Huawei AI Academy Training Materials

AI Overview

Huawei Technologies Co., Ltd.

Trademarks and Permissions


and other Huawei trademarks are trademarks of Huawei Technologies Co., Ltd.

All other trademarks and trade names mentioned in this document are the property of their respective holders.

Notice

The purchased products, services, and features are stipulated by the contract made between Huawei and the customer. All or part of the products, services, and features described in this document may not be within the purchase scope or the usage scope. Unless otherwise specified in the contract, all statements, information, and recommendations in this document are provided "AS IS" without warranties, guarantees, or representations of any kind, either express or implied.

The information in this document is subject to change without notice. Every effort has been made in the preparation of this document to ensure accuracy of the contents, but all statements, information, and recommendations in this document do not constitute a warranty of any kind, express, or implied.

Huawei Technologies Co., Ltd.

Address: Huawei Industrial Base Bantian, Longgang, Shenzhen 518129 China

Website: https://e.huawei.com

Contents

1 AI Overview

In the wave of Internet development, the emergence and rise of artificial smart (AI) is undoubtedly an extremely important part. With the continuous sinking of AI technologies, this technical concept is more and more connected with human life. Since the 1950s, with the development of related fields and the leap of software and hardware conditions, AI has been applied on a large scale in nearly a decade after several ups and downs. This chapter describes the concept, development history, and existing problems of AI.

1.1 AI Overview

1.1.1 AI in the Eyes of the Public

Person get to know AI through news, movies, and actual applications in daily life. What is AI in the eyes of the public?

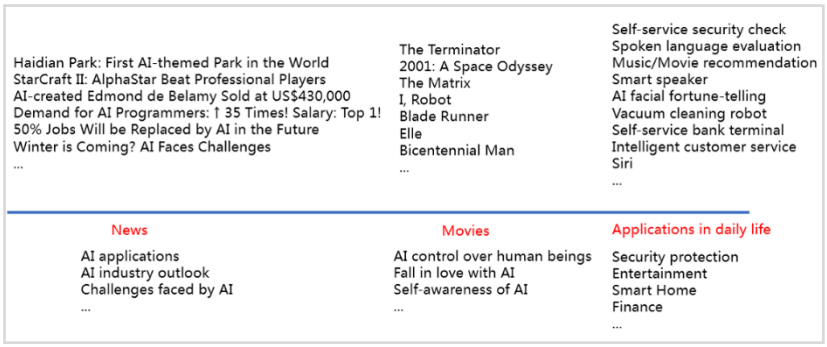

Figure 1-1 AI in the eyes of the public

As shown in Figure 1-1, the news reports AI with exaggerated titles. In movies, virtual AI was built with rich imagination. In person's daily life, AI makes it more convenient while brings privacy concerns.

"The branch of computer science concerned with making computers behave like humans." — A popular definition of AI, and an earlier one in this field proposed by John McCarthy at the Dartmouth Conference in 1956. However, it seems that this definition ignores the possibility of strong AI. According to another definition, AI is the smart (weak AI) demonstrated by artificial machines.

The following are the opinions of some scholars on AI:

"I propose to consider the question, 'Can machines think?'"

— Alan Turing in 1950

"The branch of computer science concerned with making computers behave like humans."

— John McCarthy in 1956

"The science of making machines do things that would require smart if done by men."

— Marvin Minsky in 1972

1.1.2 What Is AI?

Let's first understand what smart is before learning what AI is.

According to the theory of multiple smarts, human smart can be divided into seven categories: verbal/linguistic, logical/mathematical, visual/spatial, bodily/kinesthetic, musical/rhythmic, Inter-personal/social, and introspection Intrapersonal/Introspective.

1.1.2.1 Linguistic Smart

It refers to the ability to express thoughts and understand others by using oral speeches or in written words, and to master speech, semantics, and grammar flexibly, with the ability to think in words, express in words as well as appreciate the deep meaning of languages. Ideal professions for person with this smart include political activists, presenters, lawyers, orators, editors, writers, journalists, and teachers.

1.1.2.2 Logical-Mathematical Smart

It refers to the ability to calculate, measure, infer, conclude, classify, and to carry out complex mathematical operations. This smart includes sensitivity to logical ways and relationships, statements and propositions, functions, and other related abstract concepts. Ideal professions for person mastering logical mathematical smart include scientists, accountants, statisticians, engineers, and computer software developers.

1.1.2.3 Spatial Smart

It refers to the ability to accurately perceive the visual space and surroundings and to present the perception in the form of graphics. Person with this smart are sensitive to colors, lines, shapes, forms, and spatial relationships. Ideal professions for person mastering spatial smart include interior designers, architects, photographers, painters, and pilots.

1.1.2.4 Bodily-Kinesthetic Smart

It refers to the ability to express thoughts and emotions with the whole body and to make or operate objects with hands flexibly. This smart includes special physical skills such as balance, coordination, agility, strength, elasticity and speed, and abilities triggered by tactile sensation. Ideal professions for person mastering bodily-kinesthetic smart include athletes, actors, dancers, surgeons, gemstones, and mechanics.

1.1.2.5 Musical Smart

It refers to the ability to perceive pitches, tones, rhythms, and timbres. Person with this smart are sensitive to rhythms, tones, melodies or timbres, and endowed with the gift of music, with a strong capability to perform, create, and think about music. Ideal

professions for person with musical smart include singers, composers, conductors, music critics, musicians.

1.1.2.6 Interpersonal Smart

It refers to the ability to understand and interact with others. Person with this smart are good at perceiving other person's moods, emotions, and feelings, and able to discern and respond appropriately to the cues of different relationships. Ideal professions for person with interpersonal smart include politicians, diplomats, leaders, counselors, public relations and marketing personnel.

1.1.2.7 Intrapersonal Smart

It refers to self-awareness and the ability to act appropriately based on self-awareness. Person with this smart can recognize their strengths and weaknesses, their inner hobbies, emotions, intentions, tempers and self-esteem, and prefer thinking independently. Ideal professions for person with intrapersonal smart include philosophers, politicians, thinkers, psychologists.

AI is a new technical science that studies and develops theories, methods, techniques, and application systems for simulating and extending human smart. In 1956, the concept of AI was first proposed by John McCarthy, who defined the subject as "science and engineering of making intelligent machines, especially intelligent computer program". The purpose of AI is to make machines intelligent and give them human thoughts. As shown in Figure 1-2, the connotation of AI so far has greatly expanded and has become an interdisciplinary course.

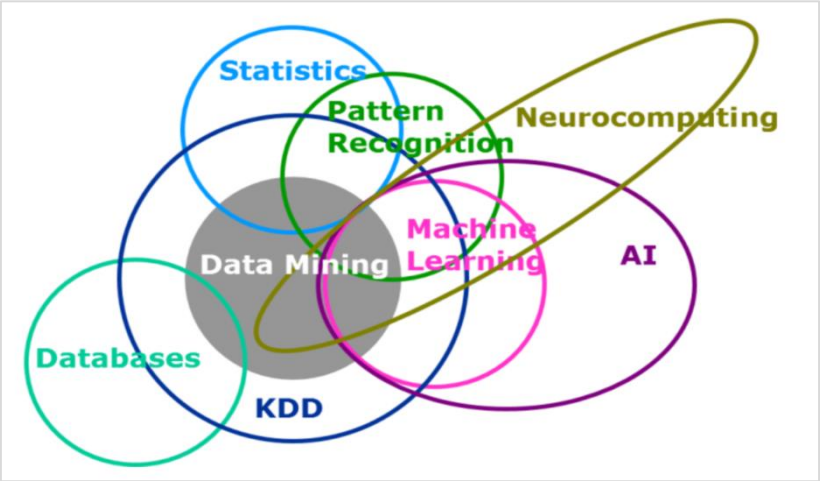

Figure 1-2 AI discipline category

Machine learning can be understood from multiple aspects. Tom Mitchell, a global machine learning scientist, provided a widely quoted definition: "A computer program is said to learn from experience *E* with respect to some class of tasks *T* and performance measure *P* if its performance at tasks in *T,* as measured by *P,* improves with experience *E*." These definitions are simple and abstract. However, as we deepen our understanding of machine learning, we will find that the connotation and extension of machine learning are changing over time. Because a variety of fields and applications are involved and

machine learning develops rapidly, it is not easy to define machine learning simply and clearly.

In general knowledge, processing systems and algorithms of machine learning are an identification mode that performs prediction by finding a hidden mode in data. Machine learning is an important subfield of AI, which also intersects with Data Mining (DM) and Knowledge Discovery in Database (KDD).

1.1.3 Relationship of AI, Machine Learning, and Deep Learning

Figure 1-3 shows the relationship among them.

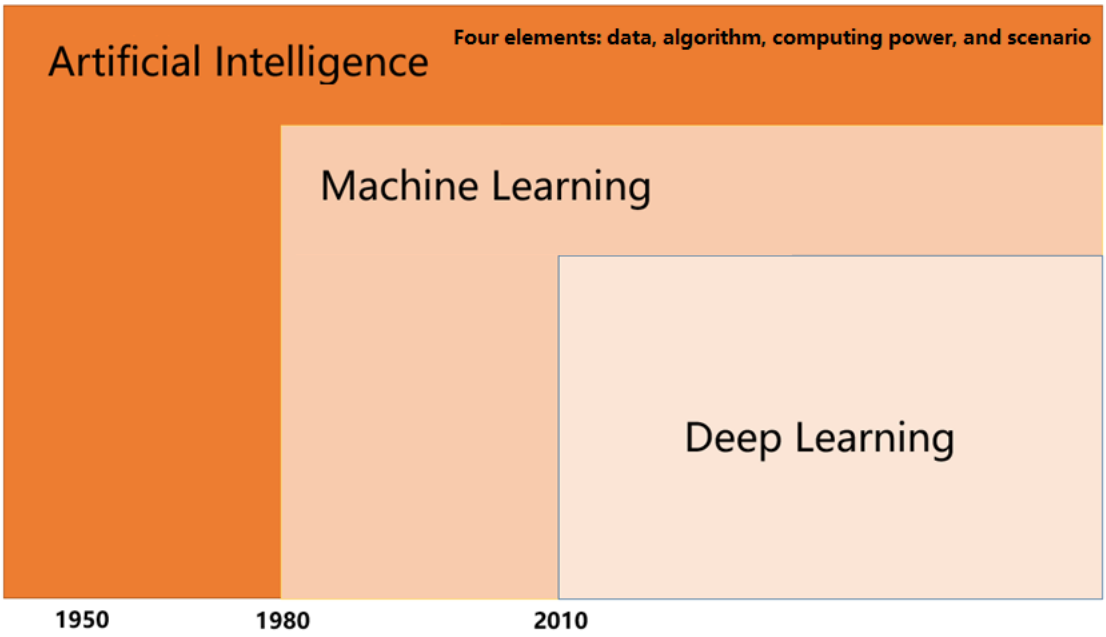

Figure 1-3 Relationship of AI, Machine Learning, and Deep Learning

Machine learning is specialized in studying how computers simulate or implement human learning behavior to acquire new knowledge or skills. The concept of Deep Learning originates from the research on Artificial Neural Networks (ANNs). Deep learning is a new field in machine learning that simulates the human brain to interpret data such as images, sounds, and texts.

Among the three, machine learning is a way or subset of AI, and deep learning is a special type of machine learning. AI can be compared to the brain. Machine learning is a process of mastering cognitive capabilities, and deep learning is an efficient teaching system in this process. AI is the goal and the result. Deep learning and machine learning are methods and tools.

1.1.4 Types of AI

AI can be classified into strong AI and weak AI.

The strong AI view holds that it is possible to create intelligent machines that can really reason and solve problems. Such machines are considered to be conscious and self-aware, can independently think about problems and work out optimal solutions to

problems, have their own system of values and world views as well as the instinct of living things, such as the needs for survival and safety. In a sense, the machine with human thoughts can be regarded as a new civilization.

