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.

This application-oriented movement has drawn in many researchers, produced many practically useful systems, attracted significant funding, and thus made important contributions to the development of the AI enterprise. However, though often profitable, these systems do not provide much insight into how the mind works. No wonder people ask, after learning how such a system works, "Where's the AI?" [Schank, 1991] — these systems look just like ordinary computer applications. Actually, many "AI systems" are indeed developed in the same way as ordinary software.

Nowadays AI researchers often complain that the field does not get the credit it deserves, since many AI research results have been used by other fields without the AI label. This seems to confirm that many people in AI are actually doing ordinary computer science and application, and therefore the results are just like the results obtained outside AI, so they do not need a fancy label. If someone insists that these works should be called AI simply because they solve problems that were previously solvable only by the human mind, then by the same token numerical calculating programs should be called AI, as well.

Beside the issues of label and credit, the real problem of this approach is that it fails to explain why ordinary computer systems are not intelligent. Many people enter AI to look for a fundamentally different way to build computer systems. To them, traditional computer systems are stupid, not because they cannot do anything (in fact, they can do many amazing things), but because they solve problems in a rigid manner. Therefore, whether a system is intelligent not only depends upon what it can do, but also upon how it does it. If an expert system is as brittle as a conventional computing system [Holland, 1986], it hardly deserves to be called "intelligent."

An interesting example is the victory of IBM's supercomputer Deep Blue alluded to earlier. While many people applaud this as a great achievement for AI [Newborn, 2002], the research team that developed the system never made such a claim [Campbell et al., 2002]. Instead, they made it clear that "although Deep Blue's speed and search capabilities enable it to play grandmaster-level chess, it is still lacking in general intelligence." [Campbell, 1997].

Human beings usually use their intelligence to play games, but it does not mean that a computer system must do the same. In theory,

it is possible to find (or invent) a game that is simple enough for a supercomputer to perform an exhaustive search to find the best move, but still too complicated for a human mind to play in like manner. Such a game can still be seen as a testing of human intelligence, because intelligent players will play better after a while (given that there are recognizable patterns in the game). However, a simple brute-force search algorithm, e.g., minimax, will be the world champion, simply because it will always find the optimum solution. In this case, should the algorithm fit the criteria of "intelligence"?

1.2.4 To carry out cognitive functions

As an attempt to generalize the various concrete behavior and capability into domain-independent form, the current AI field is often seen as studying a set of cognitive functions, including searching, recognizing, categorizing, reasoning, planning, decision making, problem solving, learning, and so on. Furthermore, for the interaction between the system and its environment (including other systems), sensorimotor and natural language processing can also be seen as cognitive functions.

Each cognitive function is typically specified as a *computation* process that starts with given input data, and after some processing generates the desired output data (plus certain side-effects inside and/or outside the system). The goal of the research is to find the most efficient *algorithm* to carry out a given function. Finally, the algorithm is implemented into a computer system, which can then handle the problem for us [Marr, 1982].

This approach has produced, and will continue to produce, information-processing 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 the claim that it is the right way to study AI. To define intelligence as a "toolbox" of cognitive functions has the following weaknesses:

- When specified in isolation, a formalized 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 [Chalmers et al., 1992].

- Having any particular cognitive function is not enough to make a system intelligent. For example, problem-solving by exhaustive search is usually not considered intelligence, and many unintelligent animals have excellent perceptual capability.

- Even if we can produce all the desired functions, it does not mean that we can easily integrate them into one system, because different functions may be developed under different assumptions, which prevent the tools from being integrated. According to the past experience in building integrated systems, "Component development is crucial; connecting the components is more crucial" [Roland and Shiman, 2002].

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 much 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 the situation may be just the opposite. When the so-called "functions" are actually phenomena produced by a complex-but-unified mechanism, reproducing all of them together (by duplicating the mechanism) is simpler than reproducing only one of them. For example, we can grow a tree, but we cannot generate a leaf *alone*, although a leaf is much simpler than a tree. Intelligence may be such a phenomenon.

As Piaget said: "Intelligence in action is, in effect, irreducible to everything that is not itself and, moreover, it appears as a total system of which one cannot conceive one part without bringing in all of it."[Piaget, 1963] This opinion does not deny that intelligence includes many distinguishable functions carried out by distinct mechanisms, but it stresses the close relations among the functions and processes, which produce intelligence as a whole. If intelligence is a toolbox, where is the hand that use the tools?

