chapter 5

# Introduction on Artificial Intelligence[1]

We call ourselves **Homo sapiens**[2]—man the wise—because our mental capacities are so important to us. For thousands of years，we have tried to understand how we think；perceive，understand，predict. The field of **artificial intelligence**，or AI，goes further still：it attempts not just to understand but also to build intelligent entities.

AI is one of the newest sciences. Work started in earnest soon after World War II. Along with molecular biology，AI is regularly cited as the "field I would most like to be in" by scientists in other disciplines. A student in physics might reasonably feel that all the good ideas have already been taken by Galileo，Newton，Einstein，and the rest. AI，on the other hand，still has openings for several full-time Einsteins.

AI currently **encompasses**[3] a huge variety of subfields，ranging from general-purpose areas，such as learning and **perception**[4] to such specific tasks as playing chess，proving mathematical theorems，writing poetry，and diagnosing diseases. AI **systematizes**[5] and automates intellectual tasks and is therefore potentially relevant to any **sphere**[6] of human intellectual activity. In this sense，it is truly a universal field.

## 批注

**artificial intelligence**：人工智能是计算机学科的一个研究分支，最初是在1956年达特茅斯学会上提出的，之后研究者们发展了众多理论和原理。总的说来，人工智能研究的一个主要目标是使机器能够胜任一些通常需要人类智能才能完成的复杂工作。除了计算机科学以外，人工智能还涉及信息论、控制论、自动化、仿生学、生物学、心理学、数理逻辑、语言学、医学和哲学等多门学科。人工智能学科研究的主要内容包括知识表示、自动推理和搜索方法、机器学习和知识获取、知识处理系统、自然语言理解、计算机视觉、智能机器人、自动程序设计等方面，应用在包括指纹识别、人脸识

[1] 本章英文内容系选自参考文献[Russell，2003]中第1章的内容，有部分删改

[2] Homo sapiens：有智慧的人

[3] encompass：*v.* 围绕，包含，涉及

[4] perception：*n.* 感知

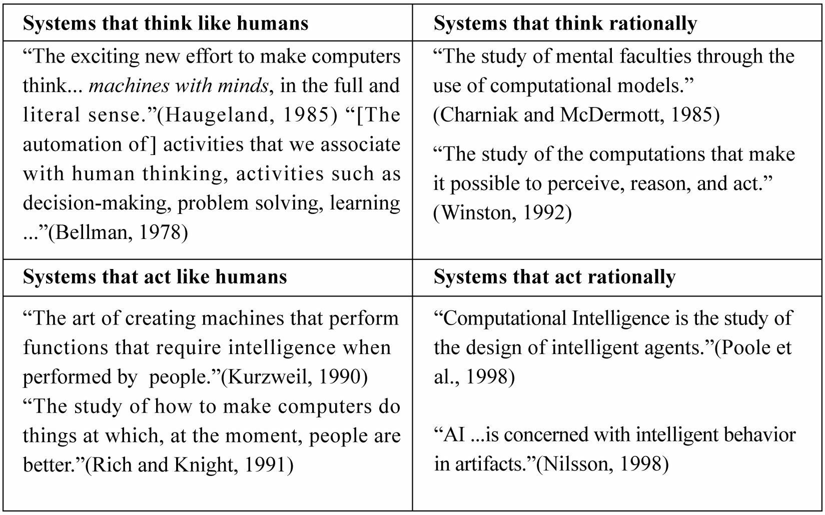
[5] systematize：*v.* 系统化

[6] sphere：*n.* 范围

别、虹膜识别、专家系统、智能检索、智能控制、语言和图像处理、定理证明、博弈、自动程序设计、机器人技术等多个领域,有广阔发展前景。

## Section 1 What is AI?

We have claimed that AI is exciting, but we have not said what it is. Definitions of artificial intelligence according to eight textbooks are shown in Figure 5-1. These definitions vary along two main dimensions. Roughly, the ones on top are concerned with thought processes and reasoning, whereas the ones on the bottom address behavior. The definitions on the left measure success in terms of fidelity to human performance, whereas the ones on the right measure against an ideal concept of intelligence, which we will call **rationality**[1]. A system is rational if it does the "right thing," given what it knows.



