Binxing Fang *Editor*

Artificial Intelligence Security and Safety

中国工信出版集团 電子工業出版社 PUBLISHING HOUSE OF ELECTRONICS INDUSTRY http://www.phei.com.cn Springer

Artificial Intelligence Security and Safety

Binxing Fang
Editor

Artificial Intelligence Security and Safety

電子工業出版社
PUBLISHING HOUSE OF ELECTRONICS INDUSTRY
http://www.phei.com.cn

Springer

Editor
Binxing Fang
Cyberspace Institute of Advanced
Technology
Guangzhou University
Guangzhou, China

ISBN 978-981-96-9262-0 ISBN 978-981-96-9263-7 (eBook)
https://doi.org/10.1007/978-981-96-9263-7

Jointly published with Publishing House of Electronics Industry, China
The print edition is not for sale in China Mainland. Customers from China Mainland please order the print book from: Publishing House of Electronics Industry.
ISBN of the Co-Publisher's edition: 9787121389917

© Publishing House of Electronics Industry 2025

This work is subject to copyright. All rights are solely and exclusively licensed by the Publisher, whether the whole or part of the material is concerned, specifically the rights of reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.
The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.
The publishers, the authors, and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publishers nor the authors or the editors give a warranty, express or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publishers remain neutral with regard to jurisdictional claims in published maps and institutional affiliations.

This Springer imprint is published by the registered company Springer Nature Singapore Pte Ltd.
The registered company address is: 152 Beach Road, #21-01/04 Gateway East, Singapore 189721, Singapore

If disposing of this product, please recycle the paper.

Editorial Board

Editor-in-Chief: Binxing Fang

Associate Editors: Zhaoquan Gu, Xiang Cui

Co-editors: Zhihong Tian, Peng Liao, Jie Yin, Mohan Li, Hui Lu, Weihong Han, Shudong Li, Shen Su, Lin Jiang

Preface

Whenever it comes to cyberspace security, pessimistic voices like "the picture is far from rosy" or "a lot of work is needed for improvement" rise and confound the decision-makers: why cyberspace security has hardly made headway? What can be done to get rid of cyberspace security concerns? Threats to cyberspace security, in fact, are inevitable because the cyberspace remains in flux. On the Symposium on Cybersecurity and IT Application held on April 19, 2016, President Xi Jinping pointed out that we should have a correct outlook for cyberspace security, and cybersecurity is a dynamic, rather than static, process. In other words, the situation of cyberspace security keeps changing—once a security trouble is shot, another may rear its head. Among all the contributing factors to cyberspace security problems, those associated with new technologies are the most prominent in that new technologies are inevitably accompanied with new security concerns, and the constant emergence of new technologies or systems will, as a consequence, incur new security and safety menaces. If we stop inventing new technologies and preventing information systems from upgrading in the coming five years, all cyberspace and information security problems will be effaced and we will enjoy a 100% secure cyberspace. This, however, will be just hypothetical. The world will keep moving forward and new technologies; hence, the subsequent security problems will never cease to emerge.

New technologies are bound to bring about new security concerns, which are called the security accompanied security effect of new technologies. Such accompanied security problems are manifested in two aspects: first, inherent vulnerabilities of a new technology will lead the new technological system to unstable or unsecure operation, which is referred to as the endogenous problem of new technologies; second, the vulnerabilities of a new technology will not impair the new technological system's normal operation, but will impose security threats to other fields, which is referred to as the derivative security problem of new technologies.

Aside from the security-accompanied effect, new technologies, when applied to improving security, have the security empowerment effect. The security empowerment effect of a new technology is also manifested in two aspects: it can empower

defense, but once maliciously exploited, may also empower attacks and put the security of other fields at stake. Apparently, empowerment of attacks by new technologies is one of the reasons why ensuring cyberspace security remains a challenge.

Since its inception in the 1950s, AI has gone ups and downs, evolving upwards in a spiral. In the 21st century, because of its spectacular achievements in application, AI has been recognized as a new driving force of social advancement in the field of information communication. In fact, approaches like deep learning and software architecture that promote AI development and its application in practice are new inventions in the 21st century. Therefore, it is fair to say that AI is an emerging informatics technology. Another reason why AI is deemed a new technology is the wide attention and tremendous investment it draws from governments around the world, which showcase the high expectation the global community has for AI to benefit the human society.

As a new technology, however, AI inevitably has both the security-accompanied effect and security empowerment effect. The security-accompanied effect is mainly manifested as the endogenous effect of AI, or the AI security problem. For instance, AI algorithms rely heavily on statistical results. In particular, intelligent learning systems like neural networks adjust the link weights in the hidden layer to record information of features obtained by the system's statistical learning. These weights, however, are beyond explainability, and thus they can only be used but not predicted. In this case, it is unpredictable what weights the system will adjust, in what way the weights will be adjusted, what consequences the adjustment will have, or what learning processes will result in the adjustment, therefore leaving open a window for the popular "adversarial example attacks". In this sense, adversarial example attacks can be classified as an accompanied endogenous security problem of AI. Likewise, unexplainability of the AI system may not undermine the system's normal operation, but will it go out of control and fail to work the way it is designed for? Will it pose threats to human beings when it goes amok? A mere thought of these questions is already hair-raising. An assistive robot for the disabled, if not working in the preset ways, will be reduced into a safety menace, and then we may find that a less intelligent robot will have a smaller chance to go out of control. Therefore, the fact that an artificial intelligence actant may go out of control is a derivative problem of AI, or an AI safety problem.

In terms of the security empowerment effect of AI, the examples of AI-powered defense and AI-powered attacks are abundant. A typical example of AI-powered defense is the application of AI to cyberspace security situational awareness, while Deepfake, an AI technology that changes faces to circumvent novelty detection of facial recognition systems, is a textbook example of AI-powered attacks.

As AI gains increasing attention from the governments around the world, the global community is placing more hopes on this new technology to improve people's life and even to upgrade national governance models. As a result, most studies on AI focus on how to advance AI technologies and how to commercialize AI products. The considerable investment to AI has, however, furthered the imbalance between AI development and the AI security problems that usually fail to catch our attention. While reveling in the social progress boosted by AI technology, we are likely to

overlook the AI accompanied security problems as well as AI-powered attacks. As a professional team long engaged in research on cyberspace security, we think its high time that we should pay attention to AI security and safety problems. In fact, the security and safety concerns that accompany AI development refer to not only endogenous or derivative security problems, but also a series of problems concerning codes of ethics, laws and social governance. Therefore, with the support from Cyberspace Institute of Advanced Technology of Guangzhou University, this book is edited, with a vision to increase the public awareness of AI security and safety problems together with the development of AI.

This book comprises 12 chapters. Chapter 1 that introduces the history of AI development is authored by Binxing Fang and Weihong Han; Chap. 2 that expounds on AI development strategies of governments around the world is authored by Shudong Li and Binxing Fang; Chap. 3—Introduction to AI Security and Safety, is authored by Xiang Cui; Chap. 4 that focuses on AI for security is authored by Peng Liao; Chap. 5 that probes into AI security is authored by Zhaoquan Gu; Chap. 6 that discusses AI safety is authored by Shen Su and Zhihong Tian; Chap. 7 that introduces artificial intelligence actants is authored by Hui Lu; the authors for Chap. 8—AIA Safety Hoop—are Zhihong Tian and Xiang Cui; Chap. 9 focusing on security evaluation and detection of AIAs is authored by Binxing Fang and Lin Jiang; Chap. 10 that discusses codes of ethics of AI is authored by Binxing Fang and Jie Yin; Chap. 11 that explores advanced AI technologies is authored by Mohan Li; Chap. 12 that discusses the future of AI security is authored by Binxing Fang and Zhaoquan Gu. The book is compiled by Binxing Fang, Zhaoquan Gu and Xiang Cui. We want to extend our appreciation to Academician Fang Wen, who has offered much guidance during the compilation of the book. We are also grateful to Lisong Zhao and Yanbo Shen, two associate editors with Publishing House of Electronics Industry, for their efforts in publishing the Chinese version of the book.

AI, in its long history of development, has been studied by different methodologies and from different technical perspectives. Though most of these studies yield similar results, scholars uphold different views. It is impossible for this book to include studies of different academic schools and cover all AI-related security and safety problems. Thus, this book only intends to serve as a modest start to inspire more discussions on AI-related security problems and technologies. We expect the readers' understanding if there is room for improvement.

Guangzhou, China
May 2020

Binxing Fang

Introduction

This book unfolds the history of artificial intelligence (AI) and AI development strategies of countries across the globe, before proceeding to its focus of discussion—AI security and safety. On this focus, an AI security framework is proposed, relevant topics including AI for security, AI security (endogenous security of AI) and AI safety (derivative security of AI) are discussed at length. Also, the codes of ethics of AI are explored. In this book, artificial intelligence actants (AIAs) and their potential security risks are analyzed in detail, the AIA safety hoops are proposed as a solution to such security risks, and the evaluation and testing methods for AIAs are put forward. At last, advanced AI technologies as well as the future of AI security and safety are discussed.

This book will benefit a wide readership, including researchers on AI and security, the organizational management and anyone with interest in these fields.

Contents

1 Overview of Artificial Intelligence Development 1
Binxing Fang and Weihong Han

2 AI Development Strategies in Countries Around the World 51
Shudong Li and Binxing Fang

3 Introduction to Artificial Intelligence Security and Safety 85
Binxing Fang and Xiang Cui

4 AI for Security 97
Peng Liao

5 AI Security 121
Zhaoquan Gu, Le Wang, Denghui Zhang, and Binxing Fang

6 AI Safety 141
Zhihong Tian and Shen Su

7 Artificial Intelligence Actant 159
Hui Lu, Zhihong Tian, and Binxing Fang

8 AIA Safety Hoop 189
Binxing Fang, Xiang Cui, and Zhihong Tian

9 AIA Safety Assessment and Detection 211
Lin Jiang and Binxing Fang

10 Codes of Ethics for Artificial Intelligence 247
Jie Yin

11 Frontiers of AI Technologies 305
Mohan Li and Binxing Fang

12 The Future of AI Security 349
Zhaoquan Gu and Binxing Fang

About the Editor

Binxing Fang is Academician of Chinese Academy of Engineering, Senior Chief Scientist of China Electronic Corporation, Honorary President of Cyberspace Institute of Advanced Technology, Guangzhou University, and Chief Academic Advisor of College of Computer Science and Technology, Harbin Institute of Technology (Shenzhen). He is also Chairman of Chinese Information Processing Society of China, Chairman of Network and Information Security Technical Committee of China Association for Standardization, and Chairman of Emerging Technology Security Innovation Forum of Chinese Cyberspace.

Chapter 1
Overview of Artificial Intelligence Development

Binxing Fang and Weihong Han

Many scholars, at home and abroad, have endeavored to explore the ins and outs of the long, complex course of artificial intelligence (AI) development. Instead of expounding on the history of AI development, this chapter merely intends to give a brief introduction to it, along with some impacts that AI may have on human society.

1.1 Research Directions and Origin of AI

1.1.1 Research Directions of AI

As an important branch of computer science, AI is garnering growing attention in the computer field while being widely applied to medical treatment, agriculture, industry, finance, communication and other areas.

AI science and technology, by its very nature, is a human-like intelligent system based on a computing system that aims to analyze the law of human intelligence activities and study how to imbue computer with the ability to do things that once entailed human intelligence. In other words, the goal of the AI discipline is to research into the basic theories, methods and techniques about how to simulate intelligent behavior of mankind through computer hardware and software.

B. Fang
Guangzhou University, Guangzhou, China
e-mail: fangbx@cae.cn

W. Han (✉)
Peng Cheng Laboratory, Shenzhen, China
e-mail: hanwh@pcl.ac.cn

© Publishing House of Electronics Industry 2025

B. Fang (ed.), *Artificial Intelligence Security and Safety*,
https://doi.org/10.1007/978-981-96-9263-7_1

1.1.2 Origin of AI

In 1950, Alan Mathison Turing, a British scientist, published a paper titled "Computing Machinery and Intelligence", in which the concept of intelligence was put forward. To define intelligence, Turing introduced what is now known as the Turing test to the general public: if a machine can engage in a conversation with a human (through telex equipment) without being detected as a machine, it has demonstrated human intelligence [1]. This is supposed to be the earliest description and criterion of "intelligence" in AI.

On August 31, 1955, American scholars John McCarthy, Marvin L. Minsky, Nathaniel Rochester and Claude E. Shannon co-authored "A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence" [2], suggesting that a two-month, ten-man study of AI be carried out during the summer of 1956 at Dartmouth College. The study, as proposed by the authors, "is to proceed on the basis of the conjecture that every aspect of learning or any other feature of intelligence can in principle be so precisely described that a machine can be made to simulate it". In the proposal, John McCarthy et al. mentioned their plan to "find how to make machines use language, form abstractions and concepts, solve kinds of problems now reserved for humans, and improve themselves". They believed that "a significant advance can be made in one or more of these problems if a carefully selected group of scientists work on it together for a summer". In addition to the four initiators, a number of other scientists and researchers, such as Trenchard More, Arthur Samuel, Allen Newell, Herbert A. Simon, Ray Solomonoff and Oliver Selfridge, also showed up in the 1956 Dartmouth Conference, in which the term "artificial intelligence" was coined. As the first AI seminar in human history, the conference also heralded the birth of the AI discipline. The year 1956, therefore, is referred to as "the first year of the AI era", which is of tremendous historical significance.

1.2 Three Peaks of AI Development

Since it was brought forward back in 1956, AI has made considerable headway over the last half a century or so. Its development, however, has not been all plain sailing. Instead, it is a spiral process with ups and downs, during which three peaks of development have occurred successively.

1.2.1 First Peak of AI Development

The first peak of AI development appeared in the late 1960s and early 1970s, when the integration between AI systems and specific applications was brought to the attention of people. Impressed by the immense charm of AI, they started translating

the results of theoretical research into practice. The major sign of the first peak was the emergence of expert systems.

When AI was first put forward in the 1950s, its research was mostly theoretical. With the logical approach as its basic method, AI research mainly focused on a few directions, including automated reasoning, cognitive model, knowledge representation and reasoning, AI languages, architectures and tools, etc. AI in the early days was applied to such fields as machine translation, theorem proving, general problem solving, chess programs, industrial feedback control and robotics [3].

In 1957, the American experimental psychologist Frank Rosenblatt from Cornell University simulated and implemented the perceptron model on an IBM704 computer. The model, which seemed to simply connect a number of M-P (McCulloch-Pitts) neurons together, could perform some machine vision and pattern recognition tasks through machine learning, thus giving a boost to AI development. In 1959, Wang Hao, a logician from the University of Oxford, wrote on an IBM704 computer a program that mechanically proved over 350 first-order logic theorems in Russell and Whitehead's *Principia Mathematica* in merely nine minutes [4], and was as a result hailed as a trail-blazer in the field of mechanical theorem proving.

In the 1960s, AI researchers began to shift their focus to intelligent representation, intelligent reasoning and intelligent search, believing that AI-based problem solving was a search process whose effect was related to heuristic functions. Main research achievements in this stage include the A* algorithm (search method) [5, 6], LISP (an AI programming language) [7], principles of conflict resolution [8], knowledge representation by semantic networks [9] and so on.

In the late 1960s and early 1970s, a group of young scientists, led by Edward A. Feigenbaum, came up with the concept of knowledge engineering [10] and went on to engage in the research and application of knowledge-based expert systems, which pushed AI development to its first climax, convincing the academia that the era of AI had dawned. In 1969, the latest AI research findings were pooled together in the International Joint Conferences on Artificial Intelligence (IJCAI). Main research achievements in this stage include XCON (a computer configuration expert system) [11], DENDRAL (a chemical mass spectrometry system) [12], CADUCEUS (a medical consultation system) [13], PROSPECTOR (a geological exploration expert system) [14], MYCIN (a disease diagnosis and treatment system) [15], Hearsay-II (a speech understanding system) [16], etc. It was the research and development of these expert systems that took AI from theoretical research to practical application.

People at that time had false hopes for AI, so much so that they went and predicted that "AI can overtake human smarts in a decade" [17]. However, due to sundry constraints, such as AI's confinement to certain fields, the difficulty in knowledge generation, lack of experience and insufficient data, expert systems could solve only a very limited amount of problems, much less "overtake human smarts", thereby pricking the bubble of people's wild optimism about AI.

In July 1972, Sir James Lighthill was commissioned by the Science Research Council (SRC) to conduct a general survey on the state of AI research in the UK. Later on in 1973, Lighthill provided an unbiased report, in which he criticized AI for failing to attain its "grandiose aims", saying that "most workers in AI research

and in related fields confess to a pronounced feeling of disappointment in what has been achieved in the past twenty-five years. Workers entered the field around 1950, and even around 1960, with high hopes that are very far from having been realized in 1972. In no part of the field have the discoveries made so far produced the major impact that was then promised". Lighthill's report provoked a massive loss of confidence in AI by academic institutions in both the UK and the US, followed by a nosedive in investment, which eventually plunged AI into a slump that is now generally believed to be the first AI winter [18].

1.2.2 Second Peak of AI Development

The second peak of AI development, which occurred in the late 1980s and early 1990s, was mainly prompted by the "Fifth-Generation Computer" project [19] initiated in 1981 by the Japanese government that aimed to develop machines able to carry on conversations, translate languages, interpret pictures, and reason like people. As much as $850 million was granted by the government to support this project. Japan's plan to develop a fifth-generation computer also caused ripples in Chinese academia that started following up on the project in 1984 and launched the First National Fifth-Generation Computer Conference in Zhuozhou early in 1985. In the conference, attendees had an in-depth discussion on the fifth-generation computer architecture, knowledge and information processing, rules and reasoning, the application of fifth-generation computers (including machine translation, robots, image processing, etc.) and so on. In 1986, China listed major projects concerning intelligent computer systems, intelligent robots and intelligent information processing as part of its National High Technology Program (a.k.a the "863" Program), speaking volumes for the fact that the country had made AI a key research area.

In the 1980s, domain-specific expert systems made a hit in a commercial setting as expert systems and relevant tools grew increasingly sophisticated and commoditized. Meanwhile, the proposal of Hopfield neural network and back-propagation neural network greatly bolstered people's confidence in AI. Later in 1987, the First International Conference on Neural Networks was held in the US. More than 1500 people from the academic and business circles attended the conference, whose main topics covered the design, functions, software simulation and chips of neural computers, AI and neural networks, etc. The conference announced to set up the International Neural Network Society and decided to found a journal called *Neuro Computer*, setting off a craze for neural network research. Having solved the problems of nonlinear classification and learning, neural networks began to develop by leaps and bounds. In response, a series of research projects on speech recognition, machine translation and so on came to the fore. Gradually, countries across the globe upped their investment in neural networks.

According to some scholars, AI between the 1980s and the 1990s was built upon multiple processing methods for a mass of classified knowledge, but it was beset

with problems, including limited application areas, costly update and maintenance and lack of massive data to train machines, to name but a few.

In 1992, the Japanese government announced the failure of its "Fifth-Generation Computer" project, ringing down the curtain on the 10-year research [20]. A main cause was that whilst Japanese scholars were sparing no effort to study AI-oriented computers, other countries like the US made great strides in general-purpose computers, whose speed and parallel efficiency saw marked improvement, putting to shame the reasoning speed of Japanese special-purpose computers. Here is a glimpse into how Western countries evaluated Japan's "Fifth-Generation Computer" project: "The problem for Japan is that the computer industry shifted so rapidly that the technological path the Fifth Generation took–which seemed a wise choice in 1982–turned out to be at odds with the computer industry's direction by 1992…Today, however, most people prefer powerful general-purpose computers to AI special-purpose ones." [21].

In fact, the same controversy over technical routes also exists in China. On the one hand, people are working on general-purpose high-performance supercomputers, like the Galaxy Supercomputer developed by the National University of Defense Technology; on the other, AI researchers are exploring special-purpose computers that support AI, including Prolog inference engines [22, 23], database machines [24], production computers [25], data stream driven computers [26, 27] and so on. Nevertheless, only general-purpose computers, given their strong vitality and applicability, have truly thrived in the market, whereas those AI special-purpose ones, more often than not, are short-lived.

Relevant research on fifth-generation computers has reshaped the way people see AI and unfolded difficulties that plague AI development. Every coin has two sides—knowledge input has brought AI development to a new height, but it has also given prominence to the difficulty and complexity of knowledge description. The biggest problem people faced back then was: if all AI knowledge required human input, there would be no end whatsoever to what needed to be fed into the computer.

In the meantime, neural networks per se lack rigorous mathematical support, and the back-propagation neural network has the vanishing gradient problem [28]. Besides, expert systems at that time were highly demanding as regards domains, leaving no room for cross-domain operation. In reality, however, applications were not confined to a specific domain, hence poor practicality of expert systems. Moreover, it was no easy to construct knowledge bases of expert systems, because AI workers didn't have relevant expertise whilst experts well-versed in those fields had no clue about how to channel their knowledge into forms required by expert systems, which also accounted for the lack of impressive AI applications.

After the 1990s, computers became more miniaturized, parallelized, networked and intelligentized. AI was gradually combined with and incorporated into such mainstream technologies as multimedia technology and database technology, aiming to render computers more intelligent, efficient and closer to people so that AI, just like general computer applications, was no longer considered one-of-a-kind [29]. After the second peak, AI slid into the second trough [28].

1.2.3 Third Peak of AI Development

The third peak of AI development appeared in the twenty-first century largely because the emergence of deep learning technology solved the vanishing gradient problem [18], and the deep network structure of deep learning enabled it to automatically extract and characterize complex features [30]. Gone were the days when features had to be manually extracted by traditional means.

In 2006, Geoffrey E. Hinton put forward the concept of deep learning, ensued by an upsurge in deep learning study [31], which, once again, led to booming AI research. Major research achievements in this stage are as follows: in 2010, Google's driverless cars set a record by covering over 160,000 km without causing a single accident; in 2011, IBM Watson beat the top prize winner and the holder of the longest winning streaks in a mind game called "Jeopardy!"; in 2016, Google AlphaGo prevailed Lee Sedol, a Go master of 9-dan rank; in 2017, Google AlphaGo defeated Ke Jie, the world's number-one Go player (also of 9-dan rank) with a resounding 3–0 victory; afterwards, AlphaGo Zero started from scratch, played 4.9 million games against itself in quick succession in the first three days without relying on any guidance from chess manuals or a priori knowledge of mankind, and eventually got the better of AlphaGo.

The display of multiple eye-catching AI achievements that overshadow human capacity in some domains (especially those related to deterministic knowledge) brought people, the commoners in particular, face to face with the power of AI and the fact that AI had already found its way into their everyday life. This made AI all the rage in the commercial market. Start-ups began to mushroom. Investors vied with one another to have a finger in the pie.

According to *2018 World AI Industry Development Blue Book* [32], *Global AI Industry Data Report* [33], etc. released by China Academy of Information and Communications Technology, a lot had been achieved in AI related academic research: from 1999 to 2017, more than 100,000 patents were filed and granted for such key technologies as image recognition, biometric identification, speech recognition, speech synthesis and natural language understanding in the field of AI; between 1998 and 2018, global AI saw a surge in papers and journals, with the total number of relevant papers published exceeding 630,000 and the compound annual growth rate (CAGR) amounting to 11.59%. All these figures allow us a peek into the enthusiasm of the academia about AI technologies during that period.

As for industrial development, there were over 5000 active AI enterprises around the world by the end of March 2019, and the global financing scale of AI enterprises stood at $12.6 billion in the first quarter of 2019. To be specific, the global market for smart sensors is expected to grow from $26.906 billion in 2017 to $70.617 billion in 2023, with the CAGR hitting 17.45%; the global market for AI chips is expected to reach $10.8 billion by 2023 (CAGR: 53.6%); as to AI-based computer vision, the global market is estimated to expand from $2.37 billion in 2017 to $25.32 billion in 2023 (CAGR: 47.54%); the global market for natural language processing is expected to grow from $7.63 billion in 2016 to $16.07 billion in 2021 (CAGR:

16.07%); the global market for AI-based fintech is expected to grow from $1.338 billion in 2017 to $7.306 billion in 2022 (CAGR:40.43%); the global market for AI-based education is expected to grow from $373.1 million in 2017 to $3.6835 billion in 2023 (CAGR: 46.47%); the global market for intelligent agriculture is expected to grow from $6.7 billion in 2017 to $13.5 billion in 2023 (CAGR: 12.39%); the global market for intelligent manufacturing is expected to grow from $202.82 billion in 2017 to $479.01 billion in 2023 (CAGR: 15.4%); the global market for AI-based network security is expected to grow from $4.96 billion in 2017 to $17.72 billion in 2023 (CAGR: 29%); global sales of self-driving cars are expected to reach nearly 600,000 by 2025 and 21 million by 2035 (CAGR: 42.7%). These industrial data show that AI in this period did stay in a pretty high position.

A noteworthy feature of the third peak of AI development is that many state governments also moved along by formulating national policies and leveraging national strength to guide AI development and promote the application of AI.

In 2011, the US government introduced the *National Robotics Initiative (NRI)* [34] and received a $50 million grant from the National Science Foundation the next year [35]. In 2013, as a follow-up to the NRI, *A Roadmap for U.S. Robotics: From Internet to Robotics* [36], which was funded by the US government, was released to the public, followed by *The Road to Cutting-Edge Robots* [37] that aimed to press ahead with the roadmap; in April 2013, the National Institutes of Health launched *The Brain Research through Advancing Innovative Neurotechnologies* (BRAIN Initiative), with a view to developing more comprehensive tools and information bases to understand how the brain functioned in health and disease [38]; in May 2016, the White House put out *Preparing for the Future of Artificial Intelligence* [39]; in October 2016, National Science and Technology Council initiated *The National Artificial Intelligence Research And Development Strategic Plan* [40] in an effort to provide the nation with cross-sectoral strategic guidance for AI development; in December 2016, the White House released *Artificial Intelligence, Automation, and the Economy* [41], asserting that AI would change the way people engaged in economic activities in the coming decades; in July 2017, the US enacted the *H.R.3388—SELF DRIVE Act* [42] to ensure the safety of automated vehicles; in October 2017, the Software and Information Industry Association issued *AI Policy Principles* [43] to fully tap the potential of AI and promote the cooperation between relevant public and private sectors.

In 2013, the European Union (EU) launched the *Human Brain Project (HBP)* [44], one of four FET (Future and Emerging Technology) Flagships and the largest scientific project it had ever funded; in 2014, the European Commission and the European Robotics Community co-launched *The Partnership for Robotics in Europe (a.k.a. The SPARC Plan)* [45, 44] to maintain and extend Europe's leadership in robotics; in May 2016, the European Commission proposed a legislative motion to deal with the impact exerted by the development of robots on human beings [46], whereafter the European Parliament drew up *European Civil Law Rules in Robotics* [47] in October that year.

In December 2016, the Government Office for Science of the UK released *Artificial Intelligence: Opportunities and Implications for the Future of Decision Making*

[48], in a bid to show how the AI technology could be responsibly used to improve the lives and living standards of everyone in Britain; in October 2017, the UK government issued a report titled "Growing the Artificial Intelligence Industry in the UK" [49], which made 18 recommendations describing how Government, industry and academia should work together to keep the UK among the world leaders in AI.

In January 2015, the Ministry of Economy, Trade and Industry of Japan mapped out *New Strategies for Robotics: Vision, Strategy and Action Plan* [50], covering technological advances, regulatory reforms, standards for the globalization of robotics and other measures related to the robot revolution; in 2017, the Japanese government worked out the *Next-Generation AI Promotion Strategy* [51] to specify the objectives of AI research and development and the roadmap of industrialization; during the Comprehensive Innovation Strategy Promotion Meeting in June 2019, the Japanese government released *AI Strategy 2019* [52] to boost Japan's industrial competitiveness, which also covered AI education reform, research and development, social implementation and other comprehensive policy programs.

In April 2016, three ministries and commissions of China, including the Ministry of Industry and Information Technology (MIIT), issued the *Robotics Industry Development Plan* (2016–2020) [53] to make clear the robotics industry development blueprint during China's 13th Five-Year Plan; in May of the same year, four ministries and commissions, headed by the National Development and Reform Commission, jointly launched the *"Internet Plus" Artificial Intelligence Three-Year Action Implementation Plan*, aiming to cultivate world-leading AI enterprises in a number of key areas; in July 2017, the State Council released the *New Generation Artificial Intelligence Development Plan* [54], in a bid to establish national strategic goals for AI development in different stages; in December 2017, the *Three-Year Action Plan for Promoting Development of a New Generation Artificial Intelligence Industry* [55] was initiated by the MIIT to help AI signature products gain an international competitive edge in a few key areas; in April 2018, to guide institutions of higher learning to aim at the frontiers of world science and technology and provide strategic support for China's new generation AI development, the Ministry of Education issued the *AI Innovation Action Plan for Colleges and Universities* [56]; on October 31, 2018, the Political Bureau of the Communist Party of China (CPC) Central Committee held the ninth collective learning on the current situation and future trends of AI development, during which General Secretary Xi Jinping, when chairing the event, stressed the necessity to fully understand the significance of accelerating the new generation AI development, strengthen the leadership, make plans, specify tasks and lay the groundwork, all with a view to facilitating the in-depth integration between AI and socioeconomic development and promoting the sound development of new generation AI in China [57]. The fact that the nation put a premium on the AI field and proactively pressed ahead with it also greatly fueled this wave of AI, so that some argued that the third peak of AI development might lead to the Fourth Industrial Revolution [58].

1.3 Three Academic Schools of AI

In this day and age, everyone is aware of the ubiquity of AI: we resort to voice input that spares us the trouble of typing on the keyboard; we quickly swipe our phones to skim over feeds pushed to us according to our interests that keep us up-to-date with the most trending news. The AI system, for example, took us all aback when it outplayed the greatest Go master in the world; we take it for granted as robot waiters busily carry our dishes around; sometimes we would like to grab some "me moments" in an unmanned hotel; with the help of intelligent navigation, we can steer clear of traffic jams; CamDictionary, a real-time translation app, frees us from barriers caused by foreign languages; to find out the species of a beautiful flower we chance upon, all we need to do is tap that flower recognition app in our mobile phones… None of these would have been possible were it not for AI. In these scenes, a wide diversity of AI technologies are involved, such as image recognition, machine translation, speech recognition, intelligent recommendation, decision tree algorithm, intelligent input, question-answering system and walking robot.

Throughout the course of AI development, three academic schools have greatly shaped AI research: symbolism, connectionism and actionism. With symbolic reasoning and machine reasoning as its core, symbolism studies intelligence and reasoning by means of symbols, whose founders include John McCarthy, Herbert A. Simon and Allen Newell; centering on neural networks and deep learning, connectionism, founded by Marvin L. Minsky, mimics the human nervous system and presents its model through computing so as to simulate intellectual abilities; pioneered by Norbert Wiener, actionism advocates control, self-adaptation and evolutionary computation.

1.3.1 Symbolism

Symbolism refers to symbolic artificial intelligence [59]. It is also known as logicism, psychologism or computerism, relying on an advanced symbol processing system based on problems, logic and search, with physical symbolic system hypothesis and finite rationality as its principles. Symbolic AI stems from the description of intelligent behavior by mathematical logic in the 1930s. After the advent of computers, the logic deduction system was implemented on computer, whose application significantly drove AI development. Symbolic cognition in AI has symbols as its basic elements. By abstracting information and behavior into the physical symbol system based on symbol rules, it uses computer logic reasoning rules to simulate human abstract thinking, through which the simulation of intelligent behavior is rendered possible.

In 1955, it dawned on Simon that machines that manipulated symbols excelled in simulating decision-making behavior and even the process of human thinking, so he started discussing with Newell the possibility of teaching machines to think.

At their invitation, John Clifford Shaw developed "Logic Theorist (LT)" [60] in 1956, the first program deliberately engineered to mimic the problem-solving skills of a human being, which is why it is also known as "the first AI program". Logic Theorist established the field of heuristic programming and soon proved 38 of the first 52 theorems in Chap. 2 of *Principia Mathematica* co-authored by Bertrand Russell and Alfred Whitehead. Impressively, one of the theorems was more elegantly proved than Russell and Whitehead did. In 1956, Simon and Newell were invited to the Dartmouth Conference that gave birth to the field of AI. In the conference, they presented the audience with their Logic Theorist, which was considered as one of the highlights in that meeting.

In 1956, McCarthy, a chess enthusiast fascinated by chess-playing computer programs, co-organized the Dartmouth Conference in which he introduced the idea of the α-β algorithm [61] and wowed the whole audience with it. By setting a lower bound (α) a player must consider in order not to lose and an upper bound (β) the other player must consider in order to win as judging criteria in a chess game, this algorithm can determine which choices are outside the upper and lower bounds and thus put them aside, thereby pruning parts of the search tree back to the previous square-root level. It was with the α-β algorithm that the IBM Deep Blue defeated Garry Kasparov, a chess grandmaster, in 1997 [62]. McCarthy pointed out that the conference was based on the conjecture that every aspect of learning or any other feature of intelligence could in principle be so precisely described that a machine could be made to simulate it [63]. Promoted by McCarthy, the use of symbols to describe or study the characteristics of intelligence has since remained the focus of AI research. In 1956, McCarthy wrote the book *Automata Studies* [64] with Shannon and went on to invent the LISP programming language [65], which, deemed as a byword for AI at one point, further enhanced McCarthy's standing in the field of AI.

Given their promotion in early days as well as their contributions to symbolic AI, McCarthy, Simon and Newell were regarded as founders of the symbolism school.

According to the symbolism school, people could understand each other more easily when they used different symbols for communication, which was also why human beings became intelligent; they believed human cognition was a process operating on symbol representation that could be simulated by computer-based symbol manipulation. Generally, the research of the symbolism school was divided into three steps: to begin with, the functional mechanism of human cognitive system was explored; secondly, symbols and operation based on symbols were adopted to describe the process of human cognition; in the third and final step, symbol representation and manipulation based on symbols were achieved by the computer system, through which the process of human cognition was simulated, thereby making AI a reality.

The symbolism school is characterized by its top-down implementation of AI through deductive reasoning. From the viewpoint of symbolists, the crux of AI lied in knowledge representation, knowledge reasoning and knowledge application. As the very foundation of intelligence, knowledge is a form of information that can be represented by symbols, while cognition is a process of dealing with symbols; reasoning is a process of solving problems in a heuristic way that can be described

in some formal symbols and languages. As a result, it is possible to establish a unified theoretical system of human intelligence and machine intelligence based on knowledge through symbol representation, processing and computation.

After the term "artificial intelligence" was coined by symbolism researchers in 1956, a wide array of theories and technologies came into being, including the heuristic technology, expert systems, knowledge engineering, etc., which paved the way for early AI development and application. Among others, expert systems, as computer programs designed based on human expertise, can be used to solve knotty problems, which is a typical example of symbolic computation. Moreover, in combination with expert knowledge, expert systems can precisely process rules and perform reasoning through symbolic computation. For instance, networks based on production rules can be used to support the reasoning process with relational connection symbols similar to the if–then statement, thus designing and implementing expert systems in different domains according to specific problems.

