelligence

Pei Wang (Ed.)

Department of Computer and Information Sciences, Temple University,
1805 N. Broad Street, Philadelphia, PA 19122, USA

Ben Goertzel (Ed.)

Novamente LLC/Biomind LLC,
1405 Bernerd Place, Rockville, MD 20851, USA

ATLANTIS PRESS

AMSTERDAM – PARIS – BEIJING

Atlantis Press

8, square des Bouleaux
75019 Paris, France

For information on all Atlantis Press publications, visit our website at: *www.atlantis-press.com*

Atlantis Thinking Machines

Volume 1: Enaction, Embodiment, Evolutionary Robotics. Simulation Models for a Post-Cognitivist Science of Mind - Marieke Rohde, Ezequiel A. Di Paolo
Volume 2: Real-World Reasoning: Toward Scalable, Uncertain Spatiotemporal, Contextual and Causal Inference - Ben Goertzel, Nil Geisweiller, Lúcio Coelho, Predrag Janicic, Cassio Pennachin
Volume 3: Integration of World Knowledge for Natural Language Understanding - Ekaterina Ovchinnikova

ISBNs

Print:	978-94-91216-61-9
E-Book:	978-94-91216-62-6
ISSN:	1877-3273

Contents

Chapter 1

Introduction: What Is the Matter Here?

Pei Wang [1] and Ben Goertzel [2]

[1] *Temple University, USA*
[2] *Novamente LLC, USA*

pei.wang@temple.edu, ben@goertzel.org

This chapter provides a general introduction to the volume, giving an overview of the AGI field and the current need for exploration and clarification of its foundations, and briefly summarizing the contents of the various chapters.

1.1 The Matter of Artificial General Intelligence

Artificial General Intelligence (AGI), roughly speaking, refers to AI research and development in which "intelligence" is understood as a general-purpose capability, not restricted to any narrow collection of problems or domains, and including the ability to broadly generalize to fundamentally new areas [4]. The precise definition of AGI is part of the subject matter of the AGI field, and different theoretical approaches to AGI may embody different slants on the very concept of AGI. In practical terms, though, the various researchers in the field share a strong common intuition regarding the core concerns of AGI – and how it differs from the "narrow AI" that currently dominates the AI field.

In the earliest days of AI research, in the middle of the last century, the objective of the field was to build "thinking machines" with capacity comparable to that of the human mind [2, 6, 9]. In the decades following the founding of the AI field, various attempts arose to attack the problem of human-level artificial general intelligence, such as the General Problem Solver [7] and the Fifth Generation Computer Systems [3]. These early attempts failed to reach their original goals, and in the view of most AI researchers, failed to make dramatic conceptual or practical progress toward their goals. Based on these experiences, the mainstream of the AI community became wary of overly ambitious research, and turned toward the study of domain-specific problems and individual cognitive functions. Some researchers view this shift as positive, arguing that it brought greater rigor to the AI field – a typical comment being that "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." [8]. However, an alternate view would be that this greater focus on narrow problems and mathematical and experimental results has come at a great cost in terms of conceptual progress and practical achievement. The practical achievements of applied AI in the last decades should not be dismissed lightly, nor should be the progress made in various specialized AI algorithms. Yet, ultimately, the mounting corpus of AI theorems and experimental results about narrow domains and specific cognitive processes has not led to any kind of clear progress toward the initial goals of the AI field. AI as a whole does not show much progress toward its original goal of general-purpose systems, since the field has become highly fragmented, and it is not easy, if possible at all, to put the parts together to get a coherent system with general intelligence [1].