The weak AI view holds that intelligent machines cannot really reason and solve problems. These machines only look intelligent, but do not have real smart or self-awareness.

Now we are in the weak AI phase. The emergence of weak AI alleviates the burden of human intellectual work, and its production principle is similar to that of advanced bionics. Both AlphaGo and robots that can write press releases and novels fall in the weak AI phase because they are better than humans only in some ways. The roles of data and computing power are self-evident in the era of weak AI, and promote the commercialization of AI. In the era of strong AI, these two factors are still critical. At the same time, the research on quantum computing by technology giants like Google and International Business Machines Corporation (IBM) also provides powerful support for humans to enter the era of strong AI.

1.1.5 AI History

1.1.5.1 Overview of AI Development

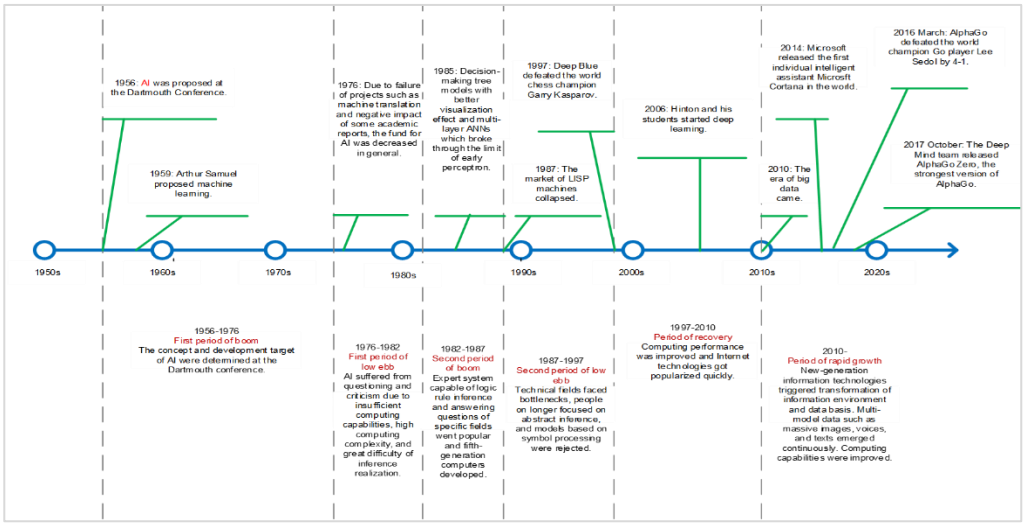

Figure 1-4 Brief development history of AI

Figure 1-4 shows the development history of AI.

The official origin of AI can back to the Turing Test proposed by Alan Mathison Turing, the father of AI, in 1950. As he envisioned, a computer is intelligent if it can talk to humans without being identified. In the same year, Turing boldly predicted the feasibility of a truly intelligent machine. However, no computer has completely passed the Turing Test so far.

Although the concept of AI has only a few decades of history, its theoretical basis and supporting technologies have been developed for a long time. The prosperity of the AI field is the result of common development of various disciplines and accumulation of generations of scientific circles.

1.1.5.2 Germination (Before 1956)

The earliest theoretical basis of AI can be back to the 4th century B.C. The famous ancient Greek philosopher and scientist Aristotle put forward the formal logic. His syllogism is still an indispensable foundation for deductive reasoning. In the 17th century, German mathematician Gottfried Wilhelm Leibniz put forward the idea of universal character and inference calculation, which laid the foundation for the generation and development of mathematical logic. In the 19th century, George Boole, a British mathematician, proposed Boolean algebra, which was the basic operation mode of computers and enabled the building of computers. Charles Babbage, the British inventor, designed a difference engine at the same time, the first computer to compute a quadratic polynomial. Although it had limited functions, it was the first time the computer really had reduced the computational pressure of the human brain. Machines began to have computational smart.

In 1945, John Mauchly and J. Presper Eckert of the Moore Group made Electronic Numerical Integrator and Computer (ENIAC), the world's first general-purpose digital computer. Although ENIAC was a milestone achievement, it still had many fatal drawbacks: large size, high power consumption, and manual input and adjustment of commands. In 1947, John von Neumann, the father of computer, designed and manufactured Mathematical Analyzer Numerical Integrator and Computer Model (MANIAC), a truly modern electronic computer device, by adapting and upgrading the device.

In 1946, American physiologist W. McCulloch built the first neural network model. His research on microcosmic AI laid an important foundation for the development of neural networks. In 1949, Donald O. Hebb put forward a neuropsychological learning paradigm, the Hebbian learning theory, which described the basic principle of synaptic plasticity. Synaptic plasticity is the continuous and repeated stimulation of presynaptic neurons to postsynaptic neurons that can lead to the increase of synaptic transmission efficiency. It has provided a theoretical basis for the establishment of the neural network model.

In 1948, Claude E. Shannon, the father of information theory, put forward the concept of "information entropy". By referring to the concept of thermodynamics, Claude E. Shannon defined the average amount of information excluding redundant information as "information entropy". This concept has had a far-reaching impact and played an extremely important role in areas such as non-deterministic inference and machine learning.

1.1.5.3 First Development (1956–1974)

At the Dartmouth Conference that lasted two months in 1956, AI was formally proposed by John McCarthy as a new discipline. This marked the birth of AI. After this conference, several AI research organizations were formed in the United States, such as the Carnegie-RAND collaboration group of Allen Newell and Herbert Alexander Simon, the Massachusetts Institute of Technology (MIT) research group of Marvin Lee Minsky and McCarthy, and Arthur Samuel's IBM Engineering Research Group.

In the next two decades, AI has developed rapidly in various fields. Researchers have been expanding the application areas of AI technologies with great enthusiasm.

1.1.5.3.1 Machine Learning

In 1956, Arthur Samuel of IBM wrote a famous checker program, which could learn an implicit model through the checkerboard state to guide the next move. After games with the program, Samuel believed that the program could reach a very high level after a certain period of learning. By using this program, Samuel rejected the model that computers could not learn patterns beyond explicit codes like humans. Since then, he has defined and explained a new word — machine learning.

1.1.5.3.2 Pattern Recognition

In 1957, Zhou Shaokang proposed to solve the pattern recognition problem by using the statistical decision theory, which promoted the rapid development of pattern recognition research from the late 1950s. In the same year, Frank Rosenblatt put forward a simplified mathematical model of simplified human brain stimulation for recognition, that is, perceptron. It initially implemented the training of the recognition system based on each sample a given category so that the system was able to correctly classify patterns of other unknown categories after learning.

1.1.5.3.3 Pattern Matching

In 1966, ELIZA, the first chat program, was developed by the Institute of Artificial Smart of the MIT. It can match patterns according to the set rules and users' questions, and select proper answers from the pre-written answer database. It was also the first software program that attempted to pass the Turing Test. ELIZA once simulated a psychotherapist talking to a patient and cheated many person when it was first used. "Dialogs are pattern matching." This is the beginning of computer natural language dialog technology.

In addition, during the first development of AI, McCarthy developed the list processing (LISP) programming language, which became the most important programming language in the AI field in the next several decades. Minsky had a more in-depth study of neural networks and found the shortcomings of simple neural networks. To overcome the limitations of neural networks, multilayer neural networks and back propagation (BP) algorithms have emerged. The expert system also started. The first industrial robot entered the production line of General Motors, and the first mobile robot capable of autonomous movement appeared.

The development of related fields also greatly promoted the progress of AI. The bionics established in the 1950s stimulated the enthusiasm of scholars for research so that simulated annealing algorithm came into being. It is a heuristic algorithm, the research foundation of search algorithms such as the ant colony optimization algorithm.

1.1.5.4 First Winter (1974–1980)

However, person's enthusiasm for AI did not last for a long time, and optimistic promises could not be fulfilled in a timely manner, causing doubts about AI technologies around the world.

The perceptron that caused a sensation in academia in 1957 was hit hard in 1969. At that time, Minsky and other scientists put forward the famous XOr problem and demonstrated the limitation of the perceptron under the linear inseparable data similar to the XOr problem. For academia, the XOr problem has almost become an insurmountable divide.

In 1973, AI was questioned by the scientific community. Many scientists thought that the seemingly ambitious goals of AI could not be achieved and that the research had completely failed. Increasing suspicions led to severe criticism and questioning of the real value of AI. Subsequently, governments and institutions have stopped or reduced their investment, and AI fell into its first winter in the 1970s.

The setback that AI encountered this time was not a coincidence. Limited by the computing capability at that time, many problems could be solved theoretically, but could not be put into actual use. At the same time, it was difficult to acquire knowledge for the algorithms of expert system at that time, leading to the failure of many projects. Researches on machine vision have started in the 1960s. The methods proposed by American scientist L. R. Roberts, such as edge detection and contour composition, are classic and have been widely used until now. However, theoretical foundations did not necessarily lead to actual output. At that time, scientists calculated that at least 1 billion instructions needed to be executed to simulate human retina vision for a computer. In 1976, the world's fastest computer Cray-1 cost millions of dollars, but the speed was less than 100 million times per second, and the computing speed of a common computer was less than 1 million times per second. Hardware conditions limited the development of AI. In addition, another major foundation for AI development is the huge database. At that time, computers and the Internet were not popularized, so large-scale data could not be obtained at all.

In this phase, the development of AI slowed down. Although the idea of BP was proposed by Linnainmaa in the 1970s as an "automatic differential reverse model", it was applied by Werbos to the multilayer perceptron until 1981. The emergence of multilayer perceptron and BP algorithm contributed to the second development of neural networks. In 1986, D.E.Rumelhart and others successfully implemented an effective BP algorithm for training a multilayer perceptron, which had a far-reaching impact.

1.1.5.5 Second Development (1980–1987)

In 1980, the XCON developed by Carnegie Mellon University was officially put into use. XCON was a comprehensive expert system that contained more than 2500 preset rules. In the following years, XCON has processed more than 80,000 orders with an accuracy of over 95%. This was a milestone in the new era. The expert system began to play a powerful role in specific fields and brought the entire AI technology into a prosperous phase.

The expert system tends to focus on a single area of expertise, simulating human experts to answer questions or provide knowledge to help staff make decisions. It limits itself to a small scope so that it avoids the difficulties of general AI and fully uses the knowledge and experience of existing experts to resolve tasks in specific work fields.

Because of the huge business success of XCON, 60% of the Fortune 500 companies began to develop and deploy their own expert systems in the 1980s. According to statistics, more than USD1 billion was invested in the AI field from 1980 to 1985, most of which was used in the AI department of enterprises, and many AI software and hardware companies emerged.

In 1986, the Bundeswehr University Munich installed computers and sensors in a Mercedes-Benz van that automatically controlled its steering wheel, accelerator and brake. It is called VaMoRs and is the first self-driving car.

In the AI field, the LISP language was mainly used at that time. To improve the transportation efficiency of various programs, many organizations began to develop specific computer chips and storage devices for running LISP programs. Although LISP machines have made some progress, personal computers (PCs) have been rising at the same time. IBM PCs and Apple computers occupied the entire computer market rapidly. Their central processing unit (CPU) frequency and speed were steadily increasing, even becoming more powerful than those expensive LISP machines.

1.1.5.6 Second Winter (1987–1993)

In 1987, the hardware market of specific LISP machines collapsed, and the AI field entered a cold winter again. The collapse of the hardware market and the fact that governments and institutions have stopped investment in AI researches have led to a trough in this field for several years, but some important achievements have also been made.

In 1988, the U.S. scientist Judea Pearl introduced the probability statistics into the inference process of AI, which greatly impacted the development of AI.

Nearly 20 years after the second winter, AI technologies have been deeply integrated with computer and software technologies. On the other hand, the progress of AI algorithm theory was slow. Many researchers could achieve groundbreaking results based on the theories of the past simply by relying on more powerful and faster computer hardware.

1.1.5.7 Stable Development (1993–2011)

In 1995, Richard S. Wallace developed Alice, a new chatbot program inspired by ELIZA. It could use the Internet to continuously add its own data sets and optimize content.