1.2.5 To develop new principles

In the cognitive sciences, especially, AI, psychology, and philosophy, there are some researchers who believe that intelligence (or cognition) are governed by a small set of general and simple principles. According to this opinion, all behaviors, capabilities, and functions of the human mind can be explained as produced by the application of these principles in concrete situations [Chater and Oaksford, 1999].

Typically, these principles are represented as some kind of "rationality," formed by the evolution process as the best adaptation strategy in a certain sense. Here are some examples:

> *Bounded rationality* [Simon, 1983]: "Within the behavioral model of bounded rationality, one doesn't have to make choices that are infinitely deep in time, that encompass the whole range of human values, and in which each problem is interconnected with all the other problems in the world."
>
> *Type II rationality* [Good, 1983]: "Type II rationality is defined as the recommendation to maximize expected utility allowing for the cost of theorizing. It involves the recognition that judgments can be revised, leading at best to consistency of *mature* judgments."
>
> *Minimal rationality* [Cherniak, 1986]: "We are in the finitary predicament of having fixed limits on our cognitive resources, in particular, on memory capacity and computing time."
>
> *General principle of rationality* [Anderson, 1990]: "The cognitive system operates at all times to optimize the adaptation of the behavior of the organization."
>
> *Limited rationality* [Russell and Wefald, 1991a]: "Intelligence was intimately linked to the ability to succeed as far as possible given one's limited computational and informational resources."

According to these ideas, an AI theory should establish a few principles to derive all the functions and behaviors, then a formal model of

intelligence should be formulated according to such a normative theory (such as logic, probability theory, and so on), which always "does the right thing," according to the underlying principles.

If such an approach works, we will eventually have a well-justified theory of AI, in which all functionalities are based on a consistent foundation.

Like the previous approaches, this approach has its problems. Though it is not a new idea that the human mind, like many other objects of scientific research, can eventually be explained by a small set of principles, none of the "principles" proposed so far has successfully achieved this goal yet. Although "haven't found such principles" does not prove "no such principles can exist," it often leaves people feeling that way. In AI, to date, there have been too many promises of success only to be followed by failure for people to believe in any new "silver bullet" that can solve the AI problem with one shot. Consequently, they would rather believe that given the complexity of the domain, AI must be treated as a collection of concrete problems that have to be handled one by one.

A much more serious challenge to this kind of approach is the existence of many well-documented psychological phenomena, showing that people often violate the proposed normative theories, such as predicate logic [Wason and Johnson-Laird, 1972] and probability theory [Tversky and Kahneman, 1974]. Any rationality-base theory must try to explain these phenomena [Anderson, 1990].

1.3 AI as a whole

1.3.1 Relations among the goals

From the previous discussion, we can see that instead of currently having one common research goal, the field of AI has a set of different, but related, goals pursued through different schools of thought. As Nilsson said: "AI shows all the signs of being in what the late Thomas Kuhn called a pre-paradigmatic, pre-normal-science stage." [Hearst and Hirsh, 2000].

Each of these goals reflects a particular aspect of our current usage of the word "intelligence," defines the term in a relatively sharp and simple way, and has been producing interesting results. In this sense, all of them guide legitimate scientific research, and contribute to our understanding of intelligence, as well as to the progress of each other. Since they have different goals, they can and should co-exist for a long time.

However, at a more general level, these schools do compete — as the best way to build a "thinking machine." There is no contradiction here. When these schools are evaluated as research goals in their own right, each of them is valuable in a particular way. But if they are evaluated as paths to the common goal of AI, as introduced at the beginning of the chapter, they are not equally good. Of course, which one is better is a controversial issue.

From the above presentation, one point I want to make is that there is no "natural" or "self-evident" definition of intelligence, and nobody in the field can escape from the responsibility of choosing a school to work within. Many people claim that they are not interested in philosophical debates, and they simply choose the natural or obvious problem to work on, but in reality they have made the choice unconsciously, guided by their intuition or non-academic factors, such as personal background, adviser expertise, practical need, grant source, publication possibility, current fashion, and so on. After the initial choices (which are typically made early in their career), they gradually get accustomed to them, and spend most of their time in solving the problems specified by the school, without considering more fundamental questions like whether they are the right problems to work on.

Some people may think that the different schools are aimed at different "parts" of intelligence, like in the parable of "The Blind Men and the Elephant," and the best way is to "integrate" them. However, here the situation is different. These schools are generally *incompatible* (though they have small overlaps here or there), and therefore cannot be fully integrated into a consistent theory on intelligence, or be satisfied together by a computer system.

