Rigid Flexibility

APPLIED LOGIC SERIES

VOLUME 34

SCOPE OF THE SERIES

Logic is applied in an increasingly wide variety of disciplines, from the traditional subjects of philosophy and mathematics to the more recent disciplines of cognitive science, computer science, artificial intelligence, and linguistics, leading to new vigor in this ancient subject. Kluwer, through its Applied Logic Series, seeks to provide a home for outstanding books and research monographs in applied logic, and in doing so demonstrates the underlying unity and applicability of logic.

The titles published in this series are listed at the end of this volume.

Rigid Flexibility

The Logic of Intelligence

by

Pei Wang
Temple University, Philadelphia, USA

Springer

A C.I.P. Catalogue record for this book is available from the Library of Congress.

ISBN-10 1-4020-5044-5 (HB)
ISBN-13 978-1-4020-5044-2 (HB)
ISBN-10 1-4020-5045-3 (e-book)
ISBN-13 978-1-4020-5045-3 (e-book)

Published by Springer,
P.O. Box 17, 3300 AA Dordrecht, The Netherlands.

www.springer.com

Printed on acid-free paper

Printed in the Netherlands

Contents

List of Tables

Preface

This book presents a research project aimed at the building of a "thinking machine," that is, a general-purpose artificial intelligence.

Artificial intelligence has a scientific and an engineering aspect. The former focuses on the explanation of "intelligence" displayed by the human mind, while the latter focuses on building a computer system that has such a nature. It is the most recently developed branch of a profound intellectual tradition, and is related to many problems studied in cognitive sciences, including philosophy, psychology, logic, linguistics, mathematics, neuroscience, and related disciplines. Ultimately, the goal is to understand notions like "intelligence," "cognition," "mind," and "thinking" well enough to reproduce them in computer systems.

Though the research field of artificial intelligence has existed for about half a century, we are still far from the goal of building a thinking machine. To a large degree, this is due to the complexity of the problem — since the mind is perhaps the most complicated phenomenon in the universe — as well as limitations of existing computer technology. However, there is also a possibility that the mainstream research in the field has been heading in the wrong direction.

The research reported in this book proposes a change in direction of the research of artificial intelligence, a "paradigm shift" *per se*. Instead of duplicating human behaviors or solving practical problems, this book proposes that the right thing to do in the research is to build systems that follow the same underlying *principle* as the human mind, that is, *to adapt to the environment and to work with insufficient knowledge and resources*.

In light of this opinion, the limitations of traditional theories are analyzed. Such theories include first-order predicate logic, model-theoretic

semantics, probability theory, computability theory, and computational complexity theory. Each of these theories makes some explicit or implicit assumptions on the sufficiency of knowledge and/or resources, and will not work when these assumptions are not satisfied. The new theory introduced in this book is designed for a situation where no traditional theory can be applied, with the belief that this is what "intelligence" is about.

This research shows that it is possible to build a formal model, which is then be implemented in a computer, by standing firm on the assumption of insufficient knowledge and resources. Furthermore, the model uses a relatively simple mechanism to uniformly reproduce many cognitive faculties, including reasoning, learning, perceiving, remembering, categorizing, planning, predicting, problem solving, and decision making.

This book consists of four parts:

Part I introduces the philosophical and methodological foundation of the *Non-Axiomatic Reasoning System*, NARS for short, project. The major schools of thought in the field are analyzed, and a new working definition of "intelligence" is proposed, according to which it is the capability of a system to adapt to its environment and to work with insufficient knowledge and resources. The choice of the reasoning system framework for this project is justified. Finally, the major components of NARS are briefly and informally introduced.

Part II describes a formal model based on the above theory. This part contains the most technical results in NARS. First, a logic system, NAL, is defined in several phases. The logic uses a categorical language, an experience-grounded semantics, and extended syllogistic inference rules. Then, the memory structure and control mechanism of the system are introduced, which let the system operate adequately while the computational resources, time and space, are in short supply.

Part III compares the above model with related approaches on several topics, and discusses the corresponding properties of NARS.

The topics include knowledge representation, semantics and interpretation, various types of inference, learning and categorization, and the control of inference processes. It is shown that an outstanding feature of NARS is that it provides a unified solution to many problems in artificial intelligence and cognitive sciences.