In 1996, IBM's Deep Blue computer played human world chess champion Kasparov, but did not win. Kasparov believed that the computer could never win the match against humans.

After that, IBM upgraded Deep Blue. The reconstructed Deep Blue has 480 specific CPUs, doubling the computing speed with 200 million times per second. It could predict the next eight moves or more and beat Kasparov.

However, this milestone match is actually a victory achieved by computers in the game with clear rules based on computing speed and enumeration. It is not AI in the real sense.

In 2006, Geoffrey Hinton published a paper in Science, opening the era of deep learning.

1.1.5.8 Prosperity (2011–present)

In 2011, Watson, also from IBM, participated in the variety show *Jeopardy!* and competed with humans. Watson beat two human champions with its outstanding natural language processing capability and powerful knowledge base. Computers at this stage were able to understand human languages, marking a major progress in the AI field.

In the 21st century, with the explosive growth of mobile Internet and cloud computing technologies and the wide use of PCs, institutions have accumulated unprecedented data volumes, providing sufficient materials and driving for the future development of AI. Deep learning became the mainstream of AI technologies. The famous Google Brain identity recognition project greatly improved the ImageNet recognition rate to 84%.

The Semantic Web was proposed in 2011, with its concept originated from the World Wide Web. Essentially, it was a massive distributed database with web data as the core and was linked by means of machine understanding and processing. The emergence of Semantic Web greatly promoted the development of knowledge representation technologies. In 2012, Google launched a search service based on knowledge graphs and proposed the concept of knowledge graphs for the first time.

In 2016 and 2017, Google launched matched between humans and the machine that caused a sensation to the world. Its AI program AlphaGo defeated two Go world champions: Lee Sedol from South Korea and Ke Jie from China.

Today, AI has penetrated into every aspect of human life. The voice assistant, represented by Apple's Siri, uses the Natural Language Processing (NLP) technology. Supported by NLP technology, computers can process human natural languages and match them with desired instructions and responses in an increasingly natural way. When browsing shopping websites, users often receive product recommendations generated by the recommendation algorithm. The recommendation algorithm can predict the products that users may purchase by analyzing historical shopping data and users' preference expressions.

1.1.6 Three Schools of Thought: Symbolism, Connectionism, and Behaviorism

1.1.6.1 Symbolism

The basic idea of symbolism is that the cognitive process of human beings is the process of inference and operation of various symbols. A human being is a physical symbol system, and so is a computer. Computers, therefore, can be used to simulate intelligent behavior of human beings. The core of AI lies in knowledge representation, knowledge inference, and knowledge application. Knowledge and concepts can be represented with symbols. Cognition is the process of symbol processing while inference refers to the process of solving problems by using heuristic knowledge and search. Symbolism lies in inference, symbolic inference and machine inference.

1.1.6.2 Connectionism

The basic idea of connectionism is that the basis of thinking is neurons instead of symbolic processing. Human brains vary from computers. A computer working mode based on connectionism is proposed to replace the one based on symbolic operation. Connectionism is derived from bionics, especially the study of the human brain model. In connectionism, a concept is represented by a set of numbers, vectors, matrices, or tensors. The concept is represented by the specific activation mode of the entire network. Each node, without specific meaning, plays its role in the representation of the concept. For example, in symbolism, the concept of a cat may be represented by a "cat node" or a set of nodes representing the cat's attributes, such as "two eyes", "four legs", and "fluffy". However, in connectionism, each node does not represent a specific concept, so it is impossible to find a "cat node" or an "eye neuron". Connectionism is based on neural networks and deep learning.

1.1.6.3 Behaviorism

The basic idea of behaviorism is that smart depends on perception and action, so the "perception-action" model of intelligent behavior is proposed. Smart requires no knowledge, representation, or inference. AI can evolve like human smart. Intelligent behavior can only be demonstrated in the real world through the constant interaction with the surrounding environment. Behaviorism concerns more about application practices and how to learn from the environment continuously to make corrections. Behaviorism is based on behavioral control, adaptation and evolutionary computing.

1.2 Overview of AI Technologies

1.2.1 Overview

As shown in Figure 1-5, AI technologies are multi-layered, covering the application, algorithm mechanism, toolchain, device, chip, process, and material layers.

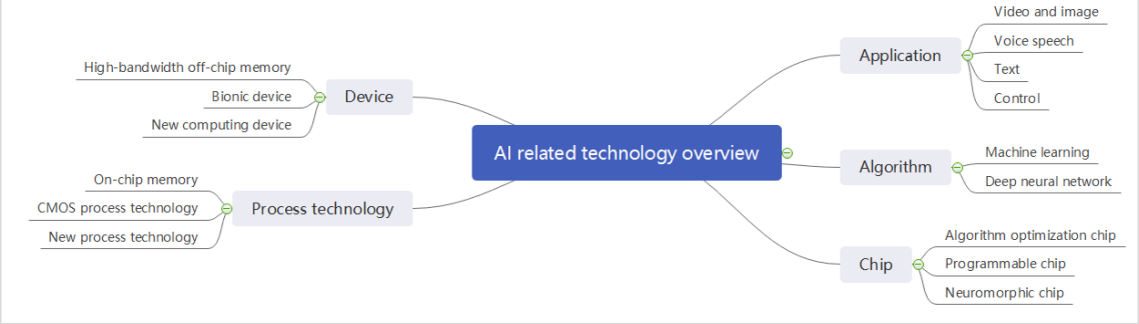

Figure 1-5 Overview of AI technologies

On one hand, the rapid development of applications and algorithms, especially deep learning and convolutional neural networks, raises performance optimization requirements for AI chips by two to three orders of magnitude, which has triggered the upsurge of AI chip R&D in recent years. On the other hand, the rapid development of new materials, processes, and components, such as 3D stacked memory and process evolution, also made significant improvements in performance and power consumption of AI chips possible. This driving came from breakthroughs in basic research. In general, the above driving have empowered rapid advancement of AI chip technologies in recent years. At each technology level, the followings are the achievements that AI technologies have made.

1.2.2 Application Layer

Video and image: facial recognition, object detection, image generation, video analysis, video content moderation, image beautification, reverse image search, AR

Voice: speech recognition, speech synthesis, voice wakeup, voiceprint recognition, music generation, smart speaker, smart navigation

Text: text analysis, language translation, man-machine dialog, reading comprehension, recommendation system

Control: autonomous driving, drone, robot, industrial automation

1.2.3 Algorithm Layer

Neural network interconnection structure: multilayer perceptron (MLP), convolutional neural network (CNN), recurrent neural network (RNN), long short-term memory (LSTM) network, and spiking neural network (SNN)

Deep neural network (DNN) structure: AlexNet, ResNet, and VGGNet

Neural network algorithms: transfer learning, reinforcement learning, one-shot learning, adversarial learning, neural Turing machine, and spike-timing-dependent plasticity (STDP)

Machine learning algorithms: support vector machine (SVM), k-nearest neighbor, Bayesian theorem, decision tree, hidden Markov model, AdaBoost, Bidirectional Encoder Representations from Transformers (BERT)

1.2.4 Chip Layer

Algorithm optimization chip: Efficiency optimization, low power consumption optimization, high-speed optimization, and flexibility optimization, such as deep learning accelerators and facial recognition chips

Neuromorphic chip: bionic brain, biological brain inspiration, brain mechanism simulation

Programmable chip: considering flexibility, programmability, algorithm compatibility, and compatibility with general software, such as digital signal processors (DSP), graphic processing unit (GPU), and field programmable gate array (FPGA)

Chip system-level structure: multi-core, many-core, Single Instruction Multiple Data (SIMD), operation array structure, memory structure, network-on-chip structure, multi-chip interconnection structure, memory interface, communication structure, and multi-level cache

Development tool chain: programming framework (TensorFlow, Caffe, and MindSpore), compiler, simulator, optimizer (quantization and tailoring), and atomic operation (network) library

1.2.5 Device Layer

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

High-speed interconnection: SerDes, optical interconnection communication

Bionic devices (artificial synapses, artificial neurons): memristors

New computing components: analog computing and in-memory computing

1.2.6 Process Technology Layer

On-chip memory (synaptic array): distributed static RAM (SRAM), resistive RAM (ReRAM), phase change RAM (PCRAM)

Complementary metal-oxide-semiconductor (CMOS) technology: process node (16, 7, 5 nm)

CMOS multilayer integration: 2.5D IC/SiP, 3D-stack technology, monolithic 3D

New technologies: 3D NAND, flash tunneling field effect transistors (FETs), ferroelectric FETs (FeFETs) and fin FETs (FinFETs).

1.2.7 Deep Learning Frameworks

The emergence of the deep learning framework lowers the threshold for getting started. You do not need to start coding from complex neural networks and BP algorithms. Instead, you can use hyperparameters of the configuration model as required. The parameters of the model are obtained through automatic training. Moreover, you can add self-defined network layers to the existing models, or select required classifiers and optimization algorithms.

A deep learning framework can be regarded as a set of building blocks. Each component in the building blocks is a model or algorithm. Therefore, developers can use components in the building blocks to assemble models that meet requirements, and do not need to start from scratch.

1.2.8 AI Processor Overview

This section describes AI processor overview, AI processor classification, AI processor status quo, comparison of mainstream AI processors, and Ascend AI Processors overview.

AI has four elements: data, algorithm, scenario, and computing power. The computing power depends on the AI processor. AI processors, also known as AI accelerators, are function modules used to process massive computing tasks in AI applications.

1.2.8.1 AI Processor Classification

AI processors can be classified by technical architectures and service applications.

AI processors can be divided into four types by technical architectures:

- CPU: It is a super-large-scale integrated circuit, the computing core and control unit of a computer. A CPU can interpret computer instructions and process computer software data.
- GPU: It is also known as display core, visual processor, and display chip. It is a microprocessor that processes images on PCs, workstations, game consoles, and some mobile devices such as tablets and smart phones.
- Application specific integrated circuit (ASIC): It is an integrated circuit designed for a specific purpose.
- FPGA: It is designed to implement functions of a semi-customized chip, that is, the hardware structure can be flexibly configured and changed in real time according to requirements.

From the perspective of service applications, there are two types: training and inference.

- In the training phase, a complex DNN model needs to be trained through a large number of data inputs or an unsupervised learning method such as enhanced learning. The training process requires massive training data and a complex DNN structure. The huge computing amount requires ultra-high performance including computing power, precision, and scalability of processors, Common GPUs include NVIDIA GPUs, Google tensor processing units (TPUs), and Huawei neural-network processing units (NPUs).
- In the inference phase, inferences are made by using trained models and new data. For example, a device uses the background DNN model to recognize a captured face. Although the calculation amount of the inference is much less than that of training, a large number of matrix operations are involved. In the inference process, GPU, FPGA and ASIC are also useful.

1.2.8.2 Status Quo of AI Processors

1.2.8.2.1 CPU

The performance of early computers was improved mainly by Moore's Law. Person impose increasingly high requirements on computer performance, while performance improvement mostly depends on advancement of underlying hardware, which accelerates upper-layer application software. In recent years, improvement brought by the Moore's Law has slowed down. Hardware development gradually encounters physical bottlenecks. Limits on heat dissipation and power consumption make it difficult to further improve the performance of serial programs in the traditional CPU architecture. The current situation drives the industry to constantly look for an architecture and the corresponding software framework more suitable to the post-Moore's Law era.

Multi-core processors are developed to improve computer performance by increasing the number of cores. Multi-core processors better meet the hardware requirements of software. For example, Intel® Core® i7 series processors use the parallel instruction processor cores constructed by four independent kernels based on the x86 instruction set. This improves the processor running speed to some extent, but also increases the power consumption and cost. The number of kernels cannot increase infinitely, and most traditional CPU programs are written by serial programming. Therefore, a large number of programs cannot be accelerated.