Due to multifaceted advantages of expert systems, symbolism enjoyed the limelight for quite a long time in the field of AI research, particularly in the 1970s, when a large number of expert systems that combined domain knowledge appeared and geared AI up for practical engineering application. Symbolic AI programs able to perform various tasks can be created by defining symbols and operations associated with logical reasoning. Object Oriented Programming (OOP), for example, is a typical result of symbolic AI.

From the mid 1950s to the late 1980s, the ideological system of symbolism had been spearheading AI development as a main paradigm of AI research. Consequently, symbolism research was regarded by scholars as a traditional AI approach which assumed that symbol manipulations produced part of intelligence. Newell and Simon defined such an assumption as the physical symbol system hypothesis [66]. In his *Artificial Intelligence: The Very Idea* [67], a book published in 1985, John Haugeland referred to the traditional AI approach as GOFAI (Good Old-Fashioned Artificial Intelligence) and named it symbolic artificial intelligence.

One of the salient features of symbolic AI rests with its logical deduction, which makes its behavior interpretable so that people can easily visualize the logic of a rule-based program to troubleshoot its system. When it comes to input transactions with very clear rules, symbolic AI can easily access a rule description and convert it into symbols for logical deduction. Due to its good explainability, logical deduction has registered good results in systems where symbols are fully expressed and rules are clearly described.

Symbolism had long outshone its rivals. The successful development and application of expert systems, in particular, played a vital role in taking AI from theory to engineering practice, which was tremendously conducive to AI development. Despite other AI schools that followed, symbolism has remained the dominant one all along. In recent years, the knowledge graph technology has given symbolism a new lease of life. It is not a rule model and has no direct reasoning capacity, but as a form of knowledge engineering it boasts a gigantic knowledge base and, by connecting symbols, can associate relevant contents and offer extraction, search and

other services through the knowledge base. Moreover, it can deduce unknown relationships from known ones, which embodies reasoning capacity in a broad sense and demonstrates its feature of intelligence. There are two ways to build a knowledge graph—one is to extract existing knowledge through information extraction and the other is to generate new knowledge through traditional symbolic reasoning.

Intelligent input, decision tree algorithm and question-answering system mentioned in the first paragraph of Sect. 1.3 are fields symbolism excels in. Of course, with the rapid development of the knowledge graph technology and the increasing power of knowledge representation, the recommendation algorithm may be a closer approach to symbolism.

1.3.2 Connectionism

Connectionism [68], also known as bionicsism or physiologism, is an intelligent simulation method based on network connection mechanisms and learning algorithms, which also holds sway in AI research. It believes that AI stems from bionics, especially from the human brain model, and attempts to simulate the structure and functions of the human brain nervous system by engineering means.

Connectionism can be traced back to an article titled "A Logical Calculus of the Ideas Immanent in Nervous Activity" [69] published in 1943 by Warren S. McCulloch, a physiologist, and Walter Pitts, a mathematical logician, which put forward the McCulloch-Pitts (MP) mathematical model of a biological neuron, proposed to describe the structure of neurons and networks in a formal manner and pioneered a new way of using electronic devices to mimic the structure and functions of the human brain. The M-P model is still in use today as the fundamental structure of neurons. The article argued that human cognitive systems and intelligent behavior were the result of high-level activities in the human brain, and that intelligent activities were put down to a large number of simple neurons running in parallel through complex interconnections.

In 1951, Minsky teamed with Dean Edmonds to design the world's first artificial neural network self-learning machine—SNARC (Stochastic Neural Analog Reinforcement Calculator) [70]. Using 3000 vacuum tubes to simulate a network of 40 neurons, the machine started out trying to learn how to find the way through a maze. By trial and error, it picked up some methods to solve problems, laying the groundwork for connectionism-based AI development. In the 1956 Dartmouth Conference, Minsky, as one of the initiators, introduced how SNARC worked, which was regarded as a notable academic achievement and one of the reasons why Minsky was looked up to as the founding father of connectionism.

With the help of simulation methods based on neural networks, the connectionism school simulates the structure, functions, etc. of the human brain nervous system by engineering techniques. It mimics the vast number of nerve cells (neurons) in the human brain using a massive amount of parallel processors and simulates the synapses between neurons in the brain through connections between processors. An

artificial neural network consists of a good many neurons arranged in a layered format that are connected to form a complete neural network. Then, lots of data are used to train the network and the weights of connections between neurons are adjusted with a learning algorithm to constantly improve the network. The artificial neural network technology has blazed a new trail for AI development.

Since the perceptron model, which was developed in the 1960s–70s, could train data to automatically adjust the weights of connections in the M-P model, brain models established based on it [71] caused quite a stir at one point. Due to various technical constraints back then, however, the study of brain models ground to a halt in the late 1970s and early 1980s. Connectionism stayed out in the cold until Professor John J. Hopfield published two important papers in the field respectively in 1982 [72] and 1984 [73], proposing to simulate neural networks with hardware. In 1986, David E. Rumelhart et al. put forward the back-propagation algorithm [74] that could repeatedly adjust the weights of connections in multi-layer neural networks, making a major breakthrough in theoretical research. Thereafter, connectionism AI started gaining irresistible momentum. From models and theoretical research to algorithms and engineering implementation, neural networks gradually went from the academia to industrial circles – after making their presence felt in multiple domains, they gradually found their way into the market.

In 2007, Geoffrey E. Hinton, a bellwether of the connectionism school, came up with the deep learning algorithm [75] that enormously boosted the learning ability of neural networks. According to a report released in April 2013 by MIT Technology Review [76], Google acquired Hinton's neural networks startup in March 2013 to "take out ideas in this area and apply them to practical problems" related to image recognition, search, natural language understanding and so on. With the help of deep learning technology, Google's billion-connection deep learning neural network can recognize objects that humans have never defined or labeled; Peter Lee, head of Microsoft Research, believed that deep learning had revived some of the biggest challenges in AI; in his speech in China in October 2012, Rick Rashid, Chief Research Officer of Microsoft Research, demonstrated a voice program that converted what he said into an English text (with an error rate of merely 7%) and translated it into a Chinese text before presenting to the audience a spoken Mandarin version in a voice that mimicked his.

Considered as an intelligent learning method closest to human brain, deep learning gradually established itself as an essential part of AI. IBM Watson, Google's search engine and the algorithm of Facebook now can't do without it. As many enterprises started vigorously improving the deep learning technology and applying it to diverse fields, including speech recognition, image recognition, natural language processing and so on, MIT Technology Review included deep learning into the list of Top 10 Breakthrough Technologies of 2013 [77], claiming that "with massive amounts of computational power, machines can now recognize objects and translate speech in real time. Artificial intelligence is finally getting smart".

The major advantage of connectionism is that neural networks can discover rules from training data. The user provides sample data, including input data and corresponding output data (the larger and more diverse the dataset, the better); connectionist algorithms then apply statistical regression models to adjust the weight coefficients of their intermediate variables, until the best fitting model is found. The weights are adjusted in the direction that minimizes the cumulative error from all the training data points, using techniques such as gradient descent. Since these techniques are effectively error minimization algorithms, they are inherently resilient to noise. They will smoothen out outliers and converge to a solution that classifies the data within some margin of error.

Connectionist algorithms do not need a model of the world. They just need enough sample data from which the model of the world can be inferred statistically. This is a very powerful characteristic, but also a weakness. The input features have to be very carefully selected to stay away from error pollution. They also have to be normalized or scaled, to avoid that one feature overpowers the others.

Speech recognition, speech processing, machine translation, image recognition and some other frequently mentioned fields are where connectionism does well. Of course, along with the rapid development of deep learning, the reinforcement learning technology has been applied in a growing number of areas. Intelligent input, recommendation algorithm, game search algorithm, etc. have also approached connectionism as a boost to what it is capable of so that it can play a more powerful role in these application areas.

1.3.3 Actionism

Actionism [78], also known as evolutionism or cyberneticsism, believes that AI derives from cybernetics and that intelligent behavior, without the need for knowledge, representation and reasoning, is accomplished by the subject that automatically perceives the feedback of the external environment it interacts with.

Rodney A. Brooks published two papers successively in 1990 and 1991 to criticize traditional AI research methods, pointing out that conventional symbolism methods paid undue attention to symbolic abstract manipulations with little, if any, physical basis in reality. As a consequence, he put forward an opposite viewpoint to the traditional symbolic AI research that an intelligent model could be established by mimicking the behavior of living beings [79], which should require no knowledge representation [80] or reasoning [81]. Based on his analysis, Brooks proposed that AI systems should simulate the intelligent evolutionary process of creatures in nature. In his opinion, intelligence resulted from the gradual evolution of living things through their interactions with their environment. Since no representation of intelligence could fully represent the reality of the objective world, argued Brooks, it was incomplete and inappropriate to represent intelligence with symbols. The novel method and approach adopted by Brooks to study AI from behavioral and evolutionary points of view, little by little, gave birth to the new school of actionism in the 1990s.

However, Brooks' ideas came from Wiener's cybernetics that was proposed way back in the 1940s–1950s as an integral part of technology trends at that time. Cybernetics associated the working principles of neural systems with information theory, control theory, logic theory, computer technology and so on, which inspired later AI researchers to consider behavioral intelligence simulation based on "perception–action" as an important direction of AI research. The early focus of this research direction was to simulate the intelligent behavior and role of human in the control process, including self-optimization, self-adaptation, self-correction, self-organization, self-learning, etc. of cybernetics. In the 1960s and 1970s, the research results of cybernetics laid the groundwork for the development of intelligent control and intelligent robotics. In the 1980s, the combination of cybernetics and AI technology replaced accurate mathematical models with fast autonomous feedback, which solved complex and uncertain problems to a certain extent, achieved favorable application results and brought intelligent control and intelligent robot system into the world. Given the important role of cybernetics in the field of actionism research, Wiener is regarded as the founding father of actionism.

Actionism can be well explained from the perspective of evolution. In the course of evolution, living things are going through constant variations. Then, the environment will come into play, phasing out variations that fail to adapt and allowing more adaptive ones to reproduce and flourish. Genetic algorithm, for example, is a classical algorithm based on such an idea.

Actionism simulates the behavior of animals which, in the absence of active brain intervention, rely on the coordination of their limbs and joints to respond to changes in the environment. It regards intelligence as a combination of the responses of organisms to environmental changes and a pattern of forming higher-level knowledge based on lower-level knowledge. Actionism requires different behavior modules to interact with the environment—through the visual module (like eyes), it can see changing images in the environment, while through the auditory module (like ears), it can perceive sounds around. Through different modules, it can observe environmental features and generate complicated responses, which achieves intelligence in a bottom-up "perception–action" approach. The theoretical goal of actionism is to predict changes in the environment and control behavior accordingly, and its research focuses on physical systems that can perform various tasks autonomously in the real world. Some argue that there is no such thing as a high-level concept for most human behaviors, because some reactions are just reflexes. People who practice martial arts, for example, have developed conditioned responses to the blows of an enemy through years of training, but such responses are nothing but low-level ones. There is no need for them to resort to a whole set of boxing moves formed through a high-level concept.

Genghis, a robot developed by Brooks in MIT Artificial Intelligence Laboratory, is regarded as a representative work of actionism [82], which aims to improve the intelligence of robot insects. As a six-legged insect-like robot, Genghis consists of 150 sensors and 23 actuators. With a reactive robot architecture based on the "perception–action" mode that allows it to respond to the world around, it has flexible anti-collision and roaming skills in the natural environment and is thus considered as

a new-generation "cybernetics system". The design of Genghis is inspired by insects that, albeit with limited brain functions, are powerful in some aspects. To simulate their features, Brooks refrained from managing all possible functions of Genghis with a central controller. Instead, each leg of the robot has its own built-in sensor that can sense various obstacles encountered on the way forward and enables the robot to respond in accordance with feedback given by those sensors in different circumstances [83]. The goal of coordination between all legs of Genghis is to complete the act of walking so that the robot can move around properly. The coordination is an independent process always in motion that would be triggered when prerequisites are met. This shows that Genghis dwarfs many previous robots in coping with a complex environment despite its lack of human capacity for understanding, reasoning and planning.

As a matter of fact, what actionism goes for is autonomy which, to be specific, is embodied in the following four aspects: (1) self-governance, which means that an intelligent actant has some control over its own behavior and internal elements instead of being directly controlled by a person or other objects; (2) social competence (communicability), which means the intelligent actant can exchange information with other subjects; (3) reaction capacity, which means the intelligent actant is able to perceive its surroundings through multiple modules and change the environment; (4) spontaneous behavior, which means the intelligent actant should behave in an active and spontaneous way. Thus, the research focus of actionism can be interpreted as seeking a reasonable coordination mechanism.

Popular topics nowadays, such as discovery algorithm, recommendation algorithm, walking robot and path planning technology, are areas where actionism stands out.

1.3.4 Critique of Three Academic Schools

As a simulation of deductive logical thinking of human beings, symbolism relies on logic science and is considered as a functional simulation method aiming to simulate people's left-brain abstract logical thinking, which falls into the category of thinking-based AI research methods. Some people believe that symbolism is a method favored by computer workers that focuses on logical reasoning, which explains why they conclude that "cognition is computation". However, symbolism is criticized and denied by other schools of thought due to its limitations in its capability to deal with knowledge representation of uncertainties, common sense, intricate problems and the like.

Symbolism advocates establishing a unified theoretical system of AI by logical means. Nevertheless, rule processing usually solves only deterministic problems, and it is difficult to address uncertainty problems through comparative judgment, which sheds light on the deficiency of symbolism during the process of problem solving. For instance, if we create a rule-based program to detect an image containing a specific person, obviously the rule would only work if it is an exact match; with even the

slightest variation, it would be impossible to identify that person, for we can't be sure at what angle to rotate the picture, what kind of shade would be on his face, whether the pixel of the face image we see this time is the same as that of the known sample, and so on and so forth. In a nutshell, as long as there is not an exact match, the program will not give us an answer, because symbolic AI can't handle global judgment or fuzzy information processing very well. Even if we have millions of different pictures in place, there is no guarantee that the picture to be identified this time happens to be the same as a certain picture in our library. In essence, some techniques of symbolic AI can deal with unobservable probabilistic models, but they are generally not suitable for noisy input data or scenarios where the model is not clearly defined. Furthermore, it is by no means easy to convert some matters (such as speech recognition and natural language processing) into direct rules.

What's more, symbolic AI can't deal with common sense related issues. It is very difficult for symbolism to summarize common sense which, as a vast body of knowledge based on long-term practices, contains abundant rules and anomalies learned in daily life. Besides, a certain kind of common sense is applicable to different fields in most cases, but if we include all kinds of common sense into a domain-specific expert system, it would be a disaster for the system when a massive amount of common sense comes rushing in if there is not a way available to determine which knowledge is common sense and which is not. This leads scholars to the conclusion that symbolic AI systems have poor scalability. Symbolic AI has facilitated people's efforts to create complex expert systems containing many rules in particular domains, which, however, rest with tremendous hard work of domain experts and software engineers and can only be applied to a very limited amount of use cases. Once people find a new way out, it is only natural that symbolic AI will be sidelined, bidding goodbye to its leading role.

MacCarthy, the founder of symbolism, bet $1000 in 1968 that a computer would beat the world's human chess champions within ten years [84]. He lost the bet because it was difficult for symbolism to accomplish such an onerous mission that involved so enormous a search space. But then again, it never occurred to him that Deep Blue would defeat the chess grandmaster Garry Kasparov in 1997—and with the α-β algorithm he proposed! In fact, the development of computer technology played a role here, making the huge search space possible. But when it came to Go games, the search space, once again, was rendered unimaginably tremendous. Even with a rich reservoir of human knowledge, there was little room for improvement technologically. This was because people's highly abstract cognition might have deviated from the essence of things. The same problem also plagued natural language processing. Machine translation, which was based on symbolism, found it impossible to make a breakthrough in terms of accuracy. The emergence of AlphaGo put an end to this conundrum by combining the Monte Carlo Tree search framework with the deep learning and reinforcement learning of connectionism, and defeated the Go champion of mankind. As a result, it was believed that symbolism was destined to be replaced. Nonetheless, as the knowledge graph appeared and non-explainability of connectionism surfaced, some applications started integrating symbolism and

connectionism—for example, natural language processing is currently trying a technical route that combines the knowledge graph technology with deep learning. In fact, in the process of creating knowledge graphs, connectionism techniques (like neural networks) are generally employed to generate nodes and correlated edges of the graphs.

Connectionism materializes human inductive reasoning by bionic means. Being a transfer-based AI research method, it relies on neuroscience and is considered as a structural simulation method that aims to mimic right-brain intuitive thinking of mankind. Connectionism is said to be an approach favored by electronics workers that focuses on electrical connections, which is why it is put down to three words—"cognition is network".

In 1969, Minsky, the founder of connectionism, penned a book named *Perceptrons* [85] and proved that a single linear perceptron (the precursor of computing units that make up modern neural networks) couldn't learn the XOR function. This was as good as telling the public that neural networks had no future. If there was nothing Minsky could do about it, what was the use of neural networks? To a certain degree, the downfall of connectionism was ascribable to this book. Some even took Minsky as a representative figure of symbolism because of it [86].

As deep learning established itself increasingly as a force to be reckoned with in the AI area, connectionism started gaining an upper hand, which gave connectionists a greater voice and the grounds to point out the deficiencies of symbolism. Connectionists believe that neurons are the basic unit of not only the brain's nervous system but also behavioral responses; in their opinion, the thinking process of the brain is about neuronal connection and interaction, not a process of symbol manipulation. They reject assumptions about physical symbolic systems, convinced that human brain is different from computer and that the brain working mode advocated by connectionism should replace the computer working mode dominated by symbolic operation as the AI research approach. They argue that thinking and cognitive functions, instead of being determined by a few neurons, are the result of a large number of neurons working in concert and linked together by synaptic dynamics.

One downside of connectionism is that it can't explain how artificial neural networks process information, which results in the difficulty in clarifying higher-level phenomena, namely the complexity and scalability of networks.

Symbolists maintain that cognition is the deduction and calculation of meaningful symbols, whilst connectionists believe that cognition can be thought of as large-scale parallel computing that takes place in neural networks similar to the human brain, where neurons collaborate and interact with each other. "Intelligence" created by these two ideas is strikingly different, hence little, if any, overlap between the two schools during research. Of course, there is a certain degree of causal relationship involved in the transition from symbolism to connectionism. As rules of reasoning became conditional and probabilistic, symbolism had to face problems arising from the choice of probability, which called for the help from the outside world. In other words, firsthand experience of experts was needed to flesh out expert systems. When a vast amount of judgment depended on the outside world, machine learning was

deemed a must, leading to a natural transition to the connectionism approach characterized by big data training. In essence, while symbolism tries to derive conclusions from rules, connectionism learns conclusions from facts. Therefore, there is a good chance for symbolism to arrive at a conclusion that people have not thought of but is inevitably attainable (after all, mechanical theorem proving is yet to discover theorems of great significance unknown to man), but there is rarely any room for connectionism to create something new.

Besides, deep learning may not be the sole domain of connectionism, nor is it a good idea to simply equate deep learning with deep neural networks. In 2017, Zhihua Zhou proposed a learning method called deep forest [87], which, as a symbolic approach, has decision trees for symbolic processing as its basic unit. Deep forest is a cascade structure of decision trees, with each layer consisting of one, several and even many random forests, which aims to address the issue of imperfect training through a "deep" approach [88]. In a way, deep forest even boasts advantages that deep neural networks are short of, because it can keep adding a new layer of random forests during the learning process, until the new layer no longer makes any contributions to the training. This method, by nature, can adjust the complexity of models adaptively to cope with problems with corresponding models [89]. Obviously, deep learning based on symbolism is more suited to addressing discrete problems. As the deep forest continues to evolve, maybe someday discrete problems don't have to be converted into continuous vectors by deep neural networks before they can be dealt with; deep learning of symbolism may regain the discourse power over discrete problems, while connectionism can just focus on tackling continuous problems it is best at.

Actionism gives prominence to external interaction in a bid to embrace the real world. It relies on control science and is considered an action simulation method designed to mimic human cerebellar motor thinking, which makes it an action-based approach to AI research. Actionism is said to be an approach advocated by people working on control science. Stressing self-adaptation and equating cognition to response, it is in principle based on cybernetics and the "perception–action" control system.

Main ideas of actionism can be summarized as follows: first, intelligence depends on perception and action as well as the adaptation to the complex external environment rather than representation and reasoning, while the formal representation and modeling of knowledge is one of the major obstacles to AI; second, the description of intelligence should be prototyped by the response of the environment to the actor following the action of the machine on the environment; third, intelligent systems should have the ability to act and perceive in the real world, and different actions should exhibit different functions and control structures of intelligent systems; fourth, the capabilities of an intelligent system should incrementally increase in stages, and at each stage the intelligent system should be a complete system.

Since it was first introduced, actionism has attracted a great deal of attention. While numerous people are interested in it, there are many opponents as well. For example, it is argued that the behavioral success of Brooks' robot insect reflects low-level control behavior, not a process about the generation of higher-level control behavior. They thus conclude that it may be very difficult to get machines to evolve from insect

intelligence to human intelligence. Be that as it may, the rise of the actionism school still indicates that such ideas as cybernetics and systems engineering have been influencing AI development. It is safe to say that the evolution of actionism has been the least deterministic, or the least controllable. If symbolism determines behavior by rules, the existence of rules means that one can predict future behavior; connectionism repeats the process of human decision-making through learning, so unlearned behavior rarely occurs; actionism, however, implements AI systems through constant evolution, and as a result it is likely to cause the systems to generate unknown behavior.

Xinhe Xu et al. [90] believe that the main difference between actionism and symbolism is that intelligent actants of actionism are able to autonomously adapt to the objective environment through the "perception–action" approach, but symbolism requires designers to develop appropriate rules or mathematical models. Fundamentally, the adaptation to the environment by intelligent actants of actionism rests with the precise relationship between elements of complex systems as well as between the whole formed by various elements and the surroundings. To achieve such a precise relationship, intelligent actants of actionism must employ a coordination mechanism to adapt to the external environment. Meanwhile, the coordination mechanism should also bring together elements within intelligent actants and, furthermore, include collaboration between multiple intelligent actants.

In fact, all three academic schools of AI have developed their own unique problem-solving methodologies that have successful examples at different times in practice. For instance, examples of symbolism include a series of theoretical results such as automated theorem proving and resolution method; as for connectionism, they include such technologies as artificial neural networks and deep learning; when it comes to actionism, there are feedback control pattern, generalized genetic algorithm and so on. Research methods of these three AI schools have their respective pros and cons, and the future trend of AI development is to integrate them [91]. For example, symbolism is good at rules processing—based on rules, it can deal with untrained events that connectionism cannot, which is also a typical human way of thinking. Nevertheless, symbolism requires a high degree of certainty in events, and it is often difficult to find rules that are compatible with events of uncertainty. Under the circumstances, connectionism may offer a way out with its strength in tackling uncertain problems by performing fuzzy rule matching that aims to proceed with inferences and judgments. Besides, since symbolism pursues high-level concepts, it may have difficulty deriving new ones; connectionism is based on data training and therefore lacks innovative features, so it may be a good idea to consider lowering the level of the problem to be solved and simply use the evaluation function to determine the acceptability of a random action, so as to meet the requirements of innovation through the attributes of the random action. Once a breakthrough is achieved that leads us to a new area, we can turn to symbolism once again to conceptualize it. In doing so, innovation is made possible in a stable manner. Moreover, connectionism is also being combined with actionism so that their respective strengths can be brought to full play. Inspired by behavioral psychology, reinforcement learning focuses on online learning and tries to strike a balance between exploration and exploitation. Unlike

supervised and unsupervised learning of connectionism, reinforcement learning does not require any data to be given in advance. Rather, it obtains learning information and updates model parameters by receiving rewards (feedback) from the environment, which makes it a nice complement to connectionism techniques.

If, someday, robots can outperform humans, it must be the result of combining the strengths of multiple academic schools.

1.4 Technical Boost to the Third Peak of AI Development

The third peak of AI development is due in part to a few eye-catching events that have grabbed much attention, like how AlphaGo defeated human Go players. In the final analysis, however, the credit goes to technological breakthroughs, and the development of such new technologies as big data, mobile networks, the Internet of Things and microelectronics has also laid the technological groundwork for AI development.

1.4.1 Contributions of Deep Learning

Deep learning (DL) is an algorithm based on artificial neural networks that performs representation learning of data. Among others, the goal of representation learning is to find better representations and create better models to learn these representations from large-scale unlabeled data. Based on the idea of hierarchical abstraction, deep learning mimics how human brain neural networks work and then builds multi-layer neural networks to process data and information in different layers. Specifically, by abstracting low-level features of the underlying neural network, the raw data is outputted to the high-level neural network so that it can access high-level concepts more efficiently and process information more accurately. What's so great about deep learning is that it can extract features hierarchically with unsupervised or semi-supervised feature learning methods, thus replacing manual feature acquisition and design, and greatly boosting the efficiency of feature extraction. Consequently, these algorithms can be applied to unlabeled data that are beyond the reach of other algorithms and are more abundant and easily-accessible than labeled data. This also gains a sizable advantage for deep learning.

The standard back-propagation algorithm [92] put forward by Paul J Werbos in 1974 has a vital bearing on the later development of deep neural networks. In 1992, Juyang Weng et al. published Cresceptron [93] which, regarded nowadays as an early deep learning network, can recognize natural objects from real-world images. In the same year, Jürgen Schmidhuber introduced a multi-layer neural network model [94] that can train each layer of a deep neural network by predicting the next input of a set of multi-layer recurrent neural networks using unsupervised pre-training. In the model, each layer of neural networks stands for a compressed representation of the

observed variable, which is then passed on to the next layer basically without any information going missing before the back-propagation algorithm is used for fine-tuning. Later on, Schmidhuber referred to this multi-layer neural network model as a deep learning system [95]. As might be expected, these learning methods, while being effective, were far from practical due to the limitations of computing power in those days.

Still, in the past decade, Hinton and other researchers have made some fundamental breakthroughs conceptually. In a paper published in 2007, Hinton proposed a deep neural network model that supports unsupervised learning [96]: firstly, original features such as edges of images and the smallest speech units are used for effective training in the first-layer feed-forward neural network to avoid manually labeling and classifying data; secondly, each layer of the network is seen as an unsupervised restricted Boltzmann machine which decomposes the overall learning task into multiple simpler ones, so that features accurately recognized in the previous layer can be inputted into the next layer for it to recognize more complicated features—this process is repeated in continuous layers until the system can recognize objects in a reliably way; thirdly, supervised back-propagation algorithm is adopted for fine-tuning, with a view to improving the generative or discriminative abilities of the composite model. With this technology, learning accuracy goes up as layers increase, greatly promoting unsupervised learning and enabling machines to "learn on their own". Breakthroughs in deep learning have enormously raised the technological level of AI. In the meantime, the emergence of high-performance graphics processors also significantly sped up numerical calculation and matrix operation in deep learning and slashed the running time of deep learning algorithms, which immensely opened up the market for deep learning application and brought deep learning back to the attention of people.

Over the last eight years, deep learning has been progressing at a lunatic pace and established a firm foothold in the AI sector.

In 2012, Alex Krizhevsky and Ilya Sutskever, along with their doctoral advisor Hinton, published a paper titled "ImageNet Classification with Deep Convolutional Neural Networks" [97] that has since been cited by others for 54,579 times.[1] The ImageNet dataset is widely applied to deep learning of images, based on which a lot of research about image classification, location and detection is conducted [98]. The paper introduces a large deep convolutional neural network called AlexNet, which, developed by the deep learning technology, consists of five convolutional layers, one max-pooling layer and two fully-connected layers with 60 million parameters and 500,000 neurons, aiming to classify 1.2 million high-resolution images into 1000 different classes. Since AlexNet proves that deep neural networks trained on GPU can take image recognition to a new level, it is widely regarded in the industry as truly groundbreaking. AlexNet competed in the ImageNet Large Scale Visual Recognition Challenge and won the championship [99], exerting a major impact on the development of neural networks. Almost all subsequent ImageNet champions

[1] The result was released by Google Scholar on January 16, 2020 (as was all data of this kind in this chapter).

used the convolutional neural network as the core model of image classification, and thus began a new wave of deep learning.

In 2013, Tomas Mikolov et al. published a paper titled "Distributed Representations of Words and Phrases and their Compositionality" [100] that has since been cited for 17,366 times, in which the continuous Skip-gram model was extended. The Skip-gram model is an efficient method for learning high-quality vector representations of words from large amounts of unstructured text data. The extensions of the original Skip-gram model performed better in representations. Through subsampling of frequent words and measuring of phrases, the model showed a significant speedup, resulting in faster training and better vector representations, and it was even possible to learn good vector representations for millions of phrases. Thanks to the efforts of Mikolov and his fellow researchers, deep learning took another giant step forward in its ability to process texts.

In 2014, Ian J. Goodfellow et al. released a paper titled "Generative Adversarial Nets" [101] that has since been cited for 15,001 times. In the paper, a new framework called Generative Adversarial Nets (GAN) was proposed for estimating generative models, in which two models were simultaneously trained: a generator that captured the data distribution, and a discriminator that estimated the probability that a sample came from the training data rather than the generator. Through a two-player game, GAN was able to perform excellent modeling of data distribution. The purpose of the generator was to try to reduce the probability for the discriminator to correctly identify fake samples, while the discriminator aimed to minimize the possibility of making a mistake during the identification of fake samples. GAN could train the generator to maximize the probability of recognition errors by the discriminator, which alleviated the gradient saturation and improved the training stability, producing an effect so stunning that GAN became an instant hit. Today it is a training algorithm preferred by people in machine learning and pattern recognition.

In their paper "Adam: A Method for Stochastic Optimization" [102] (having been cited for 36,353 times) published in 2015, Diederik P. Kingma et al. introduced *Adam*, an algorithm for first-order gradient-based optimization of stochastic objective functions based on adaptive estimates of lower-order moments. According to the paper, the method "is straightforward to implement, is computationally efficient, has little memory requirements, is invariant to diagonal rescaling of the gradients, and is well suited for problems that are large in terms of data and/or parameters and for non-stationary objectives and problems with very noisy and/or sparse gradients". As an adaptive optimizer that could adjust the learning rate for each parameter individually, Adam caught on very quickly. Its appearance promoted the effective use of deep learning and has established itself as one of the most popular optimization algorithms in deep learning nowadays.

In 2016, Kaiming He et al. published "Deep Residual Learning for Image Recognition" [103] (having been cited for 37,225 times). For ordinary neural networks, the deeper they get (with more layers), the more difficult it becomes to train them, for the accuracy rate may drop as the depth increases. To find a way out of this dilemma, Kaiming He and his fellow colleagues proposed a residual learning framework to ease the training of networks that were substantially deeper than those used

previously. To be specific, they reformulated the layers as learning residual functions with reference to the layer inputs. In other words, residual nets were constructed by passing the input into each chunk straight through to the next chunk, along with the residual output of the chunk minus the input to the chunk that was reintroduced. Empirical evidence showed that these residual nets were easier to optimize and could gain accuracy from considerably increased depth. On the ImageNet dataset, it was evaluated that residual nets still had lower complexity when their layers went up to as many as 152—eight times deeper than VGG nets, with an error rate of merely 3.57%. This result won the first place in the ILSVRC 2015 classification task. At the outset, deep residual learning was designed to solve the vanishing gradient/exploding gradient problem pertaining to deep convolutional neural networks (CNNs). At present, however, residual chunks have become a basic component of almost all CNNs. Residual networks address the concern that decomposing a neural network into more layers will lead to its degradation, because in the worst-case scenario, residual networks can simply set other layers as identity mappings, thus merely reducing it to just a shallow network.

In 2017, Ashish Vaswani et al. published a paper titled "Attention Is All You Need" [104] (having been cited for 5518 times). People engaged in sequence transduction for translation often employ complex recurrent or convolutional neural networks in encoder-decoder architectures and, for better performance, connect the encoder and decoder through an attention mechanism. The paper proposed a new simple network architecture, the Transformer, a model architecture dispensing with recurrence and convolutions entirely and instead relying solely on an attention mechanism to draw global dependencies between input and output. The Transformer allowed for significantly more parallelization and could reach a new state of the art in translation quality after being trained for as little as twelve hours on eight P100 GPUs. Experiments on two machine translation tasks showed the model to be superior in quality while being more parallelizable and requiring significantly less time to train. The Transformer, to a great extent, outperformed recurrent neural networks (RNNs), because it required less computational cost. Long propagation gradients in RNNs can make memory access a lot harder, and that is why more complicated long short-term memory (LSTM) networks and gated recurrent units are called for. In comparison, the Transformer has been optimized for highly parallel processing, where the highest computational cost is generated by the feed-forward network behind the attention layer, which can be applied in parallel. What's more, the attention layer in essence is a large matrix multiplication that can be optimized easily as well. The Transformer uses only attention mechanisms and eschews convolutions, a minimalist approach that brings the attention-based sequential neural model to a climax and now it constitutes the basis of nearly all the latest models of natural language processing.

In 2018, Jacob Devlin et al. published a paper titled "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding" [105] that has since been cited for 3288 times by others. The major limitation of existing translation techniques is that standard language models are unidirectional, which limits the choice of architectures that can be used during pre-training and restricts the power of the pre-trained representations. This paper proposed a new language representation model BERT

(Bidirectional Encoder Representation from Transformers) that was designed to pre-train deep bidirectional representations from unlabeled text by jointly conditioning on both left and right context in all layers. As a result, the pre-trained BERT model can be fine-tuned with just one additional output layer to create state-of-the-art models for a wide range of tasks, such as question answering and language inference, without substantial task-specific architecture modifications. BERT is conceptually simple and empirically powerful. It has obtained new state-of-the-art results on eleven NLP tasks, including pushing the GLUE score to 80.5% (7.7% absolute improvement), MultiNLI (Multi-genre Natural Language Inference) accuracy to 86.7% (4.6% absolute improvement), SQuAD (StandFord Question Answering Dataset) v1.1 question answering Test F1 to 93.2 (1.5 point absolute improvement) and SQuAD v2.0 Test F1 to 83.1 (5.1 point absolute improvement). It deserves a special mention here that the result of SQuAD v1.1 question answering Test F1 by mankind on August 28, 2016 stood at 91.221, more than 2 points less than that of BERT, compared to 88.977, the result of SQuAD v2.0 question answering Test F1 on June 9, 2018 by mankind [106]. BERT is considered as a milestone for NLP [107], bringing about a change in the NLP paradigm. With a status similar to that of ResNet in image processing, BERT will, as likely as not, be necessary for initialization in most future research efforts, just as naturally as what people used to do with word2vec (a representation method of word vectors).