| **Systems that think like humans** | **Systems that think rationally** |
|---|---|
| "The exciting new effort to make computers think... *machines with minds*, in the full and literal sense."(Haugeland, 1985) "[The automation of] activities that we associate with human thinking, activities such as decision-making, problem solving, learning ..."(Bellman, 1978) | "The study of mental faculties through the use of computational models." (Charniak and McDermott, 1985)<br>"The study of the computations that make it possible to perceive, reason, and act." (Winston, 1992) |
| **Systems that act like humans** | **Systems that act rationally** |
| "The art of creating machines that perform functions that require intelligence when performed by people."(Kurzweil, 1990)<br>"The study of how to make computers do things at which, at the moment, people are better."(Rich and Knight, 1991) | "Computational Intelligence is the study of the design of intelligent agents."(Poole et al., 1998)<br>"AI ...is concerned with intelligent behavior in artifacts."(Nilsson, 1998) |

**Figure 5-1 Some definitions of artificial intelligence, organized into four categories**

A human-centered approach must be an **empirical**[2] science, involving hypothesis and experimental confirmation. A rationalist approach involves a combination of mathematics and engineering. Let us look at the four approaches in more detail.

### 1. Acting humanly: the Turing test approach

The Turing test, proposed by **Alan Turing**, was designed to provide a satisfactory operational definition of intelligence. Rather than proposing a long and perhaps **controversial**[3]

[1] rationality: *n.* 理性的

[2] empirical: *adj.* 实验的, 经验主义的

[3] controversial: *adj.* 有争议的

list of qualifications required for intelligence, he suggested a test based on **indistinguishability**[1] from undeniably intelligent entities-human beings. The computer passes the test if a human **interrogator**[2], after posing some written questions, cannot tell whether the written responses come from a person or not. For now, we note that programming a computer to pass the test provides plenty to work on. The computer would need to possess the following capabilities:

- natural language processing: to enable it to communicate successfully in English;
- knowledge representation: to store what it knows or hears;
- automated reasoning: to use the stored information to answer questions and to draw new conclusions;
- machine learning: to adapt to new circumstances and to detect and **extrapolate**[3] patterns.

Turing's test deliberately avoided direct physical **interaction**[4] between the interrogator and the computer, because physical simulation of a person is unnecessary for intelligence. However, the so-called total Turing Test includes a video signal so that the interrogator can test the subject's perceptual abilities, as well as the opportunity for the interrogator to pass physical objects "through the hatch." To pass the total Turing Test, the computer will need:

- computer vision: to perceive objects, and
- robotics: to manipulate objects and move about.

These six disciplines compose most of AI, and Turing deserves credit for designing a test that remains relevant 50 years later. Yet AI researchers have devoted little effort to passing the Turing test, believing that it is more important to study the underlying principles of intelligence than to duplicate an **exemplar**[5]. **The quest for "artificial flight" succeeded when the Wright brothers and others stopped imitating birds and learned about aerodynamics. Aeronautical engineering texts do not define the goal of their field as making "machines that fly so exactly like pigeons that they can fool even other pigeons."**[i]

### 2. Thinking humanly: The cognitive[6] modeling approach

If we are going to say that a given program thinks like a human, we must have some way of determining how humans think. We need to get inside the actual workings of human

[1] indistinguishability：*n*. 不可分辨

[2] interrogator：*n*. 被访者，被测试者

[3] extrapolate：*v*. 推断，外推

[4] interaction：*n*. 交互作用

[5] exemplar：*n*. 样品，样本

[6] cognitive：*adj*. 认知的

minds. There are two ways to do this: through **introspection**[1] (trying to catch our own thoughts as they go by) and through psychological experiments. **Once we have a sufficiently precise theory of the mind, it becomes possible to express the theory as a computer program**[ii]. If the program's input/output and timing behaviors match corresponding human behaviors, that is evidence that some of the program's mechanisms could also be operating in humans. For example, **Allen Newell and Herbert Simon**, who developed GPS, the "General Problem Solver"[Newell and Simon, 1961], were not **content**[2] to have their program solve problems correctly. They were more concerned with comparing the trace of its reasoning steps to **traces**[3] of human subjects solving the same problems. The **interdisciplinary**[4] field of cognitive science brings together computer models from AI and experimental techniques from **psychology**[5] to try to construct precise and testable theories of the workings of the human mind.

