Ben Goertzel
Cassio Pennachin (Eds.)

Artificial General Intelligence

Springer

Copyrighted material

Ben Goertzel
Cassio Pennachin (Eds.)

Artificial General Intelligence

With 42 Figures and 16 Tables

Springer

Copyrighted material

Editors:

Ben Goertzel
Cassio Pennachin

AGIRI – Artificial General Intelligence Research Institute
1405 Bernerd Place
Rockville, MD 20851
USA
ben@agiri.org
cassio@agiri.org

Managing Editors:

Prof. Dov M. Gabbay
Augustus De Morgan Professor of Logic
Department of Computer Science, King's College London
Strand, London WC2R 2LS, UK

Prof. Dr. Jörg Siekmann
Forschungsbereich Deduktions- und Multiagentensysteme, DFKI
Stuhlsatzenweg 3, Geb. 43, 66123 Saarbrücken, Germany

Library of Congress Control Number: 2006937159

ACM Computing Classification (1998): F.1, F.4, H.5, I.2, I.6

ISSN 1611-2482
ISBN-10 3-540-23733-X Springer Berlin Heidelberg New York
ISBN-13 978-3-540-23733-4 Springer Berlin Heidelberg New York

This work is subject to copyright. All rights are reserved, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilm or in any other way, and storage in data banks. Duplication of this publication or parts thereof is permitted only under the provisions of the German Copyright Law of September 9, 1965, in its current version, and permission for use must always be obtained from Springer. Violations are liable for prosecution under the German Copyright Law.

Springer is a part of Springer Science+Business Media
springer.com

© Springer-Verlag Berlin Heidelberg 2007

The use of general descriptive names, registered names, trademarks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

Cover Design: KünkelLopka, Heidelberg
Typesetting: by the Editors
Production: LE-TEX Jelonek, Schmidt & Vöckler GbR, Leipzig

Printed on acid-free paper 45/3100/YL 5 4 3 2 1 0

Copyrighted material

Contents

Copyrighted material

Copyrighted material

Copyrighted material

Copyrighted material

Copyrighted material

Copyrighted material

Copyrighted material

Contemporary Approaches to Artificial General Intelligence

Cassio Pennachin and Ben Goertzel

AGIRI – Artificial General Intelligence Research Institute
1405 Bernerd Place, Rockville, MD 20851, USA
cassio@agiri.org, ben@agiri.org - http://www.agiri.org

1 A Brief History of AGI

The vast bulk of the AI field today is concerned with what might be called "narrow AI" – creating programs that demonstrate intelligence in one or another specialized area, such as chess-playing, medical diagnosis, automobile-driving, algebraic calculation or mathematical theorem-proving. Some of these narrow AI programs are extremely successful at what they do. The AI projects discussed in this book, however, are quite different: they are explicitly aimed at artificial *general* intelligence, at the construction of a software program that can solve a variety of complex problems in a variety of different domains, and that controls itself autonomously, with its own thoughts, worries, feelings, strengths, weaknesses and predispositions.

Artificial General Intelligence (AGI) was the original focus of the AI field, but due to the demonstrated difficulty of the problem, not many AI researchers are directly concerned with it anymore. Work on AGI has gotten a bit of a bad reputation, as if creating digital general intelligence were analogous to building a perpetual motion machine. Yet, while the latter is strongly implied to be impossible by well-established physical laws, AGI appears by all known science to be quite possible. Like nanotechnology, it is "merely an engineering problem", though certainly a very difficult one.

The presupposition of much of the contemporary work on "narrow AI" is that solving narrowly defined subproblems, in isolation, contributes significantly toward solving the overall problem of creating real AI. While this is of course true to a certain extent, both cognitive theory and practical experience suggest that it is not so true as is commonly believed. In many cases, the best approach to implementing an aspect of mind in isolation is very different from the best way to implement this same aspect of mind in the framework of an integrated AGI-oriented software system.

The chapters of this book present a series of approaches to AGI. None of these approaches has been terribly successful yet, in AGI terms, although several of them have demonstrated practical value in various specialized domains (narrow-AI style). Most of the projects described are at an early stage of engineering development, and some are still in the design phase. Our aim is not to present AGI as a mature field of computer science – that would be

Copyrighted material

impossible, for it is not. Our goal is rather to depict some of the more exciting ideas driving the AGI field today, as it emerges from infancy into early childhood.

In this introduction, we will briefly overview the AGI approaches taken in the following chapters, and we will also discuss some other historical and contemporary AI approaches not extensively discussed in the remainder of the book.

1.1 Some Historical AGI-Related Projects

Generally speaking, most approaches to AI may be divided into broad categories such as:

- symbolic;
- symbolic and probability- or uncertainty-focused;
- neural net-based;
- evolutionary;
- artificial life;
- program search based;
- embedded;
- integrative.

This breakdown works for AGI-related efforts as well as for purely narrow-AI-oriented efforts. Here we will use it to structure a brief overview of the AGI field. Clearly, there have been many more AGI-related projects than we will mention here. Our aim is not to give a comprehensive survey, but rather to present what we believe to be some of the most important ideas and themes in the AGI field overall, so as to place the papers in this volume in their proper context.

The majority of ambitious AGI-oriented projects undertaken to date have been in the symbolic-AI paradigm. One famous such project was the General Problem Solver [42], which used heuristic search to solve problems. GPS did succeed in solving some simple problems like the Towers of Hanoi and *crypto-arithmetic*,[1] but these are not really general problems – there is no learning involved. GPS worked by taking a general goal – like solving a puzzle – and breaking it down into subgoals. It then attempted to solve the subgoals, breaking them down further into even smaller pieces if necessary, until the subgoals were small enough to be addressed directly by simple heuristics. While this basic algorithm is probably necessary in planning and goal satisfaction for a mind, the rigidity adopted by GPS limits the kinds of problems one can successfully cope with.

[1] Crypto-arithmentic problems are puzzles like `DONALD + GERALD = ROBERT`. To solve such a problem, assign a number to each letter so that the equation comes out correctly.

Copyrighted material

Probably the most famous and largest symbolic AI effort in existence today is Doug Lenat's CYC project.[2] This began in the mid-80's as an attempt to create true AI by encoding all common sense knowledge in first-order predicate logic. The encoding effort turned out to require a large effort, and soon Cyc deviated from a pure AGI direction. So far they have produced a useful knowledge database and an interesting, highly complex and specialized inference engine, but they do not have a systematic R&D program aimed at creating autonomous, creative interactive intelligence. They believe that the largest subtask required for creating AGI is the creation of a knowledge base containing all human common-sense knowledge, in explicit logical form (they use a variant of predicate logic called CycL). They have a large group of highly-trained knowledge encoders typing in knowledge, using CycL syntax.

We believe that the Cyc knowledge base may potentially be useful eventually to a mature AGI system. But we feel that the kind of reasoning, and the kind of knowledge embodied in Cyc, just scratches the surface of the dynamic knowledge required to form an intelligent mind. There is some awareness of this within Cycorp as well, and a project called CognitiveCyc has recently been initiated, with the specific aim of pushing Cyc in an AGI direction (Stephen Reed, personal communication).

Also in the vein of "traditional AI", Alan Newell's well-known SOAR project[3] is another effort that once appeared to be grasping at the goal of human-level AGI, but now seems to have retreated into a role of an interesting system for experimenting with limited-domain cognitive science theories. Newell tried to build "Unified Theories of Cognition", based on ideas that have now become fairly standard: logic-style knowledge representation, mental activity as problem-solving carried out by an assemblage of heuristics, etc. The system was by no means a total failure, but it was not constructed to have a real autonomy or self-understanding. Rather, it's a disembodied problem-solving tool, continually being improved by a small but still-growing community of SOAR enthusiasts in various American universities.

The ACT-R framework [3], though different from SOAR, is similar in that it's an ambitious attempt to model human psychology in its various aspects, focused largely on cognition. ACT-R uses probabilistic ideas and is generally closer in spirit to modern AGI approaches than SOAR is. But still, similarly to SOAR, many have argued that it does not contain adequate mechanisms for large-scale creative cognition, though it is an excellent tool for the modeling of human performance on relatively narrow and simple tasks.

Judea Pearl's work on Bayesian networks [43] introduces principles from probability theory to handle uncertainty in an AI scenario. Bayesian networks are graphical models that embody knowledge about probabilities and dependencies between events in the world. Inference on Bayesian networks is possible using probabilistic methods. Bayesian nets have been used with

[2] See `www.cyc.com` and [38].

[3] See `http://ai.eecs.umich.edu/soar/` and [37].

Copyrighted material

success in many narrow domains, but, in order to work well, they need a reasonably accurate model of the probabilities and dependencies of the events being modeled. However, when one has to *learn* either the structure or the probabilities in order to build a good Bayesian net, the problem becomes very difficult [29].

Pei Wang's NARS system, described in this volume, is a very different sort of attempt to create an uncertainty-based, symbolic AI system. Rather than using probability theory, Wang uses his own form of uncertain logic – an approach that has been tried before, with fuzzy logic, certainty theory (see, for example, [50]) and so forth, but has never before been tried with such explicit AGI ambitions.

Another significant historical attempt to "put all the pieces together" and create true artificial general intelligence was the Japanese 5th Generation Computer System project. But this project was doomed by its pure engineering approach, by its lack of an underlying theory of mind. Few people mention this project these days. In our view, much of the AI research community appears to have learned the wrong lessons from the 5th generation AI experience – they have taken the lesson to be that integrative AGI is bad, rather than that integrative AGI should be approached from a sound conceptual basis.

