Chapter 1
A General Introduction to Artificial Intelligence

The emergence and rise of artificial intelligence undoubtedly played an important role during the development of the Internet. Over the past decade, with extensive applications in the society, artificial intelligence has become more relevant to people's daily life. This chapter introduces the concept of artificial intelligence, the related technologies, and the existing controversies over the topic.

1.1 The Concept of Artificial Intelligence

1.1.1 What Is Artificial Intelligence?

Currently, people mainly learn about artificial intelligence (AI) through news, movies, and the applications in daily life, as shown by Fig. 1.1.

A rather widely accepted definition of AI, also a relatively early one, was proposed by John McCarthy at the 1956 Dartmouth Conference, which outlined that artificial intelligence is about letting a machine simulate the intelligent behavior of humans as precisely as it can be. However, this definition seemingly ignores the possibility of strong artificial intelligence (which means the machine that has the ability or intelligence to solve problems by reasoning).

Before explaining what "artificial intelligence" is, we had better clarify the concept of "intelligence" first.

According to the theory of multiple intelligences, human intelligence can be categorized into seven types: Linguistic, Logical-Mathematical, Spatial Bodily-Kinesthetic, Musical, Interpersonal and Intrapersonal intelligence.

1. Linguistic Intelligence

   Linguistic intelligence refers to the ability to effectively express one's thoughts in spoken or written language, understand others' words or texts, flexibly master the phonology, semantics, and grammar of a language, manage

Huawei Technologies Co., Ltd., *Artificial Intelligence Technology*,
https://doi.org/10.1007/978-981-19-2879-6_1

Fig. 1.1 Social cognition of AI

verbal thinking, and convey or decode the connotation of linguistic expressions through the verbal thinking. For the people with strong linguistic intelligence, the ideal career choices could be politician-activist, host, attorney, public speaker, editor, writer, journalist, teacher, etc.

2. Logical-Mathematical Intelligence

   Logical-mathematical intelligence designates the capability to calculate, quantify, reason, summarize and classify effectively, and to carry out complicated mathematical operations. This capability is characterized by the sensitivity to abstract concepts, such as logical patterns and relationships, statements and claims, and functions. People who are strong in logic-mathematical intelligence are more suitable to work as scientists, accountants, statisticians, engineers, computer software developers, etc.

3. Spatial Intelligence

   Spatial intelligence features the potential to accurately recognize the visual space and things around it, and to represent what they perceived visually in paintings and graphs. People with strong spatial intelligence are very sensitive to spatial relationships such as color, line, shape, and form. The jobs suitable for them are interior designer, architect, photographer, painter, pilot and so on.

4. Bodily-Kinesthetic Intelligence

   Bodily-kinesthetic intelligence indicates the capacity to use one's whole body to express thoughts and emotions, and to use hands and other tools to fashion products or manipulate objects. This intelligence demonstrates a variety of particular physical skills such as balance, coordination, agility, strength, suppleness and speed, and tactile abilities. Potential careers for people with strong body-kinesthetic intelligence include athlete, actor, dancer, surgeon, jeweler, mechanic and so on.

5. Musical Intelligence

   Musical intelligence is the ability to discern pitch, tone, melody, rhythm, and timbre. People having relatively high musical intelligence are particularly sensitive to pitch, tone, melody, rhythm or timbre, and are more competitive in

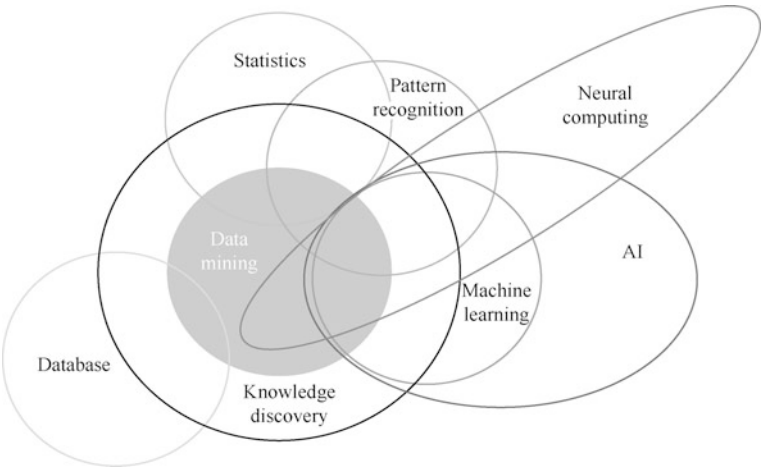

Fig. 1.2 Fields covered by artificial intelligence

performing, creating and reflecting on music. Their recommended professions include singer, composer, conductor, music critic, the piano tuner and so on.

6. Interpersonal Intelligence

   Interpersonal intelligence is the capability to understand and interact effectively with others. People with strong interpersonal intelligence can better recognize the moods and temperaments of others, empathize with their feelings and emotions, notice the hidden information of different interpersonal relationships, and respond appropriately. The professions suitable for them include politician, diplomat, leader, psychologist, PR officer, salesmen, and so on.