Part IV summarizes the conclusions reached in this research so far, and outlines the plan for the next stage of the project.

This organization is not perfect. Usually, a new technical idea should be introduced together with the problem it aims, as well as with comparisons to other related ideas. In this book, however, each idea introduced in Part II is not fully discussed until Part III, after the whole system is described. This is because that each idea in NARS is typically motivated by more than one consideration, and the solution of a problem is typically provided by the cooperation of multiple components of the system. Consequently, to fully understand part of the system without mentioning the other parts is very hard, if not impossible. Given this nature, one possible reading strategy is to read Part II quickly in the first pass, just to get a big picture of the system, then to come back to the technical details during or after reading the remaining part of the book.

This research is highly interdisciplinary. Its theoretical foundation is rooted in philosophical and psychological studies; the formal model is mainly about logic and artificial intelligence; the implementation is carried out with the tools provided by computer science.

This book is aimed at general readers interested in mind, thinking, and the computer. The readers are expected to be moderately familiar with artificial intelligence, formal logic, computer science, and cognitive sciences.

NARS is an on-going project. For up-to-date information about its progress, please visit the project website,[1] which contains on-line demonstrations, working examples, related publications, as well as additional materials and links.

[1]Currently at `http://www.cogsci.indiana.edu/farg/peiwang/NARS/`.

Acknowledgment

I became interested in artificial intelligence (AI) in the early 1980s, when I was an undergraduate student in Peking University. At that time AI was a new topic in China, so few faculty was working on it, and no course was offered on the subject. As a result, I simply read what I could find on AI and related topics. This fact partially explains why I did not follow any established approach, but tried to put together a theory by myself from the beginning of my research. The atmosphere in Peking University in those years was extraordinary — just out of the "cultural revolution", the young generation was enthusiastically, seriously, and bravely challenging traditional theories and values in all domains, and attempting to build their own. The influence of the so-called "Spirit of Peking University" was decisive to me; without it, I would not have had the confidence to work on my ideas for more than twenty years without major signs of acceptance from the AI community. Therefore, to a large extent, this research is a product of the atmosphere of "Peking University in the 1980s" — I don't think I would pursue the same path at a different place or at a different time.

As I had very limited access to established researchers and reference materials, many of my ideas were inspired by discussions with friends. Among them I want to thank Chen Gang, Cai Shan, Sun Yongping, and in particular, Bai Shuo.[2]

After developing some preliminary ideas in my thesis for my bachelor's degree, I decided to pursue them further when I became a graduate student in the same university. Professor Xu Zhuoqun (also known as Hsu Cho-Chun) found my ideas promising, and agreed to be my adviser.

[2]All Chinese names in this acknowledgment are written in the traditional Chinese order, with the family name followed by the given name.

Without his support, it would have been difficult for me to finish my earliest research as a Master Thesis. It was also from his Operating System course that I got my idea of resource management. Another member of my advisory committee, Professor Sun Huaimin of Beihang University (BUAA), was the one who triggered my interest in symbolic logic. I also obtained help from the other committee members and faculty of several universities, including Li Wei, Lin Jianxiang, Ma Xiwen, Shi Chunyi, Shi Zhongzhi, and Wang Yutian.

After I got my master's degree in July 1986, I was offered the position of Lecturer by the department. I accepted it, because Peking University was so much like a home for me. During the summer vacation that year, a fellow student, Yan Yong, persuaded me to join the translation project of Douglas Hofstadter's *Gödel, Escher, Bach: an Eternal Golden Braid*. The translation process turned out to be very exciting, and I found many of my own ideas were presented in the book in a beautiful way. As a consequence, I wrote to Professor Hofstadter and then we exchanged some research papers, at which point both of us noticed a surprising degree of overlap in our research projects. Consequently, he arranged for me to join his research group at Indiana University, and I was also accepted by the Ph.D program in computer science and cognitive science at IU.

I left Peking University (after about twelve years), and arrived in the USA in May 1991. This time I was lucky again to have an open-minded adviser. Though Professor Hofstadter did not share all my opinions about intelligence, he still provided me the support I needed to continue my research. His opinions have always been stimulating to my thought. Thanks to his help, the four and half years I spent on the peaceful Bloomington campus were very productive and pleasant. The other members of the Center for Research on Concepts and Cognition were very helpful. They include David Moser, Robert French, Terry Jones, David Chalmers, Gary McGraw, Jim Marshall, and John Rehling. Helga Keller gave me enormous help in administrative affairs. Jeff Logan played a major role in the improvement of my English. Carol Hofstadter's kindness to my family is unforgettable for us. I also want to thank the other members of my advisory committee — David Leake, Gregory Rawlins, and James Townsend — for their valuable comments and suggestions.