The neural net approach has not spawned quite so many frontal assaults on the AGI problem, but there have been some efforts along these lines. Werbos has worked on the application of recurrent networks to a number of problems [55, 56]. Stephen Grossberg's work [25] has led to a host of special neural network models carrying out specialized functions modeled on particular brain regions. Piecing all these networks together could eventually lead to a brain-like AGI system. This approach is loosely related to Hugo de Garis's work, discussed in this volume, which seeks to use evolutionary programming to "evolve" specialized neural circuits, and then piece the circuits together into a whole mind. Peter Voss's a2i2 architecture also fits loosely into this category – his algorithms are related to prior work on "neural gasses" [41], and involve the cooperative use of a variety of different neural net learning algorithms. Less biologically oriented than Grossberg or even de Garis, Voss's neural system net does not try to closely model biological neural networks, but rather to emulate the sort of thing they do on a fairly high level.

The evolutionary programming approach to AI has not spawned any ambitious AGI projects, but it has formed a part of several AGI-oriented systems, including our own Novamente system, de Garis's CAM-Brain machine mentioned above, and John Holland's classifier systems [30]. Classifier systems are a kind of hybridization of evolutionary algorithms and probabilistic-symbolic AI; they are AGI-oriented in the sense that they are specifically oriented toward integrating memory, perception, and cognition to allow an AI system to act in the world. Typically they have suffered from severe performance problems, but Eric Baum's recent variations on the classifier system theme seem to have partially resolved these issues [5]. Baum's Hayek systems were tested on a simple "three peg blocks world" problem where any disk may be placed

Copyrighted material

on any other; thus the required number of moves grows only linearly with the number of disks, not exponentially. The chapter authors were able to replicate their results only for n up to 5 [36].

The artificial life approach to AGI has remained basically a dream and a vision, up till this point. Artificial life simulations have succeeded, to a point, in getting interesting mini-organisms to evolve and interact, but no one has come close to creating an Alife agent with significant general intelligence. Steve Grand made some limited progress in this direction with his work on the Creatures game, and his current R&D efforts are trying to go even further [24]. Tom Ray's Network Tierra project also had this sort of ambition, but seems to have stalled at the stage of the automated evolution of simple multicellular artificial lifeforms.

Program search based AGI is a newer entry into the game. It had its origins in Solomonoff, Chaitin and Kolmogorov's seminal work on algorithmic information theory in the 1960s, but it did not become a serious approach to practical AI until quite recently, with work such as Schmidhuber's OOPS system described in this volume, and Kaiser's dag-based program search algorithms. This approach is different from the others in that it begins with a formal theory of general intelligence, defines impractical algorithms that are provably known to achieve general intelligence (see Hutter's chapter on AIXI in this volume for details), and then seeks to approximate these impractical algorithms with related algorithms that are more practical but less universally able.

Finally, the integrative approach to AGI involves taking elements of some or all of the above approaches and creating a combined, synergistic system. This makes sense if you believe that the different AI approaches each capture some aspect of the mind uniquely well. But the integration can be done in many different ways. It is not workable to simply create a modular system with modules embodying different AI paradigms: the different approaches are too different in too many ways. Instead one must create a unified knowledge representation and dynamics framework, and figure out how to manifest the core ideas of the various AI paradigms within the universal framework. This is roughly the approach taken in the Novamente project, but what has been found in that project is that to truly integrate ideas from different AI paradigms, most of the ideas need to be in a sense "reinvented" along the way.

Of course, no such categorization is going to be complete. Some of the papers in this book do not fit well into any of the above categories: for instance, Yudkowsky's approach, which is integrative in a sense, but does not involve integrating prior AI algorithms; and Hoyes's approach, which is founded on the notion of 3D simulation. What these two approaches have in common is that they both begin with a maverick cognitive science theory, a bold new explanation of human intelligence. They then draw implications and designs for AGI from the respective cognitive science theory.

None of these approaches has yet proved itself successful – this book is a discussion of promising approaches to AGI, not successfully demonstrated

Copyrighted material

ones. It is probable that in 10 years a different categorization of AGI approaches will seem more natural, based on what we have learned in the interim. Perhaps one of the approaches described here will have proven successful, perhaps more than one; perhaps AGI will still be a hypothetical achievement, or perhaps it will have been achieved by methods totally unrelated to those described here. Our own belief, as AGI researchers, is that an integrative approach such as the one embodied in our Novamente AI Engine has an excellent chance of making it to the AGI finish line. But as the history of AI shows, researchers' intuitions about the prospects of their AI projects are highly chancy. Given the diverse and inter-contradictory nature of the different AGI approaches presented in these pages, it stands to reason that a good percentage of the authors have got to be significantly wrong on significant points! We invite the reader to study the AGI approaches presented here, and others cited but not thoroughly discussed here, and draw their own conclusions. Above all, we wish to leave the reader with the impression that AGI is a vibrant area of research, abounding with exciting new ideas and projects – and that, in fact, it is AGI rather than narrow AI that is properly the primary focus of artificial intelligence research.

2 What Is Intelligence?

What do we mean by *general intelligence*? The dictionary defines intelligence with phrases such as "The capacity to acquire and apply knowledge", and "The faculty of thought and reason." General intelligence implies an ability to acquire and apply knowledge, and to reason and think, in a variety of domains, not just in a single area like, say, chess or game-playing or languages or mathematics or rugby. Pinning down general intelligence beyond this is a subtle though not unrewarding pursuit. The disciplines of psychology, AI and control engineering have taken differing but complementary approaches, all of which are relevant to the AGI approaches described in this volume.

2.1 The Psychology of Intelligence

The classic psychological measure of intelligence is the "g-factor" [7], although this is quite controversial, and many psychologists doubt that any available IQ test really measures human intelligence in a general way. Gardner's [15] theory of multiple intelligences argues that human intelligence largely breaks down into a number of specialized-intelligence components (including linguistic, logical-mathematical, musical, bodily-kinesthetic, spatial, interpersonal, intra-personal, naturalist and existential).

Taking a broad view, it is clear that, in fact, human intelligence is not all that general. A huge amount of our intelligence is focused on situations that have occurred in our evolutionary experience: social interaction, vision processing, motion control, and so forth. There is a large research literature

Copyrighted material

in support of this fact. For instance, most humans perform poorly at making probabilistic estimates in the abstract, but when the same estimation tasks are presented in the context of familiar social situations, human accuracy becomes much greater. Our intelligence is general "in principle", but in order to solve many sorts of problems, we need to resort to cumbersome and slow methods such as mathematics and computer programming. Whereas we are vastly more efficient at solving problems that make use of our in-built specialized neural circuitry for processing vision, sound, language, social interaction data, and so forth. Gardner's point is that different people have particularly effective specialized circuitry for different specializations. In principle, a human with poor social intelligence but strong logical-mathematical intelligence could solve a difficult problem regarding social interactions, but might have to do so in a very slow and cumbersome over-intellectual way, whereas an individual with strong innate social intelligence would solve the problem quickly and intuitively.

Taking a somewhat different approach, psychologist Robert Sternberg [53] distinguishes three aspects of intelligence: componential, contextual and experiential. Componential intelligence refers to the specific skills people have that make them intelligent; experiential refers to the ability of the mind to learn and adapt through experience; contextual refers to the ability of the mind to understand and operate within particular contexts, and select and modify contexts.

Applying these ideas to AI, we come to the conclusion that, to roughly emulate the nature of human general intelligence, an artificial general intelligence system should have:

- the ability to solve general problems in a non-domain-restricted way, in the same sense that a human can;
- most probably, the ability to solve problems in particular domains and particular contexts with particular efficiency;
- the ability to use its more generalized and more specialized intelligence capabilities together, in a unified way;
- the ability to learn from its environment, other intelligent systems, and teachers;
- the ability to become better at solving novel types of problems as it gains experience with them.

These points are based to some degree on human intelligence, and it may be that they are a little too anthropomorphic. One may envision an AGI system that is so good at the "purely general" aspect of intelligence that it doesn't need the specialized intelligence components. The practical possibility of this type of AGI system is an open question. Our guess is that the multiple-specializations nature of human intelligence will be shared by any AGI system operating with similarly limited resources, but as with much else regarding AGI, only time will tell.

Copyrighted material

One important aspect of intelligence is that it can only be achieved by a system that is capable of learning, especially autonomous and incremental learning. The system should be able to interact with its environment and other entities in the environment (which can include teachers and trainers, human or not), and learn from these interactions. It should also be able to build upon its previous experiences, and the skills they have taught it, to learn more complex actions and therefore achieve more complex goals.

The vast majority of work in the AI field so far has pertained to highly specialized intelligence capabilities, much more specialized than Gardner's multiple intelligence types – e.g. there are AI programs good at chess, or theorem verification in particular sorts of logic, but none good at logical-mathematical reasoning in general. There has been some research on completely general non-domain-oriented AGI algorithms, e.g. Hutter's AIXI model described in this volume, but so far these ideas have not led to practical algorithms (Schmidhuber's OOPS system, described in this volume, being a promising possibility in this regard).