7. Intrapersonal Intelligence

   Intrapersonal intelligence is about self-recognition, which means the capability to understand oneself and then act accordingly based on such knowledge. People with strong intrapersonal intelligence are able to discern their strengths and weaknesses, recognize their inner hobbies, moods, intentions, temperaments and self-esteem, and they like to think independently. Their suitable professions include philosopher, politician, thinker, psychologist and so on.

8. Naturalist Intelligence

   Naturalist intelligence refers to the ability to observe the various forms of nature, identify and classify the objects, and discriminate the natural and artificial systems.

However, AI is a new type of technological science that investigates and develops the theories, methods, technologies and application systems to simulate, improve and upgrade the human intelligence. The AI is created to enable machines to reason like human being and to endow them with intelligence. Today, the connotation of AI has been greatly broadened, making it an interdisciplinary subject, as shown by Fig. 1.2.

Machine learning (ML) is apparently one of the major focuses of this interdisciplinary subject. According to the definition by Tom Mitchell, the so-called "the godfather of global ML", machine learning is described as: with respect to certain type of tasks T and performance P, if the performance of a computer program at tasks in T improves with experience E as measured by P, then the computer program is deemed to learn from experience E. It is a relatively simple and abstract definition. However, as our perception on the concept deepened, we may find that the connotation and denotation of machine learning will also change accordingly. It is not easy to define machine learning that precisely in only one or two sentences, not only because that it covers a wide span of fields in terms of theory and application, but also it is developing and transforming quite rapidly.

Generally speaking, the processing system and algorithms of machine learning make predictions mainly by identifying the hidden patterns from data. It is an important sub-field of AI, and AI is intertwined with data mining (DM) and knowledge discovery in database (KDD) in a broader sense.

1.1.2 The Relationship Between AI, Machine Learning, and Deep Learning

The study of machine learning aims at enabling computers to simulate or perform human learning ability and acquire new knowledge and skills. Deep learning (DL) derives from the study of artificial neural networks (ANN). As a new subfield of machine learning, it focuses on mimicking the mechanisms of human brain in interpreting data like images, sound, and text.

The relationship between AI, machine learning, and deep learning is shown in Fig. 1.3.

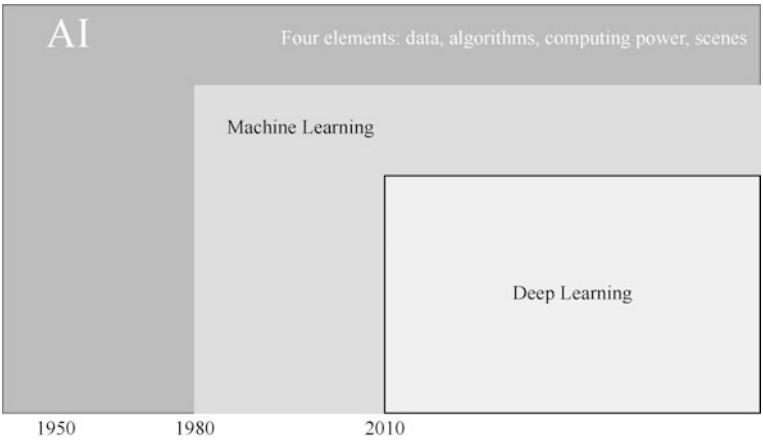

Fig. 1.3 The relationship between artificial intelligence, machine learning, and deep learning

Among the three concepts, machine learning is an approach or a subset of AI, and deep learning is one of ML's special forms. If we take AI as the brain, then machine learning is the process of the acquisition of cognitive abilities, and deep learning is a highly efficient self-training system that dominates this process. Artificial intelligence is the target and result while deep learning and machine learning are methods and tools.

1.1.3 Types of AI

AI can be divided into two types: strong artificial intelligence and weak artificial intelligence.

Strong artificial intelligence is about the possibility to create the intelligent machines that can accomplish reasoning problem-solving tasks. The machines of this kind are believed to have consciousness and self-awareness and be able to think independently and come up with the best solutions to the problems. Strong AI also has its distinctive values and worldview, and is endowed with instincts, such as the needs of survival and safety, just like all the living beings. In a certain sense, strong AI is a new civilization.

Weak artificial intelligence depicts the circumstance when it is not able to make machines that can truly accomplish reasoning and problem-solving. These machines may look smart, but they do not really have intelligence or self-awareness.

We are currently in the era of weak artificial intelligence. The introduction of weak artificial intelligence reduces the burden of intellectual work by functioning in a way similar to the advanced bionics. Whether it is AlphaGo, or the robot who writes news report and novels, they all belong to weak artificial intelligence and outperform humans in only certain fields. In the era of weak artificial intelligence, it is undeniable that data and computing power are both crucial, as they can facilitate the commercialization of AI. In the coming age of strong artificial intelligence, these two factors will still be two decisive elements. Meanwhile, the exploration on quantum computing by companies like Google and IBM also lays the foundation for the advent of the strong artificial intelligence era.

1.1.4 The History of AI

Figure 1.4 presents us a brief history of AI.

