Introduction: Aspects of Artificial General Intelligence

Pei WANG and Ben GOERTZEL

Introduction

This book contains materials that come out of the Artificial General Intelligence Research Institute (AGIRI) Workshop, held in May 20-21, 2006 at Washington DC. The theme of the workshop is "Transitioning from Narrow AI to Artificial General Intelligence."

In this introductory chapter, we will clarify the notion of "Artificial General Intelligence", briefly survey the past and present situation of the field, analyze and refute some common objections and doubts regarding this area of research, and discuss what we believe needs to be addressed by the field as a whole in the near future. Finally, we will briefly summarize the contents of the other chapters in this collection.

1. What is meant by "AGI"

"Artificial General Intelligence", AGI for short, is a term adopted by some researchers to refer to their research field. Though not a precisely defined technical term, the term is used to stress the "general" nature of the desired capabilities of the systems being researched -- as compared to the bulk of mainstream Artificial Intelligence (AI) work, which focuses on systems with very specialized "intelligent" capabilities. While most existing AI projects aim at a certain aspect or application of intelligence, an AGI project aims at "intelligence" as a whole, which has many aspects, and can be used in various situations. There is a loose relationship between "general intelligence" as meant in the term AGI and the notion of "g-factor" in psychology [1]: the g-factor is an attempt to measure general intelligence, intelligence across various domains, in humans.

The notion of "intelligence" itself has no universally accepted definition, and the chapter following this one surveys a variety of definitions found in various parts of the research literature. So, "general intelligence" as we use it here is an imprecise variation on an imprecise concept. However, such imprecise concepts are what guide the direction of research, including research into the precise formulation of concepts. We believe that "general intelligence" and AGI are important concepts to pursue, in terms of both theory and software implementation.

Modern learning theory has made clear that the only way to achieve *maximally* general problem-solving ability is to utilize infinite computing power. Intelligence given limited computational resources is always going to have limits to its generality. The human mind/brain, while possessing extremely general capability, is best at solving the types of problems which it has specialized circuitry to handle (e.g. face recognition, social learning, language learning; see [2] for a summary of arguments in

this regard). However, even though no real intelligence can display total generality, it still makes sense to distinguish systems with general scope from highly specialized systems like chess-playing programs and automobile navigation systems and medical diagnosis systems. It is possible to quantify this distinction in various ways (see [3] and [4]; [5]; [6], for example), and this sort of quantification is an active area of research in itself, but for our present purposes drawing the qualitative distinction will suffice.

An AGI system, when fully implemented, will very likely be similar to the human brain/mind in various senses. However, we do not impose this as part of the definition of AGI. Nor do we restrict the techniques to be used to realize general intelligence, which means that an AGI project can follow a symbolic, connectionist, evolutionary, robotic, mathematical, or integrative approach. Indeed, the works discussed in the following chapters are based on very different theoretical and technical foundations, while all being AGI, as far as the goal and scope of the research is concerned. We believe that at the current stage, it is too early to conclude with any scientific definiteness which conception of "intelligence" is the "correct" one, or which technical approach is most efficient for achieving such a goal. It will be necessary for the field to encourage different approaches, and leave their comparison and selection to individual researchers.

However, this inclusiveness regarding methodology does not mean that all the current AI research projects can be labeled as "AGI". Actually, we believe that most of them cannot, and that is why we favor the use of a new term to distinguish the research we are interested in from what is usually called "AI". Again, the key difference is the goal and scope of the research. For example, nobody has challenged the belief that "learning" plays an important role in intelligence, and therefore it is an issue that almost all AGI projects address. However, most of the existing "machine learning" works do not belong to AGI, as defined above, because they define the learning problem in isolation, without treating it as part of a larger picture. They are not concerned with creating a system possessing broad-scope intelligence with generality at least roughly equal to that of the human mind/brain; they are concerned with learning in much narrower contexts. Machine learning algorithms may be applied quite broadly in a variety of contexts, but the breadth and generality in this case is supplied largely by the human user of the algorithm; any particular machine learning program, considered as a holistic system taking in inputs and producing outputs without detailed human intervention, can solve only problems of a very specialized sort.

Specified in this way, what we call AGI is similar to some other terms that have been used by other authors, such as "strong AI" [7], "human-level AI" [8], "true synthetic intelligence" [9], "general intelligent system" [10], and even "thinking machine" [11]. Though no term is perfect, we chose to use "AGI" because it correctly stresses the *general* nature of the research goal and scope, without committing too much to any theory or technique.

We will also refer in this chapter to "AGI projects." We use this term to refer to an AI research project that satisfies all the following criteria:

1. The project is based on a theory about "intelligence" as a whole (which may encompass intelligence as displayed by the human brain/mind, or may specifically refer to a class of non-human-like systems intended to display intelligence with a generality of scope at least roughly equalling that of the human brain/mind).

2. There is an engineering plan to implement the above conception of intelligence in a computer system.
3. The project has already produced some concrete results, as publications or prototypes, which can be evaluated by the research community.

The chapters in this book describe a number of current AGI research projects, thus defined, and also present some AGI research ideas not tied to any particular project.

2. The past and present of AGI

A comprehensive overview of historical and contemporary AGI projects is given in the introductory chapter of a prior edited volume focused on Artificial General Intelligence [5]. So, we will not repeat that material here. Instead, we will restrict ourselves to discussing a few recent developments in the field, and making some related general observations.