After getting my Ph.D. at the end of 1995, I could not get a research position anywhere, so I went into industry. From early 1996 to Spring 1998, I worked in three companies, doing applied AI work, mostly in expert systems, and continued the work in NARS in my own time. In April 1998 I was attracted to a start-up company, Intelligenesis (renamed to Webmind later), by a recruiting advertisement that looked for people with "a passion for making computers think." After talking with the co-founder (also the Chairman and CTO), Ben Goertzel, we found enough overlap in our ideas for me to join the company. One major task I finished for the company was to design a customized version of NARS as the inference engine of a general-purpose AI system, Webmind, which integrated several techniques. Though I and Ben had been arguing many issues all the time, the cooperation was stimulating and pleasant overall. I was also benefited from discussions with Jeff Pressing, Karin Verspoor, and other colleagues in the company.

In April 2001, the company finally ran out of money in a tough economical season. In Summer 2001, with the help of Fan Wenfei, I got a teaching position at the Computer and Information Sciences Department of Temple University, where I have stayed until now. At Temple, I have been getting help from Robert Aiken, Frank Friedman, John Nosek, and other faculty members.

In the recent years, my research has benefited from the discussions with the following friends: Deng Lang, Hu Xiangen, Lin Fangzhen, Lin Yunqing, Lin Zuoquan, Ju Shier, Wang Hongbin, Yang Yingrui, Zhou Beihai, as well as with many people in various mailing lists and news groups.

I knew I needed to write such a book many years ago, simply because NARS addresses too many issues in such an interwoven manner, that it is very difficult to clearly explain any of them without touching the others. My Master Thesis [Wang, 1986] and Ph.D. Dissertation [Wang, 1995a] can be seen as immature versions of this book. The current manuscript also includes many materials in my other writings (journal articles, conference papers, book chapters, and technical reports), which are included in the Bibliography of this book. Thanks to the following publishers for their kind permission for me to use my previous publications in this book: Elsevier [Wang, 1994b, Wang, 1996a, Wang, 2004a, Wang, 2005], IEEE [Wang, 1996b], Shaker Publishing

[Wang, 2001b], Springer Science and Business Media [Wang, 1993a, Wang, 1994a, Wang, 2000c, Wang, 2004c, Wang, 2006b], and World Scientific Publishing [Wang, 1995b, Wang, 2004b, Wang, 2004d].

The first version of this book was finished in the Summer of 2003. Since then, it had been rejected by several publishers until it finally got into the hands of Dov Gabbay of the Applied Logic Series, and got favorable evaluations from two anonymous reviewers. Without them, I do not know how much longer it would take for this manuscript to be published. During this process, Jane Spurr of the Applied Logic Series and Lucy Fleet of Springer provide abundant help.

Thanks to Kevin Copple, Paul Fidika, Jordan Fultz, Edward Heflin, J. W. Johnston, Klaus Witzel, and Brad Wyble for making many comments and English corrections on earlier versions of this manuscript.

Finally, I would like to thank the help from my family members, especially, from my dear wife, Sun Hongyuan, with her understanding, support, and love.

Part I

Theoretical Foundation

Chapter 1

The Goal of Artificial Intelligence

Generally speaking, Artificial Intelligence (AI) is the creation of intelligence, as displayed by the human mind, in an artificial entity, especially, a computer system.

This chapter surveys the current state of the field of AI, albeit through my personal perspective.

1.1 To define intelligence

1.1.1 The field of AI

A key characteristic that distinguishes the human being from other currently known entities (animals, machines, and so on) is "intelligence" (similar terms include "mind," "cognition," and "thinking"). Whether this capability can be understood and reproduced in machines is a question that has been considered for a long time by philosophers, mathematicians, scientists, engineers, as well as by writers and movie makers. However, it is the modern digital computer that makes it possible to seriously test various answers to this question.