2.2 The Turing Test

Next, no discussion of the definition of intelligence in an AI context would be complete without mention of the well-known Turing Test. Put loosely, the Turing test asks an AI program to *simulate a human in a text-based conversational interchange*. The most important point about the Turing test, we believe, is that it is a *sufficient* but not *necessary* criterion for artificial general intelligence. Some AI theorists don't even consider the Turing test as a sufficient test for general intelligence – a famous example is the Chinese Room argument [49].

Alan Turing, when he formulated his test, was confronted with people who believed AI was impossible, and he wanted to prove the existence of an intelligence test for computer programs. He wanted to make the point that intelligence is defined by behavior rather than by mystical qualities, so that if a program could act like a human it should be considered as intelligent as a human. This was a bold conceptual leap for the 1950's. Clearly, however, general intelligence does not necessarily require the accurate simulation of human intelligence. It seems unreasonable to expect a computer program without a human-like body to be able to emulate a human, especially in conversations regarding body-focused topics like sex, aging, or the experience of having the flu. Certainly, humans would fail a "reverse Turing test" of emulating computer programs – humans can't even emulate pocket calculators without unreasonably long response delays.

2.3 A Control Theory Approach to Defining Intelligence

The psychological approach to intelligence, briefly discussed above, attempts to do justice to the diverse and multifaceted nature of the notion of intelli-

Copyrighted material

gence. As one might expect, engineers have a much simpler and much more practical definition of intelligence.

The branch of engineering called control theory deals with ways to cause complex machines to yield desired behaviors. Adaptive control theory deals with the design of machines which respond to external and internal stimuli and, on this basis, modify their behavior appropriately. And the theory of intelligent control simply takes this one step further. To quote a textbook of automata theory [2]:

> *[An] automaton is said to behave "intelligently" if, on the basis of its "training" data which is provided within some context together with information regarding the desired action, it takes the correct action on other data within the same context not seen during training.*

This is the sense in which contemporary artificial intelligence programs are intelligent. They can generalize within their limited context; they can follow the one script which they are programmed to follow. Of course, this is not really general intelligence, not in the psychological sense, and not in the sense in which we mean it in this book.

On the other hand, in their treatise on robotics, [57] presented a more general definition:

> *Intelligence is the ability to behave appropriately under unpredictable conditions.*

Despite its vagueness, this criterion does serve to point out the problem with ascribing intelligence to chess programs and the like: compared to our environment, at least, the environment within which they are capable of behaving appropriately is very predictable indeed, in that it consists only of certain (simple or complex) patterns of arrangement of a very small number of specifically structured entities. The *unpredictable conditions* clause suggests the experiential and contextual aspects of Sternberg's psychological analysis of intelligence.

Of course, the concept of appropriateness is intrinsically subjective. And unpredictability is relative as well – to a creature accustomed to living in interstellar space and inside stars and planets as well as on the surfaces of planets, or to a creature capable of living in 10 dimensions, our environment might seem just as predictable as the universe of chess seems to us. In order to make this folklore definition precise, one must first of all confront the vagueness inherent in the terms "appropriate" and "unpredictable".

In some of our own past work [17], we have worked with a variant of the Winkless and Browning definition,

> *Intelligence is the ability to achieve complex goals in complex environments.*

Copyrighted material

In a way, like the Winkless and Browning definition, this is a subjective rather than objective view of intelligence, because it relies on the subjective identification of what is and is not a complex goal or a complex environment. Behaving "appropriately", as Winkless and Browning describe, is a matter of achieving organismic goals, such as getting food, water, sex, survival, status, etc. Doing so under unpredictable conditions is one thing that makes the achievement of these goals complex.

Marcus Hutter, in his chapter in this volume, gives a rigorous definition of intelligence in terms of algorithmic information theory and sequential decision theory. Conceptually, his definition is closely related to the "achieve complex goals" definition, and it's possible the two could be equated if one defined *achieve*, *complex* and *goals* appropriately.

Note that none of these approaches to defining intelligence specify any particular properties of the *internals* of intelligent systems. This is, we believe, the correct approach: "intelligence" is about what, not how. However, it is possible that what implies how, in the sense that there may be certain structures and processes that are necessary aspects of any sufficiently intelligent system. Contemporary psychological and AI science are nowhere near the point where such a hypothesis can be verified or refuted.

2.4 Efficient Intelligence

Pei Wang, a contributor to this volume, has proposed his own definition of intelligence, which posits, basically, that "Intelligence is the ability to work and adapt to the environment with insufficient knowledge and resources." More concretely, he believes that an intelligent system is one that works under the *Assumption of Insufficient Knowledge and Resources* (AIKR), meaning that the system must be, at the same time,

A finite system The system's computing power, as well as its working and storage space, is limited.

A real-time system The tasks that the system has to process, including the assimilation of new knowledge and the making of decisions, can arrive at any time, and all have deadlines attached with them.

An ampliative system The system not only can retrieve available knowledge and derive sound conclusions from it, but also can make refutable hypotheses and guesses based on it when no certain conclusion can be drawn.

An open system No restriction is imposed on the relationship between old knowledge and new knowledge, as long as they are representable in the system's interface language.

A self-organized system The system can accommodate itself to new knowledge, and adjust its memory structure and mechanism to improve its time and space efficiency, under the assumption that future situations will be similar to past situations.

Copyrighted material

Wang's definition[4] is not purely behavioral: it makes judgments regarding the internals of the AI system whose intelligence is being assessed. However, the biggest difference between this and the above definitions is its emphasis on the limitation of the system's computing power. For instance, Marcus Hutter's AIXI algorithm, described in this volume, assumes infinite computing power (though his related AIXItl algorithm works with finite computing power). According to Wang's definition, AIXI is therefore unintelligent. Yet, AIXI can solve any problem at least as effectively as any finite-computing-power-based AI system, so it seems in a way unintuitive to call it "unintelligent".

We believe that what Wang's definition hints at is a new concept, that we call *efficient intelligence*, defined as:

> *Efficient intelligence is the ability to achieve intelligence using severely limited resources.*

Suppose we had a computer IQ test called the CIQ. Then, we might say that an AGI program with a CIQ of 500 running on 5000 machines has more intelligence, but less efficient-intelligence, than a machine with a CIQ of 100 that runs on just one machine.

According to the "achieving complex goals in complex environments" criterion, AIXI and AIXItl are the most intelligent programs described in this book, but not the ones with the highest efficient intelligence. According to Wang's definition of intelligence, AIXI and AIXItl are not intelligent at all, they only emulate intelligence through simple, inordinately wasteful program-search mechanisms.

As editors, we have not sought to impose a common understanding of the nature of intelligence on all the chapter authors. We have merely requested that authors be clear regarding the concept of intelligence under which they have structured their work. At this early stage in the AGI game, the notion of intelligence most appropriate for AGI work is still being discovered, along with the exploration of AGI theories, designs and programs themselves.

3 The Abstract Theory of General Intelligence

One approach to creating AGI is to formalize the problem mathematically, and then seek a solution using the tools of abstract mathematics. One may begin by formalizing the notion of intelligence. Having defined intelligence, one may then formalize the notion of computation in one of several generally-accepted ways, and ask the rigorous question: How may one create intelligent computer programs? Several researchers have taken this approach in recent years, and while it has not provided a panacea for AGI, it has yielded some

[4] In more recent work, Wang has modified the details of this definition, but the theory remains the same.

Copyrighted material

very interesting results, some of the most important ones are described in Hutter's and Schmidhuber's chapters in this book.

From a mathematical point of view, as it turns out, it doesn't always matter so much exactly how you define intelligence. For many purposes, any definition of intelligence that has the general form *"Intelligence is the maximization of a certain quantity, by a system interacting with a dynamic environment"* can be handled in roughly the same way. It doesn't always matter exactly what the quantity being maximized is (whether it's "complexity of goals achieved", for instance, or something else).

Let's use the term "behavior-based maximization criterion" to characterize the class of definitions of intelligence indicated in the previous paragraphs. Suppose one has some particular behavior-based maximization criterion in mind – then Marcus Hutter's work on the AIXI system, described in his chapter here, gives a software program that will be able to achieve intelligence according to the given criterion. Now, there's a catch: this program may require infinite memory and an infinitely fast processor to do what it does. But he also gives a variant of AIXI which avoids this catch, by restricting attention to programs of bounded length l and bounded time t. Loosely speaking, the AIXItl variant will provably be as intelligent as any other computer program of length up to l, satisfying the maximization criterion, within a constant multiplicative factor and a constant additive factor.

Hutter's work draws on a long tradition of research in statistical learning theory and algorithmic information theory, mostly notably Solomonoff's early work on induction [51, 52] and Levin's [39, 40] work on computational measure theory. At the present time, this work is more exciting theoretically than pragmatically. The "constant factor" in his theorem may be very large, so that, in practice, AIXItl is not really going to be a good way to create an AGI software program. In essence, what AIXItl is doing is searching the space of all programs of length L, evaluating each one, and finally choosing the best one and running it. The "constant factors" involved deal with the overhead of trying every other possible program before hitting on the best one!

A simple AI system behaving somewhat similar to AIXItl could be built by creating a program with three parts:

- the data store;
- the main program;
- the meta-program.

The operation of the meta-program would be, loosely, as follows:

- At time t, place within the data store a record containing the complete internal state of the system, and the complete sensory input of the system.