What has been defined above as "AGI" is very similar to the original concept of "AI". When the first-generation AI researchers started their exploration, they dreamed to eventually build computer systems with capabilities comparable to those of the human mind in a wide range of domains. In many cases, such a dream remained in their minds throughout their whole career, as evidenced for instance by the opinions of Newell and Simon [12]; [13], Minsky [14], and McCarthy [8]. And, at various points in the history of the field, large amounts of resources were invested into projects aimed at AGI, as exemplified by the Fifth Generation Computer Systems project.

However, in spite of some initial successes and the high expectations they triggered, the attempts of the first wave of AI researchers did not result in functional AGI systems. As a consequence, the AI community to a large extent has abandoned its original dream, and turned to more "practical" and "manageable" problems. After half a century, "AI has evolved to being a label on a family of relatively disconnected efforts" [9]. Though many domain-specific problems have been solved, and many special-purpose tools have been built, not many researchers feel that these achievements have brought us significantly closer to the goal of AGI. What makes the situation worse is the fact that AGI research is not even encouraged. For a long time, the "AI dream" was rarely mentioned within the AI community, and whoever pursued it was effectively committing career suicide, since few people took such attempts seriously.

In recent years, several forces seem to be turning this unfortunate trend around.

First, the year of 2006 is the fiftieth anniversary of the AI discipline, and 2005 the twenty-fifth anniversary of AAAI. Many people have taken this opportunity to reassess the field, surveying its past, present, and future. Among all the voices, a recurring one calls for the field to return to its original goal [8]; [9]; [15].

Second, some long-term AGI projects have survived and made progress. For example, Cyc recently released its open source version; Soar has been adding functionality into the system, and extending its application domain. Though each of these techniques has its limitations, they nevertheless show that AGI research can be fruitful.

Finally, after decades of work at the margin or outside of the AI community, a new generation of AGI projects has matured to the extent of producing publications and preliminary results. More or less coincidentally, several books have appeared within

the last few years, presenting several AGI projects, with theoretical and technical designs with various levels of detail [2]; [16]; [3]; [17]; [5]; [6]. Though each of these projects points to a quite different direction for AGI, they do introduce new ideas into the field, and show that the concrete, near-term possibilities of AGI are far from being exhaustively explored.

These factors have contributed to the recent resurgence of interest in AGI research. Only in the year of 2006, there have been several AGI-related gatherings in various conferences:

- Integrated Intelligent Capabilities (AAAI Special Track)
- A Roadmap to Human-Level Intelligence (IEEE WCCI Panel Session)
- Building and Evaluating Models of Human-Level Intelligence (CogSci Symposium)
- The AGIRI workshop, of which this book is a post-proceedings volume

Considering the fact that there were hardly any AGI-related meetings at all before 2004, the above list is quite remarkable.

3. Objections to AGI

Though the overall atmosphere is becoming more AGI-friendly, the researchers in this field remain a very small minority in the AI community. This situation is partially caused by various misunderstandings about AGI. As Turing did in [11], in the following paragraphs we will analyze and reject some common objections or doubts about AGI research.

3.1. "AGI is impossible"

Since the very beginning of AI research, there have been claims regarding the impossibility of truly intelligent computer systems. The best known arguments include those from Lucas [18], Dreyfus [19], and Penrose [20]. Since there is already a huge literature on these arguments [21], we will not repeat them here, but will simply remark that, so far, none of these arguments has convinced a majority of scientists with relevant expertise. Therefore, AGI remains possible, at least in theory.

3.2. "There is no such a thing as general intelligence"

There has been a lasting debate in the psychological research of human intelligence on whether there is a "general intelligence factor" ("g factor") that can be used to explain the difference in intellectual capabilities among individual human beings. Even though there is evidence supporting the existence of such a factor, as noted above there is also evidence suggesting that human intelligence is domain dependent, so is not that "general" at all.

In AI, many people have argued against ideas like the "General Problem Solver" [12], by stating that intelligent problem solving heavily depends on domain-specific knowledge. Guided by this kind of belief, various types of expert systems have been developed, whose power mostly come from domain knowledge, without which the system has little capability.

The above opinions do not rule out the possibility of AGI, for several reasons.

When we say a computer system is "general purpose", we do not require a single factor in the system to be responsible for all of its cross-domain intelligence. It is possible for the system to be an integration of several techniques, so as to be general-purpose without a single g-factor.

Also, AGI does not exclude individual difference. It is possible to implement multiple copies of the same AGI design, with different parameters and innate capabilities, and the resulting systems grew into different "experts", such as one with better mathematical capability, with another is better in verbal communication. Even in this case, the design is still "general" in the sense that it allows all these potentials. Just as the human brain/mind has a significant level of generality to its intelligence, even though some humans are better at mathematics and some are better at basketball.

Similarly, a general design does not conflict with the usage of domain-specific knowledge in problem solving. Even when an AGI system depends on domain-specific knowledge to solve domain-specific problems, its overall knowledge-management and learning mechanism may still remain general. The key point is that a general intelligence must be able to master a variety of domains, and learn to master new domains that it never confronted before. It does not need to have equal capability in all domains – humans will never be as intuitively expert at quantum physics as we are at the physics of projectiles in Earth's atmosphere, for example. Our innate, domain-specific knowledge gives us a boost regarding the latter; but, our generality of intelligence allows us to handle the former as well, albeit slowly and awkwardly and with the frequent need of tools like pencils, paper, calculators and computers.

In the current context, when we say that the human mind or an AGI system is "general purpose", we do not mean that it can solve all kinds of problems in all kinds of domains, but that it has the *potential* to solve any problem in any domain, given proper experience. Non-AGI systems lack such a potential. Even though Deep Blue plays excellent chess, it cannot do much other than that, no matter how it is trained.