In addition, AI performance can be improved by adding instructions (modifying the architecture). For example, Intel (complex instruction set computer architecture) adds instructions such as AVX-512, and adds the vector computing module (FMA) to the arithmetic logic unit (ALU) computing module. Advanced reduced instruction set computing machine ARM (reduced instruction set computer architecture) is added to the Cortex-A instruction set and is planned to be upgraded continuously.

The performance can also be increased by frequency, but the improvement space is limited. In addition, a high dominant frequency may cause excessive power consumption and overheating of the processor.

1.2.8.2.2 GPU

CPUs focus on logic control in instruction execution, while GPUs have outstanding advantages in large-scale, intensive, and parallel data computing. Program optimization requires collaboration of CPUs and GPUs.

GPUs deliver remarkable performance in matrix computing and parallel computing and play a key role in heterogeneous computing. It was first introduced to the AI field as an acceleration chip for deep learning. At present, the GPU ecosystem has matured.

NVIDIA inherits the GPU architecture and focuses on three aspects in deep learning scenarios: 1. Enriched the ecosystem: It launched the accelerated NVIDIA CUDA® Deep Neural Network library (cuDNN) for deep learning to improve its usability and optimize the GPU underlying architecture. 2. Improved customization: Multiple data types, such as INT8, are supported in addition to FP32. 3. Dedicated deep learning modules (such as TensorCore V100, an improved architecture with tensor cores) are added.

The main problems of GPUs include high costs, low energy consumption, and high input and output latency.

1.2.8.2.3 TPU

Since 2006, Google has sought to apply the design concept of ASICs to the neural network field and released the TPU, a customized AI processor that supports TensorFlow, an open-source deep learning framework. The TPUs use large-scale systolic arrays and large-capacity on-chip storage to efficiently accelerate the most common convolutional operations in the deep neural network (DNN). The systolic arrays can be used to optimize matrix multiplication and convolutional operations to provide higher computing power and lower energy consumption.

1.2.8.2.4 FPGA

Using the hardware description language (HDL) programmable mode, FPGAs are highly flexible, reconfigurable and re-programmable, and customizable. Multiple FPGAs can be used to load the DNN model on the chips to realize low-latency computing. FPGAs outperform GPUs in terms of computing performance. However, the optimal performance cannot be achieved due to continuous erasing and programming. In addition, redundant transistors and cables, logic circuits with the same functions occupy a larger chip area. Thanks to the reconfigurable structure, the supply and Research and development (R&D) risks are low. The cost is relatively free depending on the purchase quantity. The design and tape-out processes are decoupled. The development period is long, generally half a year. The entry barrier is high.

1.2.8.3 Design Comparison of GPUs and CPUs

GPUs are designed for massive data of the same type independent from each other and pure computing environments that do not need to be interrupted. CPUs are required to process different data types in a universal manner, perform logic judgment, and introduce massive branch jumps and interrupt processing, as shown in Figure 1-6.

Figure 1-6 Structure comparison between CPUs and GPUs

Each GPU comprises several large-sized parallel computing architectures consisting of thousands of smaller cores designed to handle multiple tasks simultaneously. A CPU is composed of several cores optimized for sequential serial processing.

GPUs are designed based on large throughput. There are many ALUs and few caches, different from the objective of CPU, to improve the service for the thread. Caches are combined to access the DRAM, which causes the latency problem. The controller unit performs combined access. A large number of ALUs implement a large number of threads to mask the delay issue.

CPUs are designed based on low latency. A CPU has powerful ALU and can complete computing in a short clock cycle. A large number of caches can reduce the latency. The clock frequency is high. With complex logic controller units, the latency of multi-branch programs can be reduced through the branch prediction capability. For some instructions that depend on the previous instruction results, the logic units determine the position of the instructions in the pipeline to implement fast data forwarding.

GPUs are good at computing-intensive and easy-to-parallel programs. CPUs are good at logic control and serial computing.

CPUs focus on logic control in instruction execution, while GPUs have outstanding advantages in large-scale, intensive, and parallel data computing. Program optimization requires collaboration of CPUs and GPUs.

1.2.8.4 Huawei Ascend 910 AI Processor

Neural-network processing unit (NPU): It uses a deep learning instruction set to process a large number of human neurons and synapses simulated at the circuit layer. One instruction is used to process a group of neurons.

The NPU is a processor that is specially designed for neural network computing. Its performance is much higher than that of a CPU and GPU in processing neural network tasks. Typical NPUs include Huawei's Ascend AI Processors (Ascend), Cambricon, and IBM's TrueNorth.

Huawei provides two types of Ascend AI Processors: Ascend 310 and Ascend 910. Ascend 910 is mainly used in training scenarios and is mostly deployed in data centers. Ascend 310 is mainly used in inference scenarios, covering all device-edge-cloud deployment scenarios.

Ascend 910 is the world's most powerful AI processor with the fastest training speed. Its computing power is twice that of the world's top AI processor, equivalent to 50 latest strongest CPUs. Table 1-1 lists the parameters related to Ascend 310 and Ascend 910.

Table 1-1 Parameters related to Ascend 310 and Ascend 910

Ascend 310	Ascend 910
Ascend-Mini Architecture: Da Vinci Half precision (FP16): 8 TFLOPS	Ascend-Max Architecture: Da Vinci Half precision (FP16): 256 TFLOPS

Ascend 310	Ascend 910
Integer precision (INT8): 16 TOPS 16-channel full HD video decoder: H.264/265 1-channel full-HD video encoder: H.264/265 Maximum power consumption: 8 W 12 nm FFC	Integer precision (INT8): 512 TOPS 128-channel full HD video decoder: H.264/265 Maximum power consumption: 350 W 7 nm

1.2.9 AI Industry Ecosystem

In the past 50 years, we have experienced three AI upsurges, which were represented by man-machine games. The first one occurred in 1962 when the checkers program developed by Arthur Samuel from IBM beat the best checkers player in the United States. The second one occurred in 1997 when IBM Deep Blue beat Gary Kasparov, the world champion of chess, at 3.5:2.5. The third one broke out in 2016 when AlphaGo, a robot developed by Google DeepMind defeated the Go world champion Lee Sedol who is a player of 9 dan rank in South Korea.

In the future, AI will penetrate into various industries, including automobile, finance, consumer goods and retail, healthcare, education, manufacturing, communications, energy, tourism, culture and entertainment, transportation, logistics, real estate, and environmental protection.

For example, autonomous driving is a big stage for AI technologies to implement their capabilities. AI can assist in driving and decision-making. In this way, emergencies can be handled by person, simple operations can be automatically processed by the system, and some operations can be semi-automatically processed until the highest level of fully automated driving is achieved. It can greatly reduce fatigue driving and improve driving safety. Intelligent driving is a huge market. It can well feed back researches on intelligent technologies in this field and form a healthy cycle. It is the high-quality foundation for developing AI technologies.

A large amount of data is accumulated in the financial sector. AI can implement intelligent asset management, intelligent investment, and more reasonable financial decision-making. AI can also solve the problem of financial fraud, anti-fraud, anti-money laundering, and how to infer the reliability of transactions from various clues, determine the flow of funds and the periodicity of occurrence.

In the medical field, AI can also be widely used. For example, AI can be used to accurately interpret images at the geometric level and perform a large amount of data training to determine the problems reflected by image features, providing effective assistance for doctors. Training can be done on classification jobs such as the distinguishment between normal cells and cancer cells.

According to statistics from Chinese Association for Artificial Smart (CAAI) and other organizations, the market scale of AI is expected to exceed USD3 trillion by 2025, as shown in Figure 1-7.

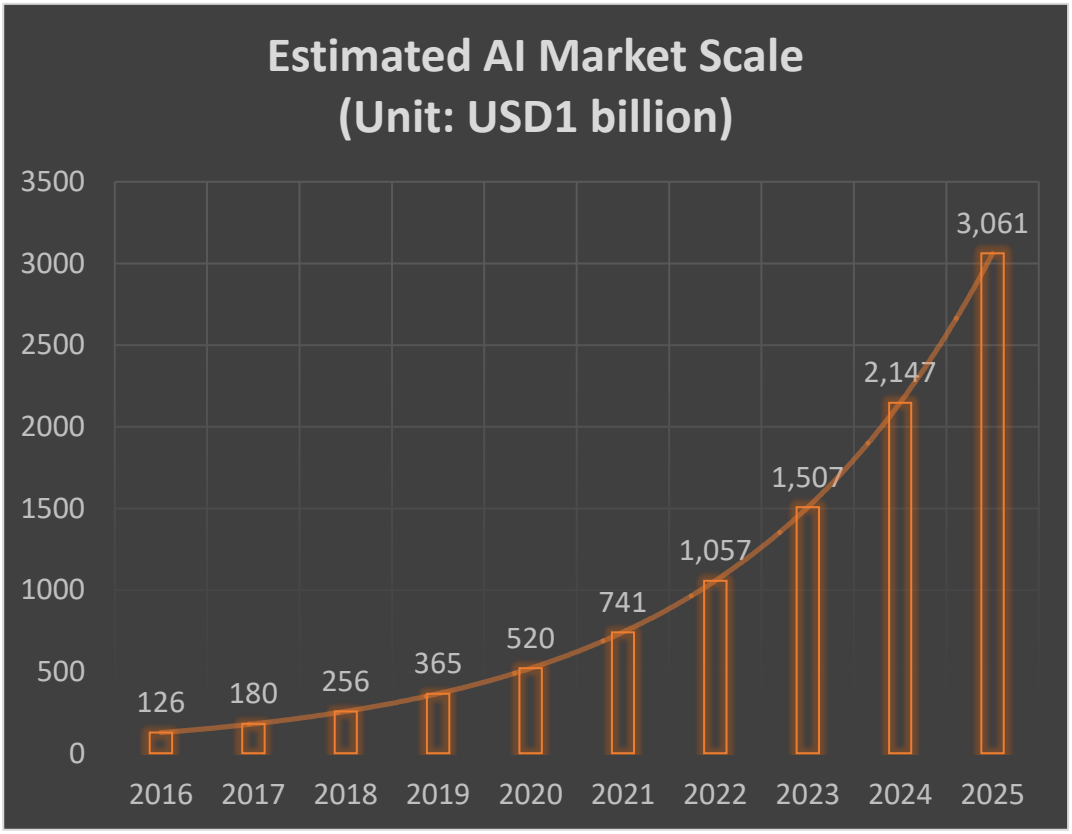

Figure 1-7 Estimated AI market scale

We can see that the AI applications have huge market potential. As mentioned in the previous section, AI has three cornerstones: data, algorithm, and computing power. However, it is not enough to implement AI only with these three elements. Application scenarios must be added to the three elements. Data, algorithms, and computing power describe the development of AI from the technical perspective. However, if there is no actual application scenario, the technological breakthrough is only a digital change. To meet the preceding application conditions, AI must be combined with cloud computing, big data, and Internet of Things (IoT) to form the platform architecture of AI applications, as shown in Figure 1-8.