Copyrighted material

- Search the space of all programs P of size $|P| < l$ to find the one that, based on the data in the data store, has the highest expected value for the given maximization criterion.[5]
- Install P as the main program.

Conceptually, the main value of this approach for AGI is that it solidly establishes the following contention:

> ***If*** *you accept any definition of intelligence of the general form "maximization of a certain function of system behavior,"*
> ***then*** *the problem of creating AGI is basically a problem of dealing with the issues of space and time efficiency.*

As with any mathematics-based conclusion, the conclusion only follows if one accepts the definitions. If someone's conception of intelligence fundamentally can't be cast into the form of a behavior-based maximization criterion, then these ideas aren't relevant for AGI as that person conceives it. However, we believe that the behavior-based maximization criterion approach to defining intelligence is a good one, and hence we believe that Hutter's work is highly significant.

The limitations of these results are twofold. Firstly, they pertain only to AGI in the "massive computational resources" case, and most AGI theorists feel that this case is not terribly relevant to current practical AGI research (though, Schmidhuber's OOPS work represents a serious attempt to bridge this gap). Secondly, their applicability to the physical universe, even in principle, relies on the Church-Turing Thesis. The editors and contributors of this volume are Church-Turing believers, as are nearly all computer scientists and AI researchers, but there are well-known exceptions such as Roger Penrose. If Penrose and his ilk are correct, then the work of Hutter and his colleagues is not necessarily informative about the nature of AGI in the physical universe.

For instance, consider Penrose's contention that non-Turing quantum gravity computing (as allowed by an as-yet unknown incomputable theory of quantum gravity) is necessary for true general intelligence [44]. This idea is not refuted by Hutter's results, because it's possible that:

- AGI is in principle possible on ordinary Turing hardware;
- AGI is only pragmatically possible, given the space and time constraints imposed on computers by the physical universe, given quantum gravity powered computer hardware.

The authors very strongly doubt this is the case, and Penrose has not given any convincing evidence for such a proposition, but our point is merely that in spite of recent advances in AGI theory such as Hutter's work, we have

[5] There are some important details here; for instance, computing the "expected value" using probability theory requires assumption of an appropriate prior distribution, such as Solomonoff's universal prior.

Copyrighted material

no way of ruling such a possibility out mathematically. At points such as this, uncertainties about the fundamental nature of mind and universe rule out the possibility of a truly definitive theory of AGI.

From the perspective of computation theory, most of the chapters in this book deal with ways of achieving *reasonable degrees of intelligence given reasonable amounts of space and time resources*. Obviously, this is what the human mind/brain does. The amount of intelligence it achieves is clearly limited by the amount of space in the brain and the speed of processing of neural wetware.

We do not yet know whether the sort of mathematics used in Hutter's work can be made useful for defining practical AGI systems that operate within our current physical universe – or, better yet, on current or near-future computer hardware. However, research in this direction is proceeding vigorously. One exciting project in this area is Schmidhuber's OOPS system [48], which is a bit like AIXItl, but has the capability of operating with realistic efficiency in some practical situations. As Schmidhuber discusses in his first chapter in this book, OOPS has been applied to some classic AI problems such as the Towers of Hanoi problem, with highly successful results.

The basic idea of OOPS is to run all possible programs, but interleaved rather than one after the other. In terms of the "meta-program" architecture described above, here one has a meta-program that doesn't run each possible program one after the other, but rather lines all the possible programs up in order, assigns each one a probability, and then at each time step chooses a single program as the "current program", with a probability proportional to its estimated value at achieving the system goal, and then executes one step of the current program. Another important point is that OOPS freezes solutions to previous tasks, and may reuse them later.

As opposed to AIXItl, this strategy allows, in the average case, brief and effective programs to rise to the top of the heap relatively quickly. The result, in at least some practical problem-solving contexts, is impressive. Of course, there are many ways to solve the Towers of Hanoi problem. Scaling up from toy examples to real AGI on the human scale or beyond is a huge task for OOPS as for other approaches showing limited narrow-AI success. But having made the leap from abstract algorithmic information theory to limited narrow-AI success is no small achievement.

Schmidhuber's more recent Gödel Machine, which is fully self-referential, is in principle capable of proving and subsequently exploiting performance improvements to its own code. The ability to modify its own code allows the Gödel Machine to be more effective. Gödel Machines are also more flexible in terms of the utility function they aim to maximize while searching.

Lukasz Kaiser's chapter follows up similar themes to Hutter's and Schmidhuber's work. Using a slightly different computational model, Kaiser also takes up the algorithmic-information-theory motif, and describes a program search problem which is solved through the combination of program construction

Copyrighted material

and the proof search – the program search algorithm itself, represented as a directed acyclic graph, is continuously improved.

4 Toward a Pragmatic Logic

One of the primary themes in the history of AI is formal logic. However, there are strong reasons to believe that classical formal logic is not suitable to play a central role in an AGI system. It has no natural way to deal with uncertainty, or with the fact that different propositions may be based on different amounts of evidence. It leads to well-known and frustrating logical paradoxes. And it doesn't seem to come along with any natural "control strategy" for navigating the combinatorial explosion of possible valid inferences.

Some modern AI researchers have reacted to these shortcomings by rejecting the logical paradigm altogether; others by creating modified logical frameworks, possessing more of the flexibility and fluidity required of components of an AGI architecture.

One of the key issues dividing AI researchers is the degree to which logical reasoning is fundamental to their artificial minds. Some AI systems are built on the assumption that basically every aspect of mental process should be thought about as a kind of logical reasoning. Cyc is an example of this, as is the NARS system reviewed in this volume. Other systems are built on the premise that logic is irrelevant to the task of mind-engineering, that it is merely a coarse, high-level description of the results of mental processes that proceed according to non-logical dynamics. Rodney Brooks' work on subsumption robotics fits into this category, as do Peter Voss's and Hugo de Garis's neural net AGI designs presented here. And there are AI approaches, such as Novamente, that assign logic an important but non-exclusive role in cognition – Novamente has roughly two dozen cognitive processes, of which about one-fourth are logical in nature.

One fact muddying the waters somewhat is the nebulous nature of "logic" itself. *Logic* means different things to different people. Even within the domain of formal, mathematical logic, there are many different kinds of logic, including forms like fuzzy logic that encompass varieties of reasoning not traditionally considered "logical". In our own work we have found it useful to adopt a very general conception of logic, which holds that logic:

- has to do with forming and combining estimations of the (possibly probabilistic, fuzzy, etc.) truth values of various sorts of relationships based on various sorts of evidence;
- is based on incremental processing, in which pieces of evidence are combined step by step to form conclusions, so that at each stage it is easy to see which pieces of evidence were used to give which conclusion

This conception differentiates logic from mental processing in general, but it includes many sorts of reasoning besides typical, crisp, mathematical logic.

Copyrighted material

The most common form of logic is predicate logic, as used in Cyc, in which the basic entity under consideration is the *predicate*, a function that maps argument variables into Boolean truth values. The argument variables are quantified universally or existentially. An alternate form of logic is term logic, which predates predicate logic, dating back at least to Aristotle and his notion of the syllogism. In term logic, the basic element is a subject-predicate statement, denotable as $A \rightarrow B$, where $\rightarrow$ denotes a notion of inheritance or specialization. Logical inferences take the form of *syllogistic rules*, which give patterns for combining statements with matching terms, such as the deduction rule

$$(A \rightarrow B \wedge B \rightarrow C) \Rightarrow A \rightarrow C.$$

The NARS system described in this volume is based centrally on term logic, and the Novamente system makes use of a slightly different variety of term logic. Both predicate and term logic typically use variables to handle complex expressions, but there are also variants of logic, based on combinatory logic, that avoid variables altogether, relying instead on abstract structures called "higher-order functions" [10].

There are many different ways of handling uncertainty in logic. Conventional predicate logic treats statements about uncertainty as predicates just like any others, but there are many varieties of logic that incorporate uncertainty at a more fundamental level. Fuzzy logic [59, 60] attaches fuzzy truth values to logical statements; probabilistic logic [43] attaches probabilities; NARS attaches degrees of uncertainty, etc. The subtle point of such systems is the transformation of uncertain truth values under logical operators like AND, OR and NOT, and under existential and universal quantification.

And, however one manages uncertainty, there are also multiple varieties of speculative reasoning. Inductive [4], abductive [32] and analogical reasoning [31] are commonly discussed. Nonmonotonic logic [8] handles some types of nontraditional reasoning in a complex and controversial way. In ordinary, monotonic logic, the truth of a proposition does not change when new information (axioms) is added to the system. In nonmonotonic logic, on the other hand, the truth of a proposition may change when new information (axioms) is added to or old information is deleted from the system. NARS and Novamente both use logic in an uncertain and nonmonotonic way.

Finally, there are special varieties of logic designed to handle special types of reasoning. There are temporal logics designed to handle reasoning about time, spatial logics for reasoning about space, and special logics for handling various kinds of linguistic phenomena. None of the approaches described in this book makes use of such special logics, but it would be possible to create an AGI approach with such a focus. Cyc comes closest to this notion, as its reasoning engine involves a number of specialized reasoning engines oriented toward particular types of inference such as spatial, temporal, and so forth.