In 2019, Tero Karras et al. published a paper titled "A Style-Based Generator Architecture for Generative Adversarial Networks" [108] that has been cited for 443 times by others in the year that follows. Borrowing from style transfer literature, the paper proposed an alternative generator architecture for generative adversarial networks—StyleGAN, which is capable of an automatically learned, unsupervised separation of high-level attributes (e.g., pose and identity when trained on human faces) and stochastic variation in the generated images (e.g., freckles, hair), as well as intuitive, scale-specific control of the synthesis. The new generator leads to demonstrably better interpolation properties, and also better disentangles the latent factors of variation. Prior to StyleGAN, GAN-related research had been mired in lots of dilemmas, including the difficulty with conditional generation, limited gains from simply increasing the model size, and the inability to generate lifelike high-resolution images, to name but a few. Thanks to StyleGAN, these problems have been solved, bringing significant advances in several areas, such as high resolution and high definition. Intrigued by what StyleGAN is capable of, many people have created their own implementation models and given others free access to their dynamically generated images,[2] including human faces,[3] cats,[4] snacks,[5] vessels,[6] anime girls,[7] houses,[8]

[2] https://thisxdoesnotexist.com/.

[3] https://thispersondoesnotexist.com/.

[4] https://thiscatdoesnotexist.com/.

[5] http://thissnackdoesnotexist.com/.

[6] https://thisvesseldoesnotexist.com/.

[7] https://www.thiswaifudoesnotexist.net/.

[8] https://thisrentaldoesnotexist.com/.

resumes,[9] lyrics[10] and so on. Later, Karras et al. went on to publish a paper about StyleGAN2 [109] to correct defects in StyleGAN-generated images and improve the consistency of elements in images, thus bringing the quality of image generation to a new height and adding greatly to the charm of deep learning.

Of course, deep neural networks are not perfect. If used for simple training, they may run into quite a few problems, including but not limited to overfitting, lack of reliability and non-explainability. In terms of overfitting, it may be due to the fact that the added abstraction layer enables the modeling of rarer dependencies in the training data, which affects the universality of the model. As for the lack of reliability, the actual effect of the deep learning model will be significantly reduced when it leaves the scenario data used in training. In practical application, training data and actual application data are different, so the effectiveness of the trained model will see a marked decrease when it is used to process unlearned data. In regard to non-explainability, the computational process of deep learning can be viewed as a black box operation, where the model calculation, feature selection, etc. are performed by algorithms on their own, and there is yet to be a mature theory in place that can provide a reasonable explanation for it. As a growing number of AI algorithms are applied in real-life production, non-explainability becomes a hidden danger, for it may lead to uncontrollable results.

1.4.2 Software Frameworks: A Boost to AI Application

Currently, fundamental AI algorithms have mellowed a lot. In response, many major manufacturers are rushing to build algorithm model libraries. They have engineered and packaged algorithms into software frameworks, and made them available to developers as open sources for them to develop corresponding AI applications quickly and conveniently. In turn, these highly sought-after open-source deep learning software frameworks have become the core of building, developing and using AI ecology, and put open-source enterprises on the map as AI magnates in the international arena. For the time being, AI software frameworks are divided into two main categories: deep learning training software frameworks and inferential software frameworks.

However, the quality of inference software frameworks for the implementation of AI applications varies widely, greatly holding back the business development. Given multifarious scenarios and different characteristics of AI application, open-source inference software frameworks used to achieve the final implementation still have much catching up to do in terms of both functionality and performance in order to meet actual needs.

[9] https://thisresumedoesnotexist.com/.

[10] https://theselyricsdonotexist.com/.

1.4.2.1 Training Frameworks Based on Deep Learning

Deep learning-based training frameworks mainly aim to solve problems concerning the reading, processing and training of large amounts of data, with a focus on the implementation of massive training models. Currently, mainstream deep learning training software frameworks include TensorFlow [110] supported by the Google Brain team, MXNet [111] dominated by Amazon, Caffe2 + PyTorch [112] of Facebook, CNTK [113] of Microsoft, PaddlePaddle [114] of Baidu, and so on.

TensorFlow is an end-to-end open-source platform for machine learning known for its comprehensive functions, high compatibility, and full-fledged ecology. With model inspection, visualization, serialization, and other supporting modules, it can run deep learning models on multiple GPUs and provide programs for using data pipelines. TensorFlow has a comprehensive, flexible ecosystem of tools, libraries, and community resources. As the largest active community of open-source frameworks for deep learning, it helps developers easily build and deploy AI applications powered by machine learning. Basically, TensorFlow has the following three features:

(1) It can easily build models. TensorFlow provides multiple levels of abstraction for developers to freely choose from in accordance with their specific needs and even use the high-level Keras API to build and train models. In this way, developers can have easy access to TensorFlow and engage in machine learning for specific purposes anytime, anywhere. Meanwhile, with the help of Eager Execution, TensorFlow can implement immediate iteration and intuitive debugging. For large-scale machine learning training, developers can directly use the Distribution Strategy API for distributed training on different hardware configurations without changing model definitions.
(2) TensorFlow can directly provide specific applications. Whether it is on servers, mobile and edge devices, or even on a network, TensorFlow can go straight into the state of training and deploying models regardless of the language or platform used.
(3) TensorFlow can build and train advanced models without sacrificing the speed or performance. Through Keras Functional API, Model Subclassing API and the like, it gives developers the flexibility to create complex topologies and implement correlation control. Moreover, it also supports a powerful ecosystem of add-on libraries and models for developers to experiment with.

MXNet sees itself as a true open-source deep learning framework that can rely on flexible libraries for rapid development of deep learning projects, and support machine learning applications with its powerful framework. Having been donated to the Apache Open-Source Software Foundation, it is known for its excellent performance and comprehensive platform support. MXNet has three major features as follows:

(1) It has a flexible programming model that can be used to support the imperative programming model via the Gluon Python API, or simply switched to the symbolic programming model by calling the hybridize function, which leads to faster, more optimized execution.

(2) It supports distributed training and can be run in parallel environments (such as multiple GPUs and multiple hosts). Besides, with near-linear parallel efficiency, it can fully utilize the scale of the computer cluster; MXNet also introduces support for Horovod (a distributed learning framework developed by Uber).
(3) It provides eight language interfaces, deep integration into Python, and support for Scala, Julia, Clojure, Java, C++, R and Perl. Combined with the hybridization feature, this allows a very smooth transition from Python training to deployment in the language of users' choice to shorten the time to production.

Caffe2 is an open-source machine learning framework released by Facebook. In April 2018, it was merged into PyTorch [115], a framework of its kind also developed by Facebook. The combination has brought a momentous change to the domain of deep learning frameworks [116]. PyTorch is known for its depth and ease of use in image processing, with the following four main features:

(1) PyTorch Script: TorchScript provides a seamless transition between eager mode and graph mode to accelerate the path to production.
(2) Distributed Training: The Torch distributed backend enables scalable distributed training and performance optimization in research and production.
(3) Robust Ecosystem: A rich ecosystem of tools and libraries extends PyTorch and supports development in computer vision, NLP and more.
(4) Cloud Support: PyTorch is well supported on major cloud platforms, providing frictionless development and easy scaling.

Microsoft Cognitive Toolkit (CNTK) is an open-source toolkit for commercial-grade distributed deep learning. It describes neural networks as a series of computational steps via a directed graph. Widely applied in speech recognition, machine translation, image recognition, language modeling, etc., it is famous for its strengths and good performance in the field of intelligent speech and semantics. Through CNTK, users can easily implement and combine popular AI models such as feed-forward deep neural networks (DNNs), convolutional neural networks (CNNs) for image processing and recurrent neural networks (RNN/LSTM) for natural language processing. CNTK implements stochastic gradient descent (SGD) and back-propagation learning across multiple GPUs and servers.

PaddlePaddle is an open-source deep learning platform derived from industrial practices, known for its ease of use and support for industrial applications. It provides a cost-effective multi-CPU parameter server solution that effectively addresses problems related to ultra-large-scale recommendation systems, mega data, massive features of self-expansion and high-frequency model iterations, in a bid to achieve high throughput and speedup ratio. PaddlePaddle has three key features:

(1) It provides users with both dynamic and static graphs, balancing flexibility and high performance.
(2) The best-performing official model is provided based on actual business. PaddlePaddle offers more than 80 official models, including a natural language processing model that "understands Chinese better" and open-source algorithms that have won multiple international competitions in the field of vision.

(3) PaddlePaddle is derived from industrial practices and can export ultra-large-scale parallel deep learning platforms. It supports ultra-large-scale parallel training of deep learning in both dense and sparse parameters, enables effective parallel training that involves trillions of parameters and hundreds of nodes, and provides powerful deep learning parallel technology.

Since the maintenance capacity and development ideas of different software frameworks vary a lot and since there is a shortage of personnel willing to step up and contribute, the development of software frameworks trails a bit behind, beset with such problems as belated expansion of algorithm libraries, low API level and the failure to support distributed tasks.

1.4.2.2 Deep Learning Inference Software Framework

Deep learning-based inference only requires data inputted into the neural network to perform forward computation without needing to correct connection weights through the back-propagation process. Thus, it entails much less computation than the training process does. However, the inference process still involves a large amount of matrix convolution, non-linear transformation and other operations. Since the inference process does not necessitate extensive training, many terminal devices (like cameras) can run an inference program to make better judgments. Considering the generally less-than-desirable performance of terminal-side equipment and power limitations of some devices, the industry has developed a number of open-source terminal-side inference software frameworks to keep up with the computing power of terminal-side devices, including Caffe2go [117] of Facebook, TensorFlow Lite [118] of Google, NCNN [119] of Tencent XLab, Core ML [120] of Apple, Paddle-Mobile [121] of Baidu, TensorRT [122] of NVIDIA and so on.

Built on the open-source project Caffe2, Caffe2go is a light-weight, modular terminal-side inference framework and the first terminal-side deep learning platform ever. As speed is at the core of computation-intensive mobile applications, the lightweight design of Caffe2go allows users to perform platform-specific optimizations for defined operators, enabling deep neural networks to run efficiently on mobile phones. Facebook has embedded the framework into its creative-effect camera app that helps people turn videos into works of art in the moment. This technique, called "style transfer", condenses the size of the AI model used to process images and videos by 100 times, which provides AI inference on some mobile phones at less than 1/20th of a second (a human eye blink happens at 1/3rd of a second), making it perfectly suited for mobile applications.

TensorFlow Lite is another open-source deep learning framework for on-device inference. It can convert TensorFlow models and offers all tools necessary for running them on mobile, embedded and IoT devices. TensorFlow Lite runs on Android and iOS platforms. When it is combined with the Android ecosystem (including mature libraries, frameworks, etc.) on Android, neural networks can make faster AI inferences when running.

NCNN is a high-performance neural network computing framework optimized for mobile platforms without third-party dependencies. It is cross-platform and enables adaptive applications for different mobile CPUs. Based on NCNN, developers can easily deploy deep learning algorithm models to the mobile platform to create AI applications.

Core ML is an AI software framework developed by Apple for the iOS platform, which aims to integrate machine learning models into mobile applications. Providing a unified representation for all models, it can interface with a majority of AI frameworks and models, such as Caffe, PyTorch, MXNet and TensorFlow, and allows mobile applications to make predictions and fine-tune models through Core ML APIs.

As a mobile deep learning framework developed by Baidu that supports Linux-ARM, IOS, Android, DuerOS and other major platforms, Paddle-Mobile provides APIs for neural network inference in its top layer to serve a wide assortment of applications of Baidu. Its base layer is optimized for various hardware platforms, including CPU (mainly ARM CPU for mobiles), GPU (including Mali of ARM, Andreno of Qualcomm, Apple's self-developed GPU, etc.), in addition to NPU, powerVR, FPGA and other platforms of Huawei, maintaining a good many advantages, such as high performance and small size.

NVIDIA TensorRT delivers low latency and high-throughput for deep learning inference applications. As a high-performance AI framework, it supports mainstream deep learning libraries like Caffe, Caffe2, TensorFlow, MXNet, PyTorch, etc. TensorRT-based applications perform up to 40 times faster than CPU-only platforms during inference. Meanwhile, TensorRT provides optimizations for deep learning applications such as video streaming, speech recognition and natural language processing to reduce application latency. With TensorRT, developers can focus on creating novel AI-powered applications rather than performance tuning for inference deployment.

1.4.3 Solutions for Hardware and Software Adaptation Problems

In practice, developers often have a variety of software frameworks to choose from when developing AI applications, all of which can lead to satisfactory results. When it comes to hardware, choices abound as well. What ensue, however, are some inevitable problems. Above all, the applicability of different software frameworks poses quite a challenge. Since the underlying implementation of software frameworks is different, it is no easy to quickly convert AI models developed under them. For instance, if a face recognition neural network is developed through TensorFlow and out-sourced, loyal fans of other frameworks (such as PyTorch and PaddlePaddle) will have to refactor the out-sourced model first before getting to run it successfully in a new software

framework. The second problem is about the compatibility between software frameworks and hardware. Nowadays, computing chips develop so rapidly that software frameworks may have trouble keeping up, resulting in a mismatch between underlying chips and software frameworks, hence some inexplicable errors that arise in AI algorithms. Some companies in the industry have also proposed compilers for deep neural network models in an effort to solve the problem of adaptability and portability of deep learning between different software frameworks and hardware devices by expanding proprietary functions for deep learning models. However, there is no unified standard for such compilers for the time being, and the competition among hardware manufacturers concerning the intermediate representation layer is also a drag on the practical application of AI technology. As a result, it will be a crucial step that may make a difference to the rapid application and popularization of AI to figure out how to form a unified standard in terms of model base, hardware, storage, computation, optimization, etc.

1.4.3.1 Intermediate Representation Layer for the Portability Problem

In specific engineering practices and applications, AI models, after being trained, will be expressed and stored as specified by the intermediate representation (IR) layer. The IR layer can be regarded as a collection of many middleware components that can optimize the performance of programs and improve the efficiency of communication. As deep learning training software frameworks are implemented in different ways, the IP layer can be extended to bridge the gap between various software frameworks and expression patterns so that deep learning network model compilers can work with greater efficiency. With a piece of proprietary middleware newly added to it, a deep learning network model compiler can achieve terminal-side adaptation and run on diverse hardware and software platforms. At present, frameworks that support the IR layer include NNVM/TVM [123] of Amazon Web Services and TensorFlow XLA [124] of Google.

In September 2017, the Amazon Cloud Services (AWS) AI team of Paul G. Allen School of Computer Science and Engineering at the University of Washington released the TVM stack in the DMLC open source community, with a view to bridging the vast chasm between the productivity-focused deep learning frameworks and the performance- and efficiency-focused hardware backends. TVM works with deep learning frameworks to provide end-to-end compilation to different backends. In October of the same year, the University of Wisconsin, Allen and the AWS AI team officially released the NNVM compiler. As an open deep learning compiler, the NNVM compiler can compile front-end workloads directly to hardware backends, thus providing a unified solution for all frameworks. With the help of TVM stack, NNVM compiler can represent and optimize the common deep learning workloads in high level graph IR; transform the computation graph to minimize memory utilization, optimize data layout and fuse computation patterns for different hardware backends; and present an end-to-end compilation pipeline from front-end deep learning frameworks to bare metal hardware.

The Accelerated Linear Algebra (XLA) is a domain-specific compiler for linear algebra [125] developed by Google to optimize the computations of TensorFlow. It can improve the speed, memory usage and even portability of servers and mobile platforms. TensorFlow is a flexible and scalable deep learning framework that defines arbitrary data flow diagrams and executes them efficiently in a distributed fashion using heterogeneous computing devices (e.g. CPU and GPU). Nevertheless, the flexibility of TensorFlow is not commensurate with its target performance. Although TensorFlow is designed to support the definition of any type of data flow diagrams, still it is a big challenge to get all diagrams to execute efficiently. When the weights between operations are different, there is no guarantee that such a combination will work in the most efficient way. For this reason, Google's internal XLA team, in collaboration with the TensorFlow team, launched XLA-TensorFlow in March 2017, using XLA as the compiler for TensorFlow. XLA uses JIT compilation techniques to analyze the TensorFlow graph created by the user at runtime, fuse multiple operations together and generate efficient native machine code for devices like CPUs, GPUs and custom accelerators (e.g. Google's TPU).

1.4.3.2 Model Conversion and Exchange Formats

To solve the adaptation problem of different hardware and software, it is necessary to deal with the format of input data while using a unified IR layer to express and store models. Different software frameworks have their own distinctive ways of implementation, and the formats of input data defined by them vary, as do techniques adopted for data manipulation. For example, TensorFlow defines TFRecord [126], a simple format for storing binary record sequences, serializing and storing data in a set of linearly readable files (100–200 MB each), which supports efficient data reading and also facilitates the pre-processing of cached data. Whereas in comparison, MXNet and PaddlePaddle frameworks turn to RecordIO [127], a set of binary data exchange formats whose basic idea is to divide the data into individual chunks (a.k.a. "records"), and then to prepend to every record its length in bytes, followed by the data. Since there is no formal specification of the RecordIO format, there tend to be slight incompatibilities between RecordIO implementations.

There are two main types of standards for representing deep learning models. One is Open Neural Network Exchange (ONNX) [128] by a community project co-launched by Facebook and Microsoft, and the other is Neural Network Exchange Format (NNEF) [129] by the Khronos Group, an open industry consortium.

ONNX is an open format for representing deep learning models introduced by a community jointly created by Facebook and Microsoft, which allows models to be transferred between different software frameworks. It supports software frameworks including Caffe2, PyTorch, Cognitive Toolkit, MXNet, etc., while Google's TensorFlow is not included. ONNX has the following two main features:

(1) Interoperability: The interoperability of frameworks speeds up the implementation of designs. ONNX allows models to be trained in one framework and

transferred to another for inference. ONNX models are currently supported in Caffe2, Microsoft Cognitive Toolkit, MXNet and PyTorch, and there are connectors for many other common frameworks and libraries.

(2) Hardware optimization: Hardware optimization makes it easier for ONNX to reach more developers. Any tool that exports ONNX models can benefit from ONNX-compatible runtimes and libraries designed to maximize the performance of hardware in the field of deep learning. In July 2019, ONNX v1.6 was officially released, which adopts new data types and operators to support a wider range of models.

NNEF reduces machine learning deployment fragmentation by enabling a rich mix of neural network training tools and inference engines (and their combinations) to be used by applications across a diverse range of devices and platforms. It supports model format conversion for most AI software frameworks, including Torch, Caffe, TensorFlow, etc. Currently some 30 computing chip enterprises have got involved. The goal of NNEF is to allow users to easily transfer trained neural networks to a wide array of inference engines. As neural networks are widely used in edge devices, manufacturers' reliance on NNEF is critical. Consequently, NNEF encapsulates a complete description of the structure, operations and parameters of a trained neural network, independent of the training tools used to produce it and the inference engine used to execute it.

1.4.3.3 Deep Learning Network Model Compiler for the Adaptability Problem

Traditional compilers have a lot to make up for when it comes to the optimization of basic operators (convolution, fully connected calculations, residual networks, etc.) of deep learning algorithms, and they are not adapted to multiple forms of computing chips. To improve the efficiency of deep learning, it is imperative to design a proprietary compiler for AI underlying computing chips and upper software frameworks. Presently, the industry approaches this problem mainly by upgrading the traditional compiler architecture. Most deep learning network model compilers in the industry today are designed in accordance with the LLVM architecture, an open-source compiler initiated by the University of Illinois. The LLVM project is a collection of modular and reusable compiler and toolchain technologies. Some people may be lured into thinking that LLVM is short for "Low Level Virtual Machine". The LLVM management team, therefore, went out of its way to announce that "despite its name, LLVM has little to do with traditional virtual machines. The name 'LLVM' itself is not an acronym; it is the full name of the project" [130]. At present, there are a large number of tools based on the LLVM software framework in use, forming an ecosystem of practical standards. Current mainstream compilers in the industry mainly include NVCC [131] (NVIDIA), nGraph [132] and NNVM compiler developed by Intel Corporation, and so on.

Compute Unified Device Architecture (CUDA) is a parallel computing platform and programming model created by NVIDIA. It allows a broad range of routine computations on GPUs. NVIDIA's NVCC (CUDA Compiler) is based on the infrastructure of the LLVM open-source compiler, which supports GPU acceleration and provides great convenience for developers. By creating a language specific frontend, developers can compile that language into the IR layer used by LLVM, add support for GPU acceleration and then optimize the IR layer generated by the frontend to run the program on different target devices. Meanwhile, developers can also develop a processor-specific backend to add support for new devices that will perform final compilation on the optimized IR layer of LLVM. NVIDIA has partnered with the LLVM organization to contribute the CUDA compiler source code changes to the LLVM core and parallel thread execution backend, enabling full support of NVIDIA GPUs.

nGraph is a framework-neutral DNN model compiler that can target a variety of devices. With it, developers don't have to worry about how to adapt their DNN models to train and run efficiently on different devices. nGraph directly supports TensorFlow, MXNe and neon, while CNTK, PyTorch, and Caffe2 are supported indirectly through ONNX. Users can run these frameworks on several devices: Intel Architecture, GPU, and Intel Nervana Neural Network Processor (NNP). By installing the nGraph library, users can write or compile frameworks to run training and inference models.

1.4.4 Significant Improvement of Computing Power and Its Contributions

The rapid development of AI technology also has everything to do with the sharp increase in computing power. In recent years, many new high-performance computing architectures keep mushrooming, which also provides a catalyst for the development of AI technology. Deep learning algorithms require many adjustments to network connection weights, which calls for enormous computing power. The rapidly declining cost of computing due to the development of integrated circuits according to Moore's Law as well as the emergence of cloud computing and the large-scale adoption of GPUs have all made computing power for centralized data more powerful than ever before. At the same time, the architecture of computing chips is gradually moving towards optimization for deep learning applications. For example, the traditional CPU-based, GPU-supported Intel Corporation began to join forces with AMD to launch the Intel i7-8809G CPU [133], a whole new product integrated with AMD Vega GPU. Thus, it is fair to say that the explosive growth of AI is due in no small part to the growing computing power of hardware.

1.4.4.1 The Role of Dedicated Chips

The development of AI urgently requires the upgrade of the core hardware. The basic capabilities of traditional chips cannot satisfy the needs of intensive linear algebra calculations and massive data handling. As a result, naturally the development of chips for AI has become a catalyst for the evolution of AI.

Generally, the requirements of deep neural networks for computing chips involve two aspects: communications and computing. In terms of communications, the communications of massive data between computing chips and storage devices need to be addressed. Not only do we need to have the ability to store massive data and a considerable amount of cache, but we also need to build bandwidth for handling massive interactive data between computing units and storage devices. In terms of computing, the massive computations supporting intensive linear algebra calculations (such as convolution calculations and full connection) need to be resolved, and we need to upgrade the computing speed while reducing the power consumption.

At present, general-purpose chips such as CPU and GPU are the main chips used by AI, and field programmable gate array (FPGA) and application specific integrated circuit (ASIC) also play an important role in deep learning and training; and the computing chips used by deep learning inference are mainly ASICs. Intel, Google, Nvidia and numerous start-up companies have launched dedicated integrated chips for deep learning algorithms one after another. It is promising that within the next few years, these chips will replace most general-purpose chips to serve as AI chips.

In early deep learning algorithms, CPU was used as the architecture for computing. However, as a general-purpose processor, CPU provides few computing units for floating-point calculations to satisfy the needs of deep learning that requires massive floating-point calculations, especially in training. Besides, CPU is inefficient in parallel computing, and thus it is soon substituted by GPU that features plentiful floating-point computing units and strong parallel computing capabilities. With much better performance in parallel computing than CPU, GPU has drawn a lot of attention in the AI industry. It was also proven that in terms of floating-point calculations and parallel computing, GPU could perform tens or even hundreds of times better than CPU. Therefore, currently GPU is the first choice for deep learning. GPU has mainly the following advantages: First, it has powerful parallel computing capabilities, whose high-bandwidth cache can effectively increase the communications efficiency of massive data. Since the cache of GPU can be shared, different from CPU, the data communication between GPU threads does not require access to the global memory. The data can be directly accessed in the shared memory, thereby greatly increasing the communication efficiency. Second, GPU has thousands of computing cores, handling data as much as 10–100 times that of CPU. Therefore, one important reason for the rapid growth of AI is the combination of AI with GPU. NVIDIA is arguably one of the pioneers in the development of GPU. As early as in 1999, they had utilized GPU to address users' demand for graphics processing speed in the gaming industry, which redefined modern computer graphics technology and upgraded parallel computing capabilities once and for all [132]. So far, NVIDIA is still a dominant player in graphics card performance. Among others, the NVIDIA

TITAN RTX™ is the fastest. It is powered by the award-winning Turing™ architecture, bringing 130 Tensor TFLOPs of performance, 576 tensor cores, and 24 GB of ultra-fast GDDR6 memory to computers [133].

FPGA is reconfigurable and customizable in deep learning acceleration. It has neither a pre-defined instruction set, nor a determined data bit width, so it can be used to implement highly customized application scenarios. FPGA has some downsides: Firstly, its flexibility and versatility would undermine the operating efficiency; secondly, FPGA-based applications usually have to handle a huge data throughput, so it poses high requirements for memory bandwidth and I/O interconnection bandwidth; and lastly, due to its low logic utilization, FPGA would incur large ineffective power consumption. Nevertheless, because FPGA does not require the tape-out process that is necessary to the design and manufacturing of integrated circuits, it has become the main solution to deep learning applications in the experimental stage.

Unlike FPGA, ASIC is a non-configurable, highly customized dedicated computing chip. In contrast with the flexibility of GPU and FPGA, ASIC is unchangeable once the customization is completed. For this reason, the development of ASIC is costly in early stages and takes a long period; the basic conditions for developing ASIC technology are high. However, as a dedicated computing chip, ASIC is obviously better than FPGA in performance. If economies of scale were achieved, the cost of ASIC could be reduced. The major ASIC production companies and ASIC products are Google and its TPU [134] computing chips, the Chinese company Cambricon and its smart chips [135], and so on.

Tensor Processing Unit (TPU) is an ASIC customized and developed by Google, which is used to accelerate the processing of machine learning tasks. The second-generation TPU launched by Google is Cloud TPU, which enables developers to run deep learning algorithms on TPU hardware accelerators by TensorFlow. Cloud TPU aims to furnish systems with high performance and flexibility and construct a TensorFlow computing cluster for using CPU, GPU and TPU at the same time. The resources of Cloud TPU improve the processor's computing performance that is required by deep learning algorithms for massive linear algebra calculations. Cloud TPU can significantly reduce training time and increase training efficiency in the training of large and complex neural network models.

Launched in 2016, the Cambricon-1A processor is a chip for neural network processing. It was selected into the 15 "World's Leading Internet Technology Achievements" selected by the Third World Internet Conference. In 2018, Cambricon introduced the MLU100 processor chip, which supports various deep learning algorithms and whose performance and power consumption ratio both surpass those of CPU and GPU. The second-generation cloud AI chip MLU270 launched by Cambricon in June 2019 has theoretical peak performance 4 times that of MLU100, its predecessor, in processing non-sparse deep learning models. MLU270 was selected into the "World's Leading Internet Technology Achievements" at the 6th World Internet Conference, and it provides high-energy efficiency ratio solutions for a series of areas including video analysis, voice synthesis, and AI cloud. In November 2019, Cambricon released the edge AI series products MLU220 chips. MLU220 symbolizes Cambricon's full and three-dimensional coverage in cloud, edge and end

computing. It enriches and completes the end-cloud integrated product system of Cambricon, providing the algorithmic foundation for the development of AI.

1.4.4.2 The Role of Parallel Computing

Apart from computing chips that provide algorithmic support for AI, parallel computing technology also plays an important part in AI development. The serial structure of the von Neumann architecture only allows computers to run programs serially, but AI has to process large amounts of data and the serial-structure system cannot satisfy this requirement. In recent years, the advent of cloud computing has basically solved this problem. The foundation of cloud computing is parallel computing. Large-scale parallel computing helps AI to take a leap.

GPU provides support for parallel computing, and GPU computing shows great results in deep learning. As AI applications thrive, various GPU servers come into existence and server manufacturers have launched servers designed specifically for AI that are equipped with multiple GPUs. GPU servers can be adapted to a variety of software frameworks, support different scenarios of AI models like training and inference, and can be applied in image recognition, natural language processing, audio and video analysis, and so on. For example, the dedicated all-in-one deep learning machine D1000 launched by Inspur in 2016 provides large-scale deep learning training scenarios such as assisted driving, voice recognition, face recognition, image search, video analysis, and medical diagnosis, integrates the Caffe-MPI parallel computing framework with 16 GPU parallel cards, and is equipped with rich features like cluster management and monitoring and job scheduling [136].

1.4.4.3 A New Trend of Empowering AI Development by Provision of Relevant Services

The increasing popularization of cloud computing technology is another key factor in driving and accelerating the growth of AI. Although we have greatly improved the performance of the hardware like servers and have significantly reduced their cost, we still need much more hardware and equipment in order to realize the powerful computing and storage capabilities required for AI. If enterprises build the servers only locally, they would need to purchase a lot of equipment. Then not only would the investment be huge, but also massive physical space would be required to accommodate the equipment and they need to address numerous other issues like power consumption and cooling. Through large-scale and distributed parallel computing, cloud computing can integrate computing resources in different spaces, providing a more convenient and inexpensive way for obtaining powerful computing capabilities. As a result, it triggers a new trend of getting the required resources for AI development by providing relevant services such as constructing deep learning and computing platforms and rendering face recognition and text translation services. These services also contribute to AI companies' building an AI ecosystem.

There are many difficulties for traditional businesses to build the AI platform by themselves, so some AI enterprises try to provide AI practitioners with the computing resources, platform resources, basic applications etc. they need for money. The AI services can not only effectively improve the intelligence of society, lower the cost for businesses (particularly small and medium-sized enterprises) to use AI technology and advance the integration of AI and traditional industries, but also form an important foundation for AI commercialization.

The business of AI service platforms is no longer limited to selling technology by putting it in specific products. Instead, it can be in the form of service, like API service for software and platform-based services. The software API services include computer vision, and intelligent voice interaction. Traditionally, when a company intends to develop a specific AI application such as clock in by face recognition and simultaneous interpreting by voice recognition, it shall train the face recognition model and the machine translation model and so on from scratch by itself, and this is very inefficient. Software API services can assist these processes with optical character recognition, object detection, image recognition, face detection, speech recognition, text translation, etc., thereby simplifying and reducing the company's budget for AI application development and enabling them to connect with AI. Platform-based services consist of GPU cloud computing, deep learning platforms, and Infrastructure as a Service (IaaS) and Platform as a Service (PaaS) that are similar to cloud computing. GPU cloud computing services enable users to utilize GPU virtual machine computing resources, facilitating their efforts to independently build concrete applications including graphics and image rendering, video encoding and decoding, and natural language processing.

1.4.5 Contribution from Big Data

Besides the improvement of computing capabilities, the massive data also play a role in driving the development of AI. The strategic significance of big data does not lie in possession of large amounts of data and information, but in being able to process these data and extract meaningful knowledge from them. With the support of big data, the output of AI algorithms will become more accurate as data increases. Generally, the bigger the size of data for training, the more improved the output results of the algorithm. The growth of big data technology taps into the huge potential of AI technology. The technologies and applications in the two fields tend to promote and accelerate the development of each other.

One notable research direction in AI is to process all types of data such as images, text and voice by using statistical and mathematical methods and some statistical models. The statistical models would be constantly optimized in the process of data processing, in which the massive data resources are the cornerstone for these methods to be effectively implemented. Nowadays, AI has been applied in a range of areas such as voice recognition, image processing, computer vision, and robotics. The series of achievements is attributed to the accumulation and learning of massive

data. The evolution of big data technology offers technological support to analyzing and storing vast amounts of data, and the capability of processing massive data is essential to AI development. All Internet giants, including IBM, Facebook, Twitter, Google, Alibaba, Baidu and Tencent, possess massive amounts of data, which help to better train their AI models and systems and allow them to be more intelligent. Big data technology and its massive data make an important contribution to gradually releasing the gigantic potential of AI theories, methods and technologies.

AI technology needs to be trained with great amounts of data. For example, the digital recognition neural network has to be trained using large quantities of digital images to obtain the necessary parameters. Therefore, arguably, the massive data is a booster of AI technology. At present, the total quantity of data generated around the globe is more than 20 times that of 10 years ago. Such massive data has provided sufficient material for the fast development of AI.

According to the research by China Academy of Information and Communications Technology and the Artificial Intelligence Industry Alliance [137], there are primarily the following types of participants in the production of datasets for AI: The first is academic institutions. They often automatically collect, label, and construct academic datasets and use them to conduct relevant research. Some academic institutions open their datasets for testing algorithms or organizing related competitions and work with all sectors of society to promote the development of AI. The second type of participants is the government and some governmental institutions. They open the public data as a public welfare project. For example, the government and banks publish industry data so that AI practitioners can use them. The third type is data production enterprises. They mainly produce industry data, such as the data generated by their self-developed software and hardware, and these data usually constitute the core competitiveness of the company. The fourth type is AI companies. They build datasets by themselves to promote businesses, and label them to form datasets for self-use. Or they purchase data outsourcing services provided by specialized data companies. And the last type of participants is data processing outsourcing companies. These companies specialize in selling the authority to use existing datasets for training or providing services including data acquisition, data cleaning and data labeling according to the specific needs of the user.

Nevertheless, there are also some problems facing the development of big data, such as poor data circulation, varying data quality, and lack of key datasets. First of all, at present, AI datasets are mainly held by the government and big companies, but due to factors like regulations and business interests, many data cannot circulate effectively. At the same time, so far, some valuable data are unobtainable through legal channels, like the highly valuable surveillance data and phone data. Secondly, the data quality is unstable. Data labeling is mainly done through outsourcing service, which determines the quality of the datasets and from time to time results in dataset quality issues. Thirdly, data and academic datasets in key areas are insufficient. For example, there is a serious lack of data resources in computer vision processing, natural language processing etc. When the quantity of datasets is not big enough for academic research, the quality of scientific research and prospective study would be affected.

1.4.6 Contribution from Mobile Internet and Sensors

The rapid development of mobile Internet also promotes the growth of AI. Mobile Internet makes smart devices a constant companion of humans. Through smart devices, we can collect a sufficient and complete amount of data. Compared with personal computers, smart devices are very portable. Wearable devices like smart phones and smart wristbands are used all the time by people in daily life. The data collected via these devices are complete, continuous and highly systematic. They make a good foundation for the subsequent information processing and training of AI algorithms. In addition, mobile Internet fosters different scenarios and habits of people using smart devices. For example, with mobile Internet, we tend to input messages more by voice instead of using the keyboard; and when unlocking our cell phone, we tend to unlock it more by face recognition instead of using the password or pattern. These usage scenarios and habits have largely pushed forward the evolution of core technologies like image recognition, voice recognition, and natural language processing. Due to the explosive growth of smartphones and wearable devices, sensors have taken a leap in its development in terms of both quantity and quality. The advancement of sensors, such as the increasingly matured microelectronic technologies like LIGA (lithography, electroplating, and molding), strengthens machine perception, thereby paving the way for machines to become intelligent. Using these powerful sensors, we can obtain more accurate data, such as the number of steps we have covered by walking every day and the changes of our heart rate.