Cognitive science is a fascinating field, worthy of an **encyclopedia**[6] in itself[Wilson and Keil, 1999]. We will not attempt to describe what is known of human cognition. We will occasionally comment on similarities or difierences between AI techniques and human cognition. Real cognitive science, however, is necessarily based on experimental investigation of actual humans or animals.

**In the early days of AI there was often confusion between the approaches: an author would argue that an algorithm performs well on a task and that it is therefore a good model of human performance, or vice versa.**[iii] Modern authors separate the two kinds of claims; this distinction has allowed both AI and cognitive science to develop more rapidly. The two fields continue to fertilize each other, especially in the areas of vision and natural language. Vision in particular has recently made advances via an integrated approach that considers **neurophysiological**[7] evidence and computational models.

### 3. Thinking rationally: The "laws of thought" approach

The Greek philosopher **Aristotle** was one of the first to attempt to codify "right thinking," that is, **irrefutable**[8] reasoning processes. His **syllogisms**[9] provided patterns for argument structures that always yielded correct conclusions when given correct **premises**[10],

[1] introspection：*n.* 内省，反省，自省

[2] content：*v.* 满足

[3] trace：*v.* 追踪，回溯，探索

[4] interdisciplinary：*adj.* 各学科之间的，交叉学科

[5] psychology：*n.* 心理学

[6] encyclopedia：*n.* 百科全书

[7] neurophysiological：*adj.* 生物学的

[8] irrefutable：*adj.* 不可反驳的，不能驳倒的

[9] syllogisms：*n.* 三段论法，推理法，演绎

[10] premise：*n.* 前提

for example, "Socrates is a man; all men are **mortal**[1]; therefore, Socrates is mortal." These laws of thought were supposed to govern the operation of the mind; their study initiated the field called logic.

Logicians in the 19th century developed a **precise notation for statements**[2] about all kinds of things in the world and about the relations among them. By 1965, **programs existed that could, in principle, solve any solvable problem described in logical notation**[iv]. The so-called logicist tradition within artificial intelligence hopes to build on such programs to create intelligent systems.

There are two main obstacles to this approach. First, **it is not easy to take informal knowledge and state it in the formal terms required by logical notation, particularly when the knowledge is less than 100% certain**[v]. Second, there is a big difference between being able to solve a problem "in principle" and doing so in practice. **Even problems with just a few dozen facts can exhaust the computational resources of any computer unless it has some guidance as to which reasoning steps to try first**[vi]. Although both of these obstacles apply to any attempt to build computational reasoning systems, they appeared first in the logicist tradition.

### 4. Acting rationally: The rational agent approach

An **agent** is just something that acts (agent comes from the Latin agere, to do). But computer agents are expected to have other attributes that distinguish them from mere "programs," such as operating under autonomous control, perceiving their environment, **persisting over a prolonged time period**[3], adapting to change, and being capable of taking on another's goals. A rational agent is one that acts so as to achieve the best outcome or, when there is uncertainty, the best expected outcome.

In the "laws of thought" approach to AI, the emphasis was on correct **inferences**[4]. Making correct inferences is sometimes part of being a rational agent, because one way to act rationally is to reason logically to the conclusion that a given action will achieve one's goals and then to act on that conclusion. On the other hand, correct inference is not all of rationality, because there are often situations where there is no **provably**[5] correct thing to do, yet something must still be done. **There are also ways of acting rationally that cannot be said to involve inference**[vii]. For example, **recoiling**[6] from a hot stove is a reflex action

[1] mortal: *adj.* 必死的

[2] precise notation for statement: 精确的符号命题

[3] persisting over a prolonged time period: 持续能力(prolonged: 延长的, 持久的)

[4] inference: *n.* 推论

[5] provably: *adv.* 可证明地, 可查明地

[6] recoil: *v.* 弹回, 畏缩, 退却

that is usually more successful than a slower action taken after careful **deliberation**[1].