Copyrighted material

When one gets into the details, the distinction between logical and non-logical AI systems can come to seem quite fuzzy. Ultimately, an uncertain logic rule is not that different from the rule governing the passage of *activation* through a node in a neural network. Logic can be cast in terms of semantic networks, as is done in Novamente; and in that case uncertain logic formulas are arithmetic formulas that take in numbers associated with certain nodes and links in a graph, and output numbers associated with certain other nodes and links in the graph. Perhaps a more important distinction than logical vs. non-logical is whether a system gains its knowledge experientially or via being given *expert rule* type propositions. Often logic-based AI systems are fed with knowledge by human programmers, who input knowledge in the form of textually-expressed logic formulas. However, this is not a necessary consequence of the use of logic. It is quite possible to have a logic-based AI system that forms its own logical propositions by experience. On the other hand, there is no existing example of a non-logical AI system that gains its knowledge from explicit human knowledge encoding. NARS and Novamente are both (to differing degrees) logic-based AI systems, but their designs devote a lot of attention to the processes by which logical propositions are formed based on experience, which differentiates them from many traditional logic-based AI systems, and in a way brings them closer to neural nets and other traditional non-logical AI systems.

5 Emulating the Human Brain

One almost sure way to create artificial general intelligence would be to exactly copy the human brain, down to the atomic level, in a digital simulation. Admittedly, this would require brain scanners and computer hardware far exceeding what is currently available. But if one charts the improvement curves of brain scanners and computer hardware, one finds that it may well be plausible to take this approach sometime around 2030-2050. This argument has been made in rich detail by Ray Kurzweil in [34, 35]; and we find it a reasonably convincing one. Of course, projecting the future growth curves of technologies is a very risky business. But there's very little doubt that creating AGI in this way is physically possible.

In this sense, creating AGI is "just an engineering problem." We know that general intelligence is possible, in the sense that humans – particular configurations of atoms – display it. We just need to analyze these atom configurations in detail and replicate them in the computer. AGI emerges as a special case of nanotechnology and in silico physics.

Perhaps a book on the same topic as this one, written in 2025 or so, will contain detailed scientific papers pursuing the detailed-brain-simulation approach to AGI. At present, however, it is not much more than a futuristic speculation. We don't understand enough about the brain to make detailed simulations of brain function. Our brain scanning methods are improving

Copyrighted material

rapidly but at present they don't provide the combination of temporal and spatial acuity required to really map thoughts, concepts, percepts and actions as they occur in human brains/minds.

It's still possible, however, to use what we know about the human brain to structure AGI designs. This can be done in many different ways. Most simply, one can take a neural net based approach, trying to model the behavior of nerve cells in the brain and the emergence of intelligence therefrom. Or one can proceed at a higher level, looking at the general ways that information processing is carried out in the brain, and seeking to emulate these in software.

Stephen Grossberg [25, 28] has done extensive research on the modeling of complex neural structures. He has spent a great deal of time and effort in creating cognitively-plausible neural structures capable of spatial perception, shape detection, motion processing, speech processing, perceptual grouping, and other tasks. These complex brain mechanisms were then used in the modeling of learning, attention allocation and psychological phenomena like schizophrenia and hallucinations.

From the experiences modeling different aspects of the brain and the human neural system in general, Grossberg has moved on to the linking between those neural structures and the mind [26, 27, 28]. He has identified two key computational properties of the structures: *complementary computing* and *laminar computing*.

Complementary computing is the property that allows different processing streams in the brain to compute complementary properties. This leads to a hierarchical resolution of uncertainty, which is mostly evident in models of the visual cortex. The complementary streams in the neural structure interact, in parallel, resulting in more complete information processing. In the visual cortex, an example of complementary computing is the interaction between the *what* cortical stream, which learns to recognize what events and objects occur, and the *where* cortical stream, which learns to spacially locate those events and objects.

Laminar computing refers to the organization of the cerebral cortex (and other complex neural structures) in layers, with interactions going bottom-up, top-down, and sideways. While the existence of these layers has been known for almost a century, the contribution of this organization for control of behavior was explained only recently. [28] has recently shed some light on the subject, showing through simulations that laminar computing contributes to learning, development and attention control.

While Grossberg's research has not yet described complete minds, only neural models of different parts of a mind, it is quite conceivable that one could use his disjoint models as building blocks for a complete AGI design. His recent successes explaining, to a high degree of detail, how mental processes can emerge from his neural models is definitely encouraging.

Steve Grand's Creatures [24] are social agents, but they have an elaborate internal architecture, based on a complex neural network which is divided into several lobes. The original design by Grand had explicit AGI goals, with

Copyrighted material

attention paid to allow for symbol grounding, generalization, and limited language processing. Grand's creatures had specialized lobes to handle verbal input, and to manage the creature's internal state (which was implemented as a simplified biochemistry, and kept track of feelings such as pain, hunger and others). Other lobes were dedicated to adaptation, goal-oriented decision making, and learning of new concepts.

Representing the neural net approach in this book, we have Peter Voss's paper on the a2i2 architecture. a2i2 is in the vein of other modern work on reinforcement learning, but it is unique in its holistic architecture focused squarely on AGI. Voss uses several different reinforcement and other learning techniques, all acting on a common network of artificial neurons and synapses. The details are original, but are somewhat inspired by prior neural net AI approaches, particularly the "neural gas" approach [41], as well as objectivist epistemology and cognitive psychology. Voss's theory of mind abstracts what would make brains intelligent, and uses these insights to build artificial brains.

Voss's approach is incremental, involving a gradual progression through the "natural" stages in the complexity of intelligence, as observed in children and primates – and, to some extent, recapitulating evolution. Conceptually, his team is adding ever more advanced levels of cognition to its core design, somewhat resembling both Piagetian stages of development, as well as the evolution of primates, a level at which Voss considers there is enough complexity on the neuro-cognitive systems to provide AGI with useful metaphors and examples.

His team seeks to build ever more complex virtual primates, eventually reaching the complexity and intelligence level of humans. But this metaphor shouldn't be taken too literally. The perceptual and action organs of their initial proto-virtual-ape are not the organs of a physical ape, but rather visual and acoustic representations of the Windows environment, and the ability to undertake simple actions within Windows, as well as various *probes* for interaction with the real world through vision, sound, etc.

There are echoes of Rodney Brooks's subsumption robotics work, the well-known Cog project at MIT [1], in the a2i2 approach. Brooks is doing something a lot more similar to actually building a virtual cockroach, with a focus on the robot body and the pragmatic control of it. Voss's approach to AI could easily be nested inside robot bodies like the ones constructed by Brooks's team; but Voss doesn't believe the particular physical embodiment is the key, he believes that the essence of experience-based reinforcement learning can be manifested in a system whose inputs and outputs are "virtual."

6 Emulating the Human Mind

Emulating the atomic structure of the brain in a computer is one way to let the brain guide AGI; creating virtual neurons, synapses and activations is another. Proceeding one step further up the ladder of abstraction, one has

Copyrighted material

approaches that seek to emulate the overall architecture of the human brain, but not the details by which this architecture is implemented. Then one has approaches that seek to emulate the human mind, as studied by cognitive psychologists, ignoring the human mind's implementation in the human brain altogether.

Traditional logic-based AI clearly falls into the "emulate the human mind, not the human brain" camp. We actually have no representatives of this approach in the present book; and so far as we know, the only current research that could fairly be described as lying in the intersection of traditional logic-based AI and AGI is the Cyc project, briefly mentioned above.

But traditional logic-based AI is far from the only way to focus on the human mind. We have several contributions in this book that are heavily based on cognitive psychology and its ideas about how the mind works. These contributions pay greater than zero attention to neuroscience, but they are clearly more mind-focused than brain-focused.

Wang's NARS architecture, mentioned above, is the closest thing to a formal logic based system presented in this book. While it is not based specifically on any one cognitive science theory, NARS is clearly closely motivated by cognitive science ideas; and at many points in his discussion, Wang cites cognitive psychology research supporting his ideas.

Next, Hoyes's paper on 3D vision as the key to AGI is closely inspired by the human mind and brain, although it does not involve neural nets or other micro-level brain-simulative entities. Hoyes is not proposing to copy the precise wiring of the human visual system in silico and use it as the core of an AGI system, but he is proposing that we should copy what he sees as the basic architecture of the human mind. In a daring and speculative approach, he views the ability to deal with changing 3D scenes as the essential capability of the human mind, and views other human mental capabilities largely as offshoots of this. If this theory of the human mind is correct, then one way to achieve AGI is to do as Hoyes suggests and create a robust capability for 3D simulation, and build the rest of a digital mind centered around this capability.

Of course, even if this speculative analysis of the human mind is correct, it doesn't intrinsically follow that 3D simulation centric approach is the only approach to AGI. One could have a mind centered around another sense, or a mind that was more cognitively rather than perceptually centered. But Hoyes' idea is that we already have one example of a thinking machine – the human brain – and it makes sense to use as much of it as we can in designing our new digital intelligences.

Eliezer Yudkowsky, in his chapter, describes the conceptual foundations of his AGI approach, which he calls "deliberative general intelligence" (DGI). While DGI-based AGI is still at the conceptual-design phase, a great deal of analysis has gone into the design, so that DGI essentially amounts to an original and detailed cognitive-science theory, crafted with AGI design in mind. The DGI theory was created against the backdrop of Yudkowsky's futurist thinking, regarding the notions of:

Copyrighted material

- a *Seed AI*, an AGI system that progressively modifies and improves its own codebase, thus projecting itself gradually through exponentially increasing levels of intelligence; [58]
- a *Friendly AI*, an AGI system that respects positive ethics such as the preservation of human life and happiness, through the course of its progressive self-improvements.