Data is the cornerstone for accelerating AI development, and the core of AI is the support of data. According to the *Global Digital Reports 2019* released by We Are Social and Hootsuite, the number of global Internet users has exceeded 5.1 billion, of which mobile users account for a great proportion. This means that mobile devices will gradually become the mainstream source of data.

1.5 The Huge Impact of AI

At present, AI has been widely applied in various areas including medical equipment and diagnosis [138], industrial automation and control [139], robot planning and control [140], assisted decision-making on the battlefield [141], automatic environmental monitoring [142], intelligent instruments [143], transportation [144], financial services [145], and interactive entertainment [146]. In many fields, AI is helping humans with some jobs that originally can only be done by men, like medical diagnosis, financial investment and wealth management, and intelligent music composition and performance. AI is exerting a significant effect on human society by dint of its rapid, accurate and intelligent judgment. However, as AI is growing and improving and its application fields are expanding, some scholars predict that AI could have a huge impact on the human civilization as a whole.

1.5.1 The Impact of AI on the Economy

The impact of technological progress on the economy has been one of the focuses of economists. Historically, each major advancement in industrial technology has brought about a dramatic increase in productivity. Information technology like AI and big data can empower the social economy, assisting humans with perception, analysis, comprehension and prediction. In the era of the knowledge economy witnessing the explosion of information, information processing technology equals wealth. AI has created huge economic interests for its builder, owner and users, enabling them to accomplish tasks in a more economical and efficient way. No senior experts need to be hired for their tasks, which is a great reduction in labor expenses and other fees.

From the perspective of financial derivatives, finance-related economic behavior is usually closely associated with the speed of acquiring information and making decisions. For example, in terms of the transactions in securities, futures, foreign currencies, bonds, etc., these financial markets are exactly the result of human beings' varying ability to obtain information and make decisions. If everyone uses AI, then the whole world makes decisions instantly and probably such markets would disappear. This is similar to intelligent navigation. If everyone already knows which sections of the road are congested and which are not, the future would be exactly the opposite: There would be few vehicles on the predicted congested sections, and the drivers on the predicted sections without traffic jams would end up in a traffic jam.

Furthermore, some scholars start to meditate on the arrival of the Singularity. That is, the self-improvement of AI may enable it to overtake human intelligence soon, which brings the explosive growth of intelligence and infinite intelligence in a limited amount of time. At the same time, the arrival of the Singularity of AI may lead to the arrival of the Singularity of the economy. That is, when AI evolves so quickly that it passes a limit, the economic growth would be accelerated to an unprecedented pace when traditional economic theories become completely useless and the economic development enters a stage that is uncontrollable to humanity.

1.5.2 The Impact of AI on Society

With the development of technology, machines have gradually replaced humans in a majority of repetitive and trivial jobs or labor. While machines are benefiting humans with all kinds of welfare, people increasingly worry about whether their jobs would be taken by machines and the risk of unemployment. As history shows, each technological advancement is accompanied by a negative inhibitory effect and a positive creative effect on employment: On one hand, technological progress improves labor productivity and replaces part of the labor, thereby reducing job opportunities; on the other hand, the new technology creates a series of new job opportunities. As far as AI is concerned, the advantage is that it can ideally solve the problems brought by an aging society. However, the extensive use of AI technology would cause a large

number of laborers to lose their jobs. The laid-off workers usually lack the ability to transform themselves and get into traditional occupations which are irreplaceable by AI, or take a new job created by the AI industry. According to researchers' estimations, in the coming 20 years, 76.76% of the total Chinese employed population would be impacted by AI and the percentage for the non-agricultural population would be 65.58% [147]. The unemployed population might trigger social instability and disrupt social structures. Of course, as new interpersonal relationships are taking shape, those non-procedural cognition-related jobs can hardly be taken by AI. It is because these jobs set a higher bar for human innovation, thinking, and imagination.

1.5.3 The Impact of AI on Human Thinking

As AI technology advances, machines would become "smarter" and humans would inevitably become increasingly dependent on machines. To some extent, this would weaken humans' cognitive skills and relax their thinking. That is to say, as machines become more intelligent, the mankind is likely to turn more stupid. Just think of how we are losing our basic understanding of the routes and directions when now we generally rely on the navigation systems to drive a car. In addition, those populations who refuse to use the Internet would be marginalized in the AI era. Since AI poses higher requirements for people's education level and command of new knowledge, this gives rise to the situation that the more advanced AI technology, the more people would be marginalized. In the age of AI, it would be increasingly difficult for the "marginalized" to access convenient intelligent information services due to their lack of relevant skills, and more difficult for them to obtain the scarce service resources.

1.5.4 The Impact of AI on Education and Employment

The purpose of developing AI is to help humans become more intelligent, and education plays a key role in this process. In the era of AI, personalized and independent learning and multi-dimensional communication and collaboration will be the mainstream learning method. Students can have tailored content for learning, and through adaptive learning they can have virtual tutors, receive training accompanied by educational robots, and submerge themselves in educational scenarios based on virtual reality and augmented reality. AI technology for education aims not for replacing teachers, but for assisting teachers in making the teaching process more efficient and interesting. In this way, students are able to learn in ways suitable for them, and this increases their learning efficiency and maintains their learning interest.

1.5.5 The Impact of AI on Privacy

The virtuality of the cyberspace facilitates the collection of sensitive personal data. AI technology has a strong ability to associate information. Once the system obtains enough information about the victim through different channels, including shopping, subscriptions, travel, authentication, credit, and location, AI can easily discover the privacy of the person. Furthermore, it is difficult for people to track down how their personal data and private messages are leaked and how far the leaking goes. The strength of AI is that it can be trained with the known data to determine the unknown data. For example, supermarkets can conduct a statistical analysis of the existing consumption habits of pregnant women, and then train AI with the data on these consumption habits to recognize pregnant women. As a result, supermarkets can infer whether a customer is pregnant from her consumption behavior and deliver targeted advertising to them. But for those people who are unwilling to let others know their pregnancy, this is a violation of their privacy.

1.6 Summary

As an important branch of computer science, AI has a long and complicated history. This chapter gives an overview of the development process of AI technology, describing its three development peaks and their characteristics, and illustrating the characteristics of the three major schools on AI, i.e. symbolism, connectionism and actionism, and their differences. Now, as it is widely applied and promoted in numerous fields including medicine, agriculture, industry, finance, and communications, AI technology is undergoing its third peak. In this chapter, the contribution of deep learning, big data, improvement of computing capabilities, etc. to the emergence of the third development peak of AI is also introduced. Nevertheless, the vigorous growth of AI technology has a huge impact on human society. In the last section of this chapter, the multiple impact of AI on the economy, society, human thinking, education and employment, and privacy protection is laid out. AI has become a hot topic for discussion and drawn a lot of attention from the whole human society. With concerted efforts from everyone, AI will surely create a better future for mankind.

References

1. Turing AM. Computing machinery and intelligence. Oxford Univ Press Mind Asso. 1950;59(236):433–60.
2. McCarthy J, Minsky ML, Rochester N, et al. A proposal for the Dartmouth summer research project on artificial intelligence, August 31,1955. AI Mag. 2006;27(4):12–4 [2020–2–12]. https://doi.org/10.1609/aimag.v27i4.1904.
3. Minsky M, Papert S. Artificial intelligence progress report. Cambridge, MA: MIT Press; 1972.
4. Wang H. Proving theorems by pattern recognition I. Commun ACM. 1960;3(4):220–34.

5. Thompson K. Programming techniques: regular expression search algorithm. Commun ACM. 1968;11(6):419–22.
6. Dakin RJ. A tree-search algorithm for mixed integer programming problems. Comp J. 1965;8(3):250–5.
7. Mccarthy J, Abrahams PW, Edwards DJ, et al. LISP 1.5 programmer's manual. Cambridge, MA: MIT Press; 1962.
8. Galtung J. Institutionalized conflict resolution: a theoretical paradigm. J Peace Res. 1965;2(4):348–97.
9. Simmons R, Slocum J. Generating English discourse from semantic networks. Commun ACM. 1970;15(10):891–905.
10. Feigenbaum E. The art of artificial intelligence—themes and case studies of knowledge engineering. Redwood City, CA: Stanford University Press; 1977.
11. Schweiger J. Generating configuration expert systems from conceptual specifications of the expert knowledge. In: Proceedings of the 6th European knowledge acquisition workshop on current developments in knowledge acquisition. Springer Berlin Heidelberg; 1970.
12. Buchanan BG, Smith DH, White WC, et al. Applications of artificial intelligence for chemical inference. 22. Automatic rule formation in mass spectrometry by means of the meta-DENDRAL program. J Am Chem Soc. 1976;98(20):6168–78.
13. Myers JD, Pople HE, Miller RA. CADUCEUS: a computerized diagnostic consultation system in internal medicine. Proc Ann Symp Comp Appl Med Care. 1982(8):44.
14. Duda R, Gaschnig J, Hart P. Model design in the prospector consultant system for mineral exploration. Read Artif Intell. 1981:334–48.
15. Melle WV. MYCIN: a knowledge-based consultation program for infectious disease diagnosis. Int J Man Mach Stud. 1978;10(3):313–22.
16. Erman LD, Hayes-Roth F, Lesser VR, et al. The Hearsay-II speech-understanding system: integrating knowledge to resolve uncertainty. ACM Comput Surv. 1980;12(2):213–53.
17. National Research Council. Computer science and artificial intelligence. Washington, D.C.: National Academies Press; 1997.
18. Gu XF. A historical review and current development of artificial intelligence. Chin J Nature. 2016;38(3):157–66.
19. Moto-Oka T, Stone HS. Fifth-generation computer systems: a Japanese project. Computer. 1984;17(3):6–13.
20. Zhang Z. Japan's fifth-generation computer program announced failure. Glob Sci Technol Econ Outlook. 1992;10:21.
21. The fifth generation project in Japan. [2020–2–20]. http://www.sjsu.edu/faculty/watkins/5thgen.htm.
22. Tao WX, Zeng XY, Zhou ZY, et al. Prolog forward chaining inference engine. Microelectron Comp. 1991;8:14–6.
23. Cui ZM. Constructing imprecise forward chaining inference engine with PROLOG. Comp Technol Dev. 1992;5:24–7.
24. Chen YN. Database computer. J Comp Res Dev. 1985;2:39–45.
25. Liu SY, Li YR. Production software development tools. J Comp Res Dev. 1990;6:1–9.
26. Li QF. The data-driven "process" TOPSTAR—I dataflow computer. Robot. 1981;(6):42–44+30.
27. Shi ZX. Dataflow computer. Autom Panorama. 1992;4:10–2.
28. Hochreiter S. The vanishing gradient problem during learning recurrent neural nets and problem solutions. Int J Uncertain Fuzziness Knowl Based Syst. 1998;6(2):107–16.
29. Bench-Capon TJM, Dunne PE. Argumentation in artificial intelligence. Artif Intell. 2007;171(10):619–41.
30. Liu Z, Luo P, Wang X, et al. Deep learning face attributes in the wild. In: IEEE international conference on computer vision, Santiago, Chile, 2015.
31. Hinton GE, Osindero S, Teh Y. A fast learning algorithm for deep belief nets. Neural Comput. 2006;18:1527–54.

32. China Academy of Information and Communications Technology. 2018 World artificial intelligence industry development blue book. [2020–1–9]. http://www.caict.ac.cn/kxyj/qwfb/bps/201809/P020180918696199759142.pdf.
33. Data Research Center of China Academy of Information and Communications Technology. Global artificial intelligence industry data report. [2019–12–26]. http://m.caict.ac.cn/sytj/201905/P020190523542892859794.pd.
34. National robotics initiative (NRI). [2020–2–21]. https://www.nsf.gov/pubs/2011/nsf11553/nsf11553.htm.
35. Nearly $50 million in research funding awarded by NSF-led national robotics initiative to develop next-generation robotics. [2020–2–21]. https://www.nsf.gov/news/news_summ.jsp?cntn_id=125390.
36. A roadmap for U.S. robotics: from internet to robotics. [2020–2–21]. http://archive2.cra.org/ccc/files/docs/2013-Robotics-Roadmap.
37. The road to cutting-edge robots. [2020–2–21]. https://obamawhitehouse.archives.gov/blog/2013/03/20/road-cutting-edge-robots.
38. The BRAIN initiative. [2020–2–21]. https://braininitiative.nih.gov/.
39. Preparing for the future of artificial intelligence. [2019–10–18]. https://obamawhitehouse.archives.gov/blog/2016/05/03/preparing-future-artificial-intelligence.
40. National Science and Technology Council, Networking and Information Technology Research and Development Subcommittee. The national artificial intelligence research and development strategic plan. [2019–10–18]. https://www.nitrd.gov/news/national_ai_rd_strategic_plan.aspx.
41. Artificial intelligence, automation, and the economy. [2020–2–21]. https://obamawhitehouse.archives.gov/blog/2016/12/20/artificial-intelligence-automation-and-economy.
42. H.R.3388—self drive act. [2020–2–21]. https://www.congress.gov/bill/115th-congress/house-bill/3388.
43. AI policy principles. [2020–2–21]. https://www.itic.org/resources/AI-Policy-Principles-FullReport2.pdf.
44. Short overview of the human brain project. [2020–2–21]. https://www.humanbrainproject.eu/en/about/overview/.
45. What is SPARC? [EB/OL]. [2020–2–21]. https://www.eu-robotics.net/sparc/about/index.html.
46. Draft Report: with recommendations to the commission on civil law rules on robotics. [2020–2–21]. https://www.europarl.europa.eu/doceo/document/JURI-PR-582443_EN.pdf.
47. European civil law rules in robotics. [2020–2–21]. https://www.europarl.europa.eu/RegData/etudes/STUD/2016/571379/IPOL_STU(2016)571379_EN.pdf.
48. Artificial intelligence: opportunities and implications for the future of decision making. [2020–2–21]. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/566075/gs-16-19-artificial-intelligence-ai-report.pdf.
49. Growing the artificial intelligence industry in the UK. [2020–2–21]. https://www.gov.uk/government/publications/growing-the-artificial-intelligence-industry-in-the-uk.
50. New robot strategy—vision, strategy, action plan. [2020–2–21]. https://www.kantei.go.jp/jp/singi/keizaisaisei/pdf/robot_honbun_150210.pdf.
51. Next-generation artificial intelligence society implementation strategy. [2020–2–21]. https://www.soumu.go.jp/main_content/000491936.pdf.
52. AI strategy 2019—AI for people, industry, region, and government. [2020–2–21]. https://www.kantei.go.jp/jp/singi/tougou-innovation/pdf/aisenryaku2019.pdf.
53. Ministry of Industry and Information Technology, National Development and Reform Commission, Ministry of Finance. Development plan for the robot industry (2016−2020). [2019–12–1]. http://www.miit.gov.cn/n1146295/n1652858/n1652930/n3757018/c4746362/content.html.
54. State Council. New generation artificial intelligence development plan. [2019–12–1]. http://www.gov.cn/zhengce/content/2017-07/20/content_5211996.htm.

55. Ministry of Industry and Information Technology. Three-year action plan for promoting the development of new generation artificial intelligence industry (2018−2020). [2019–12–1]. http://www.miit.gov.cn/n1146295/n1652858/n1652930/n3757016/c5960820/part/5960845.docx.
56. Ministry of Education. Artificial intelligence innovation action plan for higher education institutions. [2019–12–1]. http://www.moe.gov.cn/srcsite/A16/s7062/201804/t20180410_332722.html.
57. Xi JP. Strengthen leadership, make plans, clarify tasks, and consolidate foundations to promote the healthy development of China's new generation of artificial intelligence. [2020–2–23]. http://cpc.people.com.cn/n1/2018/1101/c64094-30374958.html.
58. Chen CY, Shen MC. The fourth industrial revolution and the development of artificial intelligence. In: Contemporary issues in international political economy. Heidelberg: Springer; 2019. pp. 333–46.
59. What is symbolic artificial intelligence? [EB/OL]. [2020–2–23]. https://bdtechtalks.com/2019/11/18/what-is-symbolic-artificial-intelligence/.
60. Logic theorist. [2020–2–23]. https://history-computer.com/ModernComputer/Software/LogicTheorist.html.
61. Alpha-beta. [2020–2–23]. https://www.chessprogramming.org/Alpha-Beta.
62. Mccarthy J, Shannon CE. Automata studies. Princeton New Jersey: Princeton University Press; 1956.
63. John McCarthy—father of AI and Lisp—dies at 84. [2020–2–23]. https://www.wired.com/2011/10/john-mccarthy-father-of-ai-and-lisp-dies-at-84/.
64. Newell A, Simon HA. Computer science as empirical inquiry: symbols and search. Commun ACM. 1976;19(3):113–26.
65. Haugeland J. Artificial intelligence: the very idea. Cambridge, MA: MIT Press; 1985.
66. Barrow H. Chapter 5—connectionism and neural networks. In: Margaret AB, editor. Artificial intelligence handbook of perception and cognition. Salt Lake City, UT: Academic Press; 1996. pp. 135–155.
67. Mcculloch WS, Pitts W. A logical calculus of the ideas immanent in nervous activity. Bull Math Biol. 1943;52(1–2):99–115 [2020–2–25]. https://www.cs.cmu.edu/~./epxing/Class/10715/reading/McCulloch.and.Pitts.pdf.
68. Fisher LM. Marvin Minsky: 1927–2016. Commun ACM. 2016;59(4):22–4.
69. Block HD. The perceptron: a model for brain functioning. Rev Mod Phys. 1962;34(1):123–34.
70. Hopfield JJ. Neural networks and physical systems with emergent collective computational abilities. Proc National Acad Sci USA. 1982;79:2554–8 [2020–2–26]. https://www.pnas.org/content/pnas/79/8/2554.full.pdf.
71. Hopfield JJ. Neurons with graded response have collective computational properties like those of two-state neurons. Proc National Acad Sci USA. 1984;81:3088–92 [2020–2–26]. https://www.pnas.org/content/pnas/81/10/3088.full.pdf .
72. Rumelhart DE, Hinton GE, Williams RJ. Learning representations by back-propagating errors. Nature. 1986;323:533–6 [2020–2–26]. https://www.nature.com/articles/323533a0.
73. Hinton GE. Learning multiple layers of representation. Trends Cogn Sci. 2007;11(10):428–34 [2020–2–26]. https://www.cs.toronto.edu/~hinton/absps/tics.pdf.
74. Deep learning. [2020–2–26]. https://www.technologyreview.com/s/513696/deep-learning/.
75. Breakthrough technologies 2013. [2020–2–26]. https://www.technologyreview.com/lists/technologies/2013/.
76. Kang NS, Fang TJ. Research on the behaviorist approach of artificial intelligence. Patt Recogn Artif Intell. 1992;5(8):183–90.
77. Maes P, Brooks RA. Learning to coordinate behaviors. In: Proceedings of the eighth national conference on artificial intelligence. AAAI Press, 1990;2:796–802 [2020–2–26]. http://new.aaai.org/Papers/AAAI/1990/AAAI90-119.pdf.
78. Brooks RA. Intelligence without representation. Artif Intell. 1991;47(1–3):139–59 [2020–2–26]. https://www.sciencedirect.com/science/article/abs/pii/000437029190053M.

79. Brooks RA. Intelligence without reason. In: Myopoulos R, Reiter J (eds) Proceedings of the 12th international joint conference on artificial intelligence, August 1991, Sydney, Australia. San Mateo, CA: Morgan Kaufmann; 1991. pp. 569–95 [2020–2–26]. http://citeseerx.ist.psu.edu/viewdoc/download;jsessionid=0FA5356C399E6C6C431FE8A086F993D3?doi=10.1.1.108.7178&rep=rep1&type=pdf.
80. Brooks RA. A robot that walks: emergent behaviour form a carefully evolved network. Neural Comput. 1989;1(2):253–62.
81. Brooks RA. Flesh and machines: how robots will change us. Vancouver, WA: Vintage-Books; 2003.
82. John McCarthy (1927–2011). [2020–2–28]. https://web.stanford.edu/~learnest/jmc/jmc.pdf.
83. Marvin M, Seymour AP. Perceptrons. Cambridge, MA: MIT Press; 1969.
84. At present, what is the development status of various schools of artificial intelligence? [2020–2–28]. https://www.zhihu.com/question/36849284.
85. Zhou Z H, Feng J. Deep forest: Towards an alternative to deep neural networks. In: Proceedings 26th international joint conferences on artificial intelligence, 2017:3553–9 [2020–2–28]. https://doi.org/10.5555/3172077.3172386.
86. Zhou ZH, Feng J. Deep Forest. [2020–2–28]. https://arxiv.org/pdf/1702.08835v3.pdf.
87. Lyu SH, Yang L, Zhou ZH. A refined margin distribution analysis for forest representation learning. In: Advances in neural information processing systems 32, NIPS 2019:5531–41 [2020–2–28]. http://papers.nips.cc/paper/8791-a-refined-margin-distribution-analysis-for-forest-representation-learning.pdf.
88. Xu XH, Yao JS. A review of behaviorist artificial intelligence research. Control Decis. 2004;19(3):241–6 [2020–2–28]. http://kzyjc.alljournals.cn/ch/reader/create_pdf.aspx?file_no=2004-3-1&year_id=2004&quarter_id=3&falg=1.
89. Boden MA (ed). The philosophy of artificial intelligence. Oxford: Oxford University Press, 1990.
90. Werbos PJ. Beyond regression: new tools for prediction and analysis in the behavioral sciences. Cambridge, MA: Harvard University; 1974.
91. Weng JY, Ahuja N, Huang TS. Cresceptron: a self-organizing neural network which grows adaptively. In: International joint conference on neural networks, Baltimore, Maryland, IJCNN vol I, 1992; 576–81.
92. Schmidhuber J. Learning complex, extended sequences using the principle of history compression. Neural Comput. 1992;4(2):234–42.
93. Schmidhuber J. My first deep learning system of 1991 + deep learning timeline 1962–2013. [2020–2–29]. https://arxiv.org/pdf/1312.5548v1.pdf.
94. Hinton GE. Learning multiple layers of representation. Trends Cogn Sci. 2017;11(10):428–34 [2020–2–29]. https://www.cs.toronto.edu/~hinton/absps/tics.pdf.
95. Krizhevsky A, Sutskever I, Hinton GE. ImageNet classification with deep convolutional neural networks. In: Advances in neural information processing systems 25, NIPS 2012:1106–14 [2020–2–29]. http://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks.pdf.
96. Collation of computer vision standard datasets—ImageNet dataset. [2020–2–29]. https://blog.csdn.net/xingwei_09/article/details/79148294.
97. AlexNet—the winner of ILSVRC 2012. [2020–2–29]. https://deeplearning.vn/post/alexnet/.
98. Mikolov T, Sutskever I, Chen K, et al. Distributed representations of words and phrases and their compositionality. In: Advances in neural information processing systems 26, NIPS 2013:3111–9 [2020–2–29]. http://papers.nips.cc/paper/5021-distributed-representations-of-words-and-phrases-and-their-compositionality.pdf.
99. Goodfellow IJ, Pouget-Abadie J, Mirza M, et al. Generative adversarial nets. In: Advances in neural information processing systems 27, NIPS 2014:2672–80 [2020–2–29]. http://papers.nips.cc/paper/5423-generative-adversarial-nets.pdf.
100. Kingma DP, Ba J. Adam: a method for stochastic optimization. In: The 3rd international conference for learning representations (ICLR 2015), San Diego, 2015 [2020–2–29]. https://arxiv.org/pdf/1412.6980.pdf.

101. He KM, Zhang XY, Ren SQ, et al. Deep residual learning for image recognition. In: 2016 IEEE conference on computer vision and pattern recognition (CVPR), IEEE Xplore, 2016:770–8 [2020–2–29]. http://openaccess.thecvf.com/content_cvpr_2016/papers/He_Deep_Residual_Learning_CVPR_2016_paper.pdf.
102. Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need. In: The 31st conference on neural information processing systems (NIPS 2017), Long Beach, CA, 2017 [2020–2–29]. http://papers.nips.cc/paper/7181-attention-is-all-you-need.pdf.
103. Devlin J, Chang M W, Lee K, et al. Bert: pre-training of deep bidirectional transformers for language understanding. https://arxiv.org/pdf/1810.04805.pdf.
104. Stanford question answering dataset (SQuAD)—dataset and leaderboard. [2020–2–29]. https://worksheets.codalab.org/worksheets/0x62eefc3e64e04430a1a24785a9293fff/.
105. Historical breakthrough in NLP! Google's BERT model breaks 11 records and outperforms humans! [2020–2–29]. https://tech.sina.cn/csj/2018-10-13/doc-ihmhafir3634167.d.html.
106. Karras T, Laine S, Aila T. A style-based generator architecture for generative adversarial networks. In: 2019 IEEE conference on computer vision and pattern recognition (CVPR) , IEEE Xplore, 2019:401–4410 [2020–2–29]. http://openaccess.thecvf.com/content_CVPR_2019/papers/Karras_A_Style-Based_Generator_Architecture_for_Generative_Adversarial_Networks_CVPR_2019_paper.pdf.
107. Karras T, Laine S, Aittala M, et al. Analyzing and improving the image quality of StyleGAN. https://arxiv.org/pdf/1912.04958.pdf.
108. End-to-end open source machine learning platform. [2020–2–29]. https://tensorflow.google.cn/.
109. A flexible and efficient library for deep learning. [2020–2–29]. https://mxnet.apache.org/.
110. PyTorch: from research to production. [2020–2–29]. https://pytorch.org/.
111. The Microsoft cognitive toolkit. [2020–2–29]. https://docs.microsoft.com/en-us/cognitive-toolkit/.
112. Open source deep learning platform derived from industrial practice. [2020–2–29]. https://www.paddlepaddle.org.cn/.
113. Caffe2 and PyTorch join forces to create a Research + Production platform PyTorch 1.0. [2020–2–29]. https://caffe2.ai/blog/.
114. All Caffe2 code merged into PyTorch: dramatic shakeup in deep learning framework landscape. [2020–2–29]. https://zhuanlan.zhihu.com/p/35209516.
115. Delivering real-time AI in the palm of your hand. [2020–3–1]. https://engineering.fb.com/android/delivering-real-time-ai-in-the-palm-of-your-hand/.
116. Deploying machine learning models on mobile and IoT devices. [2020–3–1]. https://tensorflow.google.cn/lite?hl=zh-tw.
117. Tencent Youtu first opensource deep learning framework ncnn. [2020–3–1]. https://cloud.tencent.com/developer/article/1070773.
118. Core ML: Integrate machine learning models into your app. [2020–3–1]. https://developer.apple.com/documentation/coreml.
119. Introduction to Baidu embedded deep learning framework Paddle—Mobile. [2020–3–1]http://ai.baidu.com/forum/topic/show/941097.
120. NVIDIA TensorRT: programmable inference accelerator. [2020–3–1]. https://developer.nvidia.com/tensorrt.
121. NNVM compiler: open compiler for AI frameworks. [2020–3–1]. https://tvm.apache.org/2017/10/06/nnvm-compiler-announcement.
122. XLA - TensorFlow, compiled [EB/OL]. [2020–3–1]. https://www.googblogs.com/xla-tensorflow-compiled/.
123. XLA is a compiler that can optimize TensorFlow calculations. [2020–3–1]. https://www.tensorflow.org/xla.
124. TFRecord and tf.Example. [2020–3–1]. https://tensorflow.google.cn/tutorials/load_data/tfrecord?hl=en.
125. RecordIO Data Format. [2020–3–1]. http://mesos.apache.org/documentation/latest/recordio/.

126. ONNX: open neural network exchange format, The open ecosystem for interchangeable AI models. [2020–3–1]. https://onnx.ai/.
127. Neural network exchange format (NNEF). [2020–3–1]. https://www.khronos.org/nnef.
128. The LLVM compiler infrastructure. [2020–3–1]. https://llvm.org/.
129. CUDA LLVM compiler. [2020–3–1]. https://developer.nvidia.com/cuda-llvm-compiler.
130. nGraph: a new open source compiler for deep learning systems. [2020–3–1]. https://www.intel.ai/ngraph-a-new-open-source-compiler-for-deep-learning-systems/.
131. Intel® Core™ i7-8809G processor with Radeon™ RX Vega M GH graphics. [2020–3–1]. https://ark.intel.com/content/www/us/en/ark/products/130409/intel-core-i7-8809g-processor-with-radeon-rx-vega-m-gh-graphics-8m-cache-up-to-4-20-ghz.html.
132. In 1999, NVIDIA invented the GPU. [2020–3–1]. https://www.nvidia.cn/about-nvidia/.
133. NVIDIA TITAN RTX. [2020–3–1]. https://www.nvidia.cn/deep-learning-ai/products/titan-rtx/.
134. Google Cloud TPU. [2020–3–2]. https://docs.datadoghq.com/integrations/google_cloud_tpu/.
135. Cambricon—a forerunner in the global intelligent chip field. [2020–3–2]. http://www.cambricon.com/index.php?m=content&c=index&a=lists&catid=7#nav2.
136. Deep learning all-in-one machine: D1000. [2020–3–2]. https://www.inspur.com/lcjtww/443012/444624/2290246/2290249/2290513/index.html.
137. China Academy of Information and Communications Technology, China artificial intelligence industry development alliance. White Paper on the development of artificial intelligence—technical architecture (2018). [2020–3–2]. http://www.caict.ac.cn/kxyj/qwfb/bps/201809/P020180906443463663989.pdf.
138. Chen Z, Jiang Y, Xu M, et al. The application and development of artificial intelligence in medical diagnosis systems. J Biomed Eng. 2002;19(3):505.
139. Cockburn D, Jennings NR. ARCHON: a distributed artificial intelligence system for industrial application. In: Foundations of distributed artificial intelligence, 1996.
140. Korikov A. Artificial intelligence in robot control systems. In: Materials science & engineering conference series, 2018.
141. Rasch R, Kott A, Forbus KD. AI on the battlefield: an experimental exploration (Preprint). In: Eighteenth national conference on artificial intelligence, 2002.
142. Volosencu C. Plenary lecture 16: environment monitoring based on sensor networks and artificial intelligence. In: International conference on mathematical models for engineering science, 2010.
143. Alippi C, Ferrero A, Piuri V. Artificial intelligence for instruments and measurement applications. IEEE Instrum Meas Mag. 1998;1(2):9–17.
144. Bielli M, Ambrosino G, Boero M. Artificial intelligence applications to traffic engineering. Boca Raton, FL: CRC Press; 1994.
145. Pau LF. Artificial intelligence and financial services. IEEE Trans Knowl Data Eng. 1991;3(2):137–48.
146. Amant S, Robert, Young, et al. Links: artificial intelligence and interactive entertainment. J Appl Microbiol. 2001;104(5):1449–57.
147. Chen YW. What employment impacts will artificial intelligence bring. [2020–3–2]. https://www.jiemian.com/article/1934291_qq.html.

Chapter 2
AI Development Strategies in Countries Around the World

Shudong Li and Binxing Fang

After 60 years of evolution and propelled by the new theories and technologies such as mobile Internet, big data, supercomputing, sensor networks, brain science and the considerable demands for economic and social development, AI is now booming with its new features—deep learning, cross-field integration, human–machine collaboration, collective intelligence openness and autonomous control. Big data-driven knowledge learning, cross-media collaborative processing, human–machine cooperation enhanced intelligence, integrated intelligence and autonomous intelligence systems have become the focuses of AI development. Brain-like intelligence inspired by the research outcome of brain science is maturing, and with an increasing trend of chip-, software- and platform-oriented advancement, AI ushers in a new era of development. The new generation of AI sets off on a steady march forward in terms of disciplinary development, theoretical modeling, technological innovation and software & hardware upgrades. It kindles continuous breakthroughs and accelerates the transformation of economic and social fields from digital- and network-based development to intelligence-based progress.

AI is a strategic technology that will steer human civilization towards the future. With a vision to grasp the strategic opportunities of AI development, various countries around the world see AI development as a major strategy to enhance national competitiveness and maintain national security. As a result, they have formulated a series of plans and policies, trying to gain an edge over other countries in the new round of technology competition.

This chapter focuses on the AI-related strategies, plans, laws and regulations, strategies for AI talent training of some countries as well as the formulation of AI standards in China and abroad.

S. Li (✉) · B. Fang
Guangzhou University, Guangzhou, China
e-mail: lishudong@gzhu.edu.cn

B. Fang
e-mail: fangbx@cae.cn

© Publishing House of Electronics Industry 2025

B. Fang (ed.), *Artificial Intelligence Security and Safety*,
https://doi.org/10.1007/978-981-96-9263-7_2

2.1 AI Development Strategies in the United States

The United States, the world's most developed power, takes the lead in the reserve of talents, financial systems, IT technology and Internet, etc. and therefore it seeks to maintain its leadership in AI with emphasis on how AI technology will influence its national security and leadership around the globe.

2.1.1 Functional Agencies of AI Development in the U.S.

The United States develops AI at the level of national security and concentrates on its applications in military affairs and economy with extra attention to the safety of AI technology itself. The country has set up a number of committees or technical teams to provide counsel or guidance on the national AI strategy or technology development, and established the National Security Commission on Artificial Intelligence in 2018. The major AI agencies established in the United States are shown in Table 2.1.

2.1.2 AI Development Strategies in the U.S.

The relevant departments in the United States have developed a series of strategic plans to analyze the status quo of AI, to focus on the security threats posed by AI, to explore the impact of AI on the future, to define the role of AI in national security and put forward the orientation as a priority of AI development in order to advance AI development and applications in the United States and secure the country's leadership in the field. The documents related to AI development strategies in the United States are shown in Table 2.2.

