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.