All the skills needed for the Turing Test are there to allow rational actions. Thus, we need the ability to represent knowledge and reason with **it**[2] because this enables us to reach good decisions in a wide variety of situations. **We need to be able to generate comprehensible sentences in natural language because saying those sentences helps us get by in a complex society**[viii]. We need learning not just for **erudition**[3], but because having a better idea of how the world works enables us to generate more effective strategies for dealing with it. We need visual perception not just because seeing is fun, but to get a better idea of what an action might achieve—**for example, being able to see a tasty morsel helps one to move toward it**[ix].

For these reasons, the study of AI as rational-agent design has at least two advantages. First, it is more **general**[4] than the "laws of thought" approach, because correct inference is just one of several possible mechanisms for achieving rationality. Second, **it is more amenable to scientific development than approaches based on human behavior or human thought because the standard of rationality is clearly defined and completely general**[x]. Human behavior, on the other hand, is well-adapted for one specific environment and is the product, in part, of a complicated and largely unknown evolutionary process that still is **far from**[5] producing perfection.

## 批 注

**Alan Turing**：图灵是英国著名的数学家和逻辑学家，被称为计算机科学之父、人工智能之父，是计算机逻辑的奠基者，提出了图灵机和图灵测试等重要概念。人们为纪念其在计算机领域的卓越贡献而设立图灵奖[百度百科,2010]。

**Allen Newell and Herbert Simon**：二人是人工智能符号主义学派的创始人，是1975年的图灵奖获得者，当时是卡内基-梅隆大学的教授。他们开发的GPS是根据人在解题中的共同思维规律编制而成的，可以解11种不同类型的问题，从而使启发式程序有了更普遍的意义。

**Aristotle**：亚里士多德（前384—前322年）世界古代史上伟大的哲学家、科学家和教育家之一，据说是柏拉图的学生，亚历山大的老师。马克思曾称亚里士多德是古希腊哲学家中最博学的人物，恩格斯称他是古代的黑格尔[百度百科,2010]。

**agent**：智能体，一般有很好的自主性，具备一个知识库和通用的专家系统引擎，能够主动分析承担的任务，具有学习能力和通信、合作能力。

[1] deliberation：*n.* 熟思，商议，考虑

[2] it：这里的it指知识

[3] erudition：*n.* 博学

[4] general：*n.* 通用，概括

[5] far from：远远没有达到的

# Section 2 The History of Artificial Intelligence (Part I)

## 1. The gestation[1] of artificial intelligence (1943—1955)

The first work that is now generally recognized as AI was done by **Warren McCulloch and Walter Pitts** (1943). They drew on three sources: knowledge of the basic physiology and function of neurons in the brain; a formal analysis of **propositional**[2] logic due to Russell and Whitehead; and Turing's theory of computation. They proposed a model of artificial neurons in which each neuron is characterized as being "on" or "off," with a switch to "on" occurring in response to stimulation by a sufficient number of neighboring neurons. **The state of a neuron was conceived of as "factually equivalent to a proposition which proposed its adequate stimulus."**[xi] They showed, for example, that any computable function could be computed by some network of connected neurons, and that all the logical **connectives** (and, or, not, etc.) could be implemented by simple net structures. McCulloch and Pitts also suggested that suitably defined networks could learn. **Donald Hebb** (1949) demonstrated a simple updating rule for modifying the connection strengths between neurons. His rule, now called Hebbian learning, remains an influential model to this day.

Two undergraduate students at Harvard, **Marvin Minsky** and Dean Edmonds, built the first neural network computer in 1950. The SNARC, as it was called, used 3000 vacuum tubes and a **surplus automatic pilot mechanism**[3] from a B-24 bomber to simulate a network of 40 neurons. Later, at Princeton, Minsky studied universal computation in neural networks. His Ph. D. committee was **skeptical**[4] about whether this kind of work should be considered mathematics, but von Neumann reportedly said, "If it isn't now, it will be someday." **Minsky was later to prove influential theorems showing the limitations of neural network research.**[xii]

There were a number of early examples of work that can be characterized as AI, but it was Alan Turing who first articulated a complete vision of AI in his 1950 article "Computing Machinery and Intelligence." Therein, he introduced the Turing test, machine learning, genetic algorithms, and **reinforcement learning**.