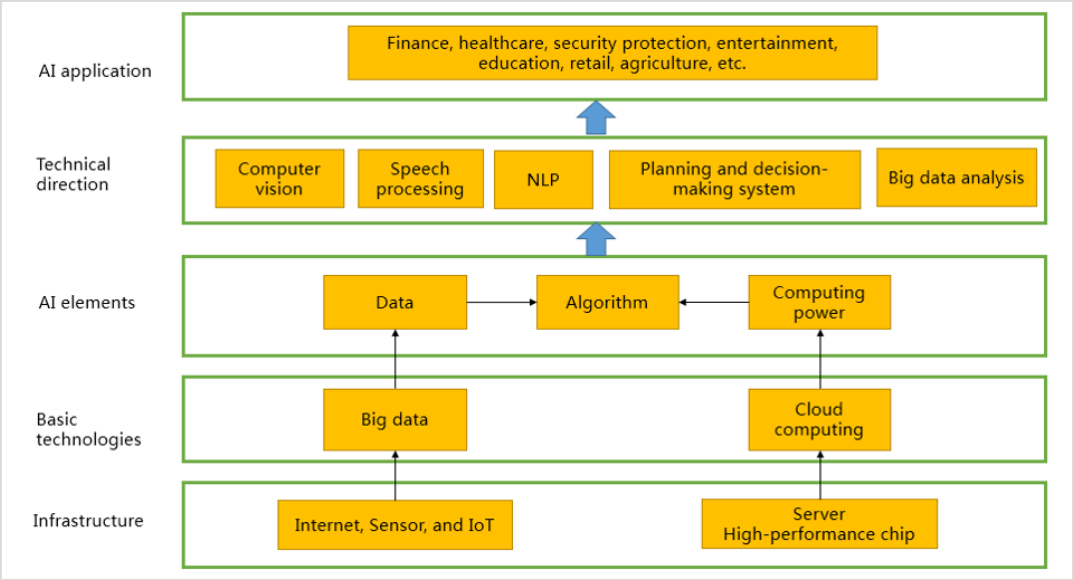

Figure 1-8 Architecture of the AI application platform

We need to combine AI with cloud computing, big data, and the IoT to build an intelligent society. The intelligent infrastructure provides computing capability support for the AI industry, including intelligent sensors, intelligent chips, and distributed computing frameworks. It is an important guarantee for the development of the AI industry. The intelligent technology service focuses on how to build an AI technology platform and provide AI-related services externally. These vendors are in a critical position in the AI industry chain. They provide key technology platforms, solutions, and services for various AI applications based on infrastructures and a large amount of data. With the acceleration of building a manufacturing power, a network power, and digital China, the demands for AI technologies and products in manufacturing, home appliance, finance, education, transportation, security protection, healthcare, and logistics will be further released. The types and forms of related intelligent products will become more and more diverse. Only the combination of infrastructure, basic elements, and specific technologies can effectively support upper-layer applications in the AI industry ecosystem.

Although AI can be widely applied, it is faced with huge challenges: AI capability development cannot meet excessive market demands. Major problems faced by AI capability development and application include:

- The prerequisites and skill requirements for AI: machine learning and deep learning knowledge, statistics knowledge, linear algebra and calculus knowledge.
- Low efficiency and long period of model training: Data collection and data cleaning, model training and optimization, and visualized experience improvement are required.
- Fragmented capabilities and experience: Data collection, data cleaning, model training and optimization, and experience improvement need to be performed again in each scenario. Capabilities cannot be directly inherited.
- Difficult to improve and enhance capabilities: It is difficult to upgrade models and obtain valid data.

At present, there is consensus in the industry that on-device AI with mobile phones as the core is the trend. More mobile phones will have built-in AI capabilities. Some consulting companies in the U.K. and the U.S. predict that 80% of mobile phones will have built-in AI capabilities by 2022 or 2023. Based on the market prospect and challenges, Huawei launched the AI capability open platform for smart devices, that is, the HiAI open platform. The purpose of HiAI is "Make it Easy for Developers: AI Connection Creates Infinite Possibilities". This platform enables developers to quickly utilize Huawei's powerful AI processing capabilities to provide better smart application experience for users.

1.2.10 HUAWEI CLOUD EI Application Platform

1.2.10.1 Overview of HUAWEI CLOUD EI

HUAWEI CLOUD Enterprise Smart (EI) is a driving for enterprises' intelligent transformation. Relying on AI and big data technologies, HUAWEI CLOUD EI provides an open, trustworthy, and intelligent platform through cloud services (in mode such as public cloud or dedicated cloud). It allows enterprise application systems to understand and analyze images, videos, languages, and texts to satisfy the requirements of different scenarios, so that more enterprises can use AI and big data services conveniently, accelerating business development and contributing to society progress.

1.2.10.2 Features of HUAWEI CLOUD EI

HUAWEI CLOUD EI has four outstanding features: industry smart, industry data, algorithm, and computing power.

- Industry smart: It has a deep understanding of the industry such as the pain points of the industry, and uses AI technologies to resolve industry pain points and drives AI implementation.
- Industry data: The industry never lacks data, so enterprises can use their own data to create a large amount of value through data processing and data mining.
- Algorithm: HUAWEI CLOUD provides enterprises with various algorithm libraries, model libraries, general AI services, and a one-stop development platform to solve problems.
- Computing power: With 30 years of experience in ICT technologies and a full-stack AI development platform, Huawei can provide enterprises with the strongest and most economical AI computing power.

1.2.10.3 Development History of HUAWEI CLOUD EI

Figure 1-9 shows the development history of HUAWEI CLOUD EI.

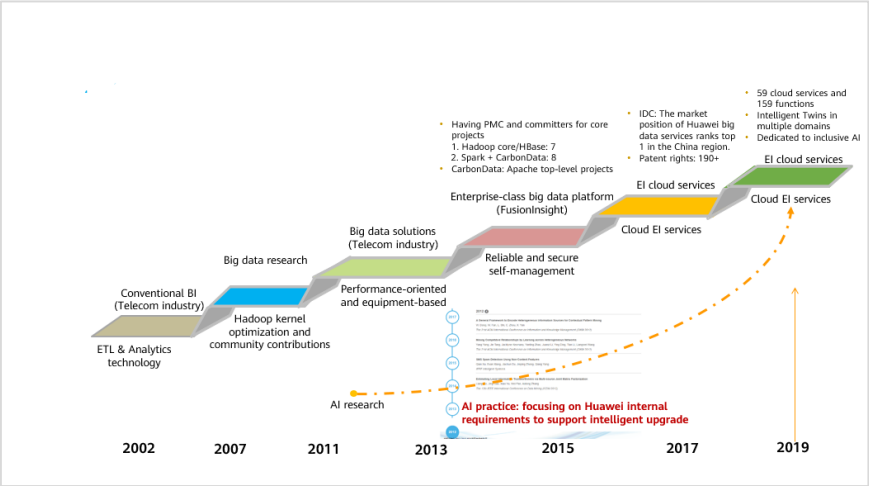

Figure 1-9 HUAWEI CLOUD EI development history

The following details these operations:

1. In 2002, Huawei started to develop data governance and analysis products for traditional Business smart (BI) services in the telecom field.
2. In 2007, Huawei started the Hadoop technology research, deployed big data technologies, and reserved a large number of talents and technical patents.
3. In 2011, Huawei applied big data technologies to telecom big data solutions for network diagnosis and analysis, network planning, and network optimization.
4. In 2013, large enterprises such as China Merchants Bank and Industrial and Commercial Bank of China started to communicate with Huawei about big data demands and started technical cooperation. In September of the same year, Huawei released FusionInsight, the enterprise-oriented big data analysis platform, at Huawei Cloud Congress (HCC), which has been widely used in various industries.
5. In 2012, Huawei officially put large-scale investment into the AI industry in and gradually started productization in 2014. In 2015, Huawei started to put AI into internal practice in finance, supply chain, engineering acceptance, e-commerce, and other products at the end of 2015, having achieved the following results:
   (1) Receipt operational cost rate (OCR) for customs declaration: import efficiency improved by 10 times.
   (2) Pickup route planning: exceptional expenses reduced by 30%.
   (3) Intelligent review: efficiency improved by six times.
   (4) Intelligent recommendations for e-commerce users: application conversion rate increased by 71%.
6. In 2017, Huawei officially started to provide cloud services and worked with more partners to provide more AI functions.

7. In 2019, HUAWEI CLOUD EI was dedicated to inclusive AI, making AI affordable, effective, and reliable. Based on the Huawei-developed Ascend chips, HUAWEI CLOUD EI provides 59 cloud services (21 platform services, 22 visual services, 12 language services, and 4 decision-making services), and 159 functions (52 platform functions, 99 application platform interface (API) functions, and 8 pre-integration solutions).
8. Huawei has invested thousands of R&D personnel in technical R&D (on productization technologies as well as cutting-edge technologies such as analysis algorithms, machine learning algorithms, and natural language processing), and actively contributed the R&D achievements to the communities.

1.3 Technical Fields and Application Fields of AI

1.3.1 AI Technology Direction

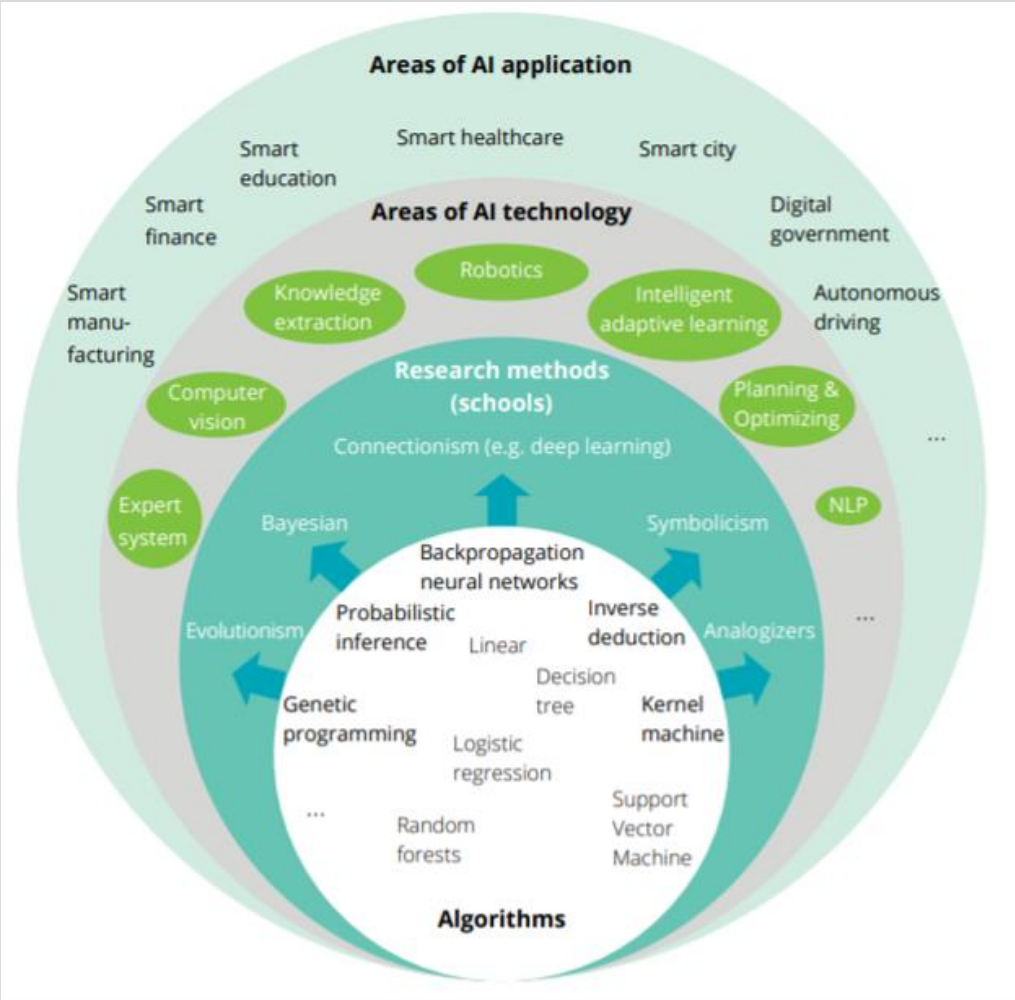

Figure 1-10 AI technology direction

Figure 1-10 shows the development trend of AI technologies. At present, application directions of AI technologies are classified into three types:

1.3.1.1 Computer Vision

Computer vision is to study how to make computers "see". Among the three AI technologies, computer vision is the most mature one, including image classification and segmentation, object detection, text recognition, and facial recognition. As shown in Figure 1-11 to Figure 1-14, the application of computer vision mainly focuses on electronic attendance, identity authentication, and image search. In the future, computer vision is expected to enter the advanced stage of autonomous understanding, analysis, decision-making, and enabling machines to "see". In scenarios such as autonomous driving and smart home, more value can be created.

