Part I

Agents in the World: What are Agents and How Can They be Built?

Chapter 1

Artificial Intelligence and Agents

> *The history of AI is a history of fantasies, possibilities, demonstrations, and promise. Ever since Homer wrote of mechanical "tripods" waiting on the gods at dinner, imagined mechanical assistants have been a part of our culture. However, only in the last half century have we, the AI community, been able to build experimental machines that test hypotheses about the mechanisms of thought and intelligent behavior and thereby demonstrate mechanisms that formerly existed only as theoretical possibilities.*
>
> – Bruce Buchanan [2005]

This book is about artificial intelligence, a field built on centuries of thought, which has been a recognized discipline for over 60 years. As Buchanan points out in the quote above, we now have the tools to test hypotheses about the nature of thought itself, as well as to solve practical tasks. Deep scientific and engineering problems have already been solved and many more are waiting to be solved. Many practical applications are currently deployed and the potential exists for an almost unlimited number of future applications. In this book, we present the principles that underlie intelligent computational agents. These principles can help you understand current and future work in AI and equip you to contribute to the discipline yourself.

1.1 What is Artificial Intelligence?

Artificial intelligence, or **AI**, is the field that studies *the synthesis and analysis of computational agents that act intelligently*. Let us examine each part of this definition.

An **agent** is something that acts in an environment; it does something. Agents include worms, dogs, thermostats, airplanes, robots, humans, companies, and countries.

We are interested in what an agent does; that is, how it **acts**. We judge an agent by its actions.

An agent acts **intelligently** when

- what it does is appropriate for its circumstances and its goals, taking into account the short-term and long-term consequences of its actions
- it is flexible to changing environments and changing goals
- it learns from experience
- it makes appropriate choices given its perceptual and computational limitations

A **computational agent** is an agent whose decisions about its actions can be explained in terms of computation. That is, the decision can be broken down into primitive operations that can be implemented in a physical device. This computation can take many forms. In humans this computation is carried out in "wetware"; in computers it is carried out in "hardware." Although there are some agents that are arguably not computational, such as the wind and rain eroding a landscape, it is an open question whether all intelligent agents are computational.

All agents are limited. No agents are omniscient or omnipotent. Agents can only observe everything about the world in very specialized domains, where "the world" is very constrained. Agents have finite memory. Agents in the real world do not have unlimited time to act.

The central **scientific goal** of AI is to understand the principles that make intelligent behavior possible in natural or artificial systems. This is done by

- the **analysis** of natural and artificial agents
- formulating and testing hypotheses about what it takes to construct intelligent agents and
- designing, building, and experimenting with computational systems that perform tasks commonly viewed as requiring intelligence.

As part of science, researchers build **empirical systems** to test hypotheses or to explore the space of possible designs. These are quite distinct from **applications** that are built to be useful for an application domain.

The definition is not for intelligent **thought** alone. We are only interested in **thinking** intelligently insofar as it leads to more intelligent **behavior**. The role of thought is to affect action.

The central **engineering goal** of AI is the **design** and **synthesis** of useful, intelligent artifacts. We actually want to build agents that act intelligently. Such agents are useful in many applications.

1.1.1 Artificial and Natural Intelligence

Artificial intelligence (AI) is the established name for the field, but the term "artificial intelligence" is a source of much confusion because artificial intelligence may be interpreted as the opposite of real intelligence.

For any phenomenon, you can distinguish real versus fake, where the fake is non-real. You can also distinguish natural versus artificial. Natural means occurring in nature and artificial means made by people.

Example 1.1 A tsunami is a large wave in an ocean. Natural tsunamis occur from time to time and are caused by earthquakes or landslides. You could imagine an artificial tsunami that was made by people, for example, by exploding a bomb in the ocean, yet which is still a real tsunami. One could also imagine fake tsunamis: either artificial, using computer graphics, or natural, for example, a mirage that looks like a tsunami but is not one.

It is arguable that intelligence is different: you cannot have *fake* intelligence. If an agent behaves intelligently, it is intelligent. It is only the external behavior that defines intelligence; acting intelligently is being intelligent. Thus, artificial intelligence, if and when it is achieved, will be real intelligence created artificially.