## 2. The birth of artificial intelligence (1956)

Princeton was home to another influential figure in AI, John McCarthy. After gradua-

[1] gestation：*n.* 酝酿

[2] propositional：*adj.* 建议的，命题的

[3] surplus automatic pilot mechanism：备用自动飞行机械装置

[4] skeptical：*adj.* 怀疑性的

tion, McCarthy moved to Dartmouth College, which was to become the official birthplace of the field. McCarthy convinced Minsky, Claude Shannon, and Nathaniel Rochester to help him bring together U. S. researchers interested in **automata**[1] theory, neural nets, and the study of intelligence. They organized a two-month workshop at Dartmouth in the summer of 1956. There were 10 attendees in all, including Trenchard More from Princeton, Arthur Samuel from IBM, and Ray Solomonoff and Oliver Selfridge from MIT.

**Two researchers from Carnegie Tech, Allen Newell and Herbert Simon, rather stole the show**[xiii]. Although the others had ideas and in some cases programs for particular applications such as checkers, Newell and Simon already had a reasoning program, the Logic Theorist (LT), **about which Simon claimed, "We have invented a computer program capable of thinking non-numerically, and thereby solved the venerable mind-body problem."**[xiv].

The Dartmouth workshop did not lead to any new **breakthroughs**[2], but it did introduce all the major figures to each other. For the next 20 years, the field would be dominated by these people and their students and colleagues at MIT, CMU, Stanford, and IBM. Perhaps the longest-lasting thing to come out of the workshop was an agreement to adopt McCarthy's new name for the field: artificial intelligence. Perhaps "computational rationality" would have been better, but "AI" has **stuck**[3].

Looking at the **proposal**[4] for the Dartmouth workshop, we can see why it was necessary for AI to become a separate field. Why couldn't all the work done in AI have taken place under the name of control theory, or **operations research**[5], or decision theory, which, after all, have objectives similar to those of AI? Or why isn't AI a branch of mathematics? The first answer is that **AI from the start embraced the idea of duplicating human faculties like creativity, self-improvement, and language use**[xv]. None of the other fields were addressing these issues. The second answer is **methodology**[6]. **AI** is the only one of these fields that is clearly a branch of computer science (although operations research does share an emphasis on computer simulations), and AI is the only field to attempt to build machines that will function **autonomously**[7] in complex, changing environments.

[1] automata: *n.* 自动操作，自动控制

[2] breakthrough: *n.* 突破

[3] stuck: *adj.* 牢固的，根深蒂固的

[4] proposal: *n.* 提议，建议

[5] operations research: 运筹学研究

[6] methodology: *n.* 方法论

[7] autonomously: *adv.* 自治地

### 3. Early enthusiasm[1], great expectations[2](1952—1969)

The early years of AI were full of successes in a limited way. Given the primitive computers and programming tools of the time, and the fact that only a few years earlier computers were seen as things that could do arithmetic and no more, it was astonishing whenever a computer did anything **remotely**[3] clever. The intellectual establishment, **by and large**[4], preferred to believe that "a machine can never do X.". **AI researchers naturally responded by demonstrating one X after another**[xvi]. **John McCarthy** referred to this period as the "Look, Ma, no hands!" era.

Newell and Simon's early success was followed up with the General Problem Solver, or **GPS**[5]. Unlike Logic Theorist, this program was designed from the start to imitate human problem-solving protocols. GPS was probably the first program to embody the "thinking humanly" approach. The success of GPS and subsequent programs as models of **cognition**[6] led Newell and Simon to formulate the famous "physical symbol system" **hypothesis**[7], which states that "**a physical symbol system has the necessary and sufficient means for general intelligent action.**[xvii]" What they meant is that any system (human or machine) exhibiting intelligence must operate by manipulating data structures composed of symbols. We will see this hypothesis has been challenged from many directions.