The electronic computer first appeared in the 1940s. Though initially the computer was used for numerical calculations, a mental activity which previously could only be accomplished by a human mind,

soon people realized that they could carry out many other mental activities by manipulating various types of symbols or codes. Naturally, people began to wonder whether all mental activities could be carried out by computers, and if not, where does the border lie?

Roughly speaking, all attempts to answer the above questions belong to the study of "Artificial Intelligence"(AI), that is, to the attempts to produce "intelligence" in artifacts, especially, computer systems.

Toward this general goal, two motivations of AI research and development coexist:

- As a science, AI attempts to establish a theory of intelligence to explain human intellectual activities and abilities.

- As a technology, AI attempts to implement a theory of intelligence in computer systems to reproduce these activities and abilities and use them to solve practical problems.

In AI, the science aspect ("What is intelligence?") and the technology aspect ("How to reproduce intelligence?") are closely related to each other. Although different researchers may focus on different aspects of the research, a complete AI project typically consists of works on the following three levels of description:[1]

1. a *theory* of intelligence, as writings in natural languages such as English or Chinese,

2. a formal *model* of intelligence based on the above theory, as formulas and expressions in formal languages like the ones used in logic or mathematics,

3. a computer *system* implementing the above model, as programs in programming languages such as Lisp or Java. Optionally, some AI projects include works on computer hardware and special devices.

[1]Similar level distinctions are made by other authors [Marr, 1982, Newell, 1990], and a summary can be found in [Anderson, 1990, page 4]. The above level distinction differs from the others in that here it is mostly determined by the *language* in which the research results are presented, and is, therefore, mostly independent of the content of the AI approach under discussion.

Roughly speaking, the mapping between descriptions of a higher level and those of a lower level is one-to-many, in the sense that one theory may be represented in more than one model (though each model only represents one theory), and that one model may be implemented in more than one way (though each implementation only realizes one model).

Because of the nature of the field, AI is closely related to other disciplines. At the top level, AI borrows concepts and theories from the disciplines that study the various aspects of the human brain and mind, including neuroscience, psychology, linguistics, and philosophy. At the middle level, AI uses tools and models developed in mathematics, logic, and computer science. At the bottom level, AI depends on components and systems provided by computer technology, like programming language, software, and hardware.

1.1.2 The need for definition

Though the previous subsection provided a brief description of the field of AI, it does not answer a key question: What is the definition of artificial intelligence?

It is generally acknowledged that the forming of AI as a research field was signified by the Dartmouth Meeting in 1956. After half a century, there is a substantial AI community with thousands of researchers all over the world, producing many books, journals, conferences, and organizations. However, the current state of AI research activities are not bounded together by a common theoretical foundation or by a set of methods, but by a group of loosely related problems.

In the acronym "AI," the "A" part is relatively easier to define — by "artificial," we mean "artifacts," especially electric computing machinery. However, the "I" part is much harder, because the debate on the essence of intelligence has been going on since the existence of the related fields like psychology and philosophy, etc, not to mention AI, and there is still no sign of consensus.

Consider what the "founding fathers" of AI had in mind about the field:

> "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." [McCarthy, 1988]
>
> Intelligence usually means "the ability to solve hard problems". [Minsky, 1985]
>
> "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." [Newell and Simon, 1976]

The above statements clearly have something in common, but there are still differences among them. The same is also true for the definitions of intelligence in AI books and articles. In fact, almost everyone in the field has a personal opinion about how the word "intelligence" should be used. These opinions in turn influence the choice of research goals and methods, as well as serve as standards for judging other researchers' results.

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 the current knowledge is doomed to be revised by future works. 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.

Inside the AI research community, the lack of a common definition of the key concept of the field is the root of many controversies and misunderstandings. Many debates can be reduced to the fact that different sides use the term "intelligence" to mean very different things, and therefore they propose very different conclusions for questions like "What is the best way to achieve AI," "How to judge whether a system is intelligent," and so on.

Outside the AI 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 a fancy label on old stuff?

More importantly, each researcher in the field needs to justify his/her research approach in accordance with such a definition. For a concept as complex as "intelligence," no direct study is possible, especially when an accurate and rigid tool, namely the computer, is used as the research medium. We have to specify the problem clearly and only then be in a position to try to solve it. In this sense, anyone who wants to work on AI is facing a two-phase problem: firstly, choosing a working definition of intelligence, and then, producing it on a computer.