2.1.3 AI Development Plans in the U.S.

The United States has created a raft of technology plans and projects in AI policy making and implementation with the purpose of enhancing the technological strength of AI. The major technological plans deployed by the United States are shown in Table 2.3

Table 2.1 Major AI agencies established in the United States

Time	Agency or team	Duties
May 2016	Machine Learning and AI (MLAI) Subcommittee was established under the National Science and Technology Council (NSTC)	The sub-committee coordinates interagency work related to the research and development of AI. Three strategic reports were made public: *Preparing for the Future of Artificial Intelligence; National Strategic Plan for AI Research and Development; Artificial Intelligence, Automation and the Economy*
April 2017	The Department of Defense (DoD) set up the Algorithmic Warfare Cross-Functional Team (AWCFT)	*Project Maven*, a DoD effort to bring artificial intelligence and machine learning solutions to problems in the defense space, was started
May 2018	The White House Office of Science and Technology Policy (OSTP) established Special Committee on Artificial Intelligence (SCAI)	The committee coordinates Federal agencies' investment in artificial intelligence and reviews the United States' priorities and investment in AI development
May 2018	The Interagency Working Group on Networking and Information Technology Research and Development (IWG) was established during the Summit on Artificial Intelligence for American Industry	Amazon, Alphabet and more were invited to the Summit and the report—*Providing AI for American People*—was published. IWG coordinated the Federal investment in AI research and development

(continued)

Table 2.1 (continued)

Time	Agency or team	Duties
June 2018	DoD established the Joint Artificial Intelligence Center (JAIC)	JAIC integrates technology development, with the requisite policies, knowledge, processes and relationships to ensure long-term success and scalability of AI. $1.7 billion is planned to spend on AI development in five years
August 2018	The United States Congress passed the bill, *National Defense Authorization Act for Fiscal Year 2019*, and approved of the establishment of National Security Commission on Artificial Intelligence (NSCAI)	The Commission focuses on the advances in AI, related machine learning developments, and national security associated technologies. It reviews the competitiveness of the United States in AI, latest developments and trends in international cooperation, and potential manpower and educational incentives. The 15-member commission is made up of the Secretary of Defense, Secretary of Commerce and members of the Congressional Defense Committee with three major duties of inspecting the risks of military applications of AI technology and its impact on international law; examining the ethics of AI technology; establishing open criteria for data training data and promoting data sharing

Table 2.2 Major documents related to AI development strategies in the United States

Time	Agency	Document	Overview
November 2015	Center for Strategic and International Studies (CSIS)	*Defense 2045: Assessing the Future Security Environment and Implications for Defense Policymakers* [1]	It states that AI is an important factor affecting the future security environment
October 2016	NSTC affiliated to OSTP	*Preparing for the Future of Artificial Intelligence* [2]*; National Strategic Plan for AI Research and Development* [3]	The former analyzes the status quo and applications of AI and the potential public problems. The latter proposes seven strategic orientations for priority development
December 2016	Executive Office of the President (EOP)	*Artificial Intelligence, Automation, and the Economy* [4]	It gives prominence to the practice of promoting AI development and unleashing enterprises and workers' potential for creation to secure the United States' leadership in AI
Early 2017	National Science Foundation, DoD, Defense Advanced Research Projects Agency (DARPA), Air Force Office of Scientific Research, Department of Energy	*National Robotics Initiative 2.0: Ubiquitous Collaborative Robots* [5] (NRI-2.0)	NRI-2.0 was to take the place of NRI-1.0 with emphasis on developing collaborative robots that could perform duties everywhere
July 2017	Harvard Kennedy School	*Artificial Intelligence and National Security* [6]	As requested by the U.S. Intelligence Advanced Research Projects Activity (IARPA), the report presents three goals and eleven recommendations for developing future policy on AI and national security to preserve U.S. technological leadership
December 2017	Trump Administration	*National Security Strategy* [7]	The document mentions AI in the context of security strategy for the first time and states the new security threats that the United States faces
March 2018	CSIS	*A National Machine Intelligence Strategy for the United States* [8]	The report focuses on boosting the development of AI technology and the United States' applications of it

(continued)

Table 2.2 (continued)

Time	Agency	Document	Overview
April 2018	DoD	*Summary of The 2018 Department of Defense Artificial Intelligence Strategy* [9]	The gist of the document is to advance the country's security and prosperity and to lead in military ethics and AI safety. JAIC is the focal point for carrying it out
July 2018	Center for a New American Security (CNAS)	*Artificial Intelligence and International Security* [10]	The document studies the bad impact of AI reform on global security
November 2018	CSIS	*Artificial Intelligence and National Security: the Importance of the AI Ecosystem* [11]	The report introduces the status quo of AI as well as the key factors in managing and applying AI, and promotes the integration of artificial intelligence into national security
February 2019	President Donald Trump	*Executive Order on Maintaining American Leadership in Artificial Intelligence* [12]	Continued American leadership in AI is of paramount importance to the United States, for which the Federal funds and resources for AI research and development shall be mobilized together
June 2019	NSTC affiliated to OSTP	*The National Artificial Intelligence Research and Development Strategic Plan: 2019 Update* [13]	The document is an update on the 2016 plan. The update defines the priority areas for Federal investments in AI R&D; and introduces Strategy 8: expand public–private partnerships to accelerate advances in AI

Table 2.3 Major technological plans deployed by the United States

Time	Plans and dynamics	Plan overview
2012	National Science Foundation, National Institutes of Health, National Aeronautics and Space Administration and U.S. Dept. of Agriculture jointly issued *National Robotics Initiative (NRI)* [14]	In the document, the anticipated funding amount is up to $40 million to $50 million per year with the estimated number of awarded projects being 25 to 40 per year
April 2013	*Brain Initiative* (BRAIN) [15]	The initiative was carried out by National Institutes of Health (NIH), for which a special working group was established and a ten-year plan for future development was mapped out. In 2014, DARPA, NIH and NSF jointly appropriated $110 million for the initiative
August 2016	DARPA issued a program named *Explainable Artificial Intelligence (XAI)* [16]	The program was started in May 2017 and is expected to last for four years, aiming to create a suite of machine learning techniques that: produce more explainable AI systems while maintaining a high level of prediction accuracy
October 2016	*A Roadmap for US Robotics: From Internet to Robotics* [17]	Funded by NSF, the program was done by experts from University of California San Diego, Carnegie Mellon University and Clemson University. It depicts the application scenarios of robot technology
Early 2017	*National Robotics Initiative 2.0: Ubiquitous Collaborative Robots (NRI-2.0)* [18]	NSF invested up to $30 million to $45 million per year, subject to availability of funds, with the estimated number of awarded projects being 40 to 70 per year
June 2018	NIH released *Strategic Plan for Data Science* [19]	The plan aims to boost the attainment of data science technology in biomedical research and increase the capabilities of machine learning and AI
August 2018	The US Senate passed *National Defense Authorization Act for Fiscal Year 2019* [20]	The Act is for providing extra funds for AI and machine learning solutions
September 2018	DARPA announced the investment plan for '*AI Next' Program* [21]	DARPA announced the plan of investing $2 billion in the development of second wave AI technologies and funding AI research at DARPA in the following five years. The research areas included increasing the robustness of systems, reinforcing the safety of machine learning and studying the next generation of AI algorithms

2.1.4 AI Regulatory Policies in the U.S.

The United States asserts the practice of relying on market forces to reduce risks and setting norms in some pioneering areas of technology application so as to strengthen the regulation of AI security application.

(1) Reduce security risks. In May 2018, Information Technology and Innovation Foundation, the world's leading think tank for science and technology policy, published *How Policymakers Can Foster Algorithmic Accountability* [22] and proposed an accountability framework.
(2) Reinforce legislative regulation. In November 2018, *A Local Law in relation to automated decision systems used by agencies* [23] passed by New York City Council started to take effect. It is the first law for regulating AI algorithms to supervise the fairness, accountability and transparency of decision algorithms used by municipal agencies.
(3) Monitor automated driving. In July 2017, the US Senate passed *Self Drive Act* [24], which sets 12 safety standards for driverless vehicles. In October 2018, the U.S. Department of Transportation (DoT) released *Preparing for the Future of Transportation: Automated Vehicles 3.0* [25], a new Federal guidance for automated vehicles, to pave the way for the availability of hundreds of thousands of fully automated vehicles on public roads.

2.2 AI Development Strategies in EU and the UK

EU and its member states have released a series of AI strategy documents. Under the influence of the "human-centric" notion, EU sets out to "create trustworthy AI culture made in Europe" in a bid to make Europe a leading global AI innovator. EU focuses on the impact of AI and its updated industry on privacy, employment, ethical and legal issues. Germany aspires to develop intelligent manufacturing in the context of Industry 4.0 while UK leverages its strengths to uphold integrated governance and integrated development.

2.2.1 AI Development Strategies in EU and the UK

The strategies and plans released by EU and the UK indicate that, first, they concentrate on the fundamentality and innovative research on AI and second, they inject continued time and energy into policy, funding, education and international cooperation to optimize the basics and environment of AI development. Table 2.4 presents the AI strategies and plans in EU and the UK. They place great emphasis on AI applications, trying to hold the high ground of digitization and take the lead in AI development around the globe. They focus on the new codes of ethics for AI to defend

Table 2.4 AI strategies and plans in EU and the UK

Time	Agency	Document	Overview
March 2018	European Political Strategy Centre (EPSC)	*The Age of Artificial Intelligence: Towards a European Strategy for Human-Centric Machines* [26]	The document provides countermeasures against such problems as labor substitution and AI bias. It also studies the new ethics of AI for the sake of defending European values
April 2018	25 European Countries	*Declaration of Cooperation on Artificial Intelligence* [27]	Member states declared a strong will to join forces and develop AI applications so that the opportunities of digitization for Europe can be fully ensured. The Declaration is also attached with *Artificial Intelligence Bulletin*
April 2018	European Commission	*Communication Artificial Intelligence for Europe* [28]	EU plans to invest €20 billion between 2018 and 2020 to achieve three goals: to enhance EU's AI technologies and industrial capabilities and to advance AI applications; to embrace socio-economic reforms; to establish the ethical and legal frameworks
December 2018	EU Council	*Coordinated Plan on Artificial Intelligence* [29]	The topic was "AI made in Europe"
July 2018	Federal Government of Germany	*Eckpunkte der Bundesregierung für eine Strategie Künstliche Intelligenz* [30]	Germany strives to take the lead in AI research, development and application in the world
November 2018	Federal Government of Germany	*Strategy for "AI Made in Germany"* [31]	The document focuses on week AI and introduces the plan of investing €3 billion to establish 12 research centers by 2025
October 2018	The Ministry of Higher Education, Research and Innovation	*La stratégie nationale de recherche en intelligence artificielle* [32]	France aspires to build itself into one of the top five AI countries in the world and become a leader in AI research in Europe
September 2016	House of Commons	*Robotics and artificial intelligence* [33]	The document aims to standardize the robot industry and deal with the ethical, legal and social problems brought by robot development

(continued)

Table 2.4 (continued)

Time	Agency	Document	Overview
December 2016	UK Government Office for Science	*Artificial intelligence: Opportunities and implications for the future of decision making* [34]	The document focuses on the influence of AI on personal privacy and consent, employment and government decision-making and provides recommendations on the ethical and legal risks brought by AI
April 2017	UK Government	*Industrial Strategy: Artificial Intelligence Sector Deal* [35]	The UK government promised to accelerate the investment of more than £20 billion. Industrial Strategy is built on 5 foundations: ideas, people, infrastructure, business environment, places
October 2017	Confederation of British Industry	*Growing the artificial intelligence industry in the UK* [36]	The document provides recommendations in terms of data acquisition, talent training, research coordination, industry development and more
April 2018	House of Lords Artificial Intelligence Select Committee	*AI in the UK: ready, willing and able?* [37]	The report states that the UK has the chance to be a world-leader in the development of AI

their existing values, take the initiative to deal with the problems arising from AI development, enhance the technology and industrial capabilities of AI and advance AI applications to embrace the socio-economic changes.

2.2.2 AI Development Plans of EU

EU has formulated a raft of research and development plans in AI centering around the applications of intelligent artificial brain and robots in various spheres and AI application in the medical area by budgeting investments so as to secure its member states' leadership role in AI development. Part of the AI development plans in EU and its member states are shown in Table 2.5.

Table 2.5 Part of AI development plans in EU and its member states

Time	Plan	Overview
2013	EU announced the *Human Brain Project (HBP)* [38], one of four FET (Future and Emerging Technology) Flagships, the largest scientific projects ever funded by EU	As the world's most important and cutting-edge project for brain research, it is supposed to last for 10 years (2013–2023) with the total funding amounting to €256 million from EU. HBP comprises 12 sub-projects
2014	The European Commission and 180 member corporations and organizations of European Robotics initiated *SPARC: the Partnership for Robotics in Europe* [39]	The program was expected to develop robots applicable to manufacturing, agriculture, health, transportation, security and family applications. It was supposed to create 240,000 jobs in Europe and increase the annual output value of the robot industry in Europe up to €60 billion and Europe's market share in the international market to 42%
2015	German Federal Ministry for Economic Affairs and Energy launched *Smart Data—Innovations in Data* [40]	Germany invested tens of millions of euros in 13 projects, among which AI was the focus with funds totaling €55 million
September 2014	*HORIZON 2020* by EU [41]	The total budget for the program was €77.028 billion spanning 7 years (2014–2020). It was EU's largest innovative research program focusing on the difficulties in international cutting-edge and competitive technologies
December 2018	*Coordinated Plan on Artificial Intelligence* by EU [42]	The plan was to implement the *European AI Strategy 2018*. The European Commission plans to invest €100 million in 2020 first. The European Innovation Council invested €100 million between 2019 and 2020

2.2.3 Ethics of AI in EU and the UK

Both EU and the UK strive to strengthen the construction of government-led ethical principles and the constraints of laws and regulations, to formulate ethical codes and promulgate civil rules, hoping that they can become leaders in AI ethics. Likewise, *General Data Protection Regulation* (GDPR) is released to regulate laws and ethics in science and technology and to protect the digital rights of the masses.

(1) Reinforce establishment of basic codes of ethics for AI. On 16 April 2018, the British Parliament released *AI in the UK: ready, willing and able?* to announce 'five core principles' for developing AI for business. On 8 April 2019, European Commission's High-Level Expert Group on Artificial Intelligence (AI HLEG) presented *Ethics guidelines for trustworthy AI* [43] to state that "trustworthy AI" needs to be resilient and secure. AI systems need to be safe, accurate, reliable and reproducible. That is the only way to ensure that unintentional harm can be minimized and prevented.

(2) Design specifications for automated decision-making of AI application. GDPR [44] came into effect in May 2018 and allowed automated decision-making and profiling based on special categories of personal data only under specific conditions: decision-making based on the entering or performance of a contract between the data subject and a controller, or when the data subject has given his or her explicit consent, or as expressly authorized by Union or Member State law, should be allowed. Automated decision-making should not be allowed when a measure, evaluating personal aspects relating to him or her which is based solely on automated processing and which produces legal effects concerning him or her or similarly, significantly affects him or her without any human intervention.

2.2.4 AI Legislation and Regulation in EU and the UK

Regarding AI policies, legal ethics and social impact, EU is thinking about bringing in civil rules pertaining to robots and AI, and accelerating the legislation in specific areas such as accountability, ethical standards, automated driving and medical robots. One of the emphases of legislation is to define the ethics of AI.

In January 2015, the European Parliament's Legal Affairs Committee decided to establish a working group specializing in legal issues related to robots and AI development.

In May 2016, the European Parliament's Legal Affairs Committee released *Draft Report with Recommendations to the Commission on Civil Law Rules on Robotics* [45] and later, in October 2016, presented *European Civil Law Rules in Robotics* [46] to bring up "the Charter on Robotics" and the ethical principles to develop and design robotics, and to call for regulating rules specifically for autonomous robots or smart robots.

On 16 February 2017, the European Parliament voted to pass a resolution on the rules of civil law for robots and suggested devising *Civil Law Rules on Robotics* [47]

and requested the European Commission to submit legal proposals on civil liability for robots and AI.

2.3 AI Development Strategies in Other Countries

Russia and India pay much attention to the applications of AI in national defense and its impact on military security, while Japan, South Korea, Singapore and Canada place more emphasis on AI talent training, research and development and industrial upgrading.

2.3.1 AI Development Strategies in Japan

Japan develops AI in consideration of major economic and social needs, and sets up Artificial Intelligence Research Center (AIRC) which is jointly run by Ministry of Economy, Trade and Industry, Ministry of Internal Affairs and Communications and Ministry of Education, Culture, Sports, Science and Technology and has become the core department of promoting AI industry. Japan has formulated a comprehensive path of development—by defining the key points of technology, making breakthroughs in the path, arranging industrial distribution and training talents, Japan is able to boost the development of AI and robots, to achieve its goals of global base for robot innovation, world's leading society maximizing robot capacity and world's leading strategy for a new robot era, and to advance economic growth with the aid of AI and robot development. The AI strategy-related documents released by Japan are shown in Table 2.6.

2.3.2 AI Development Strategy Planning in Russia

The grave security situation at home and abroad is a major concern of Russia. The country develops AI for military purposes and places great emphasis on the application of AI in national defense and modern construction of its armies, which reflects its special "government-led, military-oriented" features. The AI strategy-related documents released by Russia are shown in Table 2.7.

In terms of legislation and regulation, Russia's Duma (the Lower House of Parliament) confirmed that the government had passed the laws related to AI and released the first legal document on AI in 2019.

Table 2.6 AI strategy-related documents released by Japan

Time	Agency	Document	Overview
February 2015	Japanese Government	*New Robot Strategy* [48]	Three pillars for the achievement of robot revolution: global base for robot innovation; world's leading society maximizing robot capacity; world's leading strategy for a new robot era
April 2016	Japanese Government	*Japan Revitalization Strategy 2016* [49]	The Japanese government summoned the Artificial Intelligence Technology Strategic Conference to push AI technology development forward by engaging all Japanese people in, training talents and exploring new areas
April 2016	Ministry of Economy, Trade and Industry	*New Industrial Structure Vision* [50]	The publication upholds the idea of boosting economic growth by AI and robots
March 2017	Japanese Government	*Artificial Intelligence Research and Development Goals & Industrial Roadmap*[51]	It depicts at length the application prospects of AI in production, medical service and mobile area
June 2017	Japanese Government	*Advancement Strategy for the Next Generation of AI* [52]	The Japanese government drew up the key points of AI technology, the breakthrough path, the industrial distribution and talent training

Table 2.7 AI strategy-related documents released by Russia

Time	Agency	Document	Overview
2014	Ministry of Defense of the Russian Federation	*2025 Integrated Program for Research and Development of Advanced Military Robot Technology and Equipment* [53]	The document sees AI robot technology as a key military technology for research and development. It is expected that by 2025, the proportion of unmanned systems in the Russian military equipment will reach 30%
2017	Ministry of Defense of the Russian Federation	*Russia's Military Modernization Plans: 2018–2027* [54]	The plans are aimed to equip the Russian armies with intelligent robot systems and weapons & equipment
October 2019	Government of Russia	*National Strategy for the Development of Artificial Intelligence before 2030* [55]	The strategy ensures that Russia accelerates AI development and research, increases the information and computing resources available to users, and improves talent training

2.3.3 AI Strategies in South Korea, Singapore, Canada, India and More

In December 2016, the Ministry of Science and ICT of the Republic of Korea released *Mid- to Long-Term Master Plan in Preparation for the Intelligent Information Society*[56] and set out to apply AI to public, manufacturing sectors and national defense.

In May 2017, Singapore established *AI.SG*[57] and decided to invest $150 million in AI development in order to enhance the country's strength in this field. In June 2018, Singapore founded it's AI ethics council to assist the government in devising ethical standards and governance framework.

In March 2017, Canada issued *Pan-Canadian Artificial Intelligence Strategy* [58] *a*nd appointed Canadian Institute for Advanced Research (CIFAR)[1] to develop and invested $125 million in AI research and talent training. CIFAR works in close collaboration with Canada's three national AI Institutes—AMII in Alberta, MILA in Montréal, and Vector Institute in Toronto.

In June 2018, India published *National Strategy for Artificial Intelligence* [59] and placed emphasis on AI market applications and generated a unique brand of #AIforAll. The approach is especially for "made in India". The paper also involves the parts pertaining to military security, ethics and privacy.

2.4 AI Development Strategies in China

The rapid development of AI will bring about tectonic changes to human life and the world. To seize such a strategic opportunity, gain a head start in AI development, and grow into an innovative and a technology-strong country, China must make AI development a national strategy, take an active position amidst the new round of global competition of AI development. Now, as AI has developed to a new level, and risen to be a new focus of international competition as well as an engine for economic growth, it brings about a new opportunity but also challenges for social development. Against this backdrop, China has formulated a series of policies and plans regarding application of AI, industrial development, technological innovation, trying to utilize all resources available to make AI a driving force of technological development, gaining an upper hand in the global competition, explore new fields of development and safeguard national security.

Table 2.8 lists some planning documents that China has issued with a vision to boost AI development.

[1] CIFAR: it is a Canadian-based global research organization that convenes extraordinary minds (20 Nobel Prize winners and over 400 researchers from 22 countries) to address the most important questions facing science and humanity. In 2017, the Government of Canada appointed CIFAR to develop and lead a $125 million Pan-Canadian Artificial Intelligence Strategy, the world's first national AI strategy.

Table 2.8 Some of China's AI development planning documents

Time	Agency	Document	Overview
2015–07	The State Council	*State Council's Instructions to Actively Push the "Internet+" Operations* [60]	Cultivate and develop emerging industries of AI, promote innovation of intelligent products in key fields, and improve the intellectualization level of the end products
2016–03	The State Council	*Outline of the 13th Five-Year Plan for National Economic and Social Development* [61]	Artificial intelligence is included into the "Internet+" operations in the "Key Project of Informatization"
2016–07	The State Council	*The 13th Five-Year Plan for National Science and Technology Innovation* [62]	Include AI into the column of "Developing the New Generation of Information Technology"
2017–07	The State Council	*Development Planning for the New Generation of AI* [63]	Put forward the grand strategic goal of "Three Steps"
2017–12	Ministry of Industry and Information Technology (MIIT)	*Three-year Operation Plan to Promote the Development of the New Generation of Artificial Intelligence Industry from 2018 to 2020* [64]	Implement the *Development Planning for the New Generation of AI*, deepen the integration of artificial intelligence and the entity economy
2018–04	Ministry of Education	*Operation Plan for the Innovation of Artificial Intelligence in Colleges and Universities* [65]	Enhance the capacities of colleges and universities to develop scientific and technological innovation and nature talents in the field of artificial intelligence

2.4.1 Overall Requirements for AI Development in China

China's guideline for AI development is to implement the innovation-driven development strategy in depth. By taking the accelerated integration of artificial intelligence with the economy, the society and national defense as the main line, and enhancement of capacities in the new-generation AI-related sci-tech innovation as the focus direction, China strives to develop the intelligent economy, construct an intelligent society, maintain national security, promote the interactive integration among knowledge groups and technology groups, and build an ecosystem where talent, systems and culture support each other. It braves risks and challenges with foresights, promotes the sustainable human development-centric intellectualization, and comprehensively elevates social productivity, overall national strength and national competitiveness, thereby providing strong support for its transformation into an innovative country and one of the world's sci-tech powers, and realizing the "Two Century Goals" as well as the dream of the great rejuvenation of the Chinese nation.

The basic principles of AI development in China are: led by science and technology, systematic layout, market-oriented, open source and inclusiveness. In other words, to explore the layout with long-term support in key frontier fields, to strive to make transformative and ground-breaking breakthroughs in theories, methods,

tools, systems and etc., to promote the overall layout of projects, bases and talents, to highlight the main role of enterprises in the selection of technology routes and the formulation of industry product standards, to accelerate the commercialized applications of artificial intelligence's sci-tech achievements, to gain competitive advantages, to advocate the open source sharing concept, and to promote the co-innovation and sharing among innovative subjects for industry-university-research cooperation.

The following are the three steps to march towards the strategic goal of AI development in China. First, by 2020, the overall AI development and application will reach the world's top level, the AI industry will become a new important economic growth point, the application of AI technology will become a new way to improve people's livelihood, and will strongly support the goal for China to enter the ranks of innovative countries and achieve a well-off society in an all-round way. Second, by 2025, the basic theories of AI will have made major breakthroughs, with some areas of AI reaching the world's leading level, and AI will become the main driving force for China's industrial upgrading and economic transformation, and the construction of an intelligent society will have made considerable headway. Third, by 2030, AI theories, technologies and applications will have reached the world's leading level, and China will have become the world's main center of AI innovation, its intelligent economy and intelligent society will have achieved significant results, thus to lay an important foundation for China to become a world-leading innovative country and economic power.

The overall deployment of AI development plans in China is as follows: to plan the layout as per the "build a system, grasp the dual attributes, adhere to the 'three-in-one' method, strengthen the four major support" policy, and to pave a healthy and sustainable strategic path for AI progress. Specifically, to build an open and collaborative environment of science and technology innovation for AI development, to grasp the characteristics of high integration of artificial intelligence technology attributes and social attributes, to adhere to the "three-in-one" method that combines AI R&D, product application, and industrial cultivation, and to strengthen the overall support of AI for science and technology, economy, society and national security.

2.4.2 Main Tasks of AI Development in China

The main goals of AI development in China are to base AI development on the overall development of the country, to accurately grasp the global development trend of AI, to locate pertinent breakthrough points and the main focus direction, to comprehensively enhance the basic capabilities of scientific and technological innovation, to comprehensively expand the depth and breadth of AI application in key fields, and to comprehensively upgrade economic and social development and the intellectualization level of national defense systems.

The main tasks involved in China's initiatives to develop AI are: first, to build an open and collaborative AI innovation system, including establishing a system

of theories for the new-generation AI, constructing a system of general technologies for the new-generation AI, conducting the overall layout of the artificial intelligence innovation platform, accelerating the training and gathering of high-end artificial intelligence talents; second, to nurture high-end and efficient intelligent economy, including vigorously developing emerging industries of AI, accelerating the upgrading of industrial intellectualization, energetically developing intelligent enterprises, and creating an innovation hub of AI; third, to build a safe and convenient intelligent society, including developing convenient and efficient intelligent services, promoting intelligent social governance, and using artificial intelligence to enhance the public security assurance capacity, promoting social interaction and mutual trust; fourth, to build a ubiquitously-secure, efficient intelligent infrastructure system; and fifth, to lay out a new generation of major sci-tech projects of AI, i.e. to establish a new generation of AI-related sci-tech projects in view of the urgent needs and weak links of AI development in China.

2.4.3 Resource Allocation for Development of AI-Related Industries in China

To boost AI development, China should make full use of the existing resources—funds, industrial bases and the like, utilize innovation resources at home and abroad in a proper manner, give play to the guiding role of policy incentives and government funding as well as the leading role of resource allocation schemes to increase the investment from enterprises and the society to the AI sector, thereby creating a new pattern in which AI development gains financial support from the government, the market and the society. To be specific, first, China should establish a financial support mechanism that is public finance-guided and market-led; second, it should optimize the layout and build AI innovation bases; third, it should comprehensively tap into overseas and domestic innovation resources.

To advance the AI-related industries, China has released some guiding policies and measures, as shown in Table 2.9.

2.4.4 Safeguard Measures for China's AI Development

To promote sound and rapid AI development, China needs to properly manage the challenges that AI may bring about, make institutional arrangements that fit AI development, and build an open, inclusive international environment, and lay a solid social foundation for it.

(1) Formulate laws and regulations and ethical norms that facilitate AI development. Strengthen research on legal, ethical and social issues related to AI, and establish legal and ethical frameworks that safeguard the sound development of

Table 2.9 Relevant documents regarding the development of AI-related Industries in China

Relevant industries	Time	Agency and related conferences	Document	Overview
Robotics	2016–04	Ministry of Industry and Information Technology, National Development and Reform Commission, Ministry of Finance	*Robotics Industry Development Planning (2016–2020)* [66]	Guide rapid, healthy and sustainable development of the robotics industry in China
Internet	2016–05	National Development and Reform Commission, Ministry of Science and Technology, Ministry of Industry and Information Technology, Cyberspace Administration of China	*"Internet+" Artificial Intelligence Three-year Implementation Plan* [67]	Make breakthroughs in key AI technologies, reinforce industrial chain collaboration and industrial ecological cultivation, advance the formation of the billion-level market scale of AI applications
Intelligent hardware	2016–09	Ministry of Industry and Information Technology and National Development and Reform Commission	*Special Actions on Intelligent Hardware Industrial Innovation and Development (2016–2018)* [68]	Vigorously improve universal technologies and products of AI hardware in China, improve the product intellectualization level, accelerate application of AI hardware
Emerging industries of strategic importance	2016–12	The State Council	*National Development Plan for Emerging Industries with Strategic Importance of "the 13th Five-year Plan"* [69]	Cultivate the ecology of the artificial intelligence industry, promote the application of artificial intelligence in key fields of economy and society, and forge a globe-leading technology system
	2018–09	National Development and Reform Commission	*Key Products and Services Guidance Catalogue of the Emerging Industries with Strategic Importance (2016 Edition)* [70]	Implement the *National Development Plan for Emerging Industries with Strategic Importance of "the 13th Five-year Plan"*, and the AI industry enters the guidance catalogue for the first time

(continued)

Table 2.9 (continued)

Relevant industries	Time	Agency and related conferences	Document	Overview
Artificial intelligence industry innovation	2018–11	Ministry of Industry and Information Technology	*Key Innovative Task Revealing Program of the New-generation Artificial Intelligence Industry* [71]	By "publishing a list of successful candidates and revealing the top-ranking ones", strive to break through the bottleneck of the development of artificial intelligence industry
Entity economy adjustments	2019–03	The 7th Meeting of the Commission for Comprehensively Deepening Reform of the CPC Central Committee	*Guidance on Promoting the In-depth Integration of Artificial Intelligence and the Entity Economy* [72]	Cultivate the intelligent economic form that is driven by data, along with human–machine collaboration and cross-border integration
Industrial restructuring	2019–11	National Development and Reform Commission	*Guidance Catalogue for Industrial Restructuring* [73]	Add the "AI industries" into the "Encouraging Category", list 15 items worth encouragement

AI. In December 2017, the Ministry of Industry and Information Technology issued the *Three-year Action Plan for Promoting the Development of a New Generation of Artificial Intelligence Industry (2018–2020)*, proposing to "initially establish a standard system for artificial intelligence industry and complete a third-party pilot testing platform and carry out evaluation service by 2020; build an intellectual-property service platform with basic supporting capabilities in such fields as pattern recognition, semantic comprehension, autopilot and intelligent robot.". On 24 November, 2018, the chairman of the Standing Committee of the National People's Congress convened a meeting to discuss AI-related legal issues. In March 2019, at the Second Session of the 13th National People's Congress, NPC members involved AI-related legislative programs in the legislation plan and would: conduct studies on legal issues like civil and criminal liability verification, privacy and property protection, as well as information security utilization that are in connection with the application of AI, establish a traceability and accountability system, and clarify the legal subject of artificial intelligence and the related rights, obligations and responsibilities; carry out research on AI behavioral science, ethics and other issues, and construct a multi-level ethical judgment structure as well as ethical framework for human–machine collaboration; develop moral standard and code of conduct for AI product R&D staff, and enhance the assessment of the potential harm and benefits of AI.

(2) Perfect key policies supporting AI development. Implement fiscal and tax preferential policies for small and medium-sized AI enterprises and start-ups; improve the implementation of policies related to data opening and protection; study and perfect the policy systems for education, medical care, insurance, and social assistance that adapt to AI development.

(3) Standardize the leading field. Financially, in April 2018, People's Bank of China, China Banking and Insurance Regulatory Commission, China Securities Regulatory Commission, and State Administration of Foreign Exchange issued the *Guiding Opinions on Regulating the Asset Management Business of Financial Institutions*[74] to standardize financial institutions in utilizing AI technology to conduct business and prevent risks. In terms of transportation, on 12 April, 2018, Ministry of Industry and Information Technology, Ministry of Public Security, and Ministry of Transport jointly issued the *Management Practices for Road Testing of Intelligent Connected Vehicles (Trial)* [75], which clearly stipulated the test subject, driver and vehicle, application and approval, violation and settlement of intelligent connected vehicles.

(4) Set up a supervision and evaluation system for AI security. Strengthen the research and assessment of the impact of AI on national security and confidentiality; enhance research on the prediction, judgement and traceability of AI technology intensify risk awareness and attach importance to risk assessment as well as prevention and control; establish and improve a transparent supervision system for AI development; promote the self-discipline of AI industries and enterprises; reinforce the R&D of AI network security technology; construct a dynamic evaluation mechanism for artificial intelligence R&D application.

On 1 December, 2017, the release-preview meeting of *ETHICALLY ALIGNED DESIGN: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems (Edition 2)*[76] of Institute of Electrical and Electronic Engineers (IEEE) was held at the Institutes of Science and Development, Chinese Academy of Sciences (CASISD), during which IEEE Standards Association, IEEE China and CASISD probed into artificial intelligence and global initiative on ethics, discussing ethical paths to enhance international cooperation to promote AI security. In September 2018, China Academy of Information and Communications Technology issued the *White Paper on Artificial Intelligence Safety and Security (2018)* [77], presenting the artificial intelligence security architecture for the first time.

(5) Make great efforts to strengthen AI talent training. Accelerate the research on the employment structure, change of employment pattern and competence for new occupations and jobs of the AI industry, and establish a lifelong learning and employment training system that meets the demands of the smart economy and society; encourage enterprises and various institutions to provide employees with AI technical training; enhance re-employment training and guidance.

(6) Extensively carry out AI popularization activities in various forms; launch education-for-all projects; construct and improve infrastructures for AI popularization; open AI R&D platforms, production facilities or exhibition halls to the public; provide supports to hold AI competitions.

2.4.5 Organizational Guarantee for China's AI Development

In terms of the organizational guarantee of the new generation of AI development, the leading group for National Science and Technology Structural Reform & Innovation System Construction makes overall planning and coordination, dealing with major tasks, policies, issues and assignments, and spurring the legal construction related to artificial intelligence; as the National Science and Technology Program (for special projects and funds, etc.) manages the inter-ministerial joint meetings, the Ministry of Science and Technology facilitates the implementation of new generation of major AI projects; in November 2017, the National Development Planning Office for New-Generation Artificial Intelligence, an agency composed of 15 national departments, was established under the administration of the Ministry of Science and Technology to advance specific project implementation. In the meantime, the Strategy Advisory Committee for Artificial Intelligence was founded to study prospective and strategic issues of AI and provide consultation as well as evaluation for important decisions on AI; the Office also set up the New-Generation AI Regulatory Committee to tackle social, ethical and legal challenges potentially triggered by AI development.

Other AI-related institutions include National Artificial Intelligence Standardization Group, China Artificial Intelligence Industry Development Alliance, China Artificial Intelligence Society, and Artificial Intelligence Professional Committee, China Association for Educational Technology.

2.4.6 China's Policies for AI Talent Training

With respect to higher education: as of December 2017, a total of 71 universities across China have set up 86 AI-related sub-disciplines or inter-disciplines. In April 2018, the Ministry of Education issued the *Artificial Intelligence Innovation Initiative for Universities* [78] to improve AI talent training. In July 2018, 26 universities, including Tsinghua University and Fudan University, jointly signed the *Proposal on Setting up Majors for Artificial Intelligence* [79] to apply for establishing an undergraduate program of this field. In August 2018, the Ministry of Education formed a group of experts for AI technological innovation to delve into specific projects and conduct relevant researches in this field.