The official origin of modern AI can be traced back to the Turing Test proposed by Alan M. Turing, known as "the Father of Artificial Intelligence", in 1950. According to his assumption, if a computer can engage in dialogue with humans without being detected as a computer, then it is deemed as having intelligence. In the same year he proposed this assumption, Turing boldly predicted that creating the

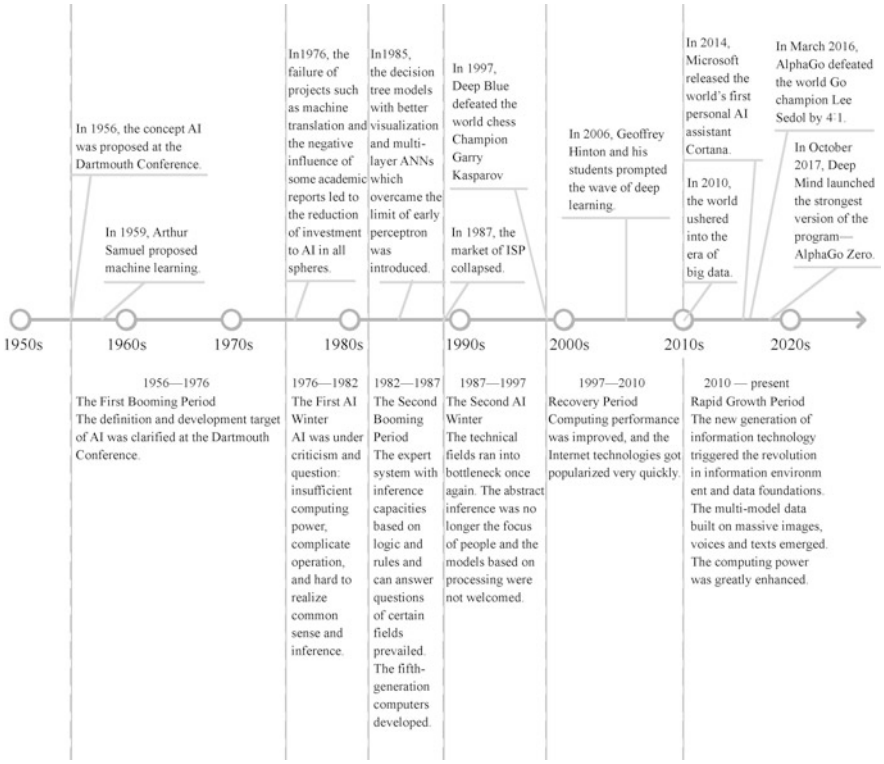

Fig. 1.4 A brief history of AI

machines with real human intelligence was possible in the future. But up to now, none of the computers has ever passed the Turing Test.

Although AI is a concept with a history of only a few decades, the theoretical foundation and supporting technology behind it have developed for a time-honored period. The current prosperity of AI is a result of the advancement of all related disciplines and the collective efforts of the scientists of all generations.

1. Precursors and Initiation Period (before 1956)

   The theoretical foundation for the birth of AI can date back to as early as the fourth century BC, when the famous ancient Greek philosopher and scientist Aristotle invented the concept of formal logic. In fact, his theory of syllogism is still working as an indispensable and decisive cornerstone for the deductive reasoning today. In the seventeenth century, the German mathematician Gottfried Leibniz advanced universal notation and some revolutionary ideas on reasoning and calculation, which laid the foundation for the establishment and development of mathematical logic. In the nineteenth century, the British mathematician George Boole developed Boolean algebra, which is the bedrock of the operation of modern computers, and its introduction makes the invention of computer

possible. During the same period, the British inventor Charles Babbage created the Difference Engine, the first computer capable of solving quadratic polynomial equations in the world. Although it only had limited functions, this computer reduced the burden of human brain in calculation per se for the first time. The machines were endowed with computational intelligence ever since.

In 1945, John Mauchly and J. Presper Eckert from a team at Moore School designed the Electronic Numerical Integrator and Calculator (ENIAC), the world's first general-purpose digital electronic computer. As an epoch-making achievement, ENIAC still had its fatal deficiencies, such as its enormous size, excessive power consumption, and reliance on manual operation to input and adjust commands. In 1947, John von Neumann, the father of modern computers, modified and upgraded on the basis of ENIAC and created the modern electronic computer in the real sense: Mathematical Analyzer Numerical Integrator and Automatic Computer (MANIAC).

In 1946, the American physiologist Warren McCulloch established the first model of neural network. His research on artificial intelligence at microscopic level laid an important foundation for the development of neural networks. In 1949, Donald O. Hebb proposed Hebbian theory, a neuropsychological learning paradigm, which states the basic principles of synaptic plasticity, namely, the efficacy of synaptic transmission will arise greatly with the repeated and persistent stimulation from a presynaptic neuron to a postsynaptic neuron. This theory is fundamental to the modelling of neural networks. In 1948, Claude E. Shannon, the founder of information theory, introduced the concept of information entropy by borrowing the term from thermodynamics, and defined information entropy as the average amount of information after the redundancy has being removed. The impact of this theory is quite far-reaching as it played an important role in fields such as non-deterministic inference and machine learning.