Outside the mainstream of AI, a small but nontrivial set of researchers has continued to pursue the perspective that intelligence should be treated as a whole. To distinguish their work from the bulk of AI work focused on highly specific problems or cognitive processes (sometimes referred to as "Narrow AI"), the phrase "Artificial General Intelligence" (AGI) has sometimes been used. There have also been related terms such as "Human-Level AI" [5]. The term AGI is meant to stress the general-purpose nature of intelligence – meaning that intelligence is a capacity that can be applied to various (though not necessarily all possible) environments to solve problems (though not necessarily being absolutely correct or optimal). Most AGI researchers believe that general-purpose intelligent systems cannot be obtained by simply bundling special-purpose intelligent systems together, but have to be designed and developed differently [11]. Though AGI projects share many problems and techniques with conventional AI projects, they are conceived, carried out, and evaluated differently. In recent years, the AGI community has significantly grown, and now has its regular conferences and publications.

1.2 The Matter of Theoretical Foundation

Like all fields of science and technology, AGI relies on a subtle interplay of theory and experiment. AGI has an engineering goal, the building of practical systems with a high level of general intelligence; and also a scientific goal, the rigorous understanding of the nature of general intelligence, and its relationship with an intelligent system's internal structures and processes, and the properties of its environment. This volume focuses on the theoretical aspect of AGI, though drawing connections between the theoretical and engineering aspects where this is useful to make the theory clearer. Even for those whose main interest is AGI engineering, AGI theory has multiple values: a good theory enables an engineer and empirical researcher to set objectives, to justify assumptions, to specify roadmaps and milestones, and to direct evaluation and comparison.

Some AGI research is founded on theoretical notions in an immediate and transparent way. Other AGI research is centered on system-building, with theory taking a back burner to building things and making them work. But every AGI project, no matter how pragmatic and empirical in nature, is ultimately based on some ideas about what intelligence is and how to realize it in artifacts. And it is valuable, as part of the process of shaping and growing an AGI project, that these ideas be clarified, justified, and organized into a coherent theory. Many times the theory associated with an AGI project is partially presented in a formal and symbolic form, to reduce the ambiguity and fuzziness in natural languages; but this is not necessarily the case, and purely verbal and conceptual theories may have value also. Some of the theories used in AGI research are inherited from other fields (such as mathematics, psychology, and computer science), and some others are specially invented for AGI. In cases where AGI theories are inherited from other fields, careful adaptations to the context of AGI are often required.

At the current stage, there is no single widely accepted theory of AGI, which is why this book uses a plural "foundations" in its title. For any AGI project, the underlying (explicit or implicit) theoretical foundation plays a crucial role, since any limitation or error in the theory will eventually show up in the project, and it is rarely possible to correct a *theoretical* mistake by a *technical* remedy. Comparing and evaluating the various competing and complementary theoretical foundations existing in the field is very important for AGI researchers, as well as for other interested individuals.

The existing AGI literature contains many discussions of AGI theory; but these are often highly technical, and they are often wrapped up together with highly specific discussions of system architecture or engineering, or particular application problems. We felt it would be valuable – especially for readers who are not intimately familiar with the AGI field – to supplement this existing literature with a book providing a broad and relatively accessible perspective on the theoretical foundations of AGI. Rather than writing a volume on our own, and hence inevitably enforcing our own individual perspectives on the field, we decided to invite a group of respected AGI researchers to write about what they considered as among the most important theoretical issues of the field, in a language that is comprehensible to readers possessing at least modest scientific background, but not necessarily expertise in the AGI field. To our delight we received many valuable contributions, which are organized in the following chapters.

These chapters cover a wide spectrum of theoretical issues in AGI research. In the following overview they are clustered into three groups: the nature of the AGI problem and the objective of AGI research, AGI design methodology and system architecture, and the crucial challenges facing AI research.

1.3 The Matter of Objective

In the broadest sense, all works in AI and AGI aim at reproducing or exceeding the general intelligence displayed by the human mind in engineered systems. However, when describing this "intelligence" using more detailed and accurate words, different researchers effectively specify different objectives for their research [10]. Due to its stress on the general and holistic nature of intelligence, the AGI field is much less fragmented than the mainstream of AI [1], with many overarching aims and recurring themes binding different AGI research programs together. But even so, substantial differences in various researchers' concrete research objectives can still be recognized.