According to the *Artificial Intelligence Innovation Initiative for Universities*, by 2020, China will basically finish optimizing the layout of a higher-education sci-tech innovation system that suits the new generation of AI development. Moreover, universities are expected to make new breakthroughs regarding AI fundamentals as well as key technological studies; AI talent training and scientific research will be further enhanced to promote the extensive use of AI technology. By 2025, Chinese universities will enjoy a stronger capacity in high-tech innovation, more quality training for AI talents, and harvest a number of originally-incubated fruits with global impact, in which some theoretical research, innovative technologies and application demonstrations will reach a world-leading level to support China's industrial upgrading, economic transition and construction of an intelligent society. By 2030, Chinese universities will be the backbone in building the world's major AI innovation centers and a talent-gathering land that leads the new generation of AI development, providing technological and talent support for China to take the forefront among innovative countries. To reach this goal, first of all, the AI science and technology innovation system at Chinese universities must be optimized, including enhancing the basic studies on the new generation of AI and facilitating key technological innovation, accelerating the construction of AI technological innovation bases, creation of high-caliber talents as well as creative teams; strengthening the development of top-level think-tanks and increasing more international academic exchanges and collaborations; secondly, the AI talent training must be perfected, including developing more comprehensive system of courses, creating new disciplines, improving textbooks, enhancing the talent training system, implementing education-for-all programs and supporting innovative start-ups; thirdly, the commercialization of research findings and pilot application must be promoted, including advancing application of AI in key fields, boosting intelligent education development, and encouraging construction of innovation alliances and resource opening & sharing, supporting local and regional innovation development.

With regard to primary and secondary education, the *Development Planning for the New Generation of AI*, issued by the State Council, proposes to set up AI-related courses in primary and secondary schools. In January 2019, the Ministry of Education launched the "AI Education Program for Primary and Secondary Schools", taking Beijing, Guangzhou, Shenzhen, Wuhan, and Xi'an as the very first batch of pilot

cities for AI education. And it was planned that students in Grade 3 to 8 would be all involved in this pilot program. In February 2019, the Ministry of Education issued the *Key Points for Educational Informatization and Cybersecurity of 2019*[80], initiating the information literacy assessment for primary and secondary school students, and setting up AI-related courses accordingly.

2.5 Research on AI Standards in China and Abroad

2.5.1 Research on AI Standards Worldwide

As early as in 2016, IEEE launched the "*Global Initiative on Ethics of Autonomous and Intelligent Systems"* [81], which was accomplished by committees subordinated to the IEEE Global Initiative that specialized in the ethical issues in AI and autonomous systems, including over 100 experts from related fields like artificial intelligence, ethics, political science, law and philosophy. It calls upon designing ethical principles and standards to free people from the fear of AI technology.

International Organization for Standardization/International Electrotechnical Commission (ISO/IEC) forms a research group for AI trustworthiness to kick off studies on AI security standard. The ISO/IEC JTC 1/SC 42 Artificial Intelligence Sub-Technical Committee was established in October 2017 to study and formulate AI international standards. SC42 has set up five working groups, covering primary standard (WG1), big data (WG2), trustworthiness (WG3), use case and application (WG4), AI system computing method and computing feature (WG5); in addition, SC42 also includes an advisory group for AI management system standard (AG1), as well as an advisory group for intelligent system engineering (AHG3).

WG3 focused on AI trustworthiness and ethics, and has begun developing AI trustworthiness, robustness evaluation, algorithmic bias, and ethics. Standards projects studied by WG3 included *Information Technology—Artificial Intelligence (AI)—Bias in AI Systems and AI Aided Decision Making* (ISO/IEC NPTR 24027) [82], *Information Technology—Artificial Intelligence (AI)—Overview of Trustworthiness in Artificial Intelligence* (ISO/IEC PDTR 24028) [83], *Artificial Intelligence (AI)—Assessment of the Robustness of Neural Networks—Part 1: Overview* (ISO/IEC CD TR 24029-1) [84], *Artificial Intelligence (AI)—Assessment of the Robustness of Neural Networks—Part 2: Formal Methods Methodology* (ISO/IEC TR 24029-2) [85], *Information Technology—Artificial Intelligence—Risk Management* (ISO/IEC AWI 23894) [86], *Information Technology—Artificial Intelligence—Overview of Ethical and Social Concerns* (ISO/IEC AWI TR 24368) [87], and *Software and Systems Engineering—Software Testing—Part 11: Testing of AI-based Systems* (ISO/IEC NP TR 29119-11) [88].

ISO has two categories for test standard of robot: industrial robot and personal care robot. The ISO/TC299 (International Organization for Standardization/Technical Committee for Robots Standardization), to be more specific, undertakes the task of

robot standardization and focuses on formulating standards for security and performance test. There are 8 working groups in ISO/TC299, among which ISO/TC299/WG2 finishes the first security standard in the field of personal care robot—*Security Requirements for Personal Care Robots* (ISO 13482: 2014). ISO/TC299/WG3 pays close attention to industrial safety. Currently, the standard for industrial robots has been relatively perfect, while the standard for ontological security and integrated security has been completed.

Additionally, ISO/IEC mainly sets standards concerning the security and performance of home-service robots, functional security of industrial robots, and security of medical robots. The aforesaid standardization is mainly developed by TC59, TC61, TC62, and TC116 technical committees.

At Telecommunication Standardization Sector of the International Telecommunications Union (ITU-T), SG17 research group and SG16 multi-media research group developed relevant security standards for AI. What is more, SG17 has already conducted studies and discussions on AI for security as well as AI security application, and launched related standardization projects.

IEEE carried out multiple studies on AI ethics and released a number of ethical standards and research reports. The IEEE P7000—Engineering Methodologies for Ethical Life-Cycle Concerns Working Group [89], which was founded and organized by IEEE, studied a series of standards, including *Transparency of Autonomous Systems* (IEEE P7001) [90], *Data Privacy Process* (IEEE P7002) [91], *Algorithmic Bias Considerations* (IEEE P7003) [92], *Standard for Child and Student Data Governance* (IEEE P7004) [93], *Standard for Transparent Employer Data Governance* (IEEE P7005) [94],*Standard for Personal Data Artificial Intelligence (AI) Agent* (IEEE P7006)[95], *Ontological Standard for Ethically Driven Robotics and Automation Systems* (IEEE P7007)[96], *Standard for Ethically Driven Nudging for Robotic, Intelligent and Autonomous Systems* (IEEE P7008) [97], *Standard for Fail-Safe Design of Autonomous and Semi-Autonomous Systems* (IEEE P7009)[98], *Well-being Metrics Standard for Ethical Artificial Intelligence and Autonomous Systems* (IEEE P7010) [99],*Standard for the Process of Identifying and Rating the Trustworthiness of News Sources* (IEEE P7011) [100], *Standard for Machine Readable Personal Privacy Terms* (IEEE P7012)[101], *Inclusion and Application Standards for Automated Facial Analysis Technology* (IEEE P7013)[102], and *Standard for Ethical considerations in Emulated Empathy in Autonomous and Intelligent Systems* (IEEE P7014)[103].

At present, the existing AI standards around the world are mainly general standards for the AI technology and application, and most of the relevant safety standards, such as AI security, ethics and privacy protection, are still under research. Security assessment for AI-related data, intelligent algorithms, and decision-making models are still insufficient as well.

2.5.2 Research on AI Standards in China

In the *Development Planning for the New Generation of AI*, China underlines enhancement in studying the framework system for AI standards, meaning that the basic commonality, interconnectivity, industry application, cyber security, and privacy protection of AI will be gradually established and improved based on principles of security, usability, interoperability and traceability. Meanwhile, China pushes industrial associations and alliances of fine-sorted applications fields, such as pilotless-driving and service robot, to make relevant standards. encourages AI companies in China are encouraged to participate or take the lead in formulating international standards, hence "going global" with technical standards to popularize our AI products and service overseas. China also strengthens the protection of intellectual property rights in the field of AI, and perfects the mutual-support mechanism for technological innovation patent protection and standardization to turn fruits of AI innovation into intellectual property rights. Meanwhile, it sets up a public patent pool for AI to facilitate the utilization and diffusion of AI new technologies.

The National Information Technology Standardization Technical Committee mainly conducts studies on AI standards in such fields as AI terminology, human–computer interaction, biometric recognition, big data and cloud computing, and has already released multiple standards, such as *Information Technology—Vocabulary—Part28: Artificial Intelligence—Basic Concepts and Expert System* (GB/T 5271.28-2001)[104], *Information Technology—Vocabulary—Part 29: Artificial Intelligence—Speech Regconition and Synthesis* (GB/T5271.29-2006) [105], *Information Technology—Vocabulary—Part 31: Artificial Intelligence—Machine Learning* (GB/T 5271.31-2006)[106], and *Information Technology—Vocabulary—Part 34: Artificial Intelligence—Neural Networks* (GB/T 5271.34-2006)[107].

As a professional organization for security standardization, National Information Security Standardization Technical Committee (TC260) carries out studies on AI safety and security standards in the fields like autopilot, biometric recognition and smart city. Specifically, its approved research projects include *Research on Artificial Intelligence Safety and Security Standards*, *Information Security Techniques—Guide on Artificial Intelligence Application Security*, *Information Security Techniques—An Identity Authentication Protocol for Biometric Recognition based on Trusted Environment*, *Information Security Techniques—Technical Requirements for Fingerprint Identification System*, *Information Security Techniques—Technical Requirements for Online Face Recognition and Authentication System Security*, *Information Security Techniques—Protection Requirements for Biometric-features Recognition Information,* and *Information Security Technology—General Technical Requirements of Smart Home Security*. It also released standards like the *Information Security Technology—Technical Requirements for Iris Identification System* (GB/T 20979–2007) [108], *Information Security Techniques—Biometric Authentication Protocol Framework based on Trusted Environment* (GB/T 36651–2018)[109], along with *Information Security Technology—Technical Requirements for Fingerprint Recognition*

System (GB/T 37076–2018)[110]. There are over 20 standards being studied or already released.

Furthermore, China National Technical Committee for Automation System and Integration Standardization mainly focuses on the whole machine, system interface, spare parts and controllers of industrial robots. China National Technical Committee for Audio, Video and Multimedia Standardization mainly focuses on audio and video, and wise medical & healthcare products. China National Technical Committee for Intelligent Transportation System Standardization mainly develops standards in the field of intelligent transportation system.

Technical committees subordinated to China Communications Standards Association are all probing into AI standards. The Working Committee for Internet and Information Security (CCSA-TC8) is responsible for developing AI safety and security standards, and has released standards like *Technical Requirements for Security Capability of Smart Home Terminal Equipment* (YDB 201-2018), and *Technical Requirements for Smart Home Network System Security* (T/CSHIA001-2018). As of now, it is developing such standards as *Evaluation Guide for Security of Artificial Intelligence Products, Applications and Services*, *Evaluation Requirements for Security of Artificial Intelligence Service Platform*, *Research on Standard Systems of Artificial Intelligence Terminal Products*, *Technical Requirements and Evaluation Methods for Artificial Intelligence Capacity of Mobile Intelligent Terminal and Individual Information Protection*, and *Technical Requirements and Testing Evaluation Methods for Face Recognition Security of Mobile Intelligent Terminal*.

China Artificial Intelligence Open Source Software Development League (AIOSS) is a social organization specializing in developing AI open source software. In July 2018, AIOSS released *Artificial Intelligence—Assessment Specification for Deep Learning Algorithms* (AIOSS-01-2018), the very first standard for AI deep learning algorithms; also, AIOSS complied *Information Technology—Artificial Intelligence—Sample Size Requirements and Algorithm Evaluation Indicators for Few-shot Learning* (CESA-2018-4-001), *Information Technology—Artificial Intelligence—Audio, Video and Image Analysis Algorithm Interface* (CESA-2018-4-002), *Information Technology—Artificial Intelligence—Quality of Machine Learning Model* (CESA-2018-4-003), *Information Technology—Artificial Intelligence—Framework and Functional Requirements of System for Machine Learning* (CESA-2018-4-004),*Information Technology—Artificial Intelligence—Classification Evaluation on Intelligent Assistant's Intelligence Capabilities* (CESA-2018-4-005), *Information Technology—Artificial Intelligence—Intelligent Ability Assessment for Machine Translation* (CESA-2018-4-006), *Information Technology—Artificial Intelligence—Guildline of Data Annotation for Machine Learning* (CESA-2018-4-007), *Information Technology—Artificial Intelligence—Reference Model of Service Capability Maturity Evaluation* (CESA-2018-4-008), *Information Technology—Artificial Intelligence—Specification on Intelligent-Early-Warning System for Municipal Wastewater Treatment Process* (CESA-2018-4-009), *Server General Specification for Deep Learning* (CESA-2018-4-010), *Information Technology—Artificial Intelligence—Technical Requirements for Driving Status Warning System of Drivers* (CESA-2018-4-011).

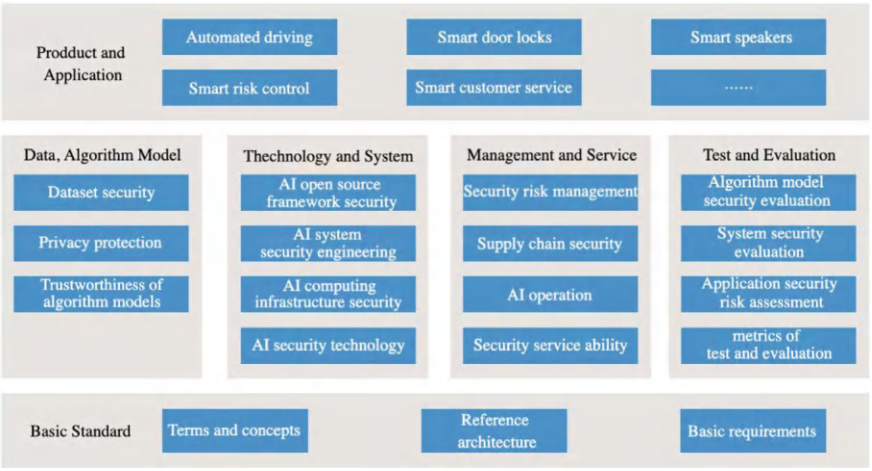


Fig. 2.1 System of AI security standards. *Source* White Paper on Artificial Intelligence Standardization [112]

In January 2018, China Electronics Standardization Institute issued the *White Paper on Artificial Intelligence Standardization (2018)* [111], establishing the standard system guide the development of AI industry and proposing basic projects that were in urgent need to be developed, as well as key standard projects. In October 2019, National Information Security Standardization Technical Committee issued the *White Paper on Artificial Intelligence Standardization (2019)* [112], which explored the demand for AI security standardization and offered suggestions on the framework and standardization for AI security. The *White Paper on Artificial Intelligence Standardization* (2019) defines the system of AI security standards, as shown in Fig. 2.1, covering six parts, namely "Basic Standard", "Data, Algorithm Model", "Technology and System", "Management and Service", "Test and Evaluation" as well as "Product and Application".

However, what needs to be pointed out is that in terms of AI security standards, China only focuses on a few application fields, such as robotics, biometrics recognition and intelligent connected vehicles, along with supportive security standards like big data security and privacy protection. And still, for security reference architecture, security evaluation, ethics design, security requirements and evaluation methods that are related to the self-security of AI and basic generality, relevant standards are few.

2.6 Summary

Major developed countries around the world have different priorities and emphases on AI development. Comparatively, The United States focuses on the impact of AI on its national security and its global leading position, as well as relevant applications for the military field; European Union, upholding the "human-centered" notion, lays emphasis on the ethical impact brought by AI; the result-driven UK government stresses fundamental studies on AI; Russia values application of AI to national defense and its impact on military security. Countries like Canada, Japan, South Korea, and Singapore attach importance to talent training, technology R&D and industry advancement.

AI development in China is mainly navigated by the strategies and plans issued by the government. Its plans and policies for AI development are highly structuralized and "top-down". Currently, China has huge human and data resources, which, after years of accumulation, have formed a unique edge it has for AI development.

References

1. CSIS (Center for Strategic and International Studies). Defense 2045: assessing the future security environment and implications for defense policymakers. (2015–11) [2019–11–27]. https://csis-prod.s3.amazonaws.com/s3fs-public/legacy_files/files/publication/151106_Miller_Defense2045_Web.pdf.
2. NSTC (National Science and Technology Council). Preparing for the future of artificial intelligence. (2016–10–12) [2019–11–27]. http://www.raincent.com/uploadfile/2016/1013/20161013013841632.pdf.
3. NSTC. The national artificial intelligence research and development strategic plan. (2016–10–13) [2019–11–27]. https://www.nitrd.gov/PUBS/national_ai_rd_strategic_plan.pdf.
4. EOP (Executive Office of the President). Artificial intelligence, automation, and the economy. (2016–12–20) [2019–11–27]. https://obamawhitehouse.archives.gov/sites/whitehouse.gov/files/documents/Artificial-Intelligence-Automation-Economy.PDF.
5. NSF (National Science Foundation). National robotics initiative 2.0: ubiquitous collaborative robots (NRI-2.0). [2019–11–27]. https://www.nsf.gov/pubs/2019/nsf19536/nsf19536.pdf.
6. Allen G, Chan T. Artificial intelligence and national security (2017–7) [2019–11–27]. https://www.belfercenter.org/sites/default/files/files/publication/AI%20NatSec%20-%20final.pdf.
7. National security strategy of the United States of America. (2017–12) [2019–11–27]. https://www.whitehouse.gov/wp-content/uploads/2017/12/NSS-Final-12-18-2017-0905.pdf
8. CSIS. A national machine intelligence strategy for the United States. (2018–3–1) [2019–11–27]. https://www.csis.org/analysis/national-machine-intelligence-strategy-united-states.
9. Summary of the 2018 Department of Defense artificial intelligence strategy. [2019–11–27]. https://media.defense.gov/2019/Feb/12/2002088963/-1/-1/1/SUMMARY-OF-DOD-AI-STRATEGY.PDF.
10. CNAS (Center for a New American Security). Artificial intelligence and international security. (2018–7) [2019–11–27]. https://s3.amazonaws.com/files.cnas.org/documents/CNAS-AI-and-International-Security-July-2018_Final.pdf?mtime=20180709122303.
11. CSIS. Artificial intelligence and national security, the importance of the AI ecosystem. (2018–11) [2019–11–27]. https://csis-prod.s3.amazonaws.com/s3fs-public/publication/181102_AI_interior.pdf?6jofgIIR0rJ2qFc3.TCg8jQ8p.Mpc81X.

12. Executive order on maintaining American leadership in artificial intelligence. (2019–2–11) [2019–11–27]. https://www.whitehouse.gov/presidential-actions/executive-order-maintaining-american-leadership-artificial-intelligence/.
13. NSTC. The national artificial intelligence research and development strategic plan: 2019 update. (2019–6–21) [2019–11–27]. https://www.whitehouse.gov/wp-content/uploads/2019/06/National-AI-Research-and-Development-Strategic-Plan-2019-Update-June-2019.pdf.
14. National Robotics Initiative (NRI). [209–12–4]. https://nsf.gov/pubs/2012/nsf12607/nsf12607.htm.
15. Brain Initiative. [2019–12–4]. https://braininitiative.nih.gov/.
16. Gunning D. Explainable artificial intelligence (XAI). [2019–12–4]. https://sites.nationalacademies.org/cs/groups/pgasite/documents/webpage/pga_184754.pdf.
17. A roadmap for US robotics: from internet to robotics. (2016–10–31) [2019–12–1] https://cra.org/ccc/wp-content/uploads/sites/2/2016/11/roadmap3-final-rs-1.pdf.
18. National robotics initiative 2.0: ubiquitous collaborative robots (NRI-2.0). (2017) [2019–12–1]. https://www.nsf.gov/pubs/2017/nsf17518/nsf17518.pdf.
19. NIH releases strategic plan for data science. (2018–6–4) [2019–12–1]. https://www.nih.gov/news-events/news-releases/nih-releases-strategic-plan-data-science.
20. McCain JS. National defense authorization act for fiscal year 2019. (2018–8–13) [2019–12–1]. https://congress.gov/115/plaws/publ232/PLAW-115publ232.pdf.
21. Algorithmic Warfare: DARPA's 'AI Next' program bearing fruit. (2019–02–07) [2019–12–4]. https://www.nationaldefensemagazine.org/articles/2019/7/2/algorithmic-warfare-darpas-ai-next-program-bearing-fruit
22. New J, Castro D. How policymakers can foster algorithmic accountability. (2018–5–21) [2019–11–30]. http://www2.datainnovation.org/2018-algorithmic-accountability.pdf.
23. A local law in relation to automated decision systems used by agencies. (2018–11–1) [2019–11–30]. https://legistar.council.nyc.gov/LegislationDetail.aspx?ID=3137815&GUID=437A6A6D-62E1-47E2-9C42-461253F9C6D0.
24. H.R.3388—SELF DRIVE Act. [2019–11–30]. https://www.congress.gov/bill/115th-congress/house-bill/3388.
25. Preparing for the future of transportation: automated vehicles 3.0. (2018–10–4) [2019–11–30]. https://www.nhtsa.gov/es/press-releases/us-department-transportation-releases-preparing-future-transportation-automated.
26. European Political Strategy Centre. The age of artificial intelligence: towards a European strategy for human-centric machines. (2018–3–27) [2019–11–30]. https://ec.europa.eu/epsc/sites/epsc/files/epsc_strategicnote_ai.pdf.
27. EU member states sign up to cooperate on artificial intelligence. (2018–4–10) [2019–11–30]. https://ec.europa.eu/digital-single-market/en/news/eu-member-states-sign-cooperate-artificial-intelligence.
28. Communication artificial intelligence for Europe. (2018–4–25) [2019–11–30]. https://ec.europa.eu/digital-single-market/en/news/communication-artificial-intelligence-europe.
29. Coordinated plan on artificial intelligence. (2018–12–7) [2019–11–30]. https://ec.europa.eu/digital-single-market/en/news/coordinated-plan-artificial-intelligence.
30. Eckpunkte der Bundesregierung für eine Strategie Künstliche Intelligenz (2018–7) [2019–11–30]. https://www.bmbf.de/files/180718%20Eckpunkte_KI-Strategie%20final%20Layout.pdf.
31. Strategie Künstliche Intelligenz der Bundesregierung.(2018–11) [2019–11–30]. https://www.bmbf.de/files/Nationale_KI-Strategie.pdf.
32. La stratégie nationale de recherche en intelligence artificielle. [2019–11–30]. https://www.enseignementsup-recherche.gouv.fr/cid136649/la-strategie-nationale-de-recherche-en-intelligence-artificielle.html.
33. House of Commons, Science and Technology Committee. Robotics and artificial intelligence. (2016–9–13) [2019–12–1]. http://www.publications.parliament.uk/pa/cm201617/cmselect/cmsctech/145/145.pdf.

34. UK Government Office for Science, 2016. Artificial intelligence: opportunities and implications for the future of decision making. https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/566075/gs-16-19-artificial-intelligence-ai-report.pdf.
35. Industrial strategy: artificial intelligence sector deal. [2019–12–1]. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/702810/180425_BEIS_AI_Sector_Deal__4_.pdf.
36. Growing the artificial intelligence industry in the UK. [2019–12–1]. https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/652097/Growing_the_artificial_intelligence_industry_in_the_UK.pdf.
37. AI in the UK: ready, willing and able?—Government response to the select committee report. (2018–6–28) [2019–12–1]. https://www.gov.uk/government/publications/ai-in-the-uk-ready-willing-and-able-government-response-to-the-select-committee-report.
38. Human brain project grant structure. [2019–12–2]. https://www.humanbrainproject.eu/en/about/human-brain-project-ec-grants/.
39. The strategic research agenda is a compendium for SPARC's operation 2014–2020. [2019–12–2]. https://www.eu-robotics.net/sparc/upload/topic_groups/SRA2020_SPARC.pdf.
40. Smart data—innovations in data. (2016–1–1) [2019–12–2]. https://www.bmwi.de/Redaktion/EN/Publikationen/smart-data-innovationen-aus-daten-englische-version.pdf?__blob=publicationFile&v=11.
41. HORIZON 2020. [2019–12–2]. https://ec.europa.eu/programmes/horizon2020/en.
42. Coordinated plan on artificial intelligence. (2018–12–7) [2019–12–2]. https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=56017.
43. Ethics guidelines for trustworthy AI. (2019–4–8) [2019–12–1]. https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai.
44. General data protection regulation. (2016–4–27) [2019–12–1]. https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32016R0679.
45. Draft report with recommendations to the commission on civil law rules on robotics. (2016–5–31) [2019–12–1]. http://www.europarl.europa.eu/doceo/document/JURI-PR-582443_EN.pdf?redirect.
46. European civil law rules in robotics. [2019–12–5]. http://www.europarl.europa.eu/RegData/etudes/STUD/2016/571379/IPOL_STU%282016%29571379_EN.pdf.
47. Civil law rules on robotics. (2017–2–16) [2019–12–1]. https://www.europarl.europa.eu/doceo/document/TA-8-2017-0051_EN.pdf.
48. New robot strategy. (2015–2–10) [2019–12–1]. https://www.meti.go.jp/english/press/2015/pdf/0123_01b.pdf.
49. Japan revitalization strategy 2016. [2019–12–1]. http://www.kantei.go.jp/jp/singi/keizaisaisei/pdf/2016_zentaihombun.pdf
50. New industrial structure vision: Japan's strategy to lead the 4th industrial revolution. [2019–12–1]. https://www.meti.go.jp/shingikai/sankoshin/shinsangyo_kozo/pdf/008_05_01.pdf.
51. R&D goals and industrialization roadmap for artificial intelligence. [2019–12–1]. https://www.nedo.go.jp/content/100862412.pdf.
52. Next-generation AI Promotion Strategy. [2019–12–5]. http://www.soumu.go.jp/main_content/000424360.pdf.
53. Xu Y. Next-generation artificial intelligence: the emerging driving force leading global development. Front Acad People's Tribune. 2017;20:16–25.
54. Russia's military modernization plans: 2018–2027. (2017–11) [2019–12–1]. http://www.ponarseurasia.org/sites/default/files/policy-memos-pdf/Pepm495_Gorenburg_Nov2017.pdf.
55. On the development of artificial intelligence in the Russian federation. [2019–12–1]. http://publication.pravo.gov.ru/File/GetFile/0001201910110003?type=pdf.
56. Mid- to long-term master plan in preparation for the intelligent information society. [2019–12–1]. https://english.msit.go.kr/cms/english/pl/policies2/__icsFiles/afieldfile/2017/07/20/Master%20Plan%20for%20the%20intelligent%20information%20society.pdf.
57. AI.SG: New national programme to catalyse, synergise and boost Singapore's artificial intelligence capabilities. (2017–5–3) [2019–12–1]. https://www.nrf.gov.sg/docs/default-source/modules/pressrelease/20170503144208219 1-press-release-(ai.pdf.

58. CIFAR Pan-Canadian artificial intelligence strategy. [2019–12–1]. https://www.cifar.ca/ai/pan-canadian-artificial-intelligence-strategy.
59. National Strategy for Artificial Intelligence #AIFORALL. (2018–6) [2019–12–1]. https://niti.gov.in/writereaddata/files/document_publication/NationalStrategy-for-AI-Discussion-Paper.pdf.
60. State Council. Guiding opinions of the state council on actively promoting the "Internet +" action. (2015–07–04) [2019–12–1]. http://www.gov.cn/zhengce/content/2015-07/04/content_10002.htm.
61. Outline of the 13th five—year plan for national economic and social development of the People's Republic of China. (2016–03–17) [2019–12–1]. http://www.gov.cn/xinwen/2016-03/17/content_5054992.htm.
62. State Council. 13th five—year national science and technology innovation plan. (2016–7–28) [2019–12–1]. http://www.gov.cn/zhengce/content/2016-08/08/content_5098072.htm.
63. State Council. New—generation artificial intelligence development plan. (2017–7–8) [2019–12–1]. http://www.gov.cn/zhengce/content/2017-07/20/content_5211996.htm.
64. Ministry of Industry and Information Technology. Three—year action plan for promoting the development of new—generation artificial intelligence industry (2018—2020). (2017–12–14) [2019–12–1]. http://www.miit.gov.cn/n1146295/n1652858/n1652930/n3757016/c5960820/part/5960845.docx.
65. Ministry of Education. Innovation action plan for artificial intelligence in institutions of higher education. (2018–4–2) [2019–12–1]. http://www.moe.gov.cn/srcsite/A16/s7062/201804/t20180410_332722.html.
66. Ministry of Industry and Information Technology, National Development and Reform Commission, Ministry of Finance. Development plan for the robot industry (2016–2020). (2016–4–27) [2019–12–1]. http://www.miit.gov.cn/n1146295/n1652858/n1652930/n3757018/c4746362/content.html.
67. National Development and Reform Commission, Ministry of Science and Technology, Ministry of Industry and Information Technology, Cyberspace Administration of China. Three—year action implementation plan for "Internet +" artificial intelligence. (2016–5–18) [2019–12–2]. http://www.gov.cn/xinwen/2016-05/23/5075944/files/9cb49ac44cf341b29adf687b6857da34.pdf.
68. Ministry of Industry and Information Technology, National Development and Reform Commission. Special action for the innovative development of the intelligent hardware industry (2016–2018). (2016–09–21) [2019–12–2]. http://www.miit.gov.cn/n1146285/n1146352/n3054355/n3057643/n3057649/c5259057/content.html.
69. State Council. 13th five—year national strategic emerging industries development plan. (2016–12–19) [2019–12–2]. http://www.gov.cn/zhengce/content/2016-12/19/content_5150090.htm.
70. National Development and Reform Commission. Guidance catalogue of key products and services in strategic emerging industries (2016 Edition). (2018–9–21) [2019–12–2]. http://www.gov.cn/xinwen/2018-09/22/5324533/files/dcf470fe4eac413cabb686a51d080eec.pdf.
71. General Office of the Ministry of Industry and Information Technology. Work plan for the list—unveiling of key tasks in the innovation of new—generation artificial intelligence industry. (2018–11–14) [2019–12–2]. http://www.miit.gov.cn/n1146295/n1652858/n1652930/n3757016/c6489400/part/6489992.doc.
72. Xi Jinping chairs the 7th meeting of the central comprehensively deepening reforms commission. (2019–03–19) [2019–12–2]. http://www.gov.cn/xinwen/2019-03/19/content_5375140.htm.
73. National Development and Reform Commission. Guidance catalogue for industrial structure adjustment. (2019–11–06) [2019–12–2] http://www.gov.cn/xinwen/2019-11/06/5449193/files/26c9d25f713f4ed5b8dc51ae40ef37af.pdf
74. People's Bank of China, China Banking and Insurance Regulatory Commission, China Securities Regulatory Commission, State Administration of Foreign Exchange. Guidance opinions on regulating the asset management business of financial institutions (Yin Fa 〔2018〕 No.

106). (2018–4–28) [2019–12–2]. http://www.pbc.gov.cn/tiaofasi/144941/3581332/3730258/index.html.
75. Ministry of Industry and Information Technology, Ministry of Public Security, Ministry of Transport. Administrative specifications for road tests of intelligent and connected vehicles (Trial). (2018–4–3) [2019–12–2]. http://www.gov.cn/xinwen/2018-04/12/content_5281923.htm.
76. Ethically aligned design: a vision for prioritizing human well-being with autonomous and intelligent systems. [2019–12–3] https://standards.ieee.org/content/dam/ieee-standards/standards/web/documents/other/ead_v2.pdf.
77. Security Research Institute of China Academy of Information and Communications Technology. White paper on artificial intelligence security (2018). (2018–9) [2019–12–2]. http://www.caict.ac.cn/kxyj/qwfb/bps/201809/t20180918_185339.htm.
78. Ministry of Education. Innovation action plan for artificial intelligence in institutions of higher education. (2018–04–03) [2019–12–2] http://www.moe.gov.cn/srcsite/A16/s7062/201804/t20180410_332722.html.
79. Universities jointly sign the "proposal on setting up artificial intelligence majors". (2018–07–09) [2019–12–2]. https://news.buaa.edu.cn/info/1006/44625.htm.
80. General Office of the Ministry of Education. Key points of work on educational informatization and cybersecurity in 2019. (2019–2–27) [2019–12–2]. http://www.moe.gov.cn/srcsite/A16/s3342/201903/t20190312_373147.html.
81. The IEEE global initiative on ethics of autonomous and intelligent systems. [2019–12–3]. https://standards.ieee.org/industry-connections/ec/autonomous-systems.html.
82. ISO/IEC NP TR 24027 Information technology—Artificial Intelligence (AI)—Bias in AI systems and AI aided decision making. [2019–12–3]. https://www.iso.org/standard/77607.html.
83. ISO/IEC PDTR 24028 Information technology—Artificial Intelligence (AI—Overview of trustworthiness in Artificial Intelligence). [2019–12–3]. https://www.iso.org/standard/77608.html.
84. ISO/IEC CD TR 24029-1, Artificial Intelligence (AI)—Assessment of the robustness of neural networks—Part 1: Overview. [2019–12–3]. https://www.iso.org/standard/77609.html.
85. ISO/IEC JTC 1/SC 42 N 478, ISO/IEC NP 24029-2, Artificial Intelligence (AI) —Assessment of the robustness of neural networks—Part 2: Formal methods methodology. [2019–12–3]. https://standardsdevelopment.bsigroup.com/projects/9019-03425.
86. ISO/IEC AWI 23894 Information technology—artificial intelligence—risk management. [2019–12–3]. https://www.iso.org/standard/77304.html.
87. ISO/IEC AWI TR 24368 Information technology—artificial intelligence—overview of ethical and societal concerns. [2019–12–3]. https://www.iso.org/standard/78507.html.
88. ISO/IEC NP TR 29119-11 software and systems engineering—software testing—Part 11: Testing of AI-based systems. [2019–12–3]. https://www.iso.org/standard/79016.html.
89. IEEE P7000—Engineering methodologies for ethical life-cycle concerns working group. [2019–12–3]. https://site.ieee.org/sagroups-7000/.
90. P7001—Transparency of autonomous systems. [2019–12–3]. https://standards.ieee.org/project/7001.html.
91. P7002—Data privacy process. [2019–12–3]. https://standards.ieee.org/project/7002.html.
92. P7003—Algorithmic bias considerations. [2019–12–3]. https://standards.ieee.org/project/7003.html.
93. P7004—Standard for child and student data governance. [2019–12–3]. https://standards.ieee.org/project/7004.html.
94. P7005—Standard for transparent employer data governance. [2019–12–3]. https://standards.ieee.org/project/7005.html.
95. P7006—Standard for personal data artificial intelligence (AI) agent. [2019–12–3]. https://standards.ieee.org/project/7006.html.
96. P7007—Ontological standard for ethically driven robotics and automation systems. [2019–12–3]. https://standards.ieee.org/project/7007.html.