However, the DGI theory also may stand alone, independently of these motivating concepts.

The essence of DGI is a functional decomposition of general intelligence into a complex supersystem of interdependent internally specialized processes. Five successive levels of functional organization are posited:

Code The source code underlying an AI system, which Yudkowsky views as roughly equivalent to neurons and neural circuitry in the human brain.

Sensory modalities In humans: sight, sound, touch, taste, smell. These generally involve clearly defined stages of information-processing and feature-extraction. An AGI may emulate human senses or may have different sorts of modalities.

Concepts *Categories* or *symbols* abstracted from a system's experiences. The process of abstraction is proposed to involve the recognition and then reification of a similarity within a group of experiences. Once reified, the common quality can then be used to determine whether new mental imagery satisfies the quality, and the quality can be imposed on a mental image, altering it.

Thoughts Conceived of as being built from structures of concepts. By imposing concepts in targeted series, the mind builds up complex mental images within the workspace provided by one or more sensory modalities. The archetypal example of a thought, according to Yudkowsky, is a human *sentence* – an arrangement of concepts, invoked by their symbolic tags, with internal structure and targeting information that can be reconstructed from a linear series of words using the constraints of syntax, constructing a complex mental image that can be used in reasoning. Thoughts (and their corresponding mental imagery) are viewed as disposable one-time structures, built from reusable concepts, that implement a non-recurrent mind in a non-recurrent world.

Deliberation Implemented by sequences of thoughts. This is the *internal narrative* of the conscious mind – which Yudkowsky views as the core of intelligence both human and digital. It is taken to include explanation, prediction, planning, design, discovery, and the other activities used to solve knowledge problems in the pursuit of real-world goals.

Yudkowsky also includes an interesting discussion of probable differences between humans and AI's. The conclusion of this discussion is that, eventually, AGI's will have many significant advantages over biological intelligences. The lack of motivational peculiarities and cognitive biases derived from an

Copyrighted material

8 The Social Nature of Intelligence

All the AI approaches discussed so far essentially view the mind as something associated with a single organism, a single computational system. Social psychologists, however, have long recognized that this is just an approximation. In reality the mind is social – it exists, not in isolated individuals, but in individuals embedded in social and cultural systems.

One approach to incorporating the social aspect of mind is to create individual AGI systems and let them interact with each other. For example, this is an important part of the Novamente AI project, which involves a special language for Novamente AI systems to use to interact with each other. Another approach, however, is to consider sociality at a more fundamental level, and to create systems from the get-go that are at least as social as they are intelligent.

One example of this sort of approach is Steve Grand's neural-net architecture as embodied in the Creatures game [24]. His neural net based creatures are intended to grow more intelligent by interacting with each other – struggling with each other, learning to outsmart each other, and so forth.

John Holland's classifier systems [30] are another example of a multi-agent system in which competition and cooperation are both present. In a classifier system, a number of rules co-exist in the system at any given moment. The system interacts with an external environment, and must react appropriately to the stimuli received from the environment. When the system performs the appropriate actions for a given perception, it is rewarded. While the individuals in Holland's system are quite primitive, recent work by Eric Baum [5] has used a similar metaphor with more complex individuals, and promising results on some large problems.

In order to decide how to answer to the perceived stimuli, the system will perform multiple rounds of competition, during which the rules bid to be activated. The winning rule will then either perform an internal action, or an external one. Internal actions change the system's internal state and affect the next round of bidding, as each rule's right to bid (and, in some variations, the amount it bids) depends on how well it matches the system's current state.

Eventually, a rule will be activated that will perform an external action, which may trigger reward from the environment. The reward is then shared by all the rules that have been active since the stimuli were perceived. The credit assignment algorithm used by Holland is called *bucket brigade*. Rules that receive rewards can bid higher in the next rounds, and are also allowed to reproduce, which results in the creation of new rules.

Another important example of social intelligence is presented in the research inspired by social insects. *Swarm Intelligence* [6] is the term that generically describes such systems. Swarm Intelligence systems are a new class of biologically inspired tools.

These systems are self-organized, relying on direct and indirect communication between agents to lead to emergent behavior. Positive feedback is

Copyrighted material

given by this communication (which can take the form of a dance indicating the direction of food in bee colonies, or pheromone trails in ant societies), which biases the future behavior of the agents in the system. These systems are naturally stochastic, relying on multiple interactions and on a random, exploratory component. They often display highly adaptive behavior to a dynamic environment, having thus been applied to dynamic network routing [9]. Given the simplicity of the individual agents, Swarm Intelligence showcases the value of cooperative emergent behavior in an impressive way.

Ant Colony Optimization [11] is the most popular form of Swarm Intelligence. ACO was initially designed as a heuristic for NP-hard problems [12], but has since been used in a variety of settings. The original version of ACO was developed to solve the famous Traveling Salesman problem. In this scenario, the environment is the graph describing the cities and their connections, and the individual agents, called *ants*, travel in the graph.

Each ant will do a tour of the cities in the graph, iteratively. At each city it will choose the next city to visit, based on a *transition rule*. This rule considers the amount of pheromone in the links connecting the current city and each of the possibilities, as well as a small random component. When the ant completes its tour, it updates the pheromone trail in the links it has used, laying an amount of pheromone proportional to the quality of the tour it has completed. The new trail will then influence the choices of the ants in the next iteration of the algorithm.

Finally, an important contribution from Artificial Life research is the *Animat* approach. Animats are biologically-inspired simulated or real robots, which exhibit adaptive behavior. In several cases [33] animats have been evolved to display reasonably complex artificial nervous systems capable of learning and adaptation. Proponents of the Animat approach argue that AGI is only reachable by embodied autonomous agents which interact on their own with their environments, and possibly other agents. This approach places an emphasis on the developmental, morphological and environmental aspects of the process of AI creating.

Vladimir Red'ko's self-organizing agent-system approach also fits partially into this general category, having some strong similarities to Animat projects. He defines a large population of simple agents guided by simple neural networks. His chapter describes two models for these agents. In all cases, the agents live in a simulated environment in which they can move around, looking for resources, and they can mate – mating uses the typical genetic operators of uniform crossover and mutation, which leads to the evolution of the agent population.

In the simpler case, agents just move around and eat virtual food, accumulating resources to mate. The second model in Red'ko's work simulates more complex agents. These agents communicate with each other, and modify their behavior based on their experience. None of the agents individually are all that clever, but the population of agents as a whole can demonstrate some interesting collective behaviors, even in the initial, relatively simplistic

Copyrighted material

implementation. The agents communicate their knowledge about resources in different points of the environment, thus leading to the emergence of adaptive behavior.

9 Integrative Approaches

We have discussed a number of different approaches to AGI, each of which has – at least based on a cursory analysis – strengths and weaknesses compared to the others. This gives rise to the idea of integrating several of the approaches together, into a single AGI system that embodies several different approaches.

Integrating different ideas and approaches regarding something as complex and subtle as AGI is not a task to be taken lightly. It's quite possible to integrate two good ideas and obtain a bad idea, or to integrate two good software systems and get a bad software system. To successfully integrate different approaches to AGI requires deep reflection on all the approaches involved, and unification on the level of conceptual foundations as well as pragmatic implementation.

Several of the AGI approaches described in this book are integrative to a certain extent. Voss's a2i2 system integrates a number of different neural-net-oriented learning algorithms on a common, flexible neural-net-like data structure. Many of the algorithms he integrated have been used before, but only in an isolated way, not integrated together in an effort to make a "whole mind." Wang's NARS-based AI design is less strongly integrative, but it still may be considered as such. It posits the NARS logic as the essential core of AI, but leaves room for integrating more specialized AI modules to deal with perception and action. Yudkowsky's DGI framework is integrative in a similar sense: it posits a particular overall architecture, but leaves some room for insights from other AI paradigms to be used in filling in roles within this architecture.

By far the most intensely integrative AGI approach described in the book, however, is our own Novamente AI approach.

The Novamente AI Engine, the work of the editors of this volume and their colleagues, is in part an original system and in part an integration of ideas from prior work on narrow AI and AGI. The Novamente design incorporates aspects of many previous AI paradigms such as genetic programming, neural networks, agent systems, evolutionary programming, reinforcement learning, and probabilistic reasoning. However, it is unique in its overall architecture, which confronts the problem of creating a holistic digital mind in a direct and ambitious way.

The fundamental principles underlying the Novamente design derive from a novel complex-systems-based theory of mind called the *psynet model*, which was developed in a series of cross-disciplinary research treatises published during 1993-2001 [17, 16, 18, 19, 20]. The psynet model lays out a series of properties that must be fulfilled by any software system if it is going to be an

Copyrighted material

autonomous, self-organizing, self-evolving system, with its own understanding of the world, and the ability to relate to humans on a *mind-to-mind* rather than a *software-program-to-mind* level. The Novamente project is based on many of the same ideas that underlay the Webmind AI Engine project carried out at Webmind Inc. during 1997-2001 [23]; and it also draws to some extent on ideas from Pei Wang's Non-axiomatic Reasoning System (NARS) [54].