This idea of intelligence being defined by external behavior was the motivation for a test for intelligence designed by Turing [1950], which has become known as the **Turing test**. The Turing test consists of an imitation game where an interrogator can ask a witness, via a text interface, any question. If the interrogator cannot distinguish the witness from a human, the witness must be intelligent. Figure 1.1 (on the next page) shows a possible dialog that Turing suggested. An agent that is not really intelligent could not fake intelligence for arbitrary topics.

There has been much debate about the usefulness of Turing test. Unfortunately, although it may provide a test for how to recognize intelligence, it does not provide a way to realize intelligence.

Recently Levesque [2014] suggested a new form of question, which he called a **Winograd schema** after the following example of Winograd [1972]:

- The city councilmen refused the demonstrators a permit because they feared violence. Who feared violence?
- The city councilmen refused the demonstrators a permit because they advocated violence. Who advocated violence?

These two sentences only differ in one word feared/advocated, but have the opposite answer. Answering such a question does not depend on trickery or lying, but depends on knowing something about the world that humans understand, but computers currently do not.

Interrogator: In the first line of your sonnet which reads "Shall I compare thee to a summer's day," would not "a spring day" do as well or better?

Witness: It wouldn't scan.

Interrogator: How about "a winter's day," That would scan all right.

Witness: Yes, but nobody wants to be compared to a winter's day.

Interrogator: Would you say Mr. Pickwick reminded you of Christmas?

Witness: In a way.

Interrogator: Yet Christmas is a winter's day, and I do not think Mr. Pickwick would mind the comparison.

Witness: I don't think you're serious. By a winter's day one means a typical winter's day, rather than a special one like Christmas.

Figure 1.1: Part of Turing's possible dialog for the Turing test

Winograd schemas have the property that (a) humans can easily disambiguate them and (b) there is no simple grammatical or statistical test that could disambiguate them. For example, the sentences above would not qualify if "demonstrators feared violence" was much less or more likely than "councilmen feared violence" (or similarly with advocating).

Example 1.2 The following examples are due to Davis [2015]:

- Steve follows Fred's example in everything. He [admires/influences] him hugely. Who [admires/influences] whom?
- The table won't fit through the doorway because it is too [wide/narrow]. What is too [wide/narrow]?
- Grace was happy to trade me her sweater for my jacket. She thinks it looks [great/dowdy] on her. What looks [great/dowdy] on Grace?
- Bill thinks that calling attention to himself was rude [to/of] Bert. Who called attention to himself?

Each of these have their own reasons why one answer is preferred to the other. A computer that can reliably answer these questions needs to know about all of these reasons, and require the ability to do **commonsense reasoning**.

The obvious naturally intelligent agent is the human being. Some people might say that worms, insects, or bacteria are intelligent, but more people would say that dogs, whales, or monkeys are intelligent (see Exercise 1.1 (page 47)). One class of intelligent agents that may be more intelligent than humans is the class of **organizations**. Ant colonies are a prototypical example of organizations. Each individual ant may not be very intelligent, but an ant

colony can act more intelligently than any individual ant. The colony can discover food and exploit it very effectively as well as adapt to changing circumstances. Corporations can be more intelligent than individual people. Companies develop, manufacture, and distribute products where the sum of the skills required is much more than any individual could master. Modern computers, from low-level hardware to high-level software, are more complicated than any human can understand, yet they are manufactured daily by organizations of humans. Human **society** viewed as an agent is arguably the most intelligent agent known.

It is instructive to consider where human intelligence comes from. There are three main sources:

Biology Humans have evolved into adaptable animals that can survive in various habitats.

Culture Culture provides not only language, but also useful tools, useful concepts, and the wisdom that is passed from parents and teachers to children.

Lifelong learning Humans learn throughout their life and accumulate knowledge and skills.

These sources interact in complex ways. Biological evolution has provided stages of growth that allow for different learning at different stages of life. Biology and culture have evolved together; humans can be helpless at birth presumably because of our culture of looking after infants. Culture interacts strongly with learning. A major part of lifelong learning is what people are taught by parents and teachers. Language, which is part of culture, provides distinctions in the world that are useful for learning.

When building an intelligent system, the designers have to decide which of these sources of intelligence need to be programmed in, and which can be learned. It is very unlikely we will be able to build an agent that starts with a clean slate and learns everything. Similarly, most interesting and useful intelligent agents learn to improve their behavior

1.2 A Brief History of Artificial Intelligence

Throughout human history, people have used technology to model themselves. There is evidence of this from ancient China, Egypt, and Greece bearing witness to the universality of this activity. Each new technology has, in its turn, been exploited to build intelligent agents or models of mind. Clockwork, hydraulics, telephone switching systems, holograms, analog computers, and digital computers have all been proposed both as technological metaphors for intelligence and as mechanisms for modeling mind.