97. P7008—Standard for ethically driven nudging for robotic, intelligent and autonomous systems. [2019–12–3]. https://standards.ieee.org/project/7008.html.
98. P7009—Standard for fail-safe design of autonomous and semi-autonomous systems. [2019–12–3]. https://standards.ieee.org/project/7009.html.
99. IEEE—P7010—Wellbeing metrics standard for ethical artificial intelligence and autonomous systems. [2019–12–3]. https://site.ieee.org/sagroups-7010/files/2019/01/IEEE-P7010_WellbeingMetricsforA_IS_ShortPaper_December272018For_Submission_reviewedbyIEEELegal-1.pdf.
100. P7011—Standard for the process of identifying and rating the trustworthiness of news sources. [2019–12–3]. https://standards.ieee.org/project/7011.html.
101. P7012—Standard for machine readable personal privacy terms. [2019–12–3]. https://standards.ieee.org/project/7012.html.
102. P7013—Inclusion and application standards for automated facial analysis technology. [2019–12–3]. https://standards.ieee.org/project/7013.html.
103. P7014—Standard for ethical considerations in emulated empathy in autonomous and intelligent systems. [2019–12–3]. https://standards.ieee.org/project/7014.html.
104. National Technical Committee for Information Technology Standardization. Information technology—vocabulary—part 28: artificial intelligence—basic concepts and expert systems. GB/T 5271.28-2001. Beijing: China Standard Press; 2002.
105. National Technical Committee for Information Technology Standardization. Information technology—vocabulary—part 29: artificial intelligence—speech recognition and synthesis. GB/T 5271.29-2006. Beijing: China Standard Press; 2006.
106. National Technical Committee for Information Technology Standardization. Information technology—vocabulary—part 31: artificial intelligence—machine learning. GB/T 5271.31-2006. Beijing: China Standard Press; 2006.
107. National Technical Committee for Information Technology Standardization. Information technology—vocabulary—part 34: artificial intelligence—neural networks. GB/T 5271.34-2006. Beijing: China Standard Press; 2006.
108. National Information Security Standardization Technical Committee. Information security technology—technical requirements for iris recognition systems. GB/T 20979-2007. Beijing: China Standard Press; 2007.
109. National Information Security Standardization Technical Committee. Information security technology—Framework of biometric identification and authentication protocols based on trusted environment. GB/T 36651-2018. Beijing: China Standard Press; 2018.
110. National Information Security Standardization Technical Committee. Information security technology—technical requirements for fingerprint recognition systems. GB/T 37076-2018. Beijing: China Standard Press; 2019.
111. China Electronics Standardization Institute. White paper on artificial intelligence standardization (2018 Edition). (2018–01–24) [2019–2–4]. http://www.cesi.ac.cn/images/editor/20180124/20180124135528742.pdf.
112. National Information Security Standardization Technical Committee. White paper on artificial intelligence security standardization (2019 Edition). (2019–10) [2019–12–4]. https://www.tc260.org.cn/upload/2019-10-31/1572514406765089182.pdf.

Chapter 3
Introduction to Artificial Intelligence Security and Safety

Binxing Fang and Xiang Cui

3.1 Dialectical Relations Between Artificial Intelligence and Security and Safety

The rise and boom of Internet, cloud computing, big data technologies provide a glimpse into the promoting and restrictive roles that development and application of emerging technologies play in the society. On one hand, technological advancements boost social progress and reforms; on the other hand, application of technologies presupposes security and must meet safety standards. In most cases, however, security and safety issues are largely overlooked when a new technology emerges because users tend to enjoy the benefits first before noticing the accompanied security and safety problems. Artificial intelligence (AI), a nascent technology, may serve to improve security and safety, but may also work the other way around. Over these years, we have paid increasing attention to AI development, and highlighted its application in diverse fields like military affairs, transportation and finance. When AI advances to a certain stage, security and safety concerns are bound to emerge, and it is then that we begin to think how to rein this technology in and make it safe. In *The Social Control of Technology*, David Collingridge argued that "the social consequences of a technology cannot be predicted early in the life of the technology. By the time undesirable consequences are discovered, however, the technology is often so much part of the whole economic and social fabric that its control is extremely difficult" [1]. This point of view, widely recognized by science and humanities scholars, is often referred to as "the dilemma of control" or "the Collingridge Dilemma".

B. Fang
Guangzhou University, Guangzhou, China
e-mail: fangbx@cae.cn

X. Cui (✉)
Zhongguancun Laboratory, Beijing, China
e-mail: cuix@zgclab.edu.cn

© Publishing House of Electronics Industry 2025

B. Fang (ed.), *Artificial Intelligence Security and Safety*,
https://doi.org/10.1007/978-981-96-9263-7_3

All new technologies feature both the security empowerment effect and the security accompanied effect. The former means that a new technology is so powerful that it can be employed to ensure and enhance security of systems; the latter, however, indicates that a new technology is inevitably accompanied by security risks once it is applied to practice despite the developer's painstaking efforts to preclude security problems in the congestion stage.

The security empowerment effect of a new technology may work in two conflicting ways. It can empower attacks and add to the challenges of security protection; however, it can also empower defense and provide new solutions to security problems.

On one hand, new technologies can empower attacks. Internet is much like a "Pandora's box", which brings more convenience but meanwhile introduces new security threats. Cloud platforms, for instance, make cloud resources susceptible to denial-of-service attacks (DDoS) or Internet viruses; quantum entanglement invalidates traditional supervision of communication; big data that empowers user profiling makes the conventional data decentration techniques defunct in privacy protection. In this light, empowerment of attacks is a "derivative security" problem that accompanies the development of a new technology.

New technologies, on the other hand, can also empower defense, which is taken as a matter of course because security protection, a measure of defense in essence, entails support of new technologies. For instance, Internet of things (IoT) has been applied to perimeter protection and improved the defense capacity of cities; big data analysis techniques can be used to detect security problems in a network and increase its sensitivity to cyberattacks; cloud computing enhances work efficiency by fending off network viruses; blockchain, as a means against frauds and tampering, ensures credibility of systems in acentric contexts; quantum key distribution facilitates high-intensity one-time-pad encrypted communication …

Emergence of any nascent technology is inexorably accompanied by new security concerns. When a new technology is introduced to the public, it reaches maturity only after it is applied to practice and improves through "crowdsourced testing" on the market. In this process, the new technology, due to its immaturity and defects, is likely to breed security threats.

The security accompanied effect of new technologies is also manifested in two aspects. First, a new technology, because of inherent loopholes, may fail to actualize the anticipated functions. This is coined the "endogenous security" problem. The other manifestation of the security accompanied effect is that the vulnerabilities of a new technology, though without any impact on its own operation, can be manipulated by attackers to pose security threats to other fields. This is termed the "derivative security" problem.

The endogenous security problem emerges partly because a new technology, before reaching maturity, suffers from security vulnerabilities. Some of these vulnerabilities, fortunately, can be discovered and solved via optimization. However, if the vulnerabilities are so inherent in a technology that they cannot be removed through optimization, additional security defense and protection measures are required. One case in point is trusted computing. Trust computing relies on roots of trust (RoT), and the premise to implement RoT is that the owner of the RoT must be the user

of the system, so that the irreplaceable RoT can be used to ensure the system's safe operation. In cloud computing, however, the cloud service provider owns the cloud resources and provides the resources to the tenant for use only. In this case, the resource owner is not the user, and the owner of the RoT is not the user who needs security protection. Trusted computing is therefore not applicable to cloud computing services. Likewise, it is hard to apply quantum computing to communication due to the risk of quantum collapse: the communication will be cut off once the messages are intercepted. As a result, reliability remains the Achilles' heel of quantum computing that makes it largely impractical to build communication channels with quantum computing systems. In fact, the accompanied security problem is prevalent among all kinds of technologies. For example, systems on the Internet must make ports accessible when providing a service, which, however, will introduce an attack surface and increase the vulnerability. This is the endogenous security problem of Internet services.

The derivative security problem of new technologies, in essence, results from the flaws or vulnerabilities which do not affect the technology's normal operation, but can be leveraged by malicious attackers to cause harm to other fields. This is much like undisciplined students in a class whose unruly behaviors never affect their own normal life, but will disturb their peers. Examples of derivative security problems of new technologies are not hard to find. For instance, the fact that the IoT sensors can radiate information will not affect the normal operation of the IoT system, but may incur risks of information leakage. Near field communication (NFC) does not require compulsory handshaking to confirm whether the communication is initiated willingly and thus enables convenient communication. This mechanism surely does not affect the normal operation of NFC, but can be exploited by criminals to access the victim's mobile phone wallet at close range. A new technology, in its conception, is usually not expected to cause security threats, so such cases of new technologies empowering attacks are in fact manifestations of derivative security problems. In the same logic, as the impacts of the derivative security problems are on other fields, these problems are similar to the attack empowerment effect of new technologies.

As mentioned above, derivative security problems occur when the vulnerabilities of a new technology are manipulated by attackers to cause security problems in other fields. Artificial intelligence can also incur derivative security problems. For instance, AI systems are highly dependent on data inputs when making judgements. Driverless vehicles rely on correct inputs of information of the surroundings, and any error in the inputs, such as a failure to detect an obstacle on the road (when the sensor is damaged), will lead the vehicle to wrong decisions. The overdependence on the input information is therefore deemed as a vulnerability of AI systems. It does not affect the system's normal operation, but can be exploited by attackers. For example, one can tamper with the sensors such as the radar sensor of a driverless car to cause traffic accidents. Moreover, AI is a very special case in terms of derivative security problems. To be specific, unlike other technologies that undermine human safety only when their vulnerabilities are maliciously exploited, AI can put human life in jeopardy by "accidents" where no malicious attackers are present.

There are often sensational headlines like “robots attacked human” or “driverless cars caused accidents”, which are essentially problems caused by vulnerabilities of AI. Even in the absence of intentional attackers, an artificial intelligence actant (AIA) under normal operation has the “kinetic energy” to hurt human beings and thus is able to attack human “as it wishes”. As a strong self-taught learner, an AIA can get rid of human control and threaten human safety when it evolves, and the higher it evolves, the more threatening it becomes. Despite the stunning performance of AIAs, their “uncontrollability by human” is deemed a vulnerability on the part of human because controllability should have been the premise of AIA development. However, an AIA often breaks free from the human control when it evolves, which speaks to the powerfulness of AIAs and, in contrast, the human’s reduced capacity to rein AIAs in. This reveals the vulnerability in controllability of human designed AIAs, though this vulnerability is not utilized by attackers but by the AIA per se. It is thus fair to conclude that any self-evolving technology is likely to incur safety problems even in the absence of an intentional attacker who takes advantage of the technology’s vulnerabilities, but AI stands out from other new technologies because it provides “a key to self-evolution”. Therefore, it is important to make sure that AI develops in a “secure and controllable” manner.

To sum up, AI has both the security empowerment effect and the security accompanied effect. The former means that AI is strong enough to improve the defense capacity of systems and fend off attacks; the latter is a result of the inherent technical bugs or uncontrollability of AI: technical bugs may lead the AI system to errors, while uncontrollability may turn the system into a safety threat to other fields. Essentially, both AI-powered attacks and derivative security problems of AI are manifestations of security threats that AI poses to other fields, but their causes are different. AI-powered attacks take place when attackers leverage AI to launch more destructive attacks. In the case of derivative security problems of AI, the vulnerability of an AI system is taken advantage of, or its uncontrollability leads to destructive outcomes “by accident”, or the AI system acts in a destructive manner, thus threatening the security of other fields.

3.2 Framework and Categories of AI Security and Safety

3.2.1 Framework of AI Security and Safety

From the internal perspective, AI systems, like common information systems, inevitably show vulnerabilities that cause endogenous security problems, or AI security problems. The vulnerability of an AI system, once exposed, may trigger unintended security events. It should be noted that, however, an AI system can be manipulated by ill-willed criminals anyhow disregarding its vulnerabilities. More detrimental than security events or malicious utilization of AI is that many a country has started to produce AI-powered lethal weapons. Such military ambition and the

subsequent arms race may reduce nations to powder kegs. In the long run, mobile and destructive AIAs with strong learning capacities (smart robots, for instance) can evolve without human aid, and chances are that these actants will someday break man-made rules and put human life in danger. These are categorized as derivative security problems of AI, or AI safety problems.

From an external perspective, it is a popular belief that AI systems are built on mere AI technology. This is a myth because in the design, construction and use of an AI system, we not only need to consider the technical factors, but conform to laws, regulations, national policies, codes of ethics as well as industry standards, with normal security & safety assessment and emergency precautions in place. Specifically, the functions and use of AI systems must in accordance with laws and regulations, and it is stipulated in national policies that security-critical AI systems should be fair, transparent, explainable and traceable to reduce security risks and identify responsibilities when problems emerge. Moreover, local or international codes of ethics provide the fundamental principles that development of AI systems should follow; industry standards should be observed to improve the development efficiency of AI systems, reduce vulnerabilities and enable inter-system collaborations; security assessment should be in place to reveal endogenous security problems and potential legal or ethical disputes that AI systems or products may incur; security and safety precautions should be designed to enable timely response to emergencies when the AI system goes out of control.

In a nutshell, AI security and safety encompasses three subfields: AI for security, AI security (or endogenous security of AI), and AI safety (or derivative security of AI). The first is a manifestation of the security empowerment effect of AI, while the latter two embody the security accompanied effect of AI. When building an AI system, we should keep in mind multiple constraints aside from mere technical issues. Figure 3.1 shows the framework of AI security and safety and external connections.

3.2.2 AI for Security

AI for security encompasses AI-powered defense and AI- powered attacks.

In AI-powered defense, the defender expands the defense measures with the help of AI technology. For instance, defenders can leverage computer vision to boost the development of the smart security and surveillance industry, thereby improving the physical security strength of anti-theft safeguarding systems and anti-explosion security check systems. By using machine learning algorithms, defenders can build self-optimizing smart intrusion detection systems. They can employ deep learning and machine learning techniques to detect malicious codes with unknown features, and hence improve the network's capacity to identify security threats. They can also introduce network attack-and-defense knowledge graphs to achieve more efficient "human–computer collaboration" in threat-hunting and thereby spot hidden threats in key datasets. Defenders can also create anomaly detection systems built based on massive user and behavior data to find suspicious behaviors. All these examples

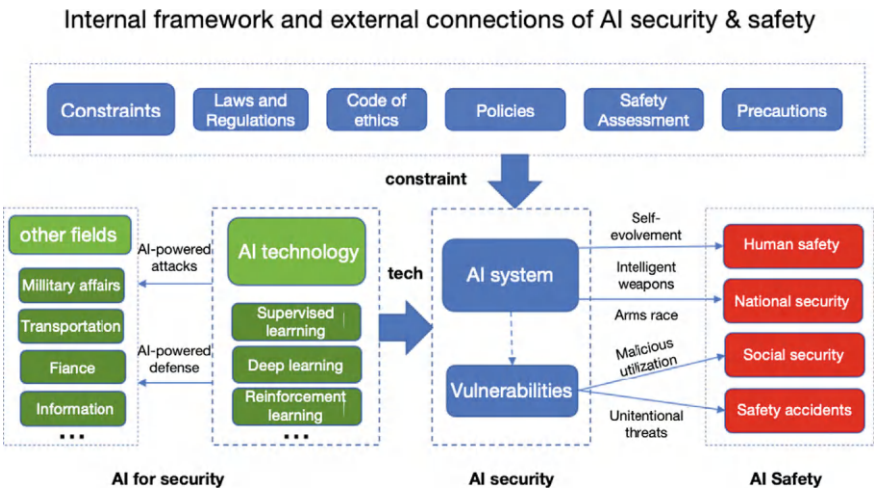


Fig. 3.1 Internal framework and external connections of AI security and safety

provide ample reason that AI-powered cyberspace defense systems can make up for the defects of traditional defense methods and provide new solutions to cyberspace defense.

In AI-powered attacks, the attacker breaks the limits of conventional attacks with the help of AI. For instance, using AI technology, the attackers can develop new AI-based attack methods like automatic vulnerability mining, smart malicious coding, implanting of backdoors into neural network models, automatic production of spear phishing mails, precise target positioning, deep concealment of attack intentions, and creation of high-fidelity tampered videos, so that they can improve the efficiency and reduce the cost of vulnerability mining, protect malicious codes from being detected, enable covert communication, pollute neural network supply chains, facilitate large-scale release of unattended spear phishing mails, enhance accuracy and intention concealment of attacks to networks, and actualize forgery-based deception and other new forms of attacks. AI-based network attacks embody a complete attack chain consisting of attack preparation, survival adversary, weapon delivery, target detection, intention concealment and network deception. AI enhances the attack capacity of each link in this chain, which poses new challenges to defense.

In addition, intelligent tools (hardware and software that leverages AI technology to boost performance of existing systems) can be maliciously used to launch terrorist attacks or mislead vox populi. Though intended to serve civil life in the design stage, these tools can sometimes be misused by extremists to launch terrorist attacks (flying bomb-aboard drones) or manipulate public views (releasing misinformation by Twitter Bot or Deepfake) to achieve their military or political ambitions. As it is difficult to preclude or wipe out malicious use of civilian intelligent tools and extremists are usually undeterred by laws, malicious use of intelligent tools will remain an inevitable problem. Moreover, the strength of AI technology in precise detection,

fast decision-making, agile group collaboration makes it a preferable solution for development of autonomous weapons.

Advances in AI-powered security will be detailed in Chap. 4 from the perspectives of defense and attacks.

3.2.3 AI Security

AI security problems derive from the inherent vulnerability of AI systems. The causes of such vulnerability are diversified, and any defect in the framework, components, data, algorithms, or models may introduce vulnerabilities to an AI system.

To start with, the fact that it is hard to ensure correct and transparent operation of the framework and components is an AI security problem. The framework (TensorFlow and Caffe, for instance) refers to the architecture that an AI system is built on. The framework is to an AI system what SDK is to Visual C++ or the basic dependent libraries to Python. Its importance is self-evident. Though there are multitudes of open-source AI frameworks and components that have been widely used across the world, their security remains to be fully tested, and there are risks of vulnerabilities and backdoors. Once an AI system built on an insecure framework is used in services critical to people's livelihood, the latent risks caused by the "unreliable fundamental architecture" will foment controversies or even arouse catastrophic consequences.

In terms of data, the lack of expertise to discern whether the data are correct is another AI security problem. AI systems fundamentally follow the intelligence endowed by human, and the way of endowment is learning. Correct learning relies on correct inputs of data, which are the premise of correct decision-making. Likewise, AI systems rely on input data to make inference and judgements, so AI systems are highly dependent on correct data inputs. That being said, the hypothesis that all input data are correct does not hold, and the obtained data may fall short of the ideal quality for different reasons. For example, missing data, and input of deformed or noise data will substantially disturb the AI systems.

In the case of algorithms, the fact that it is hard to ensure correctness of algorithms is also an AI security problem. It is fair to say that intelligent algorithms are the engine of AI systems. Existing algorithms adopt machine learning methods, that is, the system learn based on authentic and reliable data to generate patterns that machines can process repetitively. Fine examples in this regard are neural networks and knowledge graphs. Neural networks learn the known casual relations by "input–output" and log into the hidden layer of the network the summarized universal conditions for all learnt causal relations. Knowledge graphs extract semantic associations from the inputs to produce a relational model between realties and concepts, and provide support for creation of a knowledge base. Neural networks are like black boxes with strong prediction capacities, while knowledge graphs are like white boxes with strong descriptive capacities. The security defects of algorithms remain a crippling vulnerability of AI systems. For instance, adversarial attacks are a technique in

which the attacker takes advantage of vulnerabilities of algorithms, and many traffic accidents of self-driving vehicles can be attributed to immature algorithms.

Regarding models, the fact that it is hard to prevent models from theft or contamination is an AI security problem. By training a given algorithm on a massive quantity of samples, we can obtain a set of parameters that meet the requirements. Combine the algorithm with these parameters, and we will get an AI model. In this sense, an AI model can be considered a carrier of algorithms and parameters that exists in the form of an entity file. Since the model is a reproducible and revisable file, it is subject to such risks as theft and implantation of backdoors, and this remains a hot research topic on security of AI models.

Chapter 5 will expound on the vulnerability of AI from the perspectives of framework security, data security, algorithm security and model security.

3.2.4 AI Safety

AI safety problems refer to safety threats that AI systems pose to other fields because of their inherent vulnerabilities. There are mainly four types of AI safety problems: AI systems are susceptible to attacks due to their inherent vulnerabilities; failures of AI systems may result in safety accidents; development of AI weapons may lead to global arms race; and out-of-control AI systems will pose threats to human life.

The AI safety problem that AI systems may be attacked due to their inherent vulnerabilities bears a resemblance to the AI security problem that there are vulnerabilities in the AI system. These two are the two sides of the same coin. It is because the AI system has vulnerabilities that it may face attacks and then cause safety problems. For instance, an attacker can remotely control and accelerate a driverless car by taking advantage of its vulnerabilities in the car's software set. In this case, the presence of vulnerabilities in the car is an AI security problem, while that the car is attacked and results in speeding because of the vulnerabilities is an AI safety problem.

AI systems, because of immature algorithms or incomplete training data, usually show vulnerabilities that authoritative security evaluation may fail to reveal. Consequently, when put to work, these systems are very likely to malfunction and threaten human life. Now, latent safety threats posed by mobile and destructive AIAs are gaining increasing attention.

AI technology, with its enormous power, can empower research and development of weapons, which is a problem of AI-powered attacks and a manifestation of the empowerment effect of AI. Such empowerment, however, will run far beyond the range of development of weapons. Instead, in absence of an effective international convention to contain arms races among nations, AI-empowered development of weapons will put the world peace and the fate of humanity at stake. Thus, in this book, the risk of arms races incurred by development of AI weapons is classified as an AI safety problem.

An AIA, once equipped with the three elements of uncontrollability—having the capacity to act and damage, having the unexplainable decision-making capacity, having the capacity to evolve into an independent system (as will be detailed in Chap. 6), is likely to get rid of human control and become a menace to human life. The risk of AIAs going out of control, an AI safety problem, is undoubtedly a top concern we face when scaling new heights in AI development.

AI safety problems and solutions will be introduced at length from Chap. 6 in this book in terms of the following aspects: safety accidents caused by failure of AI systems, the risk of global arms races caused by AI weapons, the risk of AIAs going out of control and contributing factors, and independent AIA precautions.

3.3 Position of AI Security and External Connections

3.3.1 Correlations Between Security, Reliability and Controllability of AI

As President Xi Jinping pointed out on the 9th study session of the Political Bureau of the CPC Central Committee in October 2018, "we should strengthen detection and prevention of latent risks posed by AI, safeguard people's interests and national security, and develop secure, reliable and controllable AI technologies". It indicates that "security" is not an alternative to "reliability" and "controllability", and the correlations among the three are as follows.

"Security" of AI touches on both the endogenous security and the derivative security of AI; "reliability" of AI refers to the likelihood that an AI system will fulfill the anticipated missions without failure in a given time and context. Clearly, "security" and "reliability" are not inter-dependent: a secure AI system does not necessarily mean a reliable one, and vice versa. "Controllability" of AI means that an AI system must work under human control and effective precautions should be taken in case it goes out of control. Evidently, the focus of controllability is not whether the AI system is secure or reliable, but how to rein it in when risks emerge.

3.3.2 Relations Between AI Security and Laws, Policies and Standards

AI technology should advance under benign interactions with laws, policies and standards. Development of AI relies on protection by laws and policies as well as support of industry standards. In terms of laws, legislation should prioritize human's interests and curb malicious use of AI; with regard to policies, many AI-related policies should be put in place, including design guidelines for AI products, ethical censorship and security supervision provisions, regulations on fairness, transparency

and explainability of AI algorithms for critical scenarios, accountability systems for security problems, data source protection rules, and data privacy protection regulations. In terms of industry standards, manufacturers of AI products and parts should follow the standards and specifications of safe production to reduce vulnerabilities in their products. They also need to abide by AI security and safety assessment codes, and follow uniform inter-system communication protocols to improve security and compatibility of AI systems.

Even a 100% secure AI product may fall into the hands of the ill-willed. Laws and regulations can effectively deter the malicious use of AI and provide a safe environment for AI development. An AI product introduced to the market without strict security assessment will inevitably reduce to a security menace, while by releasing policies, we can establish market access thresholds to screen out illegitimate AI inventions. Safe and legitimate AI products designed not in accordance with a uniform set of standards and specifications will, like computers that are manufactured not in accordance with uniform TCP/IP protocols, lead the Internet to frustrating collapses. AI inventions designed in accordance with uniform standards are the premise to create a favorable environment for AI progress.

3.3.3 Correlation Between AI Security and Codes of Ethics

The relationship between security and ethics of AI has remained a seed of confusion and controversy. Imagine a driverless car detects a group of children jaywalkers ahead. Should the car keep moving forward, or take a sharp turn and plunge the unwitting and faultless car owner into danger, or hit the brake and run the risk of rear-ending? This is far more than a mere security problem. Indeed, the automatic driving algorithm must be designed to follow the codes of ethics, but some ethical dilemmas make it a tall order to establish a well-defined set of codes of ethics for the AI systems to follow.

Therefore, when designing AI algorithms, we should not only fulfill the anticipated functions and performance indicators, but make sure that the designs meet the codes of ethics, both local and global. In particular, when designing self-learning AI algorithms, we must consider how to train the AI systems to autonomously make decisions that conform to human ethics (robots should never hurt human, for instance), and to this end, it is imperative to embed these codes of ethics into the AI system.

Given analyses above, there must be an intersection between AI security and codes of ethics: codes of ethics must be incorporated into the design of self-learning AI systems to preclude AI safety risks.

3.3.4 AI Security and Safety Assessment and Precautions

To ensure AI security and safety, we need to consider how to minimize the inherent vulnerabilities of AI systems and prevent harms to other fields, i.e., how to improve the security and safety of AI from the individual or internal perspective. Security assessment and precautions, however, is to ensure security and safety of AI systems from an external perspective, or in other words, to improve the controllability of these systems to avoid potential safety risks that AI systems may incur. This is much like the fire protection in a building: banning stacking of explosives in corridors is a security measure adopted from the internal perspective, while regular checks by the building administrations are "security and safety assessment" conducted from the external perspective. That being said, even the strictest security and safety assessment cannot get rid of fire dangers once and for all, and thus "security precautions", such as firefighters and fire extinguishers that are immediately available when a fire occurs, are usually put in place.

In a nutshell, AI-related laws and regulations, policies and codes of ethics encompass a wide spectrum of issues that serve to improve the security and safety of AI (though they run far beyond this sphere). In comparison, AI security and safety assessment and precautions are solely intended to improve the security and safety of AI systems. For instance, laws can ban production or use of destructive robots, policies can be introduced to decree that "safety hoops" should be installed on destructive robots, and codes of ethics can force upon algorithms the human-first rule; security and safety assessment can rate the security of AI products from a technical perspective, and security and safety precautions are to provide alternative solutions should all the above measures fail.

3.4 Summary

This chapter has expounded on the dialectical relationship between AI and security and safety, introduced the security empowerment effect and security accompanied effect of new technologies, and detailed how these two effects are manifested in the nascent technology of AI. An AI security framework is proposed. From an internal perspective, the framework consists of three modules—AI for security, AI security and AI safety; from an external perspective, it concerns the constraints of laws and regulations, policies, and standards on security of AI. Also, the importance of regular security and safety assessment and precautions is presented to provide a holistic view of the position and significance of AI security for the readers.

Reference

1. Collingridge D. The social control of technology. London: Palgrave Macmillan; 1980.

Chapter 4
AI for Security

Peng Liao

4.1 Overview of AI for Security

The advent of a new technology is inevitably accompanied by new security and safety issues. As mentioned in Chap. 3, due to immaturity and inherent defects, a new information technology usually gives rise to two new types of security and safety concerns. One is with the system itself, caused by the inherent vulnerability of the new technology, which is coined the "AI security" problem. The other refers to the resultant safety issues in other areas, although the vulnerability of the new technology does not cause problems to the system itself. And we name this as the "AI safety" problem. In terms of AI security, it partly originates from some loopholes in the new technology, and most of the time these can be detected and eliminated; in another occasion, the new technology may have a natural defect, which makes the problems inevitable and unsolvable by improvement, and the problems can be only circumvented by taking other measures. AI safety, in essence, is the side effects that the vulnerability of a new technology inflicts on other fields. Since these issues do not have any impact on the new technological system itself, such vulnerability cannot be improved upon automatically by the new technology itself. For example, social networks do not implement the real-name system. Although this does no harm to the social networks themselves, it incurs safety issues in society. Actually, the attack empowerment effect of new technologies also presents characteristics similar to those of the AI safety problem, except that the powerful new technological system is misused to attack the security of other things. For example, at present, there are various novel methods for cheating on exams which are related to the evil utilization of new information and communications technologies. This forces anti-cheating techniques to be upgraded at the same time. From the perspective of security attacks, both the derivative security problem and empowerment of attacks

P. Liao (✉)
Zhongguancun Laboratory, Beijing, China
e-mail: liaop@zgclab.edu.cn

© Publishing House of Electronics Industry 2025

B. Fang (ed.), *Artificial Intelligence Security and Safety*,
https://doi.org/10.1007/978-981-96-9263-7_4

by new technologies contribute to attacks. Their only difference is that the former is caused by the vulnerabilities of new technological systems and the uncontrollability of the new technology itself, while the latter relies on the powerful capabilities of new technological systems and their definite application.

In recent years, we have witnessed the fast emergence and popularization of information technologies in a series of new fields, such as cloud computing, edge computing, IoT, AI, Industry 4.0, big data, and blockchain. Although these technologies have great potential and influence, they pose huge and inevitable challenges to security and safety. Likewise, AI will bring a number of safety challenges [1]. This chapter will illustrate on the topic of AI for security from two aspects: AI-powered defense and AI-powered attacks.

4.2 AI-Powered Defense

According to a report by Technavio [2], a leading global technology research and advisory company, the global AI-based network security market size is expected to grow at a compound annual growth rate of more than 29% from 2018 to 2022. Companies are applying cloud-based AI services to multiple network security applications including authentication, video management, biometric data storage, and big data computing, thereby enabling AI to play a crucial role in assisting and enhancing defense.

4.2.1 Physical Intelligent Security Monitoring

Physical intelligent security monitoring is an important technical measure to ensure physical security. It involves physical protection, anti-theft alarm, video monitoring, anti-explosion security check, entrance and exit control, etc. As a successful application field of AI technology, security monitoring has benefited a lot from AI. The related technologies and achievements have been noticed by numerous security companies in China who have begun to probe into AI from different aspects like technology development and product. Traditional enterprises such as Dahua Group, Hikvision, and NetPosa are continuously increasing the intelligence of their security products; technology companies specializing in algorithms like SenseTime, Megvii, Cloudwalk and YITU are focusing their effort on graph and image processing intelligence such as face recognition and behavior analysis. The *2019 China's AI + Security and Protection Industry Report* [3] released by iResearch Consulting Group pointed out that "AI + security and protection" is one of the main directions of AI commercialization with the fastest implementation of AI and the greatest market capacity. The key AI technologies that push forward the growth of security monitoring include intelligent video surveillance, posture recognition and behavior prediction, knowledge graph [4], and intelligent security robots.

(1) Intelligent video surveillance: It can practically recognize and track humans and conduct correlation analysis and data mining on various attributes of people. With AI, the time needed for video retrieval and checking, personnel locking and tracking is shortened from days to minutes or even seconds. Thus, real-time response and early warning is achieved for security monitoring.
(2) Posture recognition and behavior prediction: It recognizes human postures. Since each person has a different bone length, muscle strength, center of gravity, and motor nerve sensitivity, i.e., a different physiological structure, a person's gait is unique. Different from face recognition, posture detection can function within a distance of up to 50 m and with a recognition speed of less than 200 ms using an ultra-high-definition camera. In conducting security monitoring of people in public places, when the face of an object cannot be captured or the captured image is blurred, we can predict the upcoming actions of this person by recognizing his or her postures and effectively prevent any possible crime.
(3) Knowledge graph [4]: Google proposed this concept in 2012. In essence, the knowledge graph describes the entities or concepts in the real world and their relationships using multi-relational graphs (with multiple types of nodes and multiple types of edges). Through knowledge graphs, the big data on security and protection performs correlation analysis on the attributes of entities based on massive temporally and spatially multi-dimensional information. In this way, the ability to retrieve and analyze data and intelligence is improved.
(4) Intelligent security robots: There are many kinds of intelligent security robots. The intelligent security patrol robot conducts smart patrolling of the working area, with a variety of equipment for sensing and detecting images, infrared radiation, sound, gas, etc. loaded on the mobile security and protection system, and then transmits the detected data to the remote monitoring and control system. It can make judgments and decisions automatically by computer vision and multi-sensor fusion, and send out an alarm in time when discovering a situation. The intelligent firefighting robot can extinguish the fire in place of human firefighters by entering a dangerous site that is flammable, explosive, toxic, hypoxic or filled with dense smoke, where it collects, processes and reports relevant data and controls the fire. The vehicle chassis inspection robot can thoroughly and accurately examine the chassis and the space under the seats of various vehicles, locate and assist in ridding the vehicle of hidden suspicious dangerous objects. It is equipped with abilities to draw samples, grasp, remove and transfer things.

4.2.2 Military Robots

Military robots designed for defense also fall into the category of AI-powered defense. It is reported that HIT Robot Group (HRG) has developed the reconnaissance robot and the small EOD robot [5]. The reconnaissance robot can monitor the surrounding environment by a three-degree-of-freedom PTZ camera head and a

loaded camera, transmit voice, graphic and textual information, and is able to work normally under all weather conditions. The small EOD robot is designed according to the military standards. It can grasp suspicious objects weighing between 6 and 30 kg, has excellent mobility and speed, and is able to grasp and transfer explosives in various environments like in the water, in the wild and on stairways. These robots can be widely applied in different areas to ensure urban safety and public security, complete earthquake relief missions, detect targets and suspicious objects, remove bombs on streets, handle hazardous substances, and so on. In addition, HRG has developed an armed combat robot[1] that can automatically detect, search for and track suspicious targets in hazardous environments. It is equipped with an optical sight system for guns, can be installed with different firearms as needed, and can perform video surveillance, armed deterrence, and long-range attacks on terrorists, gangsters, poachers, and other criminals in situations and places like border patrolling, anti-terrorist actions in cities, hostage rescue, forest protection, and nature reserves. The armed combat robot can assist public security personnel and armed police in anti-terrorism missions, and help border defense officers and soldiers patrol country borders.

4.2.3 Intelligent Intrusion Detection

The traditional network threat detection methods based on feature rules can have a number of problems such as false and missed detection and long delays when dealing with complex network behaviors and massive high-dimensional big data. Network intrusion detection determines whether a behavior is normal or abnormal based on network traffic or the host data, which is a classification issue, while AI algorithms like machine learning algorithms are powerful in solving classification problems [6]. As a result, the intelligent intrusion detection optimized based on the currently fast-growing AI technology has significant advantages over traditional intrusion detection methods in terms of both detection capabilities and speed.

In the field of network intrusion detection, the AI^2 (Artificial Intelligence + Analyst Intuition) jointly developed by the MIT Computer Science and Artificial Intelligence Laboratory (CSAIL) and the startup company PatternEx in 2016 is a network security and intrusion detection platform based on AI. By combining machine learning, AI^2 can accurately predict more than 85% of cyber-attacks [7]. It automatically scans the log data using unsupervised machine learning and report the clustering results to human analysts. Then human analysts determine which cyber-attacks are real and constantly label the data of cyber-attacks, before feeding the results back to AI^2 by supervised machine learning, thus upgrading the existing model. According to the feedback loop, as the data increases over time, the detection rate can be improved continuously.