2. The First Booming Period (1956–1976)

Finally, in 1956, John McCarthy officially introduced AI as a new discipline at the 2-month long Dartmouth Conference, which marks the birth of AI. A number of AI research groups were formed in the United States ever since, such as the Carnegie-RAND group formed by Allen Newell and Herbert Alexander Simon, the research group the Massachusetts Institute of Technology (MIT) by Marvin Lee Minsky and John McCarthy, and Arthur Samuel's IBM engineering research group, etc.

In the following two decades, AI was developing rapidly in a wide range of fields, and thanks to the great enthusiasm of researchers, the AI technologies and applications have kept expanding.

(a) Machine Learning

In 1956, Arthur Samuel of IBM wrote the famous checkers-playing program, which was able to learn an implicit model by observing the positions on checkerboard to instruct moves for the latter cases. After played against the program for several rounds, Arthur Samuel concluded that the program could reach a very high level of performance during the course of

learning. With this program, Samuel confuted the notion that computers cannot go beyond the written codes and learn patterns like human beings. Since then, he coined and defined a new term—machine learning.

(b) Pattern Recognition

In 1957, C.K. Chow proposed to adopt statistical decision theory to tackle pattern recognition, which stimulated the rapid development of pattern recognition research since the late 1950s. In the same year, Frank Rosenblatt proposed a simplified mathematical model that imitated the recognition pattern of human brain—the perceptron, the first machine that could possibly train the recognition system by the sample of each given category, so that the system was able to correctly classify patterns of other unknown categories after learning.

(c) Pattern Matching

In 1966, ELIZA, the first conversation program in the world was invented, which was written by the MIT Artificial Intelligence Laboratory. The program was able to perform pattern matching on the basis of the set rules and user's questions, so as to give appropriate replies by choosing from the pre-written answer archives. This is also the first program try to have passed the Turing Test. ELIZA once masqueraded as a psychotherapist to talk to patients, and many of them failed to recognize it as a robot when it was firstly applied. "Conversation is pattern matching", thus this unveiled the computer natural language conversation technology.

In addition, during the first development period of AI, John McCarthy developed the LISP, which became the dominant programming language for AI IN the following decades. Marvin Minsky launched a more in-depth study of neural networks and discovered the disadvantages of simple neural networks. In order to overcome these limitations, the scientists started to introduce multilayer neural networks and Back Propagation (BP) algorithms. Meanwhile, the expert system (ES) also emerged. During this period the first industrial robot was applied on the production line of General Motors, and the world also witnessed the birth of the first mobile robot which was capable of actioning autonomously.

The advancement of relevant disciplines also contributed to the great strides of AI. The emergence of bionics in the 1950s ignited the research enthusiasm of scientists, which led to the invention of simulated annealing algorithm. It is a type of heuristic algorithm, and is the foundation for the searching algorithms, such as the ant colony optimization (ACO) algorithm which is quite popular in recent years.

3. The First AI Winter (1976–1982)

However, the AI manic did not last too long, as the over-optimistic projections failed to be fulfilled as promised, and thus incurred the doubt and suspicion on AI technology globally.

The perceptron, once a sensation in the academic world, had a hard time in 1969 when Marvin Minsky and the rest scientists advanced the famous logical

operation exclusive OR (XOR), demonstrating the limitation of the perceptron in terms of the linear inseparable data similar to the XOR problem. For the academic world, the XOR problem became an almost undefeatable challenge.

In 1973, AI was under strict questioning by the scientific community. Many scientists believed that those seemingly ambitious goals of AI were just some unfulfilled illusions, and the relevant research had been proved complete failures. Due to the increasing suspicion and doubts, AI suffered from severe criticism, and its actual value was also under question. As a consequence, the governments and research institutions all over the world withdrew or reduced funding on AI, and the industry encountered its first winter of development in the 1970s.

The setback in 1970s was no coincidence. Due to the limitation of computing power at that time, although many problems were solved theoretically, they cannot be put into practice at all. Meanwhile, there were many other obstacles, such as it was difficult for the expert system to acquire knowledge, leaving lots of projects ended in failure. The study on machine vision took off in the 1960s. And the edge detection and contour composition methods proposed by the American scientist L.R. Roberts are not only time-tested, but also still widely used today. However, having a theoretical basis does not mean actual yield. In the last 1970s, there were scientists concluded that to let a computer to imitate human retinal vision, it would need to execute at least one billion instructions. However, the calculation speed of the world's fastest supercomputer Cray-1 in 1976 (which costed millions of US dollars to make) could only register no more than 100 million times per second, and the speed of an ordinary computer could meet even no more than one million times per second. The hardware limited the development of AI. In addition, another major basis for the progress of AI is the data base. At that time, computers and the Internet were not as popular as today, so there were nowhere for the developers to capture massive data.

During this period, artificial intelligence developed slowly. Although the concept of BP had been proposed by Juhani Linnainmaa in the "automatic differential flip mode" in the 1970s, it was not until 1981 that it was applied to the multilayer perceptron by Paul J. Werbos. The invention of multilayer perceptron and BP algorithm contributed to the second leap-frogging of neural networks. In 1986, David E. Rumelhart and other scholars developed an effective BP algorithm to successfully train multilayer perceptron, which exerted a profound impact.