3.3. "General-purpose systems are not as good as special-purpose ones"

Compared to the previous one, a weaker objection to AGI is to insist that even though general-purpose systems can be built, they will not work as well as special-purpose systems, in terms of performance, efficiency, etc.

We actually agree with this judgment to a certain degree, though we do not take it as a valid argument against the need to develop AGI.

For any given problem, a solution especially developed for it almost always works better than a general solution that covers multiple types of problem. However, we are not promoting AGI as a technique that will replace all existing domain-specific AI techniques. Instead, AGI is needed in situations where ready-made solutions are not available, due to the dynamic nature of the environment or the insufficiency of knowledge about the problem. In these situations, what we expect from an AGI system are not optimal solutions (which cannot be guaranteed), but flexibility, creatively, and robustness, which are directly related to the generality of the design.

In this sense, AGI is not proposed as a competing tool to any AI tool developed before, by providing better results, but as a tool that can be used when no other tool can, because the problem is unknown in advance.

3.4. "AGI is already included in the current AI"

We guess many AI researchers may be sympathetic to our goal, but doubt the need to introduce a new subfield into the already fragmented AI community. If what we call "AGI" is nothing but the initial and ultimate goal of AI, why bother to draw an unnecessary distinction?

We do this mostly for practical reasons, rather than theoretical reasons. Even though "AGI" is indeed closely related to the *original* meaning of "AI", so that it is in a sense a new word for an old concept, it is still very different from the *current* meaning of "AI", as the term is used in conferences and publications. As mentioned previously, our observation is that the mainstream AI community has been moving away from the original goal for decades, and we do not expect the situation to change completely very soon.

We do not buy the argument that "Since X plays an important role in intelligence, studying X contributes to the study of intelligence in general", where X can be replaced by reasoning, learning, planning, perceiving, acting, etc. On the contrary, we believe that most of the current AI research works make little direct contribution to AGI, though these works have value for many other reasons. Previously we have mentioned "machine learning" as an example. One of us (Goertzel) has published extensively about applications of machine learning algorithms to bioinformatics. This is a valid, and highly important sort of research – but it doesn't have much to do with achieving general intelligence.

There is no reason to believe that "intelligence" is simply a toolbox, containing mostly unconnected tools. Since the current AI "tools" have been built according to very different theoretical considerations, to implement them as modules in a big system will not necessarily make them work together, correctly and efficiently. Past attempts in this direction have taught us that "Component development is crucial; connecting the components is more crucial" [22].

Though it is possible to build AGI via an integrative approach, such integration needs to be guided by overall considerations about the system as a whole. We cannot blindly work on "parts", with the hope that they will end up working together. Because of these considerations, we think it is necessary to explicitly identify what we call "AGI" as different from mainstream AI research. Of course, even an AGI system still needs to be built step by step, and when the details of the systems are under consideration, AGI does need to use many results from previous AI research. But this does not mean that AGI reduces to an application of specialized AI components.

3.5. "It is too early to work on AGI"

Though many people agree that AGI is indeed the ultimate goal of AI research, they think it is premature to directly work on such a project, for various reasons.

For example, some people may suggest that AGI becomes feasible only after the research results regarding individual cognitive facilities (reasoning, learning, planning, etc) become mature enough to be integrated. However, as we argued above, without the guidance of an overall plan, these "parts" may never be ready to be organized into a whole.

A similar opinion is that the design of a general-purpose system should come out of the common features of various domain-specific systems, and therefore AGI can only be obtained by generalizing the design of many expert systems. The history of AI

has not provided much support for this belief, which misses the point we make previously, that is, a general-purpose system and a special-purpose system are usually designed with very different assumptions, restrictions, and target problems.

Some people claim that truly intelligent systems will mainly be the product of more research results in brain science or innovations of hardware design. Though we have no doubt that the progress in these fields will provide us with important inspirations and tools, we do not see them as where the major AGI problems are. Few people believe that detailed emulation of brain structures and functions is the optimal path to AGI. Emulating the human brain in detail will almost surely one day be possible, but this will likely require massively more hardware than achieving an equivalent level of intelligence via other mechanisms (since contemporary computer hardware is poorly suited to emulating neural wetware), and will almost surely not provide optimal intelligence given the computational resources available. And, though faster and larger hardware is always desired, it is the AI researchers' duty to tell hardware researchers what kind of hardware is needed for AGI.

All the above objections to AGI have the common root of seeing the solution of AGI as depending on the solution of another problem. We haven't seen convincing evidence for this. Instead, AGI is more likely to be a problem that demands direct research, which it is not too early to start --- actually we think it is already pretty late to give the problem the attention it deserves.

3.6. "AGI is nothing but hype"

OK, let us admit it: AI got a bad name from unrealized predictions in its earlier years, and we are still paying for it. To make things worse, from time to time we hear claims, usually on the Internet, about "breakthrough" in AI research, which turn out to be completely groundless. Furthermore, within the research community, there is little consensus even on the most basic problems, such as what intelligence is and what the criteria are for research success in the field. Just see the example of Deep Blue: while some AI researchers take it as a milestone, some others reject it as mostly irrelevant to AI research. As a common effect of these factors, explicitly working on AGI immediately marks a researcher a possible crackpot.

As long as AGI has not been proved impossible, it remains a legitimate research topic. Given the well-known complexity of the problem, there is no reason to expect an AGI to reach its goal within a short period, and all the popular theoretical controversies will probably continue to exist even after an AGI has been completed as planned. The fact that there is little consensus in the field should make us more careful when judging a new idea as completely wrong. As has happened more than once in the history of science, a real breakthrough may come from a counter-intuitive idea.