About 400 years ago people started to write about the nature of thought and reason. Hobbes (1588–1679), who has been described by Haugeland [1985, p. 85] as the "Grandfather of AI," espoused the position that thinking was symbolic reasoning like talking out loud or working out an answer with pen and paper. The idea of symbolic reasoning was further developed by Descartes (1596–1650), Pascal (1623–1662), Spinoza (1632–1677), Leibniz (1646–1716), and others who were pioneers in the philosophy of mind.

The idea of symbolic operations became more concrete with the development of computers. Babbage (1792–1871) designed the first general-purpose computer, the **Analytical Engine**, but it was not built until 1991 at the Science Museum of London. In the early part of the twentieth century, there was much work done on understanding computation. Several models of computation were proposed, including the Turing machine by Alan Turing (1912–1954), a theoretical machine that writes symbols on an infinitely long tape, and the lambda calculus of Church (1903–1995), which is a mathematical formalism for rewriting formulas. It can be shown that these very different formalisms are equivalent in that any function computable by one is computable by the others. This leads to the **Church–Turing thesis**:

> Any effectively computable function can be carried out on a Turing machine (and so also in the lambda calculus or any of the other equivalent formalisms).

Here **effectively computable** means following well-defined operations; "computers" in Turing's day were people who followed well-defined steps and computers as we know them today did not exist. This thesis says that all computation can be carried out on a Turing machine or one of the other equivalent computational machines. The Church–Turing thesis cannot be proved but it is a hypothesis that has stood the test of time. No one has built a machine that has carried out computation that cannot be computed by a Turing machine. There is no evidence that people can compute functions that are not Turing computable. An agent's actions are a function of its abilities, its history, and its goals or preferences. This provides an argument that computation is more than just a metaphor for intelligence; reasoning *is* computation and computation can be carried out by a computer.

Once real computers were built, some of the first applications of computers were AI programs. For example, Samuel [1959] built a checkers program in 1952 and implemented a program that learns to play checkers in the late 1950s. His program beat the Connecticut state checkers champion in 1961. Wang [1960] implemented a program that proved every logic theorem (nearly 400) in *Principia Mathematica* [Whitehead and Russell, 1910, 1912, 1913]. Newell and Simon [1956] built a program, Logic Theorist, that discovers proofs in propositional logic.

In addition to work on high-level symbolic reasoning, there was also much work on low-level learning inspired by how **neurons** work. McCulloch and Pitts [1943] showed how a simple thresholding "formal neuron" could be the basis for a Turing-complete machine. The first learning for these neural networks was described by Minsky [1952]. One of the early significant works was the **perceptron** of Rosenblatt [1958]. The work on neural networks went into decline for a number of years after the 1968 book by Minsky and Papert [1988], which argued that the representations learned were inadequate for intelligent action.

The early programs concentrated on learning and search as the foundations of the field. It became apparent early that one of the main tasks was how to represent the knowledge required for intelligent action. Before learning, an agent must have an appropriate target language for the learned knowledge. There have been many proposals for representations from simple features to neural networks to the complex logical representations of McCarthy and Hayes [1969] and many in between, such as the frames of Minsky [1975].

During the 1960s and 1970s, natural language understanding systems were developed for limited domains. For example, the STUDENT program of Daniel Bobrow [1967] could solve high school algebra tasks expressed in natural language. Winograd's [1972] SHRDLU system could, using restricted natural language, discuss and carry out tasks in a simulated blocks world. CHAT-80 [Warren and Pereira, 1982] could answer geographical questions placed to it in natural language. Figure 1.2 (on the next page) shows some questions that CHAT-80 answered based on a database of facts about countries, rivers, and so on. All of these systems could only reason in very limited domains using restricted vocabulary and sentence structure. Interestingly, IBM's **Watson**, which beat the world champion in the TV game show Jeopardy!, used a similar technique to CHAT-80 [Lally et al., 2012]; see Section 13.6 (page 612).