A *working definition* is a definition that is concrete enough to allow a researcher to directly work with it. By accepting a working definition of intelligence, a researcher does not necessarily believe that it fully captures the concept "intelligence," but the researcher takes it as a goal to be sought after for the current research effort. Such a definition is not for an AI journal editor who needs a definition to decide what papers are within the field or a speaker of the AI community who needs a definition to explain to the public what is going on within the field — in those cases, what is needed is a "descriptive definition" obtained by summarizing the individual working definitions.

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 a researcher keeps a working definition, he/she cannot claim to be working on AI. It is a researcher's working definition of intelligence that relates the current research, no matter how domain-specific, to the larger AI enterprise.

By accepting a working definition of intelligence, a researcher makes important commitments on the acceptable assumptions and desired results, which bind all the concrete work that follows. Limitations 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.

To better understand the relationship between a working definition of intelligence and AI research, consider an analogy. Imagine a group

of people that want to climb a mountain. They do not have a map, and the peak is often covered by clouds. At the foot of the mountain, there are several paths leading in different directions. When you join the group, some of the paths have been explored for a while, but no one has reached the top.

If you want to get to the peak as soon as possible, what should you do? It is a bad idea to sit at the foot of the mountain until you are absolutely sure which path is the shortest, because you know it only after all paths have been thoroughly explored. You have to try some path by yourself. On the other hand, taking an arbitrary path is also a bad idea. Although it is possible that you make the right choice from the beginning, it is more advisable to use your knowledge about mountains and to study other people's reports about their explorations, so as to avoid a bad choice in advance.

There are three kinds of "wrong paths": (1) those which lead nowhere, (2) those which lead to interesting places (even to unexpected treasures) but not to the peak, and (3) those which eventually lead to the peak but are much longer than some other paths. If the only goal is to reach the peak as soon as possible, a climber should use all available knowledge to choose the most promising path to explore. Although switching to another path is always possible, it is time consuming.

AI researchers face a similar situation in choosing a working definition for intelligence. There are already many such definitions, which are different but related to each other (so hopefully we are climbing the same mountain). As a scientific community, it is important that competing approaches are developed at the same time, but it does not mean that all of them are equally justified, or will be equally fruitful.

1.1.3 Criteria of a good definition

Before studying concrete working definitions of intelligence, we need to establish the general criteria for what makes one definition better than another.

The same problem of general criteria is encountered in other areas. For example Carnap tried to clarify the concept of "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 with explicit rules for its use [Carnap, 1950].

According to Carnap, the second concept, or the *working definition* as it is called here, must fulfill the following requirements [Carnap, 1950]:

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.

Since these criteria seem suitable for our purpose, let us see what they mean concretely to the working definition of intelligence (here I change the names and order of the first two requirements):

Sharpness. The definition should draw a relatively sharp line between the systems with intelligence and the ones without it. 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*, etc., though these concepts do have close relationships with intelligence. As well, the definition needs to answer the complement question: "What is not intelligent?" — The reason is simply if everything is intelligent, then the concept becomes empty.[2]

Faithfulness. The line drawn by the definition should not be an arbitrary one. Though "intelligence" has no precise meaning in everyday language, it does have some common usage 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. For this reason, AI

[2]For this reason, to define intelligence using the recently fashionable term "agent" is also not a good idea, because the term is too vague and too outstretched.

should not be defined to have the same meaning as "computer science."

Fruitfulness. The line should not only be descriptive, but also be constructive. Given the nature of AI as both a science and a technology, the "what is it?" and the "how to do it?" parts are closely related. 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 solving fundamental problems in AI. For this reason, we want to avoid various "sterile" definitions, which sound correct, but tell us little about how to build an intelligent system.

Simplicity. Although intelligence is surely a complex mechanism, the working definition should be as simple as possible. From a theoretical point of view, a simple definition can be explored 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, judged according to the above criteria. Though there is no evidence showing that in general the requirements cannot be satisfied at the same time, the four requirements may conflict with each other when comparing proposed definitions. For example, one definition is more fruitful, while another is simpler. In such a situation, some weighing and trade-off become necessary.