On the other hand, the difficulty of the problem cannot be used as an excuse for loose discipline in research. Actually, in the AGI research field we have seen works that are as serious and rigorous as scientific results in any other area. Though the conceptions and techniques of almost all AGI projects may remain controversial in the near future, this does not mean that the field should be discredited – but rather that attention should be paid to resolving the outstanding issues through concerted research.

3.7. "AGI research is not fruitful"

Some oppositions to AGI research come mainly from practical considerations. Given the nature of the problem, research results in AGI are more difficult to obtain, more difficult to get accepted by the research community even once obtained, and more difficult to turn into practical products even once accepted. Compared to other fields currently going under the AI label, researchers in AGI are less likely to be rewarded, in terms of publication, funding, career opportunity, and so on.

These issues are all true, as the experience of many AGI researchers shows. Because of this, and also because AGI does not invalidate the other research goals currently in vogue in the AI community (as discussed previously), we are not suggesting the whole AI community to turn to AGI research. Instead, we only hope AGI research to get the recognition, attention, and respect it deserves, as an active, productive and critical aspect of the AI enterprise. Given the potential importance of this topic, such a hope should not be considered as unrealistic.

3.8. "AGI is dangerous"

This is another objection that is as old as the field of AI. Like any science and technology, AGI has the danger of being misused, but this is not a reason to stop AGI research, just as it is not a reason to stop scientific research in many other fields. The viewpoint that "AGI is fundamentally dangerous because it will inevitably lead to disaster" is usually based on various misconceptions about intelligence and AGI. For example, some version of this claim is based on the assumption that an intelligent system will eventually want to dominate the universe, which has no scientific evidence.

Like scientists and engineers in any domain, AGI researchers should be responsible for the social impacts of their work. Given the available evidence, we believe AGI research has a much larger chance to have benign consequences to the human beings than harmful ones. Therefore, we do not think AGI research should be stopped because of its possible danger, though we do agree that it is an issue that should be kept in the mind of every AGI researcher.

4. Building an AGI community

As discussed above, based on extrapolating recent trends, it can reasonably be anticipated that the AGI field will soon end its decades-long dormancy, and enter a period of awakening. Though each individual AGI approach still has many obstacles to overcome, more and more people will appreciate the value of this sort of research.

In the AGIRI workshop, a topic that was raised by many attendances is the need to develop an AGI research community. From direct personal experience, many AGI researchers strongly feel that the existing platforms of conferences and societies, as well as the channels of publication and funding, do not properly satisfy their needs for communication, coordination, cooperation, and support. As we argued above, AGI has its own issues, which have been mostly ignored by the mainstream AI community. AGI researchers have been working mostly in isolation, and finding themselves surrounded by researchers with very different research interests and agenda. As commented by an attendance of the AGIRI workshop, "I don't think in my long career (I'm getting quite

old) I've ever been to a conference or workshop where I want to listen to such a large percentage of talks, and to meet so many people."

Though communication with other AI researchers is still necessary, the crucial need of the AGI field, at the current time, is to set up the infrastructures to support the regular communication and cooperation among AGI researchers. In this process, a common language will be developed, the similarities and differences among approaches will be clarified, repeated expenses will be reduced, and evaluation criteria will be formed and applied. All these are required for the growth of any scientific discipline.

Several community-forming activities are in the planning phase, and if all goes well, will be carried out soon. Their successes require the support of all AGI researchers, who will benefit from them in the long run.

5. This collection

The chapters in this book have been written by some of the speakers at the AGIRI Workshop after the meeting; each of them is based on a workshop talk, and also takes into account the feedback and afterthought of the meeting, as well as relationships with previous publications. Rather than thoroughly summarizing the contents of the chapters, here we will briefly review each chapter with a view toward highlighting its relationships with other chapters, so as to give a feeling for how the various approaches to and perspective on AGI connect together, in some ways forming parts of an emerging unified understanding in spite of the diversity of understanding perspectives.

First of all, following this chapter, Legg and Hutter's chapter (the only chapter whose authors did not attend the AGIRI Workshop) contains a simple enumeration of all the scientifically serious, published definitions of "intelligence" that the authors could dig up given a reasonable amount of effort. This is a worthwhile exercise in terms of illustrating both the commonality and the divergence among the various definitions. Clearly, almost all the authors cited are getting at similar intuitive concept – yet there are many, many ways to specify and operationalize this concept. And, of course, the choice of a definition of intelligence may have serious implications regarding one's preferred research direction. For instance, consider two of the definitions they cite:

> "Intelligence measures an agent's ability to achieve goals in a wide range of environments." -- S. Legg and M. Hutter

> "Intelligence is the ability for an information processing system to adapt to its environment with insufficient knowledge and resources." -- P. Wang

Note that the latter refers to limitations in processing power, whereas the former does not. Not surprisingly, much of the research of the authors of the prior definition concerns the theory of AGI algorithms requiring either infinite or extremely large amounts of processing power; whereas the central research programme of the latter author aims at achieving reasonable results using highly limited computational power.