The chapter by **Nick Cassimatis** provides a natural entry to the discussion. One key goal of AGI research, in many though not all AGI research paradigms, is to build computer models of human intelligence; and thus, in many respects, AGI is not all that different from what is called "cognitive modeling" in cognitive science. Cassimatis shows the need for an "intelligence science", as well as carefully selected challenge problems that must be solved by modeling the right data.

The chapter by **Selmer Bringsjord and John Licato** addresses the question of how to define and measure artificial general intelligence, via proposing a "psychometric" paradigm in which AGI systems are evaluated using intelligence tests originally defined for humans. Since these tests have been defined to measure the "g factor", which psychologists consider a measure of human general intelligence, in a sense this automatically places a focus on general rather than specialized intelligence.

Though human-level intelligence is a critical milestone in the development of AGI, it may be that the most feasible route to get there is via climbing a “ladder of intelligence” involving explicitly nonhuman varieties of intelligence, as suggested in the chapter by **Sam Adams and Steve Burbeck**. The authors describe some interesting capabilities of octopi, comparing them to those of human beings, and argues more broadly that each of the rungs of the ladder of intelligence should be reached before trying a higher level.

The chapter by **Marcus Hutter** represents another alternative to the “human-level AGI” objective, though this time (crudely speaking) from above rather than below. Hutter’s Universal Artificial Intelligence is a formalization of “ideal rational behavior” that leads to optimum results in a certain type of environment. This project attempt “to capture the essence of intelligence”, rather than to duplicate the messy details of the human mind. Even though such an ideal design cannot be directly implemented, it can be approximated in various ways.

1.4 The Matter of Approach

Just as important as having a clear objective for one’s AGI research, is having a workable strategy and methodology for achieving one’s goals. Here the difference between AGI and mainstream AI shows clearly: while conventional AI projects focus on domain-specific and problem-specific solutions (sometimes with the hope that they will be somehow eventually connected together to get a whole intelligence), an AGI project often starts with a blueprint of a whole system, attempting to capture intelligence as a whole. Such a blueprint is often called an “architecture”.

The chapter by **Itamar Arel** proposes a very simple architecture, consisting of a perception module and an actuation module. After all, an AGI system should be able to take proper action in each perceived situation. Both modules use certain (different) types of learning algorithm, so that the system can learn to recognize patterns in various situations, as well as to acquire proper response to each situation. Unlike in mainstream AI, here the perception module and the actuation module are designed together; and the two are designed to work together in a manner driven by reinforcement learning.

Some other researchers feel the need to introduce more modules into their architectures, following results from psychology and other disciplines. The model introduced in the chapter by **Usef Faghihi and Stan Franklin** turns certain existing theories about human cognition into a coherent design for a computer system, which has a list of desired properties. This architecture is more complicated than Arel's, which can be both an advantage and a disadvantage.

The chapter by **Ben Goertzel *et al.*** provides an integrative architecture diagram that summarizes several related cognitive architectures, and a defense of this approach to architectural and paradigmatic integration. It is argued that various AGI architectures, that seem different on the surface, are actually fundamentally conceptually compatible, and differ most dramatically in which parts of cognition they emphasize. Stress is laid on the hypothesis that the dynamics of an AGI system must possess "cognitive synergy", that is, multiple processes interacting in such a way as to actively aid each other when solving problems.

There are also researchers who do not want to design a fixed architecture for the system, but stress the importance of letting an AGI system construct and modify its architecture by itself. The chapter by **Kris Thórisson** advocates a "constructivist" approach to AI, which does not depend on human designed architectures and programs, but on self-organizing architectures and self-generated code that grow from proper "seeds" provided by the designer.

Just because a system has the ability for self-modification, does not necessarily mean that all the changes it makes will improve its performance. The chapter by **Bas Steunebrink and Jürgen Schmidhuber** introduces a formal model that reasons about its own programs, and only makes modifications that can be proved to be beneficial. Specified accurately in a symbolic language, this model is theoretically optimal under certain assumptions.