4. The Second Booming Period (1982–1987)

In 1980, XCON, a complete expert system developed by the Carnegie Mello University (CMU) was officially put into use. The system contained more than 2500 set rules, and processed more than 80,000 orders featuring an accuracy of over 95% in the following years. This is considered a milestone that heralds a new era, when the expert system begun to showcase its potential in specific fields, which lifted AI technology to a completely new level of booming development.

An expert system normally attends to one specific professional field. By mimicking the thinking of human experts, it attempts to answer questions or provide knowledge to help with the decision-making by practitioners. Focusing

on only a narrow domain, the expert system avoids the challenges related to artificial general intelligence (AGI) and is able to make full use of the knowledge and experience of existing experts to solve problems of the specific domains.

The big commercial success of XCON encouraged 60% of Fortune 500 companies to embark on the development and deployment of expert systems in their respective fields in the 1980s. According to the statistics, from 1980 to 1985, more than 1 billion US dollars was invested in AI, with a majority went to the internal AI departments of those enterprises, and the market witnessed a surge in AI software and hardware companies.

In 1986, the Bundeswehr University in Munich equipped a Mercedes-Benz van with a computer and several sensors, which enabled an automatic control of the steering wheel, accelerator and brake. The installation was called VaMoRs, which proved to be the first self-driving car in the real sense in the world.

LISP was the mainstream programming language used for AI development at that time. In order to enhance the operating efficiency of LISP programs, many agencies turned to develop computer chips and storage devices designed specifically to executive LISP programs. Although LISP machines had made some progress, personal computers (PCs) were also on the rise. IBM and Apple quickly expanded the market presence in the entire computer marketplace. With a steady increase of CPU frequency and speed, the PCs were becoming even more powerful than the costly LISP machines.

5. The Second AI Winter (1987–1997)

   In 1987, along with the crash of sales market of LISP machine hardware, the AI industry once again fell into another winter. The second AI trough period lasted for years as the hardware market collapsed and governments and institutions all over the world stopped investing in AI research. But during this period, the researchers still made some important achievements. In 1988, the American scientist Judea Pearl championed the probabilistic approach to AI inference, which made a crucial contribution to the future development of AI technology.

   In the almost 20 years after the advent of the second AI winter, the AI technology became gradually and deeply integrated with computer and software technologies, while the research on artificial intelligence algorithm theory had a slow progress. The research results of many researchers were only something based on the old theories, and the computer hardware that was more powerful and faster.

6. Recovery Period (1997–2010)

   In 1995, Richard S. Wallace was inspired by ELIZA and developed a new chatbot program named A.L.I.C.E. (the Artificial Linguistic Internet Computer Entity). The robot was able to optimize the contents and enrich its datasets automatically through the Internet.

   In 1996, the IBM supercomputer Deep Blue played a chess game against the world chess champion Gary Kasparov and was defeated. Gary Kasparov believed that it was impossible for computers to defeat human in chess games ever. After the match, IBM upgraded Deep Blue. The new Deep Blue was enhanced with 480 specialized CPUs and a doubled calculation speed up to 200 million times per

second, enabling it to predict the next 8 or more moves on the chessboard. In the later rematch, the computer defeated Gary Kasparov successfully. However, this landmark event actually only marks a victory of computer over human in a game with clear rules by relying on its calculation speed and enumeration. This is not real AI.

In 2006, as Geoffrey Hinton published a paper in *Science Magazine*, AI industry entered the era of deep learning.

7. Rapid Growth Period (2010–present)

In 2011, the Watson system, also a program from IBM, participated the quiz show *Jeopardy*, competing with human players. The Watson system defeated two human champions with its outstanding natural language processing capabilities and powerful knowledge database. This time, computers can already comprehend human language, which is a big advancement in AI.

In the twenty-first century, with the widespread application of PCs and the burst of mobile Internet and cloud computing technology, the institutions are able to capture and accumulate an unimaginably huge mass of data, providing sufficient material and impetus for the ongoing development of AI. Deep learning became a mainstream of AI technology, exemplified by the famous Google Brain project, which enhanced the recognition rate of the ImageNet dataset to 84% by a large margin.

In 2011, the concept semantic network was proposed. The concept steams from the World Wide Web. It is essentially a large-scale distributed database that centers on Web data and connects Web data in the method of machine understanding and processing. The emergence of the semantic network greatly promoted the progress of technology of knowledge representation. A year later, Google first announced the concept of knowledge graph and launched a knowledge-graph-based searching service.

In 2016 and 2017, Google launched two Go competitions between human and mechanical players that shocked the world. Its AI program AlphaGo defeated two Go world champions, first Lee Sedol of South Korea and then Ke Jie of China.

Today, AI can be found in almost all aspects of people's life. For instance, the voice assistant, such as the most typical Siri of Apple, is based on the natural language processing (NLP) technology. With the support of NLP, computers can process human language and match it with the commands and responses in line with human expectation more and more naturally. When users are browsing e-commerce websites, they could possibly receive product recommendation feeds generated by a recommendation algorithm. The recommendation algorithm can predict the products that the users might want to buy by reviewing and analyzing the historical data of the users' recent purchases and preferences.