Given this interconnectedness between the specifics of the definition of intelligence chosen by a researcher, and the focus of the research pursued by the researcher, it seems best to us that, at this stage of AGI research, the definition of

intelligence be left somewhat loose and heterogeneous in the field, so as to encourage a diversity of conceptual approaches to the AGI problem. A loose analogy, in another field, might be the definition of "life" in biology. There is a clear intuitive meaning to "life," yet pinning down exactly what the term means has proven difficult – and has not really proved necessary for the progress of the field of biology. Rather, different interpretations regarding the essential nature of "life" have led to different, fruitful scientific developments; and, of course, the vast majority of research in areas as divergent as systems biology and genomics has progressed without much attention to the definitional issue. Of course, we are not suggesting that all definitions of intelligence are equally valid, or that different definitions cannot be compared – on the contrary, to identify the research goal is often the key to understand an AGI project, as discussed previously.

The next paper, "A Foundational Architecture for General Intelligence" by Stan Franklin, serves (at least) two purposes. This paper corresponds to the talk that opened up the workshop, and serves both to introduce Franklin's LIDA architecture for AGI, and also to propose a general framework for discussing and comparing various AGI systems. This latter purpose is taken up in the following chapter, entitled "Four Contemporary AGI Designs: A Comparative Treatment," in which four individuals who presented at the workshop and contributed chapters to this volume present answers to a series of 15 questions regarding their AGI architectures. These questions were mainly drawn from Franklin's article, and represent an attempt to take a first step toward a common framework for comparing different approaches to AGI.

One of the main contributions of Franklin's chapter is to systematically map connections between current understanding of the human mind, as reflected in the cognitive science literature, and the components of an AGI design. This has been done before, but Franklin does a particularly succinct and lucid job, and for those of us who have been following the field for a while, it is pleasing to see how much easier this job gets as time goes on, due to ongoing advances in cognitive science as well as AGI. Another interesting point is how similar the basic high-level "boxes and lines architecture diagrams" for various AGI architectures come out to be. Of course there is nothing like a universal agreement, but it seems fair to say that there is a rough and approximate agreement among a nontrivial percentage of contemporary AGI researchers regarding the general way that cognitive function may be divided up into sub-functions within an overall cognitive architecture. This fact is particularly interesting to the extent that it allows attention to focus less on the cognitive architecture than on the "pesky little details" of what happens inside the boxes and what passes along the lines between the boxes (of course, the phrase "pesky little details" is chosen with some irony, since many researchers believe that it is these details of learning and knowledge representation, rather than the overall cognitive architecture, that most deserve the label of the "essence of intelligence").

The following paper, by Eric Baum, seeks to focus in on this essence. Rather than giving an overall AGI architecture, Baum concentrates on what he sees as the key issue facing those who would build AGI: the "inductive bias" that he believes human brains derive from their genetic heritage. Baum's hypothesis is that the problem of learning to act like an ordinary human is too hard to be achieved by general-purpose learning algorithms of the quality embodied in the brain. Rather, he suggests, much of learning to act like a human is done via specialized learning algorithms that are tuned for the specific learning problems, such as recognizing humans face; or by means of specialized data that is fed into general learning algorithms, representing problem-

specific bias. If this hypothesis is correct, then AGI designers have a big problem: even if they get the cognitive architecture diagram right, and plug reasonably powerful algorithms into the boxes carrying out learning, memory, perception and so forth, then even so, the algorithms may not be able to carry out the needed learning, because of the lack of appropriate inductive biases to guide them on their way.

In his book *What Is Thought?* [2], this problem is highlighted but no concrete solution is proposed. In his chapter here, Baum proposes what he sees as the sketch of a possible solution. Namely, he suggests, we can explicitly program a number of small "code modules" corresponding to the inductive bias supplied by the genome. AGI learning is then viewed as consisting of learning relatively simple programs that combine these code modules in task-appropriate ways. As an example of how this kind of approach may play out in practice, he considers the problem of writing a program that learns to play the game of Sokoban, via learning appropriate programs combining core modules dealing with issues like path-finding and understanding spatial relationships.

The next chapter, by Pei Wang (one of the authors of this Introduction), reviews his AGI project called NARS (Non-Axiomatic Reasoning System) which involves both a novel formal and conceptual foundation, and a software implementation embedding many aspects of the foundational theory. NARS posits that adaptation under knowledge-resources restriction is the basic principle of intelligence, and uses an AGI architecture with an uncertain inference engine at its core and other faculties like language processing, perception and action at the periphery, making use of specialized code together with uncertain inference. Compared to Franklin's proposed AGI architecture, NARS does not propose a modular high-level architecture for the core system, but places the emphasis on an uncertain inference engine implementing the proper semantics of uncertain reasoning. Regarding Baum's hypothesis of the need to explicitly code numerous modules encoding domain-specific functionalities (considered as inductive biases), Wang's approach would not necessarily disagree, but would consider these modules as to be built by the AGI architecture, and so they do not constitute the essence of intelligence but rather constitute learned special-purpose methods by which the system may interface with the world. Since the details of NARS have been covered by other publications, such as [17], this chapter mainly focuses on the development plan of NARS, which is a common issue faced by every AGI project. Since NARS is an attempt to minimize AGI design, some functionalities included in other AGI designs are treated as optional in NARS.

Nick Cassimatis's chapter presents a more recently developed approach to AGI architecture, which focuses on the combination of different reasoning and learning algorithms within a common framework, and the need for an integrative framework that can adaptively switch between and combine different algorithms depending on context. This approach is more similar to Franklin's than Wang's in its integrative nature, but differs from Franklin's in its focus on achieving superior algorithmic performance via hybridizing various algorithms, rather than interconnecting different algorithms in an overall architecture that assigns different algorithms strictly different functional roles. Cassimatis's prior work has been conceptually critical in terms of highlighting the power of AI learning algorithms to gain abstract knowledge spanning different domains – e.g. gaining knowledge about physical actions and using this knowledge to help learn language. This brings up a key difference between AGI work and typical, highly-specialized AI work. In ordinary contemporary AI work, computational language learning is one thing, and learning about physical objects and

their interrelationships is something else entirely. In an integrated intelligent mind, however, language and physical reality are closely interrelated. AGI research, to be effective, must treat these interconnections in a concrete and pragmatic way, as Cassimatis has done in his research.