Especially, the requirement of "faithfulness" should not be understood as to mean that the working definition of intelligence should be determined according to an authoritative dictionary, or a poll among all the people. A working definition might even be counter-intuitive, if there is evidence showing that such a definition is faithful to the "deep meaning" of a concept. This is why we cannot argue that Einstein's concepts of "time" and "space" should be renamed because they conflict with our everyday usage of these terms. As Feyerabend said, "without a constant misuse of language there cannot be any discovery, any progress." [Feyerabend, 1993].

1.2 Various schools in AI research

With the above criteria in mind, we can evaluate the current AI approaches by analyzing their working definitions of intelligence. Since it is impractical to study each of the existing working definitions of intelligence one by one (there are simply too many of them), I will group them into several schools of thought and consider each school in turn. As usual, a concrete definition may belong to more than one school.

Stated previously, AI is the attempt of building computer systems that are "similar to the human mind." But in what sense are they "similar"? To different schools, the desired similarity may involve *structure, behavior, capability, function,* or *principle* of the systems. In the following, I discuss typical opinions in each of the five schools, to see where such a working definition of intelligence will lead research to.

1.2.1 To simulate the human brain

In the middle of all puzzles and problems about intelligence, there is one obvious and undoubtable fact, that is, the most typical example of intelligence we know today is produced by the human brain. Therefore, it is very natural to attempt to achieve AI by building a computer system that is as similar to a human brain as possible.

There are many researchers working on various kinds of "brain models" and "neurocomputational systems," though not all of them associate themselves with AI. However, there are people who believe that the best way to achieve AI is by looking into the brain, and some of them even argue that "the ultimate goals of AI and neuroscience are quite similar" [Reeke and Edelman, 1988]. Recent attempts in this direction include [Hawkins and Blakeslee, 2004, Hecht-Nielsen, 2005].

Though there is motive to identify AI with a *brain model*, few AI researchers take such an approach in a very strict sense. Even the "neural network" movement is "not focused on *neural modeling* (i.e., the modeling of neurons), but rather ... focused on *neurally inspired* modeling of cognitive processes" [Rumelhart and McClelland, 1986].

Why? One obvious reason is the daunting *complexity* of this approach. Current technology is still not powerful enough to simulate a huge neural network, not to mention the fact that there are still many mysteries about the brain.

Moreover, even if we were able to build a brain model at the neuron level to any desired accuracy, it could not be called a success for AI, though it would be a success for neuroscience. From the very beginning, and for a good reason, AI has been more closely related to the notion of a "model of mind", that is, a *high-level* description of brain activity in which biological concepts do not appear [Searle, 1980].

A high-level description is preferred, not because a low-level description is impossible, but because it is usually simpler and more general. When building a model, it is not always a good idea to copy the object or process to be modeled as accurately as possible, because a major purpose of modeling is often to identify the "essence" of the object or process, and to filter out unnecessary details. By ignoring irrelevant aspects, we gain insights that are hard to discern in the object or process itself. For this reason, an accurate duplication is not a model, and a model including unnecessary details is not a good model.

Intelligence (and the related notions like "thinking" and "cognition") is a complicated phenomena mainly observed only in the human brain at the current time. However, the very idea of "*artificial* intelligence" assumes that the same phenomena can be reproduced in something that is different from the human brain. This attempt to "get a mind without a brain", i.e., to describe mind in a medium-independent way, is what makes AI important and attractive. Even if we finally build an "artificial brain" which is like the human brain in all details, it still does not tell us much about intelligence and thinking in general. If one day we can build a system which is very different from the human brain in details, but we nevertheless recognize it as intelligent, then it will tell us much more about intelligence than a duplicated brain does.

If we agree that "brain" and "mind" are different notions, then a good model of the brain is not a good model of the mind, though the former is useful for its own sake, and may be helpful for the building of the latter.

1.2.2 To duplicate human behavior

For the people who believe that intelligence can be defined independently of the structure of the human brain, a natural alternative is to

define it in terms of human intellectual behavior. After all, if a system behaves like a human mind, it should deserve the title of "intelligence" for both theoretical and practical reasons. From this standpoint, whether the system's internal structure is similar to the human brain is mostly irrelevant.

This view is perhaps best captured by Turing in his famous "Imitation Game," later known as the "Turing Test" [Turing, 1950]. According to this idea, if a computer is indistinguishable from a human in a conversation (where the physical properties of the system are not directly observable), the system has intelligence.