During the 1970s and 1980s, there was a large body of work on **expert systems**, where the aim was to capture the knowledge of an expert in some domain so that a computer could carry out expert tasks. For example, **DENDRAL** [Buchanan and Feigenbaum, 1978], developed from 1965 to 1983 in the field of organic chemistry, proposed plausible structures for new organic compounds. **MYCIN** [Buchanan and Shortliffe, 1984], developed from 1972 to 1980, diagnosed infectious diseases of the blood, prescribed antimicrobial therapy, and explained its reasoning. The 1970s and 1980s were also a period when AI reasoning became widespread in languages such as **Prolog** [Colmerauer and Roussel, 1996; Kowalski, 1988].

During the 1990s and the 2000s there was great growth in the subdisciplines of AI such as perception, probabilistic and decision-theoretic reasoning, planning, embodied systems, machine learning, and many other fields. There has also been much progress on the foundations of the field; these form the frame-

Does Afghanistan border China?
What is the capital of Upper_Volta?
Which country's capital is London?
Which is the largest African country?
How large is the smallest American country?
What is the ocean that borders African countries and that borders Asian countries?
What are the capitals of the countries bordering the Baltic?
How many countries does the Danube flow through?
What is the total area of countries south of the Equator and not in Australasia?
What is the average area of the countries in each continent?
Is there more than one country in each continent?
What are the countries from which a river flows into the Black_Sea?
What are the continents no country in which contains more than two cities whose population exceeds 1 million?
Which country bordering the Mediterranean borders a country that is bordered by a country whose population exceeds the population of India?
Which countries with a population exceeding 10 million border the Atlantic?

Figure 1.2: Some questions CHAT-80 could answer

work of this book.

1.2.1 Relationship to Other Disciplines

AI is a very young discipline. Other disciplines as diverse as philosophy, neurobiology, evolutionary biology, psychology, economics, political science, sociology, anthropology, control engineering, statistics, and many more have been studying aspects of intelligence much longer.

The science of AI could be described as "synthetic psychology," "experimental philosophy," or "computational epistemology"– **epistemology** is the study of knowledge. AI can be seen as a way to study the nature of knowledge and intelligence, but with a more powerful experimental tool than was previously available. Instead of being able to observe only the external behavior of intelligent systems, as philosophy, psychology, economics, and sociology have traditionally been able to do, AI researchers experiment with executable models of intelligent behavior. Most important, such models are open to inspection, redesign, and experiment in a complete and rigorous way. Modern computers provide a way to construct the models about which philosophers have only been able to theorize. AI researchers can experiment with these models as op-

posed to just discussing their abstract properties. AI theories can be empirically grounded in implementations. Moreover, we are often surprised when simple agents exhibit complex behavior. We would not have known this without implementing the agents.

It is instructive to consider an analogy between the development of **flying machines** over the past few centuries and the development of thinking machines over the past few decades. There are several ways to understand flying. One is to dissect known flying animals and hypothesize their common structural features as necessary fundamental characteristics of any flying agent. With this method, an examination of birds, bats, and insects would suggest that flying involves the flapping of wings made of some structure covered with feathers or a membrane. Furthermore, the hypothesis could be tested by strapping feathers to one's arms, flapping, and jumping into the air, as Icarus did. An alternative methodology is to try to understand the principles of flying without restricting oneself to the natural occurrences of flying. This typically involves the construction of artifacts that embody the hypothesized principles, even if they do not behave like flying animals in any way except flying. This second method has provided both useful tools – airplanes – and a better understanding of the principles underlying flying, namely **aerodynamics**.

AI takes an approach analogous to that of aerodynamics. AI researchers are interested in testing general hypotheses about the nature of intelligence by building machines that are intelligent and that do not necessarily mimic humans or organizations. This also offers an approach to the question, "Can computers really think?" by considering the analogous question, "Can airplanes really fly?"

AI is intimately linked with the discipline of computer science because the study of computation is central to AI. It is essential to understand algorithms, data structures, and combinatorial complexity to build intelligent machines. It is also surprising how much of computer science started as a spinoff from AI, from timesharing to computer algebra systems.

Finally, AI can be seen as coming under the umbrella of **cognitive science**. Cognitive science links various disciplines that study cognition and reasoning, from psychology to linguistics to anthropology to neuroscience. AI distinguishes itself within cognitive science by providing tools to build intelligence rather than just studying the external behavior of intelligent agents or dissecting the inner workings of intelligent systems.

1.3 Agents Situated in Environments

AI is about practical reasoning: reasoning in order to do something. A coupling of perception, reasoning, and acting comprises an **agent**. An agent acts in an