1.5 Challenges at the Heart of the Matter

Though AGI differs from mainstream AI in its holistic attitude toward intelligence, the design and development of an AGI system still needs to be carried out step by step, and some of the topics involved are considered to be more important and crucial than the others. Each chapter in this cluster addresses an issue that the author(s) takes to be one, though by no means the only one, major challenge in their research toward AGI.

The chapter by **Tsvi Achler** considers recognition as the foundation of other processes that altogether form intelligence. To meet the general-purpose requirements of AGI, a more flexible recognition mechanism is introduced. While the majority of current recognition algorithms are based on a “feedforward” transformation from an input image to a recognized pattern, the mechanism Achler describes has a bidirectional “feedforward-feedback” structure, where the system’s expectation plays an important role.

Creativity is an aspect where computers are still far behind human intelligence. The chapter by **Maricarmen Martinez *et al.*** proposes analogy making and theory blending as ways to create new ideas. In this process, problem-solving theories are generalized from a source domain, then applied in a different target domain to solve novel problems. There is evidence showing that such processes indeed happen in human cognition, and are responsible for much of the creativity and generality of intelligence.

Contrary to many peoples’ assumption that “intelligence” is cold and unemotional, **Joscha Bach** argues that a model of intelligence must cover emotion and affect, since these play important roles in motivational dynamics and other processes. Emotion emerges via the system’s appraisal of situations and objects, with respect to the system’s needs and desires; and it in turn influences the system’s responses to those situations and objects, as well as its motivations and resource allocation.

Consciousness is one of the most mysterious phenomena associated with the human mind. The chapter by **Antonio Chella and Riccardo Manzotti** concludes that consciousness is necessary for general intelligence, and provides a survey of the existing attempts at producing similar phenomena in computer and robot systems. This study attempts to give some philosophical notions (including consciousness, free will, and experience) functional and constructive interpretations.

The difficulty of the problem of consciousness partly comes from the fact that it is not only a technical issue, but also a conceptual one. The chapter by **Richard Loosemore** provides a conceptual analysis of the notion of consciousness, helping us to understand what kind of answer might qualify as a solution to the problem. Such a meta-level reflection is necessary because if we get the problem wrong, there is little chance to get the solution right.

1.6 Summary

This book is not an attempt to settle all the fundamental problems of AGI, but merely to showcase and comprehensibly overview some of the key current theoretical explorations in the field. Given its stress on the generality and holistic nature of intelligence, AGI arguably has a greater demand for coherent theoretical foundations than narrow AI; and yet, the task of formulating appropriate theories is harder for AGI than for narrow AI, due to the wider variety of interdependent factors involved.

The last chapter by **Pei Wang** is an attempt to provide common criteria for the analysis, comparison, and evaluation of the competing AGI theories. It is proposed that, due to the nature of the field, a proper theory of intelligence for AGI should be *correct* according to our knowledge about human intelligence, *concrete* on how to build intelligent machines, and *compact* in its theoretical structure and content. Furthermore, these criteria should be balanced against each other.

This collection of writings of representative, leading AGI researchers shows that there is still no field-wide consensus on the accurate specification of the objective and methodology of AGI research. Instead, the field is more or less held together by a shared attitude toward AI research, which treats the problem of AI as one problem, rather than as a group of loosely related problems, as in mainstream AI. Furthermore, AGI researchers believe that it is possible to directly attack the problem of general intelligence now, rather than to postpone it to a unspecified future time.

The problems discussed in this book are not the same as those addressed by the traditional AI literatures or in AI's various sibling disciplines. As we have argued previously [11], general-propose AI has its own set of problems, which is neither a subset, nor a superset, of the problems studied in mainstream AI (the latter being exemplified in [8], e.g.). Among the problems of AGI, many are theoretical in nature, and must be solved by theoretical analysis – which in turn, must often be inspired and informed by experimental and engineering work. We hope this book will attract more attention, from both inside and outside the AGI field, toward the theoretical issues of the field, so as to accelerate the progress of AGI research – a matter which has tremendous importance, both intellectually and practically, to present-day human beings and our human and artificial successors.