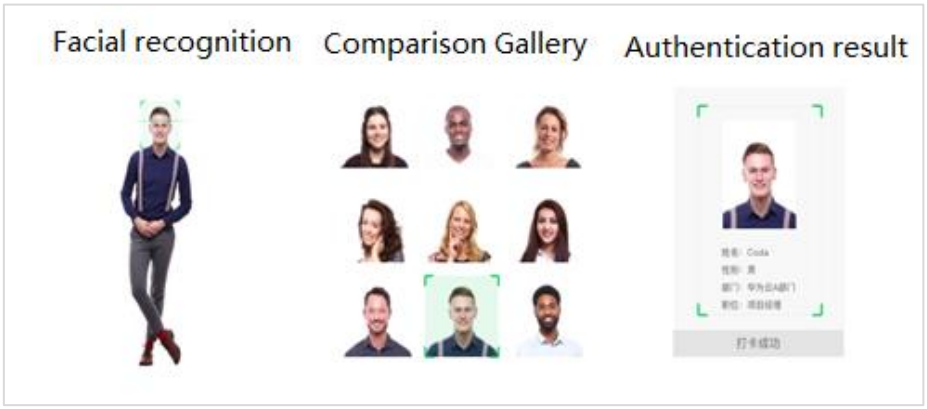

Figure 1-11 Electronic attendance

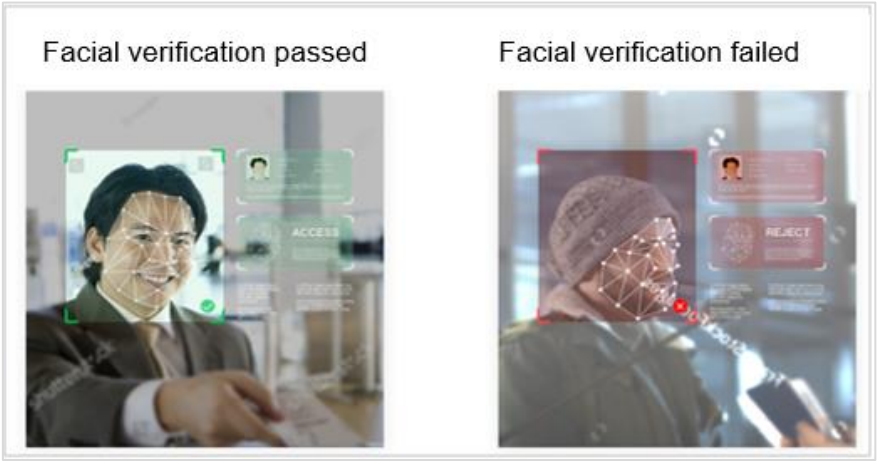

Figure 1-12 Enable identity authentication

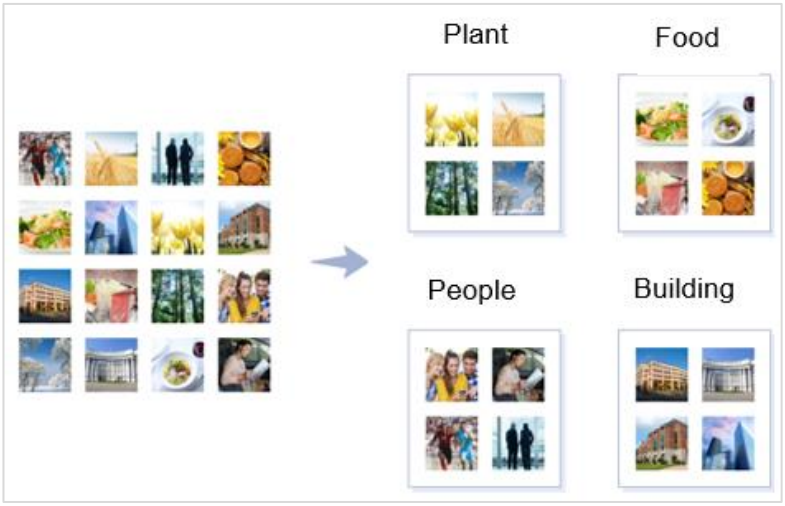

Figure 1-13 Image recognition

Figure 1-14 Image search

1.3.1.2 Speech Processing

Speech processing is a general term for various processing technologies, including the voice processing, statistical features of speech signals, speech recognition, machine-based voice synthesis, and speech perception. The main topics of voice processing research include voice recognition, voice synthesis, voice wakeup, voiceprint recognition, and audio-based incident detection. Among them, the most mature technology is speech recognition. The near field recognition in a quiet indoor environment can deliver accuracy up to 96%. As shown in Figure 1-15 and Figure 1-16, speech recognition technologies mainly focus on aspects such as speech Q&A and intelligent navigation at present.

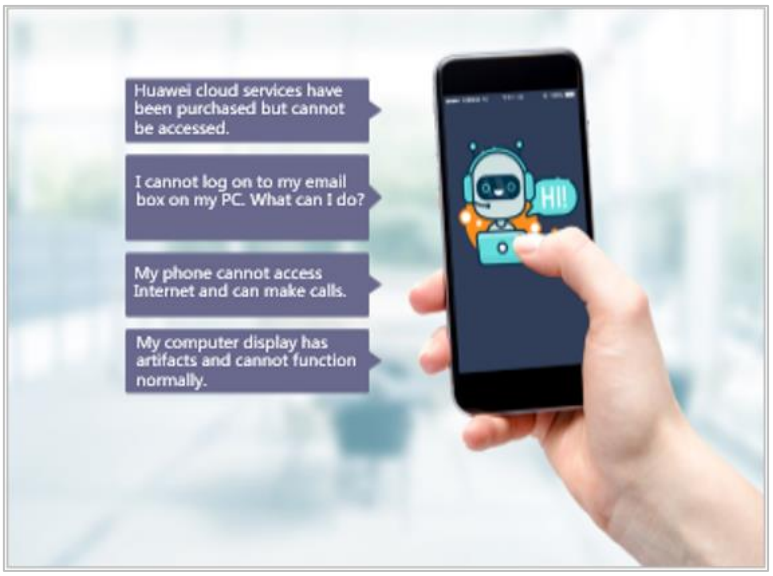

Figure 1-15 Question-Answering Bot (QABot)

Figure 1-16 Voice navigation

1.3.1.3 NLP

NLP is a discipline that uses computer technology to understand and use natural languages. It studies topics such as machine translation, text mining, and sentiment analysis. NLP imposes high requirements on technologies but confronts low technology maturity. Due to the highly complex semantics, it is difficult for the deep learning based on big data and parallel computing to think and understand things as humans. At present, NLP can only understand shallow semantics, but it will be able to automatically extract features and understand deep semantics in the future, that is, from single-purpose smart (machine learning) to hybrid smart (machine learning, deep learning, and reinforcement learning). As shown in Figure 1-17 to Figure 1-19, the NLP technology is

widely used in fields now, such as public opinion analysis, comment analysis, and machine translation.

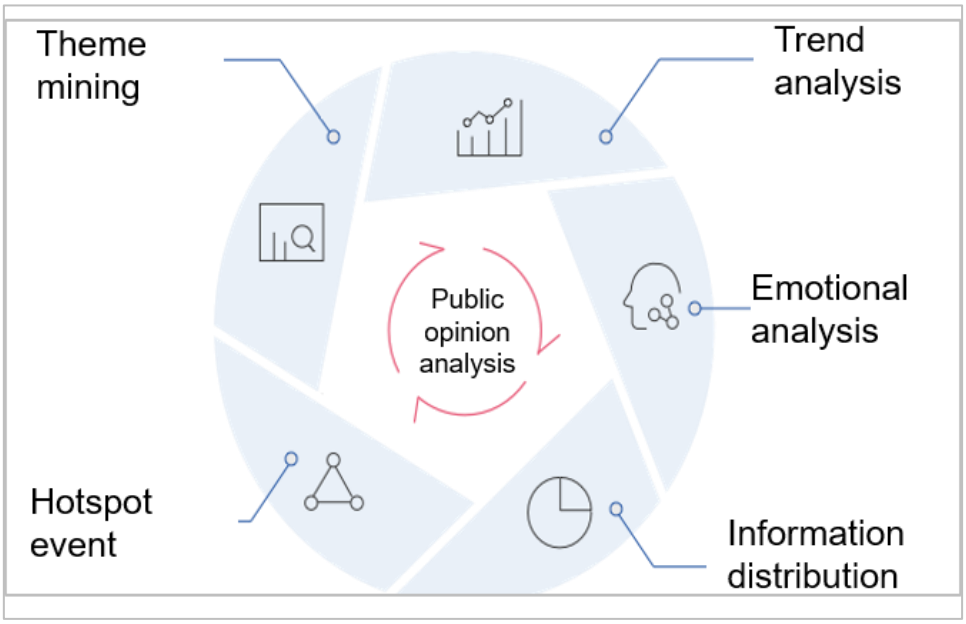

Figure 1-17 Public opinion analysis

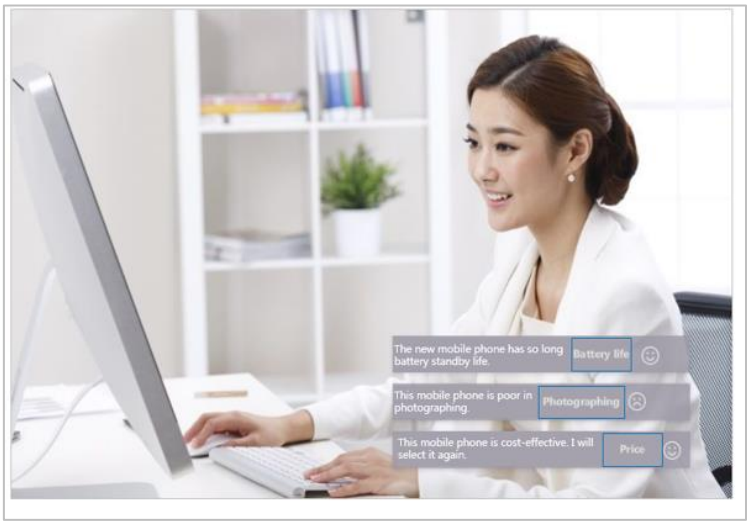

Figure 1-18 Comment analysis

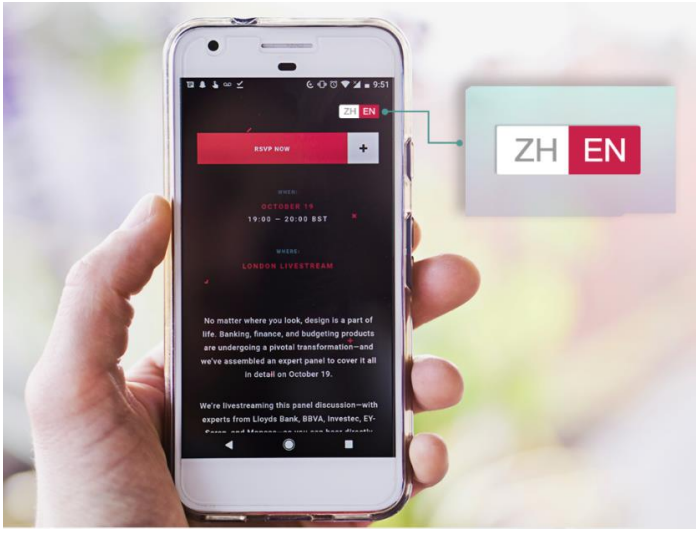

Figure 1-19 Machine translation

1.3.2 AI Application Field

1.3.2.1 Intelligent Healthcare

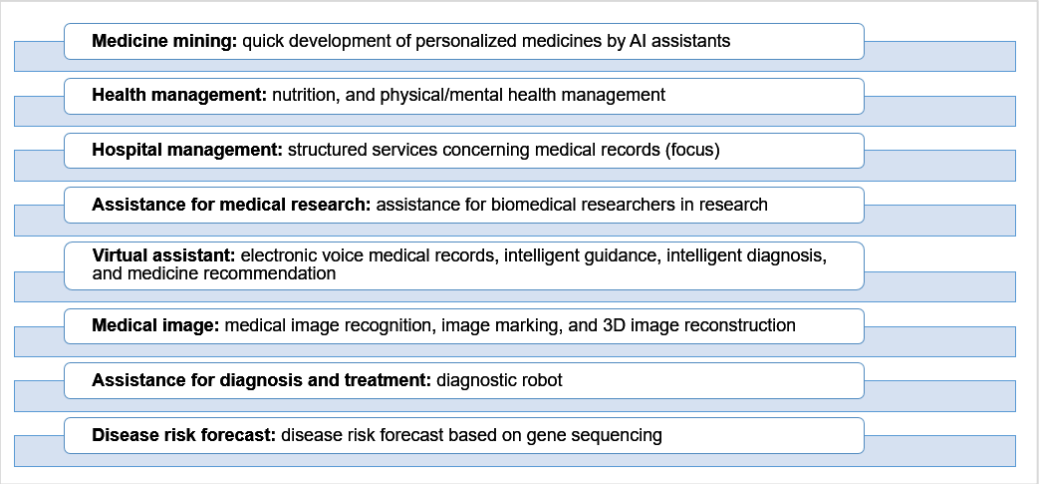

Figure 1-20 Smart healthcare

As shown in Figure 1-20, with AI technologies, we can enable AI to "learn" professional medical knowledge, "remember" numerous historical medical cases, and identify medical images with computer vision technologies to provide reliable and efficient assistance for doctors. For example, in the medical imaging technology that has been widely used today, researchers can build models based on historical data to identify existing medical images, quickly identify patients' lesions, and improve diagnosis efficiency.