There are many systems that use techniques developed in different schools of thought, but these "integrated" or "hybrid" systems are often justified by what they can do, rather than by a consistent theory on intelligence. This is the case because different schools usually make

different design decisions. For example, the most efficient way to solve a problem in a computer is often very different from the way used in the human mind. Which one is more "intelligent"? Well, it depends on that you mean by "intelligence."

It is fine to set up a "major" goal, and, at the same time, to achieve other "minor" goals as much as possible. Even in this case, school conflicts exist, necessitating compromise and trade-offs to make progress.

The multi-school nature of the current AI field causes much confusion, because people often use the requirement of one school to judge the results produced by another school, which usually does not provide a fair conclusion. Also, the answers to many general questions on AI depend on which school is referred to. Examples of these questions include "When can we get an intelligent system?" "Can an AI be more intelligent than a human being?" "Can an intelligent system be creative or original or conscious?" "Can an intelligent system run out of human control?" and so on — their answers are different in different schools.

1.3.2 Different opinions on unified AI

Should AI be addressed as one problem, or a collection of loosely related problems that can be handled one by one separately? Again, the answer to this question depends on the interpretation of the concept of "intelligence," and there are very different opinions.

The majority of the current AI community believes in a "divide-and-conquer" approach toward AI. Many researchers claim that their research will contribute to the whole AI enterprise by focusing on a particular aspect of intelligence. Usually, there is an implicit assumption under this kind of claims, that is, when all these particular solutions are finally put together, we will have an AI, a "thinking machine."

However, as mentioned previously, such an assumption is hard to justify. It may be true that a complicated problem should be cut into pieces of smaller problems and solved one by one, but if everyone cuts the problem in his/her own way, and only works on a small piece of the problem obtained in this manner, we have no reason to believe that the solutions can later be put together to form a solution to the original problem [Brooks, 1991].

This is well illustrated by considering the present state of research in the field of AI. Browsing a journal with AI in its title or attending a conference with AI in its name, it is all too common to find articles or presentations that have very little to do with each other. In fact, nowadays few people even mention the relation between their current research and the big picture of AI.

Many people seem comfortable with this situation. They think that the idea of a "thinking machine" or something like that belongs to science fiction only, and that few people are pursuing the goal only means that the field has become mature. They do not care whether their systems are really "intelligent," which is just a label that can be attached or removed according to context for convenience purposes.

Of course, there are still attempts to unify the AI field. Newell is one of the few people who actually tried to build a unified AI theory. In [Newell, 1990], he argued for the need of unified theories, and discussed what such a theory should include. Though his theory is well known, and his project, Soar, is still alive, this kind of work does not attract many followers these days.

For the people who feel uncomfortable about the fragmented status of the field, one response is to find a unified way to *describe* the problems and solutions. The most recent attempt in this category is to uniformly describe the field within the framework of *intelligent agent* [Nilsson, 1998, Russell and Norvig, 2002]. Though this effort improves the coherence of AI textbooks, it is far from enough in unifying the techniques covered under the umbrella of "agent."

People who still associate themselves to the original AI goal find the current situation disappointing. As Minsky said [Stork, 1997b]:

> The bottom line is that we really haven't progressed too far toward a truly intelligent machine. We have collections of dumb specialists in small domains; the true majesty of general intelligence still awaits our attack.
>
> We have got to get back to the deepest questions of AI and general intelligence and quit wasting time on little projects that don't contribute to the main goal.

Wolfram made a similar comment [Stork, 1997a]:

> Nobody's trying more fundamental stuff. Everyone assumes it's just too difficult. Well, I don't think there's really any evidence of that. It's just that nobody has tried to do it. And it would be considered much too looney to get funded or anything like that.

Though there is no evidence showing the impossibility of unified AI, the past experience does make the AI community turn away from such a goal. In this atmosphere, only two types of people continue to pursue the "thinking machine" dream: the well-established researchers, and the people at the margin or even the outside of the AI community. Though these two groups of people have opposite status in many attributes, they have one thing in common, that is, they don't care too much about what the others say, and they can keep their research going even if the majority of the AI community dislikes it.

1.3.3 AGI projects

The "Thinking Machine Dream" mentioned above goes with many names, such as "Unified AI," "Strong AI," "Real AI," "Hard AI," "AGI (Artificial General Intelligence)," "Human-Level Intelligence," and so on. I'll use the term AGI in this book, though this choice does not really make any difference, since few of these terms are accurately defined. The common thesis behind these terms is the belief that intelligence is a unified mechanism that should be described and developed as a whole, independent of any application domain. Even if the development must be carried out step by step, an overall plan should be drawn first to guide the process.