At the moment, a complete Novamente design has been laid out in detail [21], but implementation is only about 25% complete (and of course many modifications will be made to the design during the course of further implementation). It is a C++ software system, currently customized for Linux clusters, with a few externally-facing components written in Java. The overall mathematical and conceptual design of the system is described in a paper [22] and a forthcoming book [21]. The existing codebase implements roughly a quarter of the overall design. The current, partially-complete codebase is being used by the startup firm Biomind LLC, to analyze genetics and proteomics data in the context of information integrated from numerous biological databases. Once the system is fully engineered, the project will begin a phase of interactively teaching the Novamente system how to respond to user queries, and how to usefully analyze and organize data. The end result of this teaching process will be an autonomous AGI system, oriented toward assisting humans in collectively solving pragmatic problems.

10 The Outlook for AGI

The AGI subfield is still in its infancy, but it is certainly encouraging to observe the growing attention that it has received in the past few years. Both the number of people and research groups working on systems designed to achieve general intelligence and the interest from outsiders have been growing.

Traditional, narrow AI does play a key role here, as it provides useful examples, inspiration and results for AGI. Several such examples have been mentioned in the previous sections in connection with one or another AGI approach. Innovative ideas like the application of complexity and algorithmic information theory to the mathematical theorization of intelligence and AI provide valuable ground for AGI researchers. Interesting ideas in logic, neural networks and evolutionary computing provide both tools for AGI approaches and inspiration for the design of key components, as will be seen in several chapters of this book.

The ever-welcome increase in computational power and the emergence of technologies like Grid computing also contribute to a positive outlook for AGI. While it is possible that, in the not too distant future, regular desktop machines (or whatever form the most popular computing devices take 10 or 20 years from now) will be able to run AGI software comfortably, today's AGI prototypes are extremely resource intensive, and the growing availability of world-wide computing farms would greatly benefit AGI research. The

Copyrighted material

popularization of Linux, Linux-based clusters that extract considerable horsepower from stock hardware, and, finally, Grid computing, are seen as great advances, for one can never have enough CPU cycles.

We hope that the precedent set by these pioneers in AGI research will inspire young AI researchers to stray a bit off the beaten track and venture into the more daring, adventurous and riskier path of seeking the creation of truly general artificial intelligence. Traditional, narrow AI is very valuable, but, if nothing else, we hope that this volume will help create the awareness that AGI research is a very present and viable option. The complementary and related fields are mature enough, the computing power is becoming increasingly easier and cheaper to obtain, and AGI itself is ready for popularization. We could always use yet another design for an artificial general intelligence in this challenging, amazing, and yet friendly race toward the awakening of the world's first real artificial intelligence.

Acknowledgments

Thanks are due to all the authors for their well-written collaborations and patience during a long manuscript preparation process. Also, we are indebted to Shane Legg for his careful reviews and insightful suggestions.

References

1. Bryan Adams, Cynthia Breazeal, Rodney Brooks, and Brian Scassellati. Humanoid Robots: A New Kind of Tool. *IEEE Intelligent Systems*, 15(4):25–31, 2000.
2. Igor Aleksander and F. Keith Hanna. *Automata Theory: An Engineering Approach*. Edward Arnold, 1976.
3. J. R. Anderson, M. Matessa, and C. Lebiere. ACT-R: A Theory of Higher-Level Cognition and its Relation to Visual Attention. *Human Computer Interaction*, 12(4):439–462, 1997.
4. D. Angluin and C. H. Smith. Inductive Inference, Theory and Methods. *Computing Surveys*, 15(3):237–269, 1983.
5. Eric Baum and Igor Durdanovic. An Evolutionary Post Production System. 2002.
6. Eric Bonabeau, Marco Dorigo, and Guy Theraulaz. *Swarm Intelligence: From Natural to Artificial Systems*. Oxford University Press, 1999.
7. Christopher Brand. *The G-Factor: General Intelligence and its Implications*. John Wiley and Sons, 1996.
8. Gerhard Brewka, Jürgen Dix, and Kurt Konolige. *Nonmonotonic Reasoning: An Overview*. CSLI Press, 1995.
9. G. Di Caro and M. Dorigo. AntNet: A Mobile Agents Approach to Adaptive Routing. Technical Report Tech. Rep. IRIDIA/97-12, Université Libre de Bruxelles, 1997.
10. Haskell Curry and Robert Feys. *Combinatory Logic*. North-Holland, 1958.

Copyrighted material

The Logic of Intelligence

Pei Wang

Department of Computer and Information Sciences, Temple University
Philadelphia, PA 19122, USA
pei.wang@temple.edu - http://www.cis.temple.edu/~pwang/

Summary. Is there an "essence of intelligence" that distinguishes intelligent systems from non-intelligent systems? If there is, then what is it? This chapter suggests an answer to these questions by introducing the ideas behind the NARS (Non-axiomatic Reasoning System) project. NARS is based on the opinion that the essence of intelligence is the ability to adapt with insufficient knowledge and resources. According to this belief, the author has designed a novel formal logic, and implemented it in a computer system. Such a "logic of intelligence" provides a unified explanation for many cognitive functions of the human mind, and is also concrete enough to guide the actual building of a general purpose "thinking machine".

1 Intelligence and Logic

1.1 To Define Intelligence

The debate on the essence of intelligence has been going on for decades, but there is still little sign of consensus (this book itself is evidence of this).

In "mainstream AI", the following are some representative opinions:

> "AI is concerned with methods of achieving goals in situations in which the information available has a certain complex character. The methods that have to be used are related to the problem presented by the situation and are similar whether the problem solver is human, a Martian, or a computer program." [19]

> Intelligence usually means "the ability to solve hard problems". [22]

> "By 'general intelligent action' we wish to indicate the same scope of intelligence as we see in human action: that in any real situation behavior appropriate to the ends of the system and adaptive to the demands of the environment can occur, within some limits of speed and complexity." [23]

Maybe it is too early to define intelligence. It is obvious that, after decades of study, we still do not know very much about it. There are more questions than answers. Any definition based on current knowledge is doomed to be

Copyrighted material

revised by future work. We all know that a well-founded definition is usually the result, rather than the starting point, of scientific research. However, there are still reasons for us to be concerned about the definition of intelligence at the current time. Though clarifying the meaning of a concept always helps communication, this problem is especially important for AI. As a community, AI researchers need to justify their field as a scientific discipline. Without a (relatively) clear definition of intelligence, it is hard to say why AI is different from, for instance, computer science or psychology. Is there really something novel and special, or just fancy labels on old stuff? More vitally, every researcher in the field needs to justify his/her research plan according to such a definition. Anyone who wants to work on artificial intelligence is facing a two-phase task: to choose a working definition of intelligence, then to produce it in a computer.

A *working definition* is a definition concrete enough that you can directly work with it. By accepting a working definition of intelligence, it does not mean that you really believe that it fully captures the concept "intelligence", but that you will take it as a goal for your current research project.

Therefore, the lack of a consensus on what intelligence is does not prevent each researcher from picking up (consciously or not) a working definition of intelligence. Actually, unless you keep one (or more than one) definition, you cannot claim that you are working on artificial intelligence.

By accepting a working definition of intelligence, the most important commitments a researcher makes are on the acceptable assumptions and desired results, which bind all the concrete work that follows. The defects in the definition can hardly be compensated by the research, and improper definitions will make the research more difficult than necessary, or lead the study away from the original goal.

Before studying concrete working definitions of intelligence, we need to set up a general standard for what makes a definition better than others.

Carnap met the same problem when he tried to clarify the concept "probability". The task "consists in transforming a given more or less inexact concept into an exact one or, rather, in replacing the first by the second", where the first may belong to everyday language or to a previous stage in the scientific language, and the second must be given by explicit rules for its use [4].

According to Carnap, the second concept, or the *working definition* as it is called in this chapter, must fulfill the following requirements [4]:

1. It is *similar* to the concept to be defined, as the latter's vagueness permits.
2. It is defined in an *exact* form.
3. It is *fruitful* in the study.
4. It is *simple*, as the other requirements permit.

It seems that these requirements are also reasonable and suitable for our current purpose. Now let us see what they mean concretely to the working definition of intelligence:

Copyrighted material

Similarity (to standard usage). Though "intelligence" has no exact meaning in everyday language, it does have some common usages with which the working definition should agree. For instance, normal human beings are intelligent, but most animals and machines (including ordinary computer systems) are either not intelligent at all or much less intelligent than human beings.

Exactness (or well-definedness). Given the working definition, whether (or how much) a system is intelligent should be clearly decidable. For this reason, intelligence cannot be defined in terms of other ill-defined concepts, such as *mind, thinking, cognition, intentionality, rationality, wisdom, consciousness*, and so on, though these concepts do have close relationships with intelligence.

Fruitfulness (and instructiveness). The working definition should provide concrete guidelines for the research based on it – for instance, what assumptions can be accepted, what phenomena can be ignored, what properties are desired, and so on. Most importantly, the working definition of intelligence should contribute to the solving of fundamental problems in AI.

Simplicity. Although intelligence is surely a complex mechanism, the working definition should be simple. From a theoretical point of view, a simple definition makes it possible to explore a theory in detail; from a practical point of view, a simple definition is easy to use.

For our current purpose, there is no "right" or "wrong" working definition for intelligence, but there are "better" and "not-so-good" ones. When comparing proposed definitions, the four requirements may conflict with each other. For example, one definition is more fruitful, while another is simpler. In such a situation, some weighting and trade-off become necessary. However, there is no evidence showing that in general the requirements cannot be satisfied at the same time.