Bibliography

[1] Brachman, R. J. (2006). (AA)AI – more than the sum of its parts, 2005 AAAI Presidential Address, *AI Magazine* **27**, 4, pp. 19–34.

[2] Feigenbaum, E. A. and Feldman, J. (1963). *Computers and Thought* (McGraw-Hill, New York).

[3] Feigenbaum, E. A. and McCorduck, P. (1983). *The Fifth Generation: Artificial Intelligence and Japan's Computer Challenge to the world* (Addison-Wesley Publishing Company, Reading, Massachusetts).

[4] Goertzel, B. and Pennachin, C. (eds.) (2007). *Artificial General Intelligence* (Springer, New York).

[5] McCarthy, J. (2007). From here to human-level AI, *Artificial Intelligence* **171**, pp. 1174–1182.

[6] McCarthy, J., Minsky, M., Rochester, N. and Shannon, C. (1955). A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence,
URL: `http://www-formal.stanford.edu/jmc/history/dartmouth.html`.

[7] Newell, A. and Simon, H. A. (1963). GPS, a program that simulates human thought, in E. A. Feigenbaum and J. Feldman (eds.), *Computers and Thought* (McGraw-Hill, New York), pp. 279–293.

[8] Russell, S. and Norvig, P. (2010). *Artificial Intelligence: A Modern Approach*, 3rd edn. (Prentice Hall, Upper Saddle River, New Jersey).

[9] Turing, A. M. (1950). Computing machinery and intelligence, *Mind* **LIX**, pp. 433–460.

[10] Wang, P. (2008). What do you mean by 'AI', in *Proceedings of the First Conference on Artificial General Intelligence*, pp. 362–373.

[11] Wang, P. and Goertzel, B. (2007). Introduction: Aspects of artificial general intelligence, in B. Goertzel and P. Wang (eds.), *Advance of Artificial General Intelligence* (IOS Press, Amsterdam), pp. 1–16.

Chapter 2

Artificial Intelligence and Cognitive Modeling 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

Cognitive modelers 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.

2.1 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?

2.1.1 *Naming the problem*

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. A caveat to this is that artificial intelligence systems may not be able to perform in some situations, not for reasons of their programming, but because of issues related to their physical manifestation. For example, it would be difficult for a machine without hand gestures and facial expressions to converse as well as a human in many situations because hand gestures and facial expressions are so important to many conversations. In the long term, it may be necessary therefore to sort out exactly which situations matter and which do not. However, the current abilities of artificial systems are so far away from human-level that resolving this issue can generally be postponed for some time. One point that does follow from these reflections, though, is the inadequacy of the Turing Test. Just as the invention of the airplane was an advance in artificial flight without convincing a single person that it was a bird, it is often irrelevant whether a major step-step towards human-intelligence cons observers into believing a computer is a human.

2.1.2 *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 the 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.

2.1.3 *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 quite far from 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, called *Intelligence Science*[1], and outlines 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 are not original (see (Cassimatis, 2010) and (Shanahan, 1997)) or exhaustive. 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, or 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 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?

2.2 Existing Methods and Standards are not Sufficient

Historically, AI and cognitive science were driven in part by the goal of understanding and engineering human-level intelligence. There are many 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,

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

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 good 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.

2.2.1 *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.

2.2.2 *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.

2.2.3 *Artificial intelligence*

An important aim 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.

(1) 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.

(2) 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.

2.2.4 *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.

2.3 Cognitive Modeling: 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[2]. 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,

[2] 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.

build fine models that fit the data and accurately predict new 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[3], 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 flight 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

[3] In a talk at RPI.

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 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 the 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 where 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 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.

2.4 Artificial Intelligence and Cognitive Modeling Can Help Each Other

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 principal 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 normally 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 and elsewhere (Cassimatis *et al.*, 2008) 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