At IBM, Nathaniel Rochester and his colleagues produced some of the first AI programs. Herbert Gelernter (1959) constructed the **Geometry Theorem Prover**[8], which was able to prove theorems that many students of mathematics would find quite tricky. Starting in 1952, Arthur Samuel wrote a series of programs for checkers that eventually learned to play at a strong amateur level. Along the way, he **disproved**[9] the idea that computers can do only what they are told to: his program quickly learned to play a better game than its creator. The program was demonstrated on television in February 1956, creating a very strong impression. Like Turing, **Samuel had trouble finding computer time**[xviii]. Working at night, he used machines that were still on the testing floor at IBM's manufacturing plant.

John McCarthy moved from Dartmouth to **MIT** and there made three crucial contributions in one historic year: 1958. In MIT AI Lab Memo No. 1, McCarthy defined the high-

[1] enthusiasm: *n.* 热情，积极性

[2] expectation: *n.* 期望

[3] remotely: *adv.* 偏远地(anything remotely clever: 任何一点点聪明的事情)

[4] by and large: 大体而言

[5] GPS: 通用问题求解器

[6] cognition: *n.* 认识，此指人工智能中的认知模型

[7] hypothesis: *n.* 假设

[8] Geometry Theorem Prover: 几何定理证明机，这也是人工智能的一个重要应用

[9] disprove: *v.* 反驳，证明……为谬误

level language Lisp, which was to become the dominant AI programming language. Lisp is the second-oldest major high-level language in current use, one year younger than **FORTRAN**. With Lisp, McCarthy had the tool he needed, but access to **scarce**[1] and expensive computing resources was also a serious problem. In response, he and others at MIT invented time sharing. Also in 1958, McCarthy published a paper entitled *Programs with Common Sense*, in which he described the **Advice Taker**, a hypothetical program that can be seen as the first complete AI system. Like the Logic Theorist and Geometry Theorem Prover, McCarthy's program was designed to use knowledge to search for solutions to problems. But unlike the others, it was to embody general knowledge of the world. For example, **he showed how some simple axioms would enable the program to generate a plan to drive to the airport to catch a plane**[xix]. The program was also designed so that it could accept new **axioms**[2] in the normal course of operation, **thereby allowing it to achieve competence in new areas without being reprogrammed**[xx]. The Advice Taker thus embodied the central principles of knowledge representation and reasoning: that it is useful to have a formal, explicit representation of the world and of the way an agent's actions affect the world and to be able to manipulate these representations with **deductive**[3] processes. It is remarkable how much of the 1958 paper remains relevant even today.

1958 also marked the year that Marvin Minsky moved to MIT. His initial collaboration with McCarthy did not last, however. McCarthy stressed representation and reasoning in **formal logic**[4], whereas Minsky was more interested in getting programs to work and eventually developed an anti-logical **outlook**[5]. In 1963, McCarthy started the AI lab at Stanford. His plan to use logic to build the ultimate Advice Taker was advanced by J. A. Robinson's discovery of the **resolution**[6] method (a complete theorem-proving algorithm for first-order logic). Work at Stanford emphasized general-purpose methods for logical reasoning.

**Minsky supervised a series of students who chose limited problems that appeared to require intelligence to solve**[xxi]. These limited domains became known as microworlds. James Slagle's SAINT program was able to solve **closed-form calculus integration problems**[7] typical of first-year college courses. Tom Evans's ANALOGY program solved geometric analogy problems that appear in IQ tests. Daniel Bobrow's STUDENT program solved algebra story problems, such as the following: "**If the number of customers Tom gets is twice the square of 20 percent of the number of advertisements he**

[1] scarce：*adj.* 稀有的，稀缺的

[2] axiom：*n.* 指数学中的公理

[3] deductive：*adj.* 推论的，演绎的

[4] formal logic：形式逻辑

[5] outlook：*n.* 观点

[6] resolution：*n.* 归结

[7] closed-form calculus integration problem：封闭型的积分问题