After half a century, "passing the Turing Test" is still regarded by many people as the ultimate goal of AI [Saygin et al., 2000]. There are some research projects targeting it, sometimes under the name of "cognitive modeling." In recent years, there are also many "chatbots" developed to simulate human behavior in conversation.

On the other hand, this approach to AI has been criticized from various directions:

Is it sufficient? Searle argues that even if a computer system can pass the Turing Test, it still cannot *think*, because it lacks the *causal capacity* of the brain to produce *intentionality*, which is a biological phenomenon [Searle, 1980]. However, he does not demonstrate convincingly why thinking, intentionality, and intelligence cannot have high-level (higher than the biological level) descriptions.

Is it possible? Due to the nature of the Turing Test and the resource limitations of present computer systems, it is unlikely for the system to have stored in its memory all possible questions and proper answers in advance, and then give a convincing imitation of a human being by searching its memory upon demand. The only realistic way to imitate human behavior in a conversation is to produce the answers in real time. To do this, it needs not only cognitive faculties, but also much prior "human experience" [French, 1990]. It must, therefore, have a "body" that feels human, and all human motivations, including biological ones. Simply put, it must be an "artificial person," rather than a computer system with artificial intelligence. Furthermore, to build such a

system is not merely a technical problem, since acquiring human experience means that humans treat and interact with it as a human being.

Is it necessary? By using behavior as evidence, the Turing Test is a criterion solely for *human* intelligence, not for intelligence in general [French, 1990]. As a working definition of intelligence, such an approach can lead to good psychological models, which are valuable for many reasons, but it suffers from "human chauvinism" [Hofstadter, 1979]. We would have to say, according to this definition, that "extraterrestrial intelligence" cannot exist, simply because that human experience can only be obtained on the Earth. This strikes me as a very unnatural and unfruitful way to use concepts. Actually, Turing did not claim that passing the imitation test is a necessary condition for being intelligent. He just thought that if a machine could pass the test satisfactorily, we would not be troubled by the question [Turing, 1950]. However, this part of his idea is often ignored, and consequently his test is taken by many people as a sufficient and necessary condition of intelligence.

In summary, though "reproducing human (verbal) behavior" may still be a sufficient condition for being intelligent (as suggested by Turing), such a goal is difficult, if not impossible, to achieve presently. More importantly, it is not a necessary condition for "intelligence", if we want it to be a more general notion than "human intelligence."

1.2.3 To solve hard problems

For people whose main interest in AI is its practical application, whether a system is structured like a brain or behaves like a human does not matter at all, but what counts is what practical problems it can solve — after all, that is how the intelligence of a human being is measured. Therefore, according to this opinion, intelligence means the capability of solving hard problems.

This intuitive idea explains why early AI projects concentrated on typical and challenging intellectual activities, such as theorem proving and game playing, and why achievements on these problems are

still seen as milestones of AI progress. For example, many people, both within the AI community and among the general public, regard the victory of IBM's supercomputer Deep Blue over the World Chess Champion Kasparov as a triumph of AI.

For similar reasons, many AI researchers devote their effort to building "expert systems" in various domains, and view this as the way to general AI. The relation between these systems and the notion of intelligence seems to be obvious — experts are more intelligent in their domains than the average person. If computer systems can solve the same problems, they should deserve the title of intelligence, and whether the solutions are produced in a "human manner" has little importance. The way Deep Blue plays chess is very different from the way a human player plays chess. But as far as it wins the game, why should we care? Similarly, the intelligence of an expert system is displayed by its capability to solve problems for which it was designed.

Compared to the previously discussed goals, e.g., to model the human brain or to pass the Turing Test, this kind of goals is much easier to achieve, though still far from trivial. As today, we already have some success stories in game playing, theorem proving, and expert systems in various domains.

Though this approach toward AI sounds natural and practical, it has its own trouble.

If intelligence is defined as "the capability to solve hard problems," then the next obvious question is "Hard for whom?" If we say "hard for human beings," then most existing computer systems are already intelligent — no human manages a database as well as a database management system, or substitutes 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" [Hofstadter, 1979] and the "gee whiz view" [Schank, 1991].

The view that AI is a "perpetually expanding 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 hardness, they will not be likely to make any progress in understanding and replicating intelligence.