Alexei Samsonovich and Giorgio Ascoli, the authors of the next chapter, are also involved with developing an ambitious AGI architecture, called BICA-GMU, created with funding from DARPA. Their architecture has been described elsewhere, and bears a family resemblance to LIDA in that it uses a (very LIDA-like) high-level architecture diagram founded on cognitive science, and fills in the boxes with a variety of different algorithms. So far the focus with BICA-GMU has been on declarative rather than procedural knowledge and learning, and the focus of these authors' contribution to this volume is along these lines. The chapter is called "Cognitive Map Dimensions of the Human Value System Extracted from Natural Language," and it reports some fascinating experiments in statistical language processing, oriented toward discovering "natural conceptual categories" as clusters of words that naturally group together in terms of their contexts of occurrence in text. The categories found by the authors' automated learning method have an obvious intuitive naturalness to them, and essentially the same categories were found to emerge from analysis of text in two different languages. Of course, these results are preliminary and could particularly use validation via analysis of texts in non-Western languages; but they are nonetheless highly thought-provoking. One is reminded of Chomsky's finding of universal grammatical patterns underlying various languages, which gives rise to the question of whether these grammatical patterns are innate, evolved "inductive bias" or learned/self-organized patterns that characterize spontaneously emerging linguistic structures. Similarly, the findings in this chapter give rise to the question of whether these conceptual categories represent innate, evolved inductive bias, versus learned/self-organized patterns that spontaneously emerge in any humanly embodied mind attempting to understand itself and the world. This sort of question may of course be explored via ongoing experimentation with teaching AGI systems like BICA-GMU and some of the other AGI systems described in this book: one can experiment with such systems both with and without programmer-supplied innate conceptual categories, and see how the progress and nature of learning is affected. (More specifically: this kind of experimentation can be done only with AGI systems whose knowledge representation supports explicit importation of declarative knowledge, which is the case with most of the AGI designs proposed in this book, but is not obviously the case e.g. with neural net architectures such as the one suggested in the following chapter, by Hugo de Garis.

De Garis's chapter is somewhat different from the preceding ones, in that it doesn't propose a specific AGI architecture, but rather proposes a novel tool for building the components of AGI systems (or, as De Garis terms it, "brain building"). Although most of the authors in this book come from more of a cognitive/computer science perspective, another important and promising approach to AGI involves neural networks, computational models of brain activity at varying levels of granularity. In a sense this is the lowest-risk approach to producing AGI, since after all the human brain is the best example of an intelligent system that we know right now. So, there is some good common sense in approaching AGI by trying to emulate brain function. Now, there is also a major problem with this approach, which is that we don't currently understand human brain function very well. Some parts of the brain are understood better than others; for example, Jeff Hawkins' [16] AI architecture is closely modeled on the visual cortex, which is one of the best-understood parts of the human brain. At

the current time, rather than focusing on constructing neural net AGI systems based on neuroscience knowledge, De Garis is focused on developing tools for constructing small neural networks that may serve as components of such AGI systems. Specifically he is focused on the problem of evolutionary learning of small neural networks: i.e., given a specification of what a neural net is supposed to do, he uses evolutionary learning to find a neural net doing that thing. The principal novelty of his approach is that this learning is conducted in hardware, on a reprogrammable chip (a field-programmable gate array), an approach that may provide vastly faster learning that is achievable through software-only methods. Preliminary results regarding this approach look promising.

Loosemore's paper considers the methodology of AGI research, and the way that this is affected by the possibility that all intelligent systems must be classified as complex systems. Loosemore takes a dim view of attempts to create AGI systems using the neat, formal approach of mathematics or the informal, bash-to-fit approach of engineering, claiming that both of these would be severely compromised if complexity is involved. Instead, he suggests an empirical science approach that offers a true marriage of cognitive science and AI. He advocates the creation of novel software tools enabling researchers to experiment with different sorts of complex intelligent systems, understanding the emergent structures and dynamics to which they give rise and subjecting our ideas about AI mechanisms to rigorous experimental tests, to see if they really do give rise to the expected global system performance.

The next paper, by Moshe Looks, harks back to De Garis et al's paper in its emphasis on evolutionary learning. Like De Garis et al, Looks is concerned with ways of making evolutionary learning much more efficient with a view toward enabling it to play a leading role in AGI – but the approach is completely different. Rather than innovating on the hardware side, Looks suggests a collection of fundamental algorithmic innovations, which ultimately constitute a proposal to replace evolutionary learning with a probabilistic-pattern-recognition based learning algorithm (MOSES = Meta-Optimizing Semantic Evolutionary Search) that grows a population of candidate problem solutions via repeatedly recognizing probabilistic patterns in good solutions and using these patterns to generate new ones. The key ideas underlying MOSES are motivated by cognitive science ideas, most centrally the notion of "adaptive representation building" – having the learning algorithm figure out the right problem representation as it goes along, as part of the learning process, rather than assuming a well-tuned representation right from the start. The MOSES algorithm was designed to function within the Novamente AGI architecture created by one of the authors of this Introduction (Goertzel) together with Looks and others (and discussed in other papers in this volume, to be mentioned below), but also to operate independently as a program learning solution. This chapter describes some results obtained using stand-alone MOSES on a standard test problem, the "artificial ant" problem. More powerful performance is hoped to be obtained by synthesizing MOSES with the PLN probabilistic reasoning engine, to be described in Ikle' et al's chapter (to be discussed below). But stand-alone MOSES in itself appears to be a dramatic improvement over standard evolutionary learning in solving many different types of problems, displaying fairly rapid learning on some problem classes that are effectively intractable for GA/GP. (This, of course, makes it interesting to speculate about what could be achievable by running MOSES on the reconfigurable hardware FPGA discussed in De Garis's chapter. Both MOSES and FPGA's can massively speed up evolutionary learning – the former in a fundamental order-of-complexity sense on certain problem classes, and the latter