1.2 A Working Definition of Intelligence

Following the preparation of the previous section, we propose here a working definition of intelligence:

> *Intelligence is the capacity of a system to adapt to its environment while operating with insufficient knowledge and resources.*

The *environment* of a system may be the physical world, or other information processing systems (human or computer). In either case, the interactions can be described by the *experiences* (or *stimuli*) and *responses* of the system, which are streams of input and output information, respectively. For the system, perceivable patterns of input and producible patterns of output constitute its *interface language*.

To *adapt* means that the system learns from its experiences. It adjusts its internal structure to approach its goals, as if future situations will be similar

Copyrighted material

to past situations. Not all systems adapt to their environment. For instance, a traditional computing system gets all of its knowledge during its design phase. After that, its experience does not contribute to its behaviors. To acquire new knowledge, such a system would have to be redesigned.

Insufficient knowledge and resources means that the system works under the following restrictions:

Finite. The system has a constant information-processing capacity.
Real-time. All tasks have time requirements attached.
Open. No constraints are put on the knowledge and tasks that the system can accept, as long as they are representable in the interface language.

The two main components in the working definition, *adaptation* and *insufficient knowledge and resources*, are related to each other. An adaptive system must have some insufficiency in its knowledge and resources, for otherwise it would never need to change at all. On the other hand, without adaptation, a system may have insufficient knowledge and resources, but make no attempt to improve its capacities.

Not all systems take their own insufficiency of knowledge and resources into full consideration. Non-adaptive systems, for instance, simply ignore new knowledge in their interactions with their environment. As for artificial adaptive systems, most of them are not finite, real-time, and open, in the following senses:

1. Though all actual systems are finite, many theoretical models (for example, the Turing Machine) neglect the fact that the requirements for processor time and/or memory space may go beyond the supply capacity of the system.
2. Most current AI systems do not consider time constraints at run time. Most real-time systems can handle time constraints only if they are essentially deadlines [35].
3. Various constraints are imposed on what a system can experience. For example, only questions that can be answered by retrieval and deduction from current knowledge are acceptable, new knowledge cannot conflict with previous knowledge, and so on.

Many computer systems are designed under the assumption that their knowledge and resources, though *limited* or *bounded*, are still *sufficient* to fulfill the tasks that they will be called upon to handle. When facing a situation where this assumption fails, such a system simply panics or crashes, and asks for external intervention by a human user.

For a system to work under the assumption of insufficient knowledge and resources, it should have mechanisms to handle the following types of situation, among others:

- a new processor is required when all existent processors are occupied;
- extra memory is required when all available memory is already full;

Copyrighted material

- a task comes up when the system is busy with something else;
- a task comes up with a time requirement, so exhaustive search is not an option;
- new knowledge conflicts with previous knowledge;
- a question is presented for which no sure answer can be deduced from available knowledge.

For traditional computing systems, these types of situations usually require human intervention or else simply cause the system to refuse to accept the task or knowledge involved. However, for a system designed under the assumption of insufficient knowledge and resources, these are *normal situations*, and should be managed smoothly by the system itself. According to the above definition, intelligence is a "highly developed form of mental adaptation" [26].

When defining intelligence, many authors ignore the complementary question: what is unintelligent? If everything is intelligent, then this concept is empty. Even if we agree that intelligence, like almost all properties, is a matter of degree, we still need criteria to indicate what makes a system more intelligent than another. Furthermore, for AI to be an (independent) discipline, we require the concept "intelligence" to be different from other established concepts, because otherwise we are only talking about some well-known stuff with a new name, which is not enough to establish a *new branch of science*. For example, if every computer system is intelligent, it is better to stay within the theory of computation. Intuitively, "intelligent system" does not mean a faster and bigger computer. On the other hand, an unintelligent system is not necessarily incapable or gives only wrong results. Actually, most ordinary computer systems and many animals can do something that human beings cannot. However, these abilities do not earn the title "intelligent" for them. What is missing in these capable-but-unintelligent systems? According to the working definition of intelligence introduced previously, an *unintelligent* system is one that does not adapt to its environment. Especially, in artificial systems, an *unintelligent* system is one that is designed under the assumption that it only works on problems for which the system has sufficient knowledge and resources. An intelligent system is not always "better" than an unintelligent system for practical purposes. Actually, it is the contrary: when a problem can be solved by both of them, the unintelligent system is usually better, because it guarantees a correct solution. As Hofstadter said, for tasks like adding two numbers, a "reliable but mindless" system is better than an "intelligent but fallible" system [13].

1.3 Comparison With Other Definitions

Since it is impossible to compare the above definition to each of the existing working definitions of intelligence one by one, we will group them into several categories.

Generally speaking, research in artificial intelligence has two major motivations. As a field of science, we want to learn how the human mind, and

Copyrighted material

systems look just like ordinary computer application systems, and still suffer from great rigidity and brittleness (something AI wants to avoid).

If intelligence is defined as "the capacity to solve hard problems", then the next question is: "Hard for whom?" If we say "hard for human beings", then most existing computer software is already intelligent – no human can manage a database as well as a database management system, or substitute a word in a file as fast as an editing program. If we say "hard for computers," then AI becomes "whatever hasn't been done yet," which has been dubbed "Tesler's Theorem" [13]. The view that AI is a "perpetually extending frontier" makes it attractive and exciting, which it deserves, but tells us little about how it differs from other research areas in computer science – is it fair to say that the problems there are easy? If AI researchers cannot identify other commonalities of the problems they attack besides mere difficulty, they will be unlikely to make any progress in understanding and replicating intelligence.

To Carry out Cognitive Functions

According to this view, intelligence is characterized by a set of cognitive functions, such as reasoning, perception, memory, problem solving, language use, and so on. Researchers who subscribe to this view usually concentrate on just one of these functions, relying on the idea that research on all the functions will eventually be able to be combined, in the future, to yield a complete picture of intelligence. A "cognitive function" is often defined in a general and abstract manner. This approach has produced, and will continue to produce, tools in the form of software packages and even specialized hardware, each of which can carry out a function that is similar to certain mental skills of human beings, and therefore can be used in various domains for practical purposes. However, this kind of success does not justify claiming that it is the proper way to study AI. To define intelligence as a "toolbox of functions" has serious weaknesses.

When specified in isolation, an implemented function is often quite different from its "natural form" in the human mind. For example, to study analogy without perception leads to distorted cognitive models [5]. Even if we can produce the desired tools, this does not mean that we can easily combine them, because different tools may be developed under different assumptions, which prevents the tools from being combined.

The basic problem with the "toolbox" approach is: without a "big picture" in mind, the study of a cognitive function in an isolated, abstracted, and often distorted form simply does not contribute to our understanding of intelligence.

A common counterargument runs something like this: "Intelligence is very complex, so we have to start from a single function to make the study tractable." For many systems with weak internal connections, this is often a good choice, but for a system like the mind, whose complexity comes directly from its tangled internal interactions, the situation may be just the opposite. When the so-called "functions" are actually phenomena produced

Copyrighted material

by a complex-but-unified mechanism, reproducing all of them together (by duplicating the mechanism) is simpler than reproducing only one of them.

To Develop New Principles

According to this type of opinions, what distinguishes intelligent systems and unintelligent systems are their *postulations*, applicable *environments*, and basic *principles* of information processing.

The working definition of intelligence introduced earlier belongs to this category. As a system adapting to its environment with insufficient knowledge and resources, an intelligent system should have many cognitive *functions*, but they are better thought of as emergent phenomena than as well-defined tools used by the system. By learning from its experience, the system potentially can acquire the *capacity* to solve hard problems – actually, *hard problems* are those for which a solver (human or computer) has insufficient knowledge and resources – but it has no such built-in capacity, and thus, without proper training, no capacity is guaranteed, and acquired capacities can even be lost. Because the human mind also follows the above principles, we would hope that such a system would behave similarly to human beings, but the similarity would exist at a more abstract level than that of concrete *behaviors*. Due to the fundamental difference between human experience/hardware and computer experience/hardware, the system is not expected to accurately reproduce masses of psychological data or to pass a Turing Test. Finally, although the internal *structure* of the system has some properties in common with a description of the human mind at the subsymbolic level, it is not an attempt to simulate a biological neural network.

In summary, the *structure* approach contributes to neuroscience by building brain models, the *behavior* approach contributes to psychology by providing explanations of human behavior, the *capacity* approach contributes to application domains by solving practical problems, and the *function* approach contributes to computer science by producing new software and hardware for various computing tasks. Though all of these are valuable for various reasons, and helpful in the quest after AI, these approaches do not, in my opinion, concentrate on the *essence* of intelligence.

To be sure, what has been proposed in my definition of intelligence is not entirely new to the AI community. Few would dispute the proposition that adaptation, or learning, is essential for intelligence. Moreover, "insufficient knowledge and resources" is the focus of many subfields of AI, such as heuristic search, reasoning under uncertainty, real-time planning, and machine learning. Given this situation, what is *new* in this approach? It is the following set of principles:

1. an explicit and unambiguous definition of intelligence as "adaptation under insufficient knowledge and resources";
2. a further definition of the phrase "with insufficient knowledge and resources" as *finite*, *real-time*, and *open*;

Copyrighted material

Index

Copyrighted material

Copyrighted material

Copyrighted material

Copyrighted material

Copyrighted material