1.1.5 The Three Main Schools of AI

Currently, symbolism, connectionism, and behaviorism constitute the three main schools of AI. The following passages will introduce them in detail.

1. Symbolism

   The basic theory of symbolism believes that, the cognitive process of human being consists of the inference and processing of symbols. Human is an example of physical symbol system, and so does the computer. Therefore, computers should be able to simulate human intelligent activities. And knowledge representation, knowledge reasoning, and knowledge application are three crucial to artificial intelligence. Symbolism argues that knowledge and concepts can be represented by symbols, thus cognition is a process of processing the symbols, and reasoning is a process of solving problems with heuristic knowledge. The core of symbolism lies in reasoning, namely the symbolic reasoning and machine reasoning.
2. Connectionism

   The foundation of connectionism is that the nature of human logical thinking is neurons, rather than a process of symbol processing. Connectionism believes that the human brain is different from computers, and put forward a connectionist model imitating brain work to replace the computer working model operated by symbols. Connectionism is believed to stem from bionics, especially in the study of human brain models. In connectionism, a concept is represented by a set of numbers, vectors, matrices, or tensors, namely, by the specific activation mode of the entire network. Each node (neuron) in the network has no specific meaning, but every node all participates in the expression of overall concept. For example, in symbolism, the concept of a cat can be represented by a "cat node" or a group of nodes that feature the attributes of a cat (e.g., the one with "two eyes", "four legs" or "fluffy"). However, connectionism believes that each node does not have a specific meaning, so it is impossible to search for a "cat node" or "eye neuron". The core connectionism lies in neuron networks and deep learning.
3. Behaviorism

   The fundamental theory of behaviorism believes that intelligence depends on perception and behavior. Behaviorism introduces a "perception-action" model for intelligent activities. Behaviorism believes that intelligence has nothing to do with knowledge, representation, or reasoning. AI can evolve gradually like human intelligence, and intelligent activities can only be manifested through human's ongoing interactions with the surrounding environment in the real world. Behaviorism emphasizes application and practices and constantly learning from the environment to modify the activities. The core behaviorism lies in behavior control, adaptation and evolutionary computing.

1.2 AI-Related Technologies

AI technology is multi-layered, running through technical levels such as applications, algorithms, chips, devices, and processes, as shown in Fig. 1.5.

AI technology has achieved the following developments at all technical levels.

1. Application Level

   Video and image: face recognition, target detection, image generation, image retouching, search image by image, video analysis, video review, and augmented reality (AR).

   Speech and voice: speech recognition, speech synthesis, voice wake-up, voiceprint recognition, and music generation.

   Text: text analysis, machine translation, human-machine dialogue, reading comprehension and recommender system.

   Control: autonomous driving, drones, robots, industrial automation.

2. Algorithm Level

   Machine learning algorithms: neural network, support vector machine (SVM), K-nearest neighbor algorithm (KNN), Bayesian algorithm, decision tree, hidden Markov model (HMM), ensemble learning, etc.

   Common optimization algorithms for machine learning: gradient descent, Newton's method, quasi-Newton method, conjugate gradient, spiking timing dependent plasticity (STDP), etc.

   Deep learning is one of the most essential technologies for machine learning. The deep neural network (DNN) is a hotspot of research in this field in recent years, consisting of multilayer perceptron (MLP) and convolutional neural network (CNN), recurrent neural network (RNN), spiking neural network (SNN) and other types. While the relatively popular CNNs include AlexNet, ResNet amd VGGNet, and the popular RNNs include long short-term memory (LSTM)

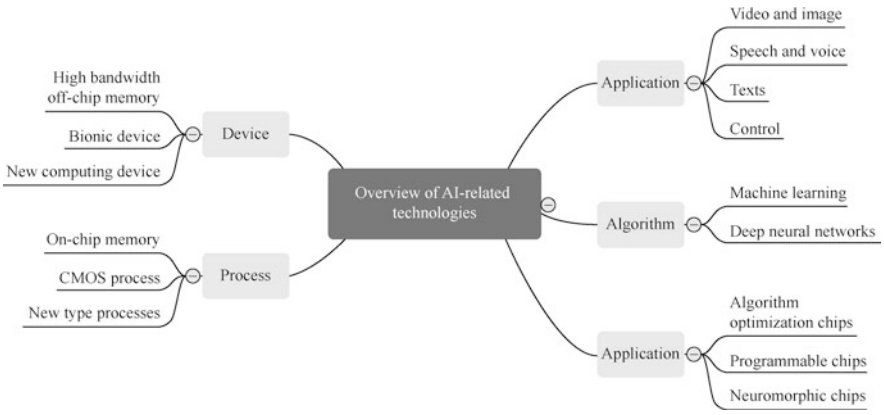

Fig. 1.5 Overview of AI-related technologies

networks and Neural Turing Machine (NTM). For instance, Google's BERT (Bidirectional Encoder Representation from Transformers) is a natural language processing pre-training technology developed on the basis of neural networks.