For such a project, one crucial issue is to have a theoretical foundation with sufficient width to support all kinds of functions and capabilities. Though there has been a very small number of people doing this kind of research, they still belong to different schools of thought, as described previously.

In the following I will discuss some representative AGI projects, though it is not an attempt of reviewing all such projects exhaustively.

To many people, the capability to solve various types of problems is at the core of intelligence. The first attempt to build a general-purpose system for this task is the General Problem Solver (GPS) [Newell and Simon, 1976]. By analyzing "protocols" observed in the problem-solving processes of human beings, Newell and Simon represented them as state-space search, and used "means-ends analysis" to lead the search process. According to this approach, problem solving is treated as finding a sequence of actions that transforms the initial state into a final state, step by step. When there are multiple alternatives at a state, the difference between the next state and a final state is used as a *heuristic* to estimate the distance from the former to the latter. Although few people still believe that all problem solving can and should be handled in this way, the search is still referred by some as "the most fundamental method of all" [Newell, 1990].

Another ambitious attempt to make a breakthrough in AI is the Fifth Generation Computer Systems (FGCS) project of Japan, initiated in the early 1980's. The belief behind the project, roughly speaking, was that the bottleneck of AI was in the von Neumann computer architecture, which was initially designed for *sequential calculation*. On the contrary, the key in AI should be *parallel inference*. To build a machine that is "as smart as a person," we should turn from "sequential calculation on data" to "parallel inference on knowledge," and to build a proper computer system to support the latter is the key of AI [Feigenbaum and McCorduck, 1983]. This approach caused quite a splash in the AI community and consequently received a lot of attention from governments and companies all around the world. Today, however, there haven't been any remarkable advances as a result of FGCS, not to mention any breakthrough in AI research, although parallel inference engines have been built as scheduled. On the contrary, now this project is rarely mentioned, as if it never existed.

No matter what the reasons are, none of the historical AGI projects has delivered the results initially promised. This partly explains why there are few such projects being actively worked on.

For these on-going projects, in the following I directly cite their project websites, with a brief description.

In the mainstream AI, there are three well-known projects that can be categorized as AGI.

Cog (`http://www.ai.mit.edu/projects/humanoid-robotics-group/cog/`)
This project is based on the belief that intelligence should come out of a robot that directly interacts with the physical world.

CYC (`http://www.cyc.com`)
This project was initially the American response to FGCS. Instead of focusing on the inference engine, this project puts most of its efforts in the building of a huge knowledge base that holds "common sense."

Soar (`http://sitemaker.umich.edu/soar`)
This system can be seen as a follow up of GPS. It attempts to provide a unified model of cognition in the framework of state-space search, which is implemented as a production system.

The next group of projects have smaller scopes in their goals, though they also are in the direction of AGI, to various degrees.

ACT-R (`http://act-r.psy.cmu.edu/`)
This is not an AI project, but a psychological model of human cognition. Nevertheless, it is still closely related to AGI. The basic architecture is also a production system (like Soar).

OSCAR (`http://www.u.arizona.edu/~pollock/`)
This is an architecture for rational agents based upon a philosophical theory of rational cognition. The core technique is defeasible inference.

SNePS (`http://www.cse.buffalo.edu/sneps/`)
This is an attempt to unify knowledge representation, reasoning, and natural-language processing, using a semantic network. The research has begun to integrate sensorimotor capability into the system.

Finally, there is a group of projects that are ambitious and take more radical paths, though they have not got much attention from the mainstream AI community.

AIXI (`http://www.idsia.ch/~marcus/ai/index.htm`)
A mathematical theory of universal induction, based on probability theory and computation theory.

a2i2 (`http://adaptiveai.com/project/index.htm`)
A connectionist AGI system, which is embodied, and learns from its interaction with the environment.

CAM-Brain (`http://www.cs.usu.edu/\%7Edegaris/cam/index.html`)
An artificial brain consisting of roughly a million modules of cellular automata based neural circuits, which grow and evolve.

Novamente (`http://www.agiri.org/engine.htm`)
An integrated AGI system with multiple techniques, include probabilistic reasoning, genetic programming, and so on.

These projects, as well as some other AGI approaches, are described in [Goertzel and Pennachin, 2006].

In summary, past research on unified artificial intelligence has not produced encouraging results; currently there is only a small number of people involved in AGI work; wherein the on-going AGI projects are based on very different opinions.