by a large constant multiplier in a less problem-class-dependent way. The combination of the two could be extremely powerful.)

The chapter by Matthew Ikle' et al, reviews aspects of Probabilistic Logic Networks (PLN) -- an AI problem-solving approach that, like MOSES, has been created with a view toward integration into the Novamente AI framework, as well as toward stand-alone performance. PLN is a probabilistic logic framework that combines probability theory, term logic and predicate logic with various heuristics in order to provide comprehensive forward and backward chaining inference in contexts ranging from mathematical theorem-proving to perceptual pattern-recognition, and speculative inductive and abductive inference. The specific topic of this chapter is the management of "weight of evidence" within PLN. Like NARS mentioned above and Peter Walley's imprecise probability theory [23], PLN quantifies truth values using a minimum of two numbers (rather than, for instance, a single number representing a probability or fuzzy membership value). One approach within PLN is to use two numbers (s,n), where s represents a probability value, and n represents a "weight of evidence" defining how much evidence underlies that probability value. Another, equivalent approach within PLN is to use two numbers (L,U), representing an interval probability, interpreted to refer to a family of probability distributions the set of whose means has (L,U) as a b% confidence interval. The chapter discusses the relationship between these representations, and the way that these two-number probabilities may be propagated through inference rules like deduction, induction, abduction and revision. It is perhaps worth noting that PLN originally emerged, in 1999-2000, as an attempt to create a probabilistic variant of the NARS uncertain logic, although it has long since diverged from these roots. Part of the underlying motivation for both NARS and PLN is the assumption that AGIs must be able to carry out a diversity of inferences involving uncertain knowledge and uncertain conclusions, and thus must possess a reasonably robust method of managing all this uncertainty. Humans are famously poor at probability estimation [24];[25], but nonetheless we are reasonably good at uncertainty management in many contexts, and both PLN and NARS (but using quite different methods) attempts to capture this kind of pragmatic uncertainty management that humans are good at. A difference between the two approaches is that PLN is founded on probability theory and attempts to harmonize human-style robust uncertainty management with precise probabilistic calculations – using the notion that the former is appropriate when data is sparse, and gradually merges into the latter as more data becomes available. On the other hand, in NARS the representation, interpretation, and processing of uncertainty do not follow probability theory in general, though agree with it on special cases. Furthermore, precise probabilistic inference would be implemented as a special collection of rules running on top of the underlying NARS inference engine in roughly the same manner that programs may run on top of an operating system.

Following up on the uncertain-logic theme, the next chapter by Stephan Vladimir Bugaj and Ben Goertzel moves this theme into the domain of developmental psychology. Piaget's ideas have been questioned by modern experimental developmental psychology, yet remain the most coherent existing conceptual framework for studying human cognitive development. It turns out to be possible to create a Piaget-like theory of stages of cognitive development that is specifically appropriate to uncertain reasoning systems like PLN, in which successive stages involve progressively sophisticated inference control: simple heuristic control (the infantile stage); inductive, history-based control (the concrete operational stage); inference-based inference control (the formal stage); and inference-based modification

of inference rules (the post-formal stage). The pragmatic implications of this view of cognitive development are discussed in the context of classic Piagetan learning problems such as learning object permanence, conservation laws, and theory of mind.

In a general sense, quite apart from the specifics of the developmental theory given in Goertzel and Bugaj's chapter, one may argue that the logic of cognitive development is a critical aspect of AGI that has received far too little attention. Designing and building AGI's is important, but once they are built, one must teach them and guide their development, and the logic of this development may not be identical or even very similar to that of human infants and children. Different sorts of AGIs may follow different sorts of developmental logic. This chapter discusses cognitive development specifically in the context of uncertain logic based AGI systems, and comparable treatments could potentially be given for different sorts of AGI designs.

The following chapter, by Ben Goertzel (one of the authors of this Introduction), discusses certain aspects of the Novamente AGI design. A comprehensive overview of the Novamente system is not given here, as several published overviews of Novamente already exist, but the highlights are touched and some aspects of Novamente that have not been discussed before in publications are reviewed in detail (principally, economic attention allocation and action selection). Commonalities between Novamente and Franklin's LIDA architecture are pointed out, especially in the area of real-time action selection. Focus is laid on the way the various aspects of the Novamente architecture are intended to work together to lead to the emergence of complex cognitive structures such as the self and the "moving bubble of attention." These ideas are explored in depth in the context of a test scenario called "iterated Easter Egg Hunt," which has not yet been experimented with, but is tentatively planned for the Novamente project in mid-2007. This scenario appears to provide an ideal avenue for experimentation with integrated cognition and the emergence of self and adaptive attention, and is currently being implemented in the AGISim 3D simulation world, in which the Novamente system controls a humanoid agent.