In addition to deep learning, transfer learning, reinforcement learning, one-shot learning and adversarial machine learning are also the important technologies to realize machine learning, and the solutions to some of the difficulties faced by deep learning.

3. Chip Level

Algorithm optimization chips: performance optimization, low power consumption optimization, high-speed optimization, and flexibility optimization-such as deep learning accelerator and face recognition chip.

Neuromorphic chips: bionic brain, biological brain-inspired intelligence, imitation of brain mechanism.

Programmable chips: taking flexibility, programmability, algorithm compatibility, and general software compatibility into consideration, such as digital signal processing (DSP) chips, graphics processing units (GPUs), field programmable gates array (FPGA).

Structure of system on chip: multi-core, many-core, single instruction, multiple data (SIMD), array structure of operation, memory architecture, on-chip network structure, multi-chip interconnection structure, memory interface, communication structure, multi-level cache.

Development toolchain: connection between deep learning frameworks (TensorFlow, Caffe, MindSpore), compiler, simulator, optimizer (quantization and clipping), atomic operation (network) library.

4. Device Level

High-bandwidth off-chip memory: high-bandwidth memory (HBM), dynamic random-access memory (DRAM), high-speed graphics double data rate memory (GDDR), low power double data rate (LPDDR SDRAN), spin-transfer torque magnetic random-access memory (STT-MRAM).

High-speed interconnection devices: serializer/deserializer (SerDes), optical interconnection communication.

Bionic devices (artificial synapses, artificial neurons): memristor.

New type computing devices: analog computing, in-memory computing (IMC).

5. Process Level

On-chip memory (synaptic array): distributed static random-access memory (SRAM), Resistive random-access memory (ReRAM), and phase change random-access memory (PCRAM).

CMOS process: technology node (16 nm, 7 nm).

CMOS multi-layer integration: 2.5D IC/SiP technology, 3D-Stack technology and Monolithic 3D.

New type processes: 3D NAND, Flash Tunneling FETs, FeFET, FinFET.

1.2.1 Deep Learning Framework

The introduction of deep learning frameworks has made deep learning easier to build. With the deep learning framework, we do not need to firstly code complex neural networks with backpropagation algorithms, but can just configure the model hyperparameters according to our demands, and the model parameters can be learned automatically from training. We can also add a custom layer for the existing model, or choose the classifier and optimization algorithm we need at the top.

We can consider a deep learning framework as a set of building blocks. Each block, or component of the set is a model or algorithm, and we can assemble the components into an architecture that meets the demands.

The current mainstream deep learning frameworks include: TensorFlow, Caffe, PyTorch and so on.

1.2.2 An Overview of AI Processor

In the four key elements of AI technology (data, algorithms, computing power, and scenarios), computing power is the one most reliant on AI processor. Also known as AI accelerator, an AI processor is a specialized functional module to tackle the large-scale computing tasks in AI applications.

1. Types of AI Processors

   AI processors can be classified into different categories from different perspectives, and here we will take the perspectives of technical architecture and functions.

   In terms of the technical architecture, AI processors can be roughly classified into four types.

   (a) CPU

   Central processing unit (CPU) is a large-scale integration circuit, which is the core of computing and control of a computer. The main function of CPU is to interpret program instructions and process data in software that it receives from the computer.

   (b) GPU

   Graphics processing unit (GPU), also known as display core (DC), visual processing unit (VPU) and display chip, is a specialized microprocessor dealing with image processing in personal computers, workstations, game consoles and some mobile devices (such as tablets and smartphones).

   (c) ASIC

   Application specific integrated circuit (ASIC) is designed for the integrated circuit product customized for a particular use.

   (d) FPGA

Field programmable gate array (FPGA) is designed to build reconfigurable semi-custom chips, which means the hardware structure can be adjusted and re-configured flexibly real-time as required.

In terms of the functions, AI processors can be classified into two types: training processors and inference processors.

(a) In order to train a complex deep neural network model, the AI training usually entails the input of a large amount of data and learning methods such as reinforcement learning. Training is a compute-intensive process. The large-scale training data and the complex deep neural network structure that the training involves put up a huge challenge to the speed, accuracy, and scalability of the processor. The popular training processors include NVIDIA GPU, Google's tensor processing unit (TPU), and Huawei's neural-network processing unit (NPU).
(b) Inference here means inferring various conclusions with new data obtained on the basis of the trained model. For instance, the video monitor can distinguish whether a captured face is the specific target by making use of the backend deep neural network model. Although inference entails much less computation than training, it still involves lots of matrix operations. GPU, FPGA and NPU are commonly used in inference processors.

2. Current Status of AI Processor

(a) CPU

The improvement of CPU performance in the early days mainly relied on the progress made by the underlying hardware technology in line with Moore's Law. In recent years, as Moore's Law seems gradually losing its effectiveness, the development of integrated circuits is slowing down, and the hardware technology has faced physical bottlenecks. The limitation of heat dissipation and power consumption restricted the CPU performance and serial program efficiency under the traditional architecture from making much progress.