1.3.2.2 Intelligent Security

security is considered a field ideal for AI implementation, and the AI application in this field is more mature than that in others. The field generates massive images and videos, laying a solid foundation for the training of AI algorithms and models. At present, AI technologies are mainly applied to the civil use and police use in the public safety field.

Civil use: card swipe based on facial recognition, warning against potential danger, and alert deployment at home

Police use: suspect identification, vehicle analysis, suspect search and comparison, and access control at key places

1.3.2.3 Smart Home

Based on IoT technologies, a smart home ecosystem consists of hardware, software, and cloud platforms, providing users with personalized life services that create a more convenient, comfortable, and secure home.

It uses voice processing to control smart home products, such as air conditioning temperature adjustment, curtain switch control, and voice control on the lighting system.

It leverages computer vision technologies to implement home security protection, such as facial or fingerprint recognition for unlocking, real-time intelligent camera, and intrusion detection.

Based on historical records of smart speakers and smart TVs, it adopts machine learning and deep learning technologies for user profiling and content recommendation.

1.3.2.4 Smart City

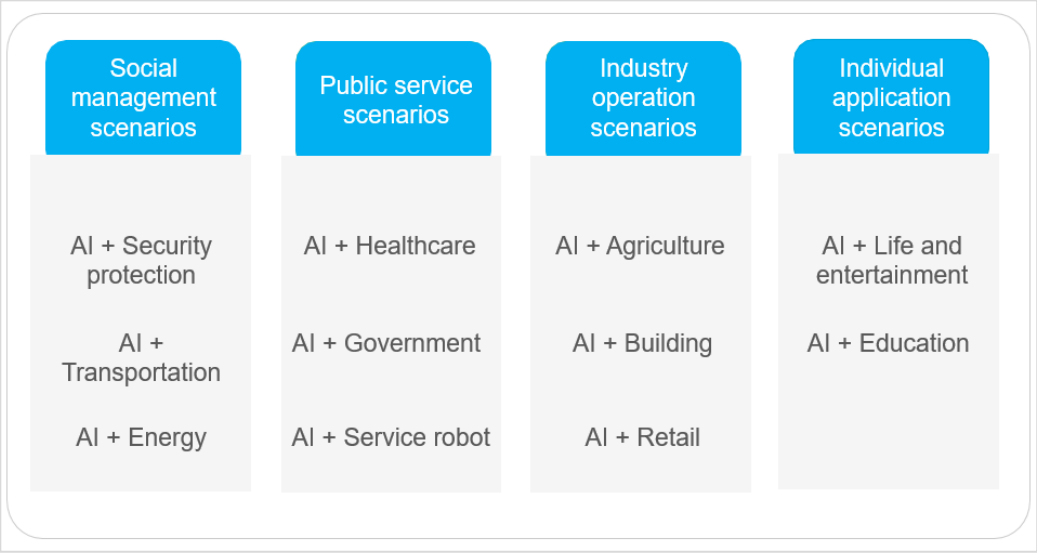

Figure 1-21 Smart city

As shown in Figure 1-21, smart city uses information and communication technology to sense, analyze, and integrate key information of the core operating system, to further make intelligent responses to various needs in livelihood, environmental protection, public safety, urban services, and industrial and commercial activities. Substantially, advanced information technologies are used to implement smart city management and

operation, create a better life for person in the cities, and promote the harmonious and sustainable city development. In the smart city scenario, AI is mainly applied to smart environment, smart economy, smart life, smart information, smart logistics, and smart government. For example, it transportation and logistics, and uses facial recognition for safety protection.

1.3.2.5 Retail

AI will completely transform the retail industry. A typical case is the fully unmanned supermarket. For example, Amazon Go, unmanned supermarket of Amazon, uses sensors, cameras, computer vision, and deep learning algorithms to completely cancel the checkout process, allowing customers to pick up goods and "just walk out".

One of the biggest challenges to unmanned supermarkets is how to charge customers correctly. So far, Amazon Go is the only successful business case and even this case involves many controlled factors. For example, only Prime members can enter Amazon Go. Other enterprises that intend to follow the example of Amazon have to build their membership system first.

1.3.2.6 Autonomous Driving

The Society of Automotive Engineers (SAE) in the U.S. defines 6 levels of driving automation ranging from 0 (fully manual) to 5 (fully autonomous). L0 indicates that the driving of a vehicle completely depends on the driver's operation. The system above L3 can implement the driver's hand-off operation in specific cases, and L5 depends on the system in all scenarios.

Now only some commercial passenger vehicle models, such as Audi A8, Tesla, and Cadillac, support L2 and L3 advanced driver-assistance systems (ADAS). It is estimated that by 2020, more L3 vehicle models will emerge with the further improvement of sensors and vehicle-mounted processors. L4 and L5 autonomous driving is expected to be first implemented on commercial vehicles in a closed campus. The popularization of advanced autonomous driving requires refined technologies, policies, and infrastructure. It is also estimated that L4 and L5 autonomous driving will not be supported on common roads until 2025 to 2030.

1.3.3 Phases of AI

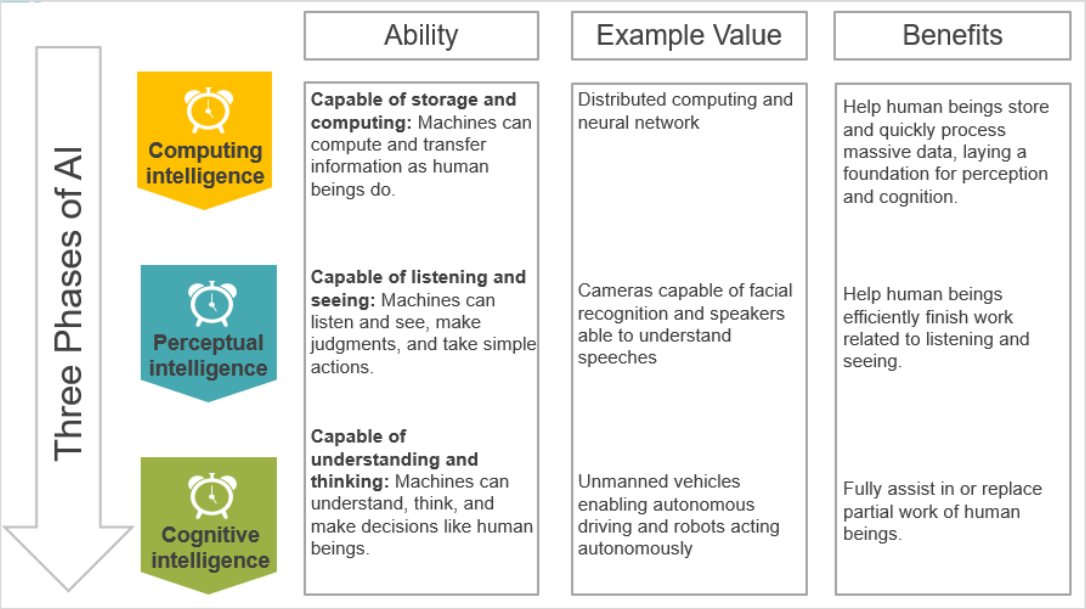

Figure 1-22 Three phases of AI

Figure 1-22 shows the three phases of AI. At present, AI is still in the initial phase of perceptive smart.

1.4 Huawei's AI Strategy

1.4.1 Huawei's Full-Stack, All-Scenario AI Portfolio

Huawei announced that it will open source the server OS on December 31, 2020, the standalone GaussDB OLTP database in June 2020, and the MindSpore all-scenario AI computing framework in the first quarter of 2020.

"Full-Stack" refers to its technical function. Huawei's full-stack portfolio includes chips, chip enablement, a training and inference framework, and application enablement.

"All-Scenario" refers to different deployment scenarios for AI, including public clouds, private clouds, edge computing in all forms, industrial IoT devices, and consumer devices.

As the cornerstone of Huawei full-stack AI solution, Atlas provides modules, cards, and servers based on the Ascend AI Processor to meet customers' computing requirements in all scenarios.

1.4.2 Huawei AI Full-Stack Direction

1.4.2.1 HUAWEI CLOUD One-Stop AI Development Platform — ModelArts

ModelArts is a one-stop development platform for AI developers. With data pre-processing, semi-automated data labeling, distributed training, automated model building, and model deployment on the device, edge, and cloud, ModelArts helps AI developers build models quickly and manage the lifecycle of AI development. It has the following features:

1. Automatic learning: It can automate model design, parameter adjustment, and model training, compression, and deployment with the labeled data. The process is code-free and requires no model development experience.

   ModelArts Pro is a professional development suite for enterprise-class AI applications. Based on the advanced algorithms and fast training capabilities of HUAWEI CLOUD, it provides workflows and models are provided to improve the development efficiency of enterprise AI applications and reduce the development difficulty. Customers can manage workflows to quickly develop, share, and release applications, build an open ecosystem, and implement AI in inclusive industries. ModelArts Pro suites include the NLP suite, text recognition suite, and vision suite, which can quickly respond to AI implementation requirements in different industries and scenarios.
2. Device-Edge-Cloud: It indicates devices, Huawei intelligent edge devices, and HUAWEI CLOUD, respectively.
3. Online inference: It is a web service that synchronously provides the inference result for each inference request.
4. Batch inference: It is a job that processes batch data for inference.
5. Ascend chips: a series of Huawei-designed AI chips with high computing performance but low power consumption.
6. The built-in AI data framework combines automatic pre-labeling and hard example labeling to improve the data preparation efficiency by over 100 folds.
7. The Huawei-developed high-performance distributed framework MoXing uses core technologies such as hybrid parallel cascade, gradient compression, and convolution acceleration, greatly reducing the model training time.
8. Models can be deployed on devices, edges, and clouds in different scenarios with one click to meet the requirements of high concurrency and lightweight deployment.
9. ModelArts allows visualized management of the AI development lifecycle, including data preparation, training, modeling, and inference. It also supports resumed training, training result comparison, and model.
10. The AI market supports data and model sharing, helping enterprises improve AI development efficiency and allowing developers to convert knowledge to value.