Novamente is an integrative architecture, in the sense that it combines a number of different learning algorithms in a highly specific way. Probabilistic logic is used as a common language binding together the various learning algorithms involved. Two prior chapters (by Looks, and Ikle' et al) reviewed specific AI learning techniques that lie at the center of Novamente's cognition (MOSES and PLN). The next two chapters discuss particular applications that have been carried out using Novamente, in each case via utilizing PLN in combination with other simpler Novamente cognitive processes.

The Heljakka et al chapter discusses the learning of some very simple behaviors for a simulated humanoid agent in the AGISim 3D simulation world, via a pure "embodied reinforcement learning" methodology. In Piagetan terms, these are "infantile-level" tasks, but to achieve them within the Novamente architecture nevertheless requires a fairly subtle integration of various cognitive processes. The chapter reviews in detail how perceptual pattern mining, PLN inference and predicate schematization (declarative-to-procedural knowledge conversion) have been used to help Novamente learn how to play the classic human-canine game of "fetch."

The last two chapters of the book are not research papers but rather edited transcriptions of dialogues that occurred at the workshop. The first of these, on the topic of the ethics of highly intelligent AGIs, was probably the liveliest and most entertaining portion of the workshop, highlighted by the spirited back-and-forth between Hugo de Garis and Eliezer Yudkowsky. The second of these was on the

practicalities of actually creating powerful AGI software systems from the current batch of ideas and designs, and included a variety of "timing estimates" for the advent of human-level AGI from a number of leading researchers. These dialogues give a more human, less formal view of certain aspects of the current state of philosophical and pragmatic thinking about AGI by active AGI researchers.

All in all, it cannot be claimed that these chapters form a balanced survey of the current state of AGI research – there are definite biases, such as a bias towards symbolic and uncertain-reasoning-based systems versus neural net type systems, and a bias away from robotics (though there is some simulated robotics) and also away from highly abstract theoretical work a la Hutter [3] and Schmidhuber [26]. However, they do present a survey that is both broad and deep, and we hope that as a collection they will give you, the reader, a great deal to think about. While we have a long way to go to achieve AGI at the human level and beyond, we do believe that significant progress is being made in terms of resolving the crucial problem of AGI design, and that the chapters here do substantively reflect this progress.

References

[1] A. R. Jensen, The G Factor: the Science of Mental Ability, Psycoloquy: 10,#2, 1999
[2] E. Baum, What is Thought? MIT Press, 2004.
[3] M. Hutter, Universal Artificial Intelligence: Sequential Decisions based on Algorithmic Probability, Springer, 2005.
[4] B. Goertzel. The Structure of Intelligence, Springer, 1993.
[5] B. Goertzel and C. Pennachin (editors), Artificial General Intelligence, Springer, 2007.
[6] B. Goertzel. The Hidden Pattern, BrownWalker, 2006.
[7] J. Searle, Minds, Brains, and Programs, Behavioral and Brain Sciences 3 (1980), 417-424.
[8] J. McCarthy, The Future of AI — A Manifesto, AI Magazine, 26(2005), Winter, 39
[9] R. Brachman, Getting Back to "The Very Idea", AI Magazine, 26(2005), Winter, 48–50
[10] P. Langley, Cognitive Architectures and General Intelligent Systems, AI Magazine 27(2006), Summer, 33-44.
[11] A. M. Turing, Computing machinery and intelligence, Mind LIX (1950), 433-460.
[12] A. Newell and H. A. Simon, GPS, a program that simulates human thought, E. A. Feigenbaum and J. Feldman (editors), Computers and Thought, 279-293, McGraw-Hill, 1963.
[13] A. Newell, Unified Theories of Cognition, Harvard University Press, 1990.
[14] D. G. Stork, Scientist on the Set: An Interview with Marvin Minsky, D. G. Stork (editor), HAL's Legacy: 2001's Computer as Dream and Reality, 15-30, MIT Press, 1997.
[15] N. J. Nilsson, Human-Level Artificial Intelligence? Be Serious! AI Magazine, 26(2005), Winter, 68–75.
[16] J. Hawkins and S. Blakeslee, On Intelligence, Times Books, 2004.
[17] P. Wang, Rigid Flexibility: The Logic of Intelligence, Springer, 2006.
[18] J. R. Lucas, Minds, Machines and Gödel, Philosophy XXXVI (1961), 112-127.
[19] H.L. Dreyfus, What Computers Still Can't Do, MIT Press, 1992.
[20] R. Penrose, The Emperor's New Mind: Concerning Computers, Minds, and the Laws of Physics, Oxford University Press, 1989.
[21] D. Chalmers, Contemporary Philosophy of Mind: An Annotated Bibliography, Part 4: Philosophy of Artificial Intelligence, http://consc.net/biblio/4.html
[22] A. Roland and P. Shiman, Strategic computing: DARPA and the quest for machine intelligence, 1983-1993, MIT Press, 2002.
[23] P. Walley: Towards a unified theory of imprecise probability. Int. J. Approx. Reasoning 24(2-3): 125-148 (2000)
[24] D. Kahneman, P. Slovic, & A. Tversky (editors), Judgment under Uncertainty: Heuristics and Biases. Cambridge, UK: Cambridge University Press, 1982.
[25] T. Gilovich,, D. Griffin & D. Kahneman (editors), Heuristics and biases: The psychology of intuitive judgment. Cambridge, UK: Cambridge University Press, 2002.
[26] J. Schmidhuber, Goedel machines: self-referential universal problem solvers making provably optimal self-improvements. In B. Goertzel and C. Pennachin (editors), Artificial General Intelligence, 2006.