The status quo of the industry prompted researchers to keep on looking for CPU architectures and the relevant software frameworks that can better adapted to the post-Moore Era. As a result, the multi-core processor came into being, which allows higher CPU performance with more cores. Multi-core processors can better meet the demands of software on hardware. For example, Intel Core i7 processors adopt instruction-level parallel processors with multiple independent kernels on the x86 instruction set, which improves the performance considerably, but also leads to higher power consumption and cost. Since the number of cores cannot be increased indefinitely, and most traditional programs are written in serial programming, this approach has limited improvements in CPU performance and program efficiency.

In addition, AI performance can be improved by adding instruction set. For example, adding instruction sets like AVX512 to the x86 complex

instruction set computer (CISC), architecture, adding the fused-multiply-add (FMA) instruction set to the arithmetic logic unit (ALU) module, and adding instruction set to the ARM reduced instruction set computer (RISC) architecture.

The CPU performance can also be improved by increasing the frequency, but there is a limit, and the high frequency will cause excessive power consumption and high temperature.

(b) GPU

GPU is very competitive in matrix computing and parallel computing and serves as the engine of heterogeneous computing. It was first introduced into the field of AI as an accelerator to facilitate deep learning and now has formed an established ecology.

With regard to the GPUs in the field of deep learning, NVIDIA made efforts mainly in the following three aspects:

- Enrich ecology: NVINIA launches the NVIDIA CUDA deep neural network horary (CUDNN), the GPU-accelerated library customized for deep learning, which optimizes the underlying architecture of GPU and ensures an easier application of GPU in deep learning.
- Improve customization: embracing multiple data types (no longer insisting on float32, and adopting int8, etc.).
- Add module specialized for deep learning (e.g., NVIDIA V100 Tensor Core GPU adopts the improved Volta architecture introducing and equipped with tensor cores).

The main challenges of current GPUs are high cost, low energy consumption ratio, and high input and output latency.

(c) TPU

Since 2016, Google has been committed to applying the concept of application-specific integrated circuits (ASIC) to the study of neural networks. In 2016, it launched the AI custom-developed processor TPU which supports the open-source deep learning framework TensorFlow. By combining large-scale systolic arrays and high-capacity on-chip memory, TPU manages to efficiently accelerate the convolutional operations that are most common in deep neural networks: systolic arrays can optimize matrix multiplication and convolutional operations, so as to increase computing power and reduce energy consumption.

(d) FPGA

FPGA uses a programmable hardware description language (HDL), which is flexible, reconfigurable, and can be deeply customized. It can load DNN model on the chips to perform low-latency operation by incorporating multiple FPGAs, contributing to a computing performance higher than GPU. But as it has to take account the constant erasing process, the performance of FPGA cannot reach the optimal. As FPGA is reconfigurable, its risk of supply and R&D is relatively low. The cost of hardware is decided by the amount of

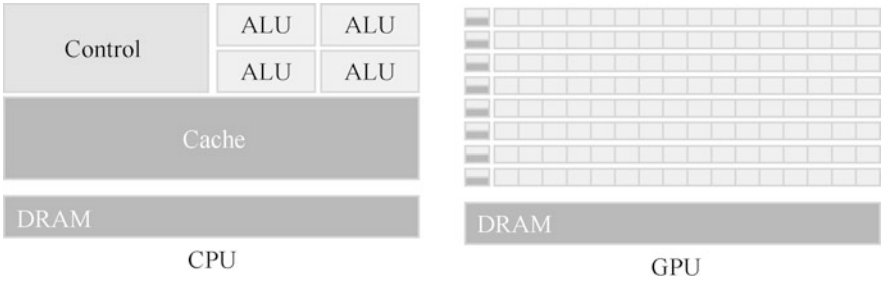

Fig. 1.6 CPU and GPU architecture

hardware purchased, so it is easy to control the cost. However, the design of FPGA and the tape-out process are decoupled, so the development cycle is long, which usually takes half a year, and has high standards.

3. Comparison Between the Design of GPU and CPU

The GPU is generally designed to tackle large-scale data that are highly unified in type and independent from each other, and deal with a pure computing environment without interruption. The CPU is designed more general-purpose, so as to process different types of data, and perform logical decisions at the same time, and it also needs to introduce a large number of branch-jump instructions and interrupt processing. The comparison between CPU and GPU architecture is shown in Fig. 1.6.

The GPU has numerous massively parallel computing architectures composed by thousands of much smaller cores (designed for simultaneous processing of multiple tasks). The CPU consists of several cores optimized for serial processing.

(a) The GPU works with many ALUs and little cache memory. Unlike the CPU, cache of the GPU serves for threads merely and plays the role of data forwarding. When multiple threads need to access the same data, the cache will coalesce these accesses, then access the DRAM, and forward the data to each thread after obtaining them, which will cause latency. However, as the large number of ALUs ensure the threads run in parallel, the latency is eased. In addition, the control units of GPUs can coalesce access.

(b) The CPU has powerful ALUs, which can complete computation in a very short clock cycle. The CPU has a large number of caches to reduce latency, and the complicated control units that can perform branch prediction and data forwarding: when a program has multiple branches, the control units will reduce latency through branch prediction; for the instructions that depend on the results of previous instructions, the control units must determine the positions of these instructions in the pipeline and forward the result of the previous instruction as quickly as they can.