1.4.2.2 MindSpore

In the intelligent era, AI applications in device-edge-cloud scenarios are booming. However, AI still faces huge challenges. Technical barriers, high development cost, and long deployment period hinder the development of the AI developer ecosystem in the entire industry. The all-scenario AI computing framework MindSpore is developed based on the principles of friendly development, efficient operation, and flexible deployment.

In terms of deep learning framework, Huawei MindSpore is the strongest challenger to TensorFlow (Google), MXNet (Amazon), PyTorch (Facebook), and CNTK (Microsoft), which are listed as the four major players.

MindSpore has been open-sourced on March 30, 2020. It is a product that competes with frameworks such as TensorFlow (Google), PyTorch (Facebook), PaddlePaddle (Baidu), and Caffe.

MindSpore provides automatic parallel capabilities. With MindSpore, senior algorithm engineers and data scientists who focus on data modeling and problem solving can run algorithms on dozens or even thousands of AI computing nodes with only a few lines of description.

The MindSpore framework supports both large-scale and small-scale deployment, adapting to independent deployment in all scenarios. In addition to the Ascend AI Processors, MindSpore also supports other processors such as GPUs and CPUs.

1.4.2.3 CANN

CANN is a chip enabling layer developed by Huawei for DNNs and Ascend AI Processors. It consists of four functional modules:

- FusionEngine: FusionEngine is an operator-level fusion engine. It fuses operators, reduces the memory transfer between operators, and improves the performance by 50%.
- CCE operator library: The optimized general operator library provided by Huawei can meet the requirements of most mainstream vision and NLP neural networks. (It is estimated that APIs of the CCE operator library will be released in the first quarter of 2020.) Requirements for timeliness, privacy and research of the customers and partners will lead to the requirements for custom operator. In this case, the third functional module is used.
- Tensor Boost Engine (TBE) is an efficient and high-performance custom operator development tool. It abstracts hardware resources as APIs, enabling customers to quickly construct required operators. (This function module is expected to be available in the fourth quarter of 2020.)
- The last module is the bottom-layer compiler that optimizes performance and supports Ascend IA Processors in all scenarios.

1.4.2.4 Ascend AI Processor

Demands for AI are soaring worldwide. However, with the market being dominated by only a few vendors, AI processors are sold at a very high price. The delivery cycle is long and the local service support is weak. Therefore, the AI requirements of many industries cannot be effectively met.

At HUAWEI CONNECT held in October 2018, Huawei unveiled its Ascend 310 processor for AI inference and Ascend 910 processor for AI training. Built upon the unique Da Vinci 3D Cube architecture, Huawei's Ascend AI Processors boast high computing power, energy efficiency, and scalability.

Ascend 310, an AI SoC with ultimate performance per watt, is designed for edge inference. It provides up to 16 TOPS of computing power, with a power consumption of only 8 watts. This makes it a perfect choice for edge computing.

The Ascend 910 AI processor delivers the industry's highest computing density on a single AI chip. It applies to AI training and delivers 512 TOPS of computing power, with a maximum power consumption of 310 watts.

1.4.2.5 Atlas AI Computing Platform

Figure 1-23 Atlas AI computing platform portfolio

As shown in Figure 1-23, powered by the Ascend AI Processors, the Huawei Atlas AI computing platform supports rich form factors, including modules, cards, edge stations, servers, and clusters. Atlas enables AI solutions for all scenarios across the device, edge, and cloud. As an important part of Huawei's full-stack AI solution, Atlas launches the training platform this year following the inference platform unveiled last year, providing the industry with a complete AI solution. Huawei will also enhance all-scenario deployment, and drive full collaboration across the device, edge, and cloud, enabling every phase of the AI industry chain.

1.5 AI Disputes

1.5.1 Algorithmic Bias

Algorithmic biases are mainly caused by data biases.

When we use AI algorithms for decision-making, the algorithms may learn to discriminate an individual based on existing data, such as making discriminatory decisions based on race, gender or other factors. Even if factors such as race or gender are

excluded from the data, the algorithms can make discriminatory decisions based on information of names and addresses.

For example, if we search with a name that sounds like an African American, an advertisement for a tool used to search criminal records may be displayed. The advertisement, however, is not likely to be displayed in other cases. Online advertisers tend to display advertisements of lower-priced goods to female users. Google's image software once mistakenly labeled an image of black person as "gorilla".

1.5.2 Privacy Issues

The existing AI algorithms are all data-driven. In this case, we need a large amount of data to train models. We enjoy the convenience brought by AI every day while technology companies, such as Facebook, Google, Amazon, and Alibaba, are obtaining an enormous amount of user data, which will reveal various aspects of our lives including politics, and gender.

In principle, technology companies can record each click, each page scrolling, time of viewing any content, and browsing history when users access the Internet.

Technology companies can know our privacy including where we, where we go, what we have done, education background, consumption capabilities, and personal preferences based on our ride-hailing and consumption records.

1.5.3 Contradiction Between Technology and Ethics

With the development of computer vision technologies, reliability of images and videos is decreasing. Fake images can be produced with technologies such as Photoshop (PS) and generative adversarial network (GAN), making it hard to identify whether images are true or not.

Taking GAN as an example, Ian Goodfellow, a machine learning researcher, proposed this concept in 2014. The reason why the model is called "generative" is that the output of the model is images rather than prediction values related to the input data. The "adversarial network" is from the model where two sets of neural networks competing with each other, just like cashiers and counterfeiters in the battle of wits. One side tries to deceive the other side into believing that it is the authentic money, while the other side tries to identify the counterfeit money.

1.5.4 AI Development = Rising Unemployment?

Looking back, human beings have always been seeking ways to improve efficiency, that is, obtain more with less resources. We used sharp stones to hunt and collect food more efficiently. We used steam engines to reduce the need for horses. Every step in achieving automation will change our life and work. In the AI era, AI will replace jobs that involve little creativity and social interaction, such as couriers, taxi drivers, and soldiers. On the other hand, writers, managers, software engineers, and other highly creative jobs are not easily replaced.

1.6 AI Development Trend

1.6.1 Development Trend of AI Technologies

- Easier-to-use development framework

  Various AI development frameworks are evolving towards ease-of-use and all-function, continuously lowering the threshold for AI development.

- Algorithms model with better performance

  In the computer vision field, GAN has been able to generate high-quality images that cannot be identified by human eyes. GAN-related algorithms have been applied to other vision-related jobs, such as semantic segmentation, facial recognition, video synthesis, and unsupervised clustering. In the NLP field, the pre-training model based on the Transformer architecture has made a significant breakthrough. Related models such as BERT, general-purpose technology (GPT), and XLNet are widely used in industrial scenarios. In the reinforcement learning field, AlphaStar of the DeepMind team defeated the top human player in StarCraft II.

- Smaller deep learning models

  A model with better performance usually has a larger quantity of parameters, and a large model has lower running efficiency in industrial applications. More and more model compression technologies are put forward to further compress the model volume, reduce the model parameters, accelerate the inference speed, and meet the requirements of industrial applications while ensuring the model performance.

- Computing power with comprehensive device-edge-cloud development

  The scale of AI chips applied to the cloud, edge devices, and mobile devices keeps increasing, further meeting the computing power demand of AI.

- More comprehensive AI basic data services

  The AI basic data service industry is maturing, and related data labeling platforms and tools are being released.

- More secure data sharing

As shown in Figure 1-24, federated learning uses different data sources to train models, further breaking data bottlenecks while ensuring data privacy and security.

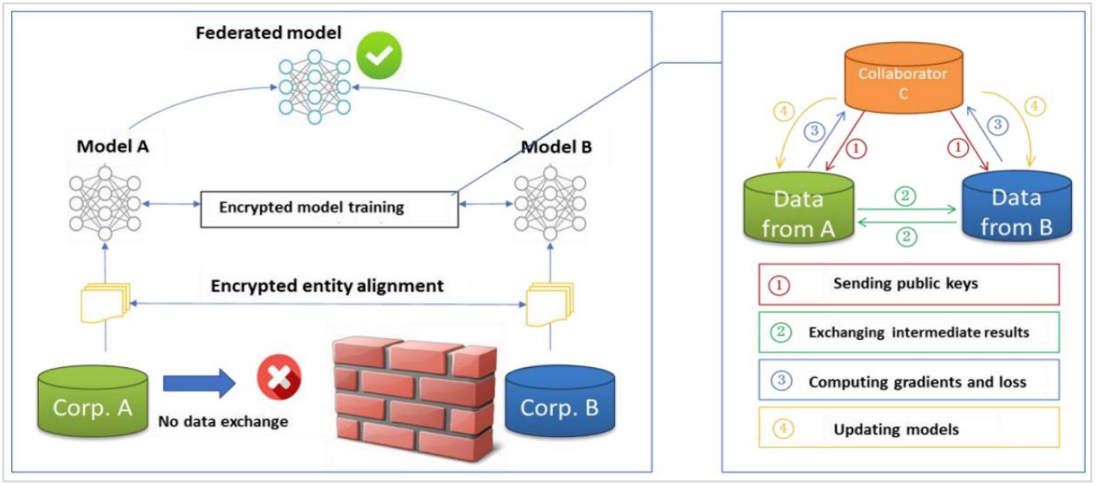

Figure 1-24 Federated learning

1.6.2 GIV 2025 — 10 Trends for 2025

- Popularization of Intelligent robots

  Intelligent robots are machines and even family members. Huawei predicts that by 2025, 14% of the world's households will have smart robots. Smart household robots will play an important role in person's lives.

- Popularization of Augmented reality (AR)/virtual reality (VR)

  The number of enterprises that use VR/AR technology will increase to 10%. The application of VR and other technologies will bring vigor and vitality to industries such as commercial display and audio-visual entertainment.

- Wide application of AI

  Man-machine collaboration: 97% of large enterprises will use AI technologies. It is mainly used in various fields, including voice smart, image recognition, facial recognition, and man-machine interaction.

- Popularization of big data applications

  Frictionless communication: The data utilization of enterprises will increase to 86%. Big Data analytics and processing will save time and improve work efficiency for enterprises.

- Weakening of search engines

  Zero search: 90% of the world's population will have personal smart device assistants. This means that chances of getting through a search portal will be greatly reduced.

- Popularization of Internet of Vehicles (IoV)

  Understand my road: cellular Vehicle-to-Everything (C-V2X) will be embedded in 15% of the global vehicles. Smart automobiles and Internet automobiles will be popularized, making driving more secure and reliable.

- Popularization of industrial robots

  Machines are engaged in "three-high" work: 103 robots will work with every 10,000 manufacturing employees. High-risk, high-precision, and high-intensity work will be assisted or completed independently by industrial robots.

- Popularization of cloud technologies and applications

  In the symbiotic economy, the usage of cloud-based applications will reach 85%. Massive applications and program collaboration will be completed at the cloud.

- Popularization of 5G

  With the acceleration of 5G, 58% of the world's population will enjoy 5G services. In the future, communications will bring a disruptive leap forward, and communication technologies and rates will be greatly improved.

- Popularization of digital economy and big data

Global digital governance: 180 ZB of data will be stored globally every year. Digital economy and blockchain technologies will be widely used in the Internet.

1.7 Summary

This chapter describes the basic concepts, development history, and application background of AI. After reading this chapter, you can understand that, as a cross-field discipline, the application and development of AI cannot be separated from the support of other disciplines. Its physical implementation depends on large-scale hardware, and its upper-layer applications depend on software design and implementation methods. As a learner, you are required to have a clear understanding of the scope and boundary of AI applications, and make improvements based on that.

1.8 Quiz

1. There are different interpretations of the concept of AI in different contexts. Please explain what AI is based on your understanding.
2. AI, machine learning, and deep learning are often mentioned at the same time. What is the relationship between them? What are the commonalities and differences of them?
3. After reading the description of AI application scenarios, please describe an AI application field and its application scenario in reality based on your life experience.
4. CANN is a chip enabling layer developed by Huawei for DNNs and Ascend AI Processors. Please describe the four modules of CANN.
5. Please describe the development directions of AI based on your knowledge and understanding.