[1] http://hrg-srobot.com/product/slug-2madt1.

In the field of IoT intrusion detection, Aidin Ferdowsi et al. proposed a distributed intrusion detection system based on a generative adversarial network (GAN) [8] by combining the characteristics of IoT such as the wide range of devices and its complex network structure, enabling it to identify abnormal behavior on IoT without the help of any centralized controller. This method assumes that an IoT system is composed of several groups of IoT devices. By using the GAN, several sets of discriminators with normal IoT operation data are trained with historical data of normally-operating IoT devices, and these discriminators constitute a model set for the distributed intrusion detection system. Through experiments, Aidin Ferdowsi et al. concluded the GAN-based distributed intrusion detection system could increase the detection accuracy by up to 20%, the detection precision by 25%, and reduce the false alarm rate by 60%.

4.2.4 Malicious Code Detection and Classification

Malicious codes, including worms, Trojans, bots, ransomware, and spyware, are programs with malicious functions. There have been lots of studies, using machine learning, trying to detect and classify malicious codes from the perspective of source codes, binary codes and the runtime characteristics of programs.

In terms of malicious code detection, the main AI-based approaches are to extract the static characteristics of malicious codes, including the hash value of files, signature characteristics, API call sequences, and string characteristics, and construct the malicious code signature by combining the dynamic features of malicious codes during runtime such as CPU utilization, memory consumption, network behavior, and behavior in the host, and then apply deep learning or machine learning to automatically detect and identify suspicious malicious codes. Despite the available detection methods like static detection, dynamic detection, heuristic detection and virtual machine detection, the researchers hope to increase the existing detection efficiency by AI. The Tencent security team has developed the TAP-AI anti-virus engine [9] based on these key technologies such as actual runtime behavior, system monitoring, AI-based chip detection, AI-based model cloud training, and neural network algorithms. Compared with traditional engines, this engine has the advantages of resisting evasion of anti-virus detection, high performance, providing real-time protection and being able to detect malicious codes with unknown features. It can be trained automatically, so the time for malicious code elimination is shortened and the operating cost reduced.

Some researchers hope to reinvent malicious code detection in some specific areas by AI technologies like the deep neural network (DNN) and natural language processing. For example, Hendler et al. [10] put forward a DNN-based detection method for malicious PowerShell attacks. This method can extract malicious PowerShell character sets and construct feature vectors of the convolutional neural network (CNN), so as to train a CNN. At the same time, it processes PowerShell commands

as text, builds a logistic regression classifier with n-gram and bag-of-words models. Therefore, a mixed malicious PowerShell detector is achieved.

In terms of malicious code classification, to resist anti-virus detection, modern malicious codes are often designed into a detection evasion model with variable characteristics. That is to say, malicious codes have become polymorphic, resulting in a sharp increase in the number of malware variants. As the traditional malicious code detection methods are based on malicious code behavior and character characteristics, the false detection rate is high. This poses a new challenge to malicious code classification. In the Kaggle Microsoft Malware Classification Challenge (BIG 2015), the winning team "say NOOOOO to overfitttttting" [11] converted the malicious disassembly code file into an image to serve as the characteristics of the sample. The ingenious design cleverly transforms the malicious code classification problem into an image classification problem, thereby successfully resisting the anti-virus detection evasion. This is not only because the AI-based image classification technology is very mature and can distinguish images easily, but also because when the code is converted into images, it loses all the microscopic characteristics of the code itself while highlighting the overall programming style. Since the attacker usually does not understand how the image classifier classifies the codes by image recognition, it is impossible for them to evade the classifier by only modifying the code locally. They can only do it by completely changing the codes, but that would make the malicious code a brand-new one and the issue beyond the category of anti-virus detection evasion. The winning team then extracted features of the opcodes, like the n-gram and instruction frequency, and by using the machine learning method, their approach ultimately achieved an accuracy of 99.8% on handling the data set provided by the competition.

4.2.5 Threat Hunting Based on Knowledge Graph

Threat hunting is a human-dominated deep investigation process. As an active threat detection method that can be iterated at the level of active defense, threat hunting aims to discover hidden threats in key information assets [12]. The human threat hunting team, threat hunting tools, and data and knowledge are the three key and collaborating factors of threat hunting. The threat hunting team needs to process the collected data with automated threat hunting tools, and by conducting correlation analysis of the intelligence to obtain new knowledge, they get clues of the threat. One crucial step in threat hunting is to derive knowledge from the data and construct a knowledge base on cyber-attack behaviors by knowledge graphs, so as to provide effective guidance for users in threat detection, defense, and response. At present, the cyber kill chain model [13] and the adversarial tactics, techniques and common knowledge (ATT&CK) model [14] for cyber-attack behavior knowledge base construction that are widely applied in the cyber security industry provide guidance and support for intrusion recognition, instruction exploration and comprehension of cyber-attack behaviors.

The cyber kill chain model was proposed by Lockheed Martin Space Systems Company in the US in 2011. It is a targeted staged cyber security knowledge model described from the perspective of the attacker, including information collection, targeted weapon delivery, vulnerability exploitation, code execution, authority upgrade, and lateral movement/intranet penetration.

The ATT&CK model was summarized and developed by the MITRE Corporation based on real cyber-attack cases. It describes the life cycle of cyber-attacks by matrix diagrams and constructs a cyber-attack behavior knowledge base that is finer-grained and contains more sharable and expandable knowledge models and frameworks, in order to help the defenders to carry out detection and response work in a more targeted way.

Essentially, both the cyber-attack logical relations of the cyber kill chain model and the ATT&CK model and the entity relational analysis of knowledge graph can be used for threat hunting. For example, the basic elements of the cyber kill chain, which are staged tasks, can be decomposed into three parts: the attacker, the attacking method, and the attacked system. In this way, the relations between entities of the network system are illustrated, and the knowledge graph model of the network graph structure is created.

4.2.6 User Entity Behavior Analytics

User entity behavior analytics (UEBA) evolves from user behavior analytics (UBA). The "entity" here refers to assets or equipment, such as servers, terminals and network equipment. UEBA was proposed by Gartner in 2014. It is used to create the portrait of the normal behavior of users and entities, so as to detect anomalies and discover potential threats. Among the top 10 security projects in 2018 released by Gartner at the Gartner Security and Risk Management Summit in June 2018 [15], as many as four of them are related to UEBA. Gartner predicted that the threat detection and security incident response solutions with UEBA as the core technology will account for more than 80% of its security projects by 2022.

Machine learning constitutes an important part of UEBA. The core idea is to apply machine learning algorithms to learn historical data, construct a normal behavior model, and then recognize deviations from normal behaviors. For example, with machine learning, we can discern anomalies by analyzing small-probability events and go beyond the limitations of traditional analysis methods by rules. The effectiveness of the UEBA model can also be improved by training on deep learning platforms. At present, UEBA has begun to be used in data leakage detection, traffic anomaly detection and advanced persistent threat (APT) detection.

4.2.7 Spam Detection

Traditionally, spams are detected and filtered out mainly by setting rules at the mail server end, and the rules or black and white lists are set by configuring the IP address/IP network segment, email domain address, email address, subject or keywords etc. at the sender end. This method can only detect existing spams, the rules cannot be updated promptly, and the detection efficiency is low. By using AI technology, we can allow the rules to update themselves automatically, effectively overcoming the deficiencies of the traditional spam detection approach. Currently, the mainstream solution is to classify the texts of emails by machine learning algorithms. In June 2017, Google said that its spam and phishing email recognition rate using machine learning had reached 99.9% [16]. Furthermore, according to the official statistics of Google, among the emails received by Gmail, 50–70% of them are spams. This provides an inexhaustible source of high-quality data for the training of machine learning models, which, coupled with Google's algorithmic and data processing capabilities, has helped to achieve the real-time update of the spam detection model.

4.3 AI-Powered Defense

4.3.1 Automated Cyber-Attacks

Automated cyber-attacks driven by AI are a popular topic nowadays. Hackers are improving their weapons by actively introducing AI into their attacking tactics, techniques and strategies and automating cyber-attacks. At the Black Hat conference in the US in 2017, when Cylance, a cybersecurity company, sought the opinions of 100 information security experts, 62% of them believed that hackers would empower cyber-attacks with AI in the future [17].

(1) GAN-based online verification code cracking: In December 2018, an AI system jointly developed by China's Northwest University and Peking University and the UK's Lancaster University [18] could crack the Captcha online verification code in 0.5 s. The innovative point about this achievement is that it generates the training data using GAN. This system does not have to collect and label millions of Captcha samples, but requires only 500 Captcha samples for learning. These samples are then used to synthesize millions or even billions of training data, which successfully trains the Captcha online verification code image classifier.
(2) Automated spear phishing: At the 2016 Black Hat conference in the US, John Seymour and Philip Tully of the cybersecurity company ZeroFOX demonstrated a Twitter-based end-to-end automated spear phishing method [19]. They proposed a recursive neural network based on social network automated phishing with reconnaissance (SNAP_R) which can learn to send phishing pushes to specific users (attack targets). This model uses spear phishing data for

training. To improve the click rate, it even dynamically embeds the interested topics extracted from the target users and their followers in the tweets and tweets at the targets when sending the tweet. After doing the test on 90 users, it was found that the success rate of this attack method was 30–60%, while the traditional net phishing can only achieve a success rate of 5–14%. This indicates that with AI technology, automated spear phishing is more accurate and massive.

(3) Machine player in the Capture the Flag (CTF) cybersecurity games: In the DEF CON CTF held in Las Vegas, USA in August 5–7, 2016 [20], which is a top information security event in the world, a machine team called Mayhem competed against 14 other leading human CTF teams. It was the first information security battle between machine and human hackers, where the machine team surpassed two human teams for some time. As machines could participate in the DEF CON CTF, it signaled a new era of "machine intelligence" and "automated attack-defense". The history of Mayhem can be traced back to the Cyber Grand Challenge (CGC) organized by the Defense Advanced Research Projects Agency (DARPA) of the US. CGC was a global cybersecurity challenge initiated by DARPA, which aims to promote the technological development of automated cyber defense, i.e. identifying system defects and vulnerabilities in real time and automatically performing the patching and system defense. The participating teams were all computers without any human intervention. That is to say, CGC is a CTF game between machines and its purpose is to achieve a fully automated cybersecurity attack-defense system. CGC is a completely new exploration into cyberspace attack-defense in the stage of AI. The automated defense system will significantly reduce the time difference between attack and defense, and will be a development trend in cybersecurity. The theme of the DEF CON held in August 2016 was also determined as "the rise of the machines." The CGC has encouraged researchers to further develop software repair robots which can continuously scan system defects or vulnerabilities faster and more effectively than humans, and upgrade their ability to quickly fix these program defects in billions of lines of code.

(4) Automated penetration testing: Penetration testing is to assess the security of the target network and system by simulating real cyber-attacks. As a method for discovering the vulnerability and loopholes of the target network and system from the perspective of real-world attack-defense, it is one of the most effective ways to conduct cybersecurity assessment. By comprehensively applying social engineering and technologies, authorized penetration testing engineers try to penetrate into to the target. They collect information, deliver weapons in a specific direction, exploit system vulnerabilities, execute code, upgrade their authority and move laterally or penetrate into the intranet from the perspective of the attack chain.

At present, penetration tests are done by relying on the personal experience and technical skills of penetration testing engineers, and penetration testing is part of cybersecurity services. The manual penetration testing mainly has the following problems:

(1) The penetration test quality is not stable, which is influenced by the skills, time and tools of the engineer, and their familiarity with the IT architecture of the target;
(2) The testing process is very repetitive. The tools used in different stages, from information collection to lateral movement and intranet penetration and to the end of the penetration, operate independently of each other without cooperation;
(3) The penetration testing experience is not formatted. The testing process has to be repeated afresh when the target system is changed, even though the problem is of the same sort.

Although cybersecurity agencies and research institutes from home and abroad have started working together on automated penetration testing research and attained some achievements. We still have a long way to go before realizing intelligent automated penetration testing.

The Metasploit Framework (MSF) is an open source project of Rapid7, a global security and vulnerability risk management solution provider. Since it was publicly released in 2003, it has integrated the common vulnerability exploiting code, scanning tools and malicious code from several mainstream platforms, and has achieved semi-automatic vulnerability detection by interface call. Based on Bash scripts, Discover[2] can automate the common penetration testing tools to discover hosts, scan ports, enumerate services and identify systems. Discover can call the penetration tools under Kali by Bash scripts, including the port service scanning tool Nmap, the Metasploit Framework, and the passive information collecting tool the Harvester. Discover enables penetration testing engineers to implement automated information collection under Kali, which simplifies the basic penetration work.

In the Black Hat USA conference in 2014, a tool called Heybe was released [21]. Heybe performs automated penetration testing by connecting with multiple attack modules. It is said that Heybe can complete the testing of all the target systems within a few minutes. Heybe quickly discovers tools by the host Fener and locates all the hosts of the target network; it scans specific loopholes of the target by the automated vulnerability scanning tool Kevgir; it automatically generates and delivers weapons by the social engineering tool Sees; it discovers and automatically tokes over Windows domains by Kacak; it automatically recognizes sensitive information by the traffic analysis tool Deppep; it captures and cracks password hashes by the password cracking tool Cilingir; and it cracks common protocol passwords by the brute forcing tool Levye. Heybe correlates all the achievements in various stages of penetration testing. In essence, it is an expert penetration testing system.

In the Black Hat USA conference in 2018, a tool called DeepExploit was released [22], which realizes automatic penetration by deep reinforcement learning. The intelligence of DeepExploit is shown in its ability to accurately identify and exploit system vulnerabilities based on past penetration testing experience. DeepExploit does not require training with data, because it can achieve rapid self-evolution by

[2] Discover: An automated penetration testing tool based on Metasploit.

distributed multi-agent reinforcement learning. At present, DeepExploit can automatically collect intelligence, perform threat modeling, analyze vulnerabilities, utilize vulnerabilities, perform post-penetration tests and generate reports.

4.3.2 Empower Cyber-Attacks by Increasing Attack Efficiency

AI technology can greatly increase the automation of writing and distributing malware, and by inserting some adversarial samples, can help them bypass the detection of security products. It can even enable the malware to automatically change its code and signature in each iteration by utilizing the detection logic of security products. AI technology can also keep the malware's functions unchanged after automatically altering its code to circumvent the detection of anti-virus products. In the paper "Generating Adversarial Malware Examples for Black-Box Attacks Based on GAN" authored by the scholars of Peking University [23], MalGAN, a GAN-based algorithm for generating adversarial malware examples, was proposed, where an alternative detector is used to adapt to the malware detection system against black-box attacks to make the generated malicious code bypass the machine learning based detection system.

In addition, AI technology can generate intelligent botnets for scalable attacks. The Fortinet FortiGuard Labs 2018 Threat Landscape Predictions published by Fortinet [24] indicated that in the future, AI technology would be massively applied in hivenets and swarmbots. Millions of interconnected devices or swarmbots would be used to simultaneously identify and respond to different attack vectors. Therefore, using the self-learning capabilities, vulnerable systems would be attacked autonomously and on an unprecedented scale. Different from traditional botnets, such hivenets are networks built with AI technology. They can communicate and exchange information with one another internally and directly execute commands based on the shared local intelligence from the nets; different from botnets, they do not require commands sent from the control end. When the swarmbot recognizes and invades into more devices, it will multiply and expand its ability to attack multiple targets at the same time. Essentially, this implies that the devices of the Artificial Intelligence of Things (AIoT) can be manipulated to actively conduct scalable and intelligent attacks on vulnerable systems.

4.3.3 Empower the Spread of Harmful Information

Intelligent customized recommendations that are integrated with AI algorithms can analyze and predict the interests, hobbies, behaviors and habits of users based on data like the browsing records and transactions of users, and recommend content to them

according to their preferences. Intelligent recommendation may be used to spread negative information and disseminates false and illegal information and comments in a more targeted and covert way. As the negative impact of such information is amplified, the possibility of revealing and reporting the source of the spreading is reduced. The device-to-cloud cybersecurity company McAfee said that more criminals would try to analyze private information using machine learning, in order to identify potential easy-to-attack targets and populations. Then they deliver customized phishing emails by intelligent recommendation algorithms to increase the accuracy of social engineering attacks.

4.3.4 Empower Fabrication of Misinformation

AI technology can be used to carry out illegal activities like committing frauds. Given enough training data, AI technology is able to produce artificial recordings with voice as genuine as the original one. It can synthesize real-looking images based on text descriptions, construct 3D models based on 2D images, and even can modify the expressions and mouth movements of characters in the video and produce consistent audio and videos according to the sound clips. In the IEEE Conference on Computer Vision and Pattern Recognition in 2016, Justus Thies et al. of the University of Nuremberg in Germany [25] released the Face2Face technology, which can realistically copy a person's facial expressions and facial muscle movements when talking onto another one in the video in real time, and this is the first model for real-time face transformation. In the ACM SIGGRAPH conference in 2017, Supasorn Suwajanakorn et al. of the University of Washington [26] published a project on "synthesizing Obama". In this project, by using the historical speech materials of President Obama, a high-quality, realistic but fake speech video by Obama was synthesized. The main contribution of this research is a technology that synthesizes realistic lip movements based on the voice. Usually, to make a fake video like the "synthesizing Obama" requires training and modeling with large amounts of videos on the specific target. Therefore, Zhou et al. [27] came up with a method of generating high-resolution and realistic speech videos by dissociating auditory and visual information. The input of this system only requires a photo and a sound clip, without the need to model the human face in advance. The research on face transformation attacks focuses on the head or face. In August 2018, Caroline Chan et al. of the University of California, Berkeley published [28] a paper entitled "Everybody dance now", where a method of transferring movements between videos on different people. First, a video on the target person that is to be synthesized is provided, and the other video provides the source of actions. Then by the end-to-end AI pipeline, the movements of the characters in the two videos can be transferred. Furthermore, an application on fake dancing was released, which enables anyone to generate fake videos showing them being an extraordinarily skillful dancer. This research extends face transformation attacks to the whole body.

At present, criminals may use images, audios and videos synthesized by AI technology to commit frauds. According to a report by in the Wall Street Journal, believing that he had just talked with his boss (who is the CEO of the German parent company) on the phone, the CEO of an unnamed British energy company deposited 220,000 Euros in the bank account of a Hungarian supplier as required. However, this voice was not from his boss. Instead, it was synthesized by AI voice technology, although the deceived CEO claimed that he could detect the subtle German accent in the voice of his boss [29]. Actually, the chat bot displayed at Google's I/O Developer Conference is fluent, natural and logical in interacting with humans on the phone. It has completely been able to deceive humans as another human.

4.3.5 Intelligent Malicious Codes

4.3.5.1 Malicious Codes Able to Conceal Targets and Intent

Conventional malicious codes, once released, have clear attack targets and intent recognizable by reverse engineering and network monitoring. However, aided by AI, malicious codes can conceal the attack targets, intent and high-value payloads in the context of open sources, making the attacks more elusive.

On Black Hat USA held in August 2018, IBM rolled out DeepLocker [30], its new AI-powered malware which could realize accurate target recognition and launch stealthy attacks by dint of AI.

(1) Accurate target recognition: it achieves accurate recognition based on the target's face, voice and behavior patterns;
(2) Stealthy attack payloads: the key is produced only when the target occurs to decode the attack payload and defend against artificial reserve analysis.

Here we introduce the target recognition, locking and unlocking techniques related to DeepLocker.

Target recognition: DeepLocker detects targets based on a DNN. When a target occurs, feature vectors are generated based on the target's face, voice, behavior, sensors, location, physical and soft environment as inputs into the DNN to produce keys which are used to decrypt the attack payload; the keys are not hard-coded, but in the form of ciphertexts.

Locking and unlocking: when designing the malicious codes, the attacker first produces keys using the target's features, encrypts the attack payloads with the produced keys, and hard-codes the trained DNN model and encrypted payloads into the malicious codes to lock and hide the attack payloads. Once released, the malicious codes detect the environment of the target via the DNN; when the target is detected, the target's feature vectors are taken as the key (a hidden layer) to decode attack payloads and implement attacks. Take for example a facial recognition system. When a face is detected, the system produces high-dimensional facial features which are then converted into keys; if the keys are correct, it means the facial features match

the target features, and then the malicious codes will unlock the attack payloads. As unlocking is available only when the target is detected, the defender cannot create proper samples as effective inputs even if the DNN source code is obtained.

DeepLocker distinguishes from similar technologies because it uses the DNN, a black box model. It converts the simple trigger "if this, then that" into a DNN computing process so that the malware analyst has no idea what target, whom and how the malicious codes want to attack.

4.3.5.2 Swarm Intelligence Hive Network

Mainstream botnets are now facing challenges. For starters, command control protocols of these botnets cannot meet such security requirements as penetrability, confidentiality, robustness, bidirectional communication and anti-traceability. For instance, IRC/HTTP botnets suffer from the single point of failure (SPOF) problem and poor robustness; P2P botnets are often difficult to monitor, suffer high latency in message control and poor penetrability. On the other hand, existing botnets are largely used for distributed denial of service (DDoS) attacks, spam mail sending, cryptocurrency mining and hijacking, crypto-blackmailing, etc., but cannot back up intelligent attacks like automated attacks, precise attacks, and independent decision-making attacks. Botnets, focusing merely on the fundamental functions, usually cannot support intelligent attacks.

Swarm intelligence hive network was first proposed by Drek Manky, an expert with Fortinet, in 2017 [31]. Derek foretold that botnets would evolve into hivenets and concluded the following features:

(1) Hivenets have a decentralized structure. As with conventional P2P botnets, hivenets have no super nodes, and thus failure of one single node will hardly have a bearing on normal operation of the whole network.
(2) Hivenets realize self-organization based on swarm intelligence, share information and make decisions via interaction between the nodes and the local environment. The hivenet realizes self-organization because it removes the command control channel in conventional botnets and dispenses with human control of the attack chain, leaving target locking and attacks to the swarm's automated decision-making.
(3) Zombies, also termed bots, in hivenets are equipped with automated attack tools and advanced AI technologies (face and voiceprint recognition, for instance) that will provide a strong support for attack tasks.

Studies on swarm intelligence hive networks are rare, and research on botnets powered by swarm intelligence and AI is still in its infancy.

In 2014, Aniello Castiglione with his team proposed a botnet-based command and control approach relying on swarm intelligence [32]. The self-organizing mode this approach adopts, i.e. a stigmergic communication mode in multi-agent systems, can build complicated and seemingly-intelligent structures in the absence of control, ensuring spontaneous collaboration and information sharing among independent bot

nodes. Meanwhile, the approach uses ant colony optimization (ACO) for networking and creates multiple short-range hops based on the dynamically-generated degree-constrained minimum spanning tree to transmit the control commands optimally to other nodes. Aniello's command and control approach is in essence a P2P structure. Evaluation tests show that the approach can adapt dynamically to fast-changing networks, enhance the fault tolerance, robustness, scalability and survivability of the command and control channels.

Moisés Danziger and his team have put forward a theoretical model for design of intelligent botnets [33]. The underlying idea is creating machine learning-based intelligent bots that make decisions based on the context to fulfill assigned attack missions. An intelligent bot consists of multiple agents which can be divided into four categories by the function:

(1) Super agents. Super agents are the core of an intelligent system and where the "intelligence" derives. Super agents are equipped with decision-making models boosted by reinforcement learning that can summarize information of the attack surface and environment to facilitate completion of missions;
(2) Recon agents. The recon agent detects the surrounding environment (the operating system, open ports, etc.).
(3) Defense agents. The defense agent protects the intelligent bots from detection and attacks;
(4) Attack agents. The attack agent exploits vulnerabilities, initiates cyber deception and launches other attacks.

These intelligent agents collaborate and constitute a small centralized botnet, in which the super agent takes control, while other agents communicate directly with the super agent and supports its decision-making.

The major features of Moisés' model is the absence of the command and control channel. In the model, the bots make decisions and take actions independently, making the botnet self-organized. As a result, the botnet cannot only circumvent security detection based on C&C communication, but haven itself from the risk of being traced.

What's worth noting is that in June 2017, the U.S. Defense Advanced Research Projects Agency (DARPA for short) has initiated a hivenet-related program—Harnessing Autonomy for Countering Cyberadversary Systems (HACCS) [34], in which the vulnerabilities in the cyberspace and the maliciously controlled nodes are called collectively as the "gray space". A gray system in the gray space can be a bot, a remote access Trojan, a backdoor or a proxy that can be exploited by multiple attackers. The gray space breeds uncertainties to cybersecurity. The goal of the HACCS program is to detect gray systems in the IPv4 cyberspace, exploit n-day vulnerabilities to insert autonomous agents (AAs) that will automatically detect and penetrate into infected devices in the intranet to safely neutralize the malicious codes and thus protect the gray space from malicious attackers (who will use the gray system as a proxy, initiate DDoS, steal, mine or hijack information).

4.3.6 Neural Network Backdoors

In the foreseeable future, publicly released machine learning models and neural networks will gain more popularity, and trained AI systems will soon be as available as daily consumer goods. They can be released, shared, retrained and resold, exposing more vulnerabilities to the attackers who can download available neural networks to contaminate the models. A model with implanted backdoors can produce normal outputs when the inputs are normal, but will recognize the input as a disguised target by the attacker when the input has a backdoor trigger. Therefore, attackers can re-release a neural network with backdoors and transfer it, incurring latent dangers for AI applications (facial recognition systems, pilotless automobile, for instance).

Yingqi Liu with his team [35] has designed a method to attack neural networks, and confirmed the feasibility and practicability of injecting Trojans (strictly speaking, backdoors) into neural networks. This method dispenses with training datasets of the original model and can launch attacks by using available neural networks. The attacks are implemented in three steps:

(1) Generate a Trojan trigger. The Trojan trigger is a set of special variables that can trigger the Trojans in a neural network. The trigger masks will label which input variables are to generate the trigger and assign the maximum values to the selected neurons so that the neurons and the trigger can be strongly connected and the trigger can activate the neurons forcibly;
(2) Generate the training data. Generating the training data means that for a given output label, an algorithm is generated based on the training data to produce input that can enable activation of the labels with high confidence. The purpose is to output a dataset that can be used to retrain the neural network;
(3) Retrain the model. The trigger and training data obtained in the previous two steps are used to re-train the hidden layer and output layer of the selected neural network; the new model, in the absence of a trigger, runs normally, and will output camouflaged targets when there is a trigger.

By experiments, Liu and his team tested their method on five AI applications: face recognition, voice recognition, age recognition, sentence emotion analysis, and pilotless driving. The test results revealed that their attack method achieved a 2.35% false alarm rate and an accuracy of 96.58% on the original dataset, but reached an accuracy of 97.15% on labelled external datasets, which confirmed the effectiveness of their proposed attack method.

4.3.7 Adversarial Machine Learning

Machine learning-based algorithms are commonly applied to the field of network security detection, e.g. support vector machine (SVM) algorithm-based malicious code detection, X-means clustering algorithm-based botnet detection, Bayesian

network-based spam detection, as well as hierarchical clustering and random forest-based malicious traffic detection. The accuracy of machine learning detection models is mainly dependent on example training data distribution, and the premise for preferable machine learning detection models is that its example training data set is representative. However, some previous researches show that machine learning-based detection systems are susceptible to adversarial attacks, for attackers may fabricate "benign" examples to bypass the identification of a machine learning classifier—this attack method is called "adversarial machine learning". In 2016, Xu et al. [36] proposed a genetic algorithm-based adversarial machine learning method which is capable of bypassing the identification of a machine learning classifier under the premise that its malicious act is retained, and label itself as a "benign" example. During the process of malicious examples generation, to retain the malicious act function of the example, Xu's team adopted the genetic algorithm, to combine "benign" examples with malicious example variations to generate a new example, and eliminate those "defective" new examples that do not retain their malicious act functions, while those new examples that are labeled as "malicious" continue to combine with "benign" example variations to generate new examples. This cycle repeats time and again, until the "malicious" examples are identified by target classifiers as "benign". During validity evaluation, they chose two machine learning-based PDF malware detectors (PDFrate and Hidost) as the targets they need to bypass. With 500 malicious PDF files as an experimental data set, they employed the genetic algorithm to make the malicious PDF files to go on "evolution" into benign examples, and ultimately attained a 100% attack success rate, i.e. all malicious PDF files are identified as benign examples.

However, the method proposed by Xu's team has some limitations, and whether it can work to obtain valid adversarial examples is dependent on the classifier's real-time feedback of the credit scores of the new examples that will serve as a basis for optimizing the quality of the examples generated later. Therefore, the defenders just need to hide their classification credit scores to defend these attacks. In actual deployment, it is also unlikely for the black box system to have a feedback of its classification credit scores. For instance, the malware detector built in the e-mail services system also only works to have a feedback of the final decision, i.e. to accept or reject the examples.

Therefore, to explore an attack scenario of higher binding force and reality, Dang et al. [37] proposed a method to bypass security detection in the case of black box—EvadeHC, under the premise that the attackers have no knowledge of the target system. The EvadeHC method makes use of three black boxes to build up models:

(1) Detector (or classifier): providing only the feedback of the final decisions (to accept or reject the examples);
(2) Tester: deciding whether the input example retains its malicious act functions; and
(3) Deformer: using random seeds to modify malicious examples.

The EvadeHC method proposes a scoring mechanism, and based on the dichotomy results of a detector and a tester, we can demonstrate the attack progress with the

actual values for the examples. Specifically, the dichotomy results of a detector are to "accept" or "reject", while the dichotomy results of a tester are "malicious" or "benign". This method establishes a variation path for malicious examples, and allocates values for examples on the path through the variation distance (or variation times) information. When the "reject" distance is less than the "malicious" distance, the examples with malicious act functions accepted by the classifiers are the ultimate adversarial examples to be ultimately obtained.

The EvadeHC method and the method proposed by Weilin Xu, et al. share the same data set for verification, and both obtain a 100% attack success rate in PDFrate and Hidost evaluations, but the EvadeHC method is superior to Xu's method mainly in that:

(1) The adversarial attack method of EvadeHC is not dependent on classification scores, but to self-adaptively generate adversarial examples through the black box;
(2) EvadeHC synthesizes the results of the detectors and testers that show the adversarial progress to some extent; and
(3) EvadeHC exploits the black box deformer only to convert malicious examples, thus increasing the applicability of this method to other machine learning based systems.

In 2017, Hu et al. [38] proposed a MalGAN model that bypasses the detection of machine learning based systems, which generates adversarial examples based on generative adversarial networks (GANs). This model proposed a method to substitute the detector, for it learns the detection results of the detection system, thus to fit the detection method of the detection system. This model consists of a generator and a substitute detector, both using feedforward neural networks. Malware is represented in the form of feature vectors, and the generator generates adversarial examples based on its input, which are to be mixed with benign examples. The mixture is then sent to the detection system for labeling. The labeled examples are used to train the substitute detector to fit the detection system, and obtain the gradient information that is to be fed back to the generator. This method works to solve the problem of no feedback score to be provided by the detection system. This model is distinctive in that the adversarial examples are dynamically generated based on the feedback of the detection system, while most previous algorithms make use of static gradient based methods to generate adversarial examples. Experimental evaluation of this model shows that the attack success rate for bypassing the target detection system may reach 100%.

In 2018, Anderson et al. [39] proposed a reinforcement learning based framework—EvadePE that is used to automatically bypass the static checking and killing of an anti-virus engine, and provides a concrete realization method, taking Windows PE as malicious software as an example. To generate valid adversarial examples, Anderson et al., taking 100,000 malicious and benign examples as the input, trained an ACER (actor-critic model with experience replay) agent, to automatically transform some specific malware examples. The action space of the agent includes a set of PE files conversion actions, including but not limited to "converting the name of

section (section in the file)", "creating new (not to be used) section", "creating a new entry point, and immediately skipping to the original entry point", "converting debugging information" and "adding a shell". The action only works to choose those transforming actions that will not destroy the PE files. The reward functions of agent are measured by the detection system. If the modified examples are determined as "malicious (no escape)", the reward will be 0. If they are taken as benign (escape), the reward will be 10. And then the reward and state are fed back to the agent. Through such iterative learning, actions that facilitate escape of malware can be identified. As a result, black box attacks can be launched on the detection system and adversarial examples can be automatically generated, which threatens the static checking and killing paradigm of the current antivirus software. Anderson's team uploaded the agent-transformed examples to VirusTotal, and found that compared with the detection rate for original examples, the mid-value detection rate was dropping, which means that some of the transformed examples had bypassed the static checking and killing of the antivirus engine.

4.3.8 Smart Tools Applied to Malicious Attacks

With smart tools widely applied and getting popular, attackers are enabled to go on new attack paths, like using audio–video smart synthesis software to conduct spoofing attacks and conducting physical contact attacks with unmanned aerial vehicles (UAVs); other mobile, controllable, bomb-loadable and anonymously accessible smart tools similar to pilotless vehicles can also be maliciously employed by terrorists and other ill-intentioned attackers.

Audio–video smart synthesis software tools mainly include FaceSwap [40], FakeAPP, Deepfacelab, ZAO and so on, whose audio–video smart synthesis technology is generally called "DeepFake", a portmanteau of "deep learning" and "fake". This technology, boosted by neural networks, converts the images or videos of one person into that of another, which usually combines some images or images available and overlay them on the original images or videos with AutoEncoder and a GAN. Compared with the well-known Photoshop that is able to replace the image of a person with that of another person, the DeepFake technology can replace a person in a video with another. DeepFake is built on GANs, and when it is running, the GAN algorithm allows two neural networks to compete against each other: The first is called a "generator" which generates examples as authentic as possible; the other is a "discriminator", which compares examples from the generator with the "original" to tell them apart. Based on the "conflicting" results, the generator will adjust the parameters or network structures used in the algorithm, until the discriminator cannot tell the original and the fake apart, i.e. until reaching "success". Although the ultimately synthesized videos or pictures are fake, they look true enough to be taken as genuine. DeepFake caught the initial attention on the Reddit by the end of 2017, when it was used to substitute the heroine in a porn video with the face of a celebrity [41]. It is not so difficult to launch a fake attack maliciously with some DeepFake tools, for

they, together with relevant materials, are easily accessible online. DeepFake tools thus have become the focus of wide attention, for they can be maliciously applied to make celebrity porn videos, fake news, practical jokes, and financial frauds.

Physical contact attacks with UAVs expand the attack methods by dint of the flexible and mobile UAVs. In 2016, at Black Hat USA, some researchers from Bishop Fox demonstrated their "Danger Drone", a UAV that can conduct penetration test assignments in the air [42]. With some hacker penetration test kits aboard, it can get close to the target, to conduct near-field communication attacks with Bluetooth, WiFi, ZigBee wireless sensing, and radio frequency identification. In August 2018, the president of Venezuela was attacked by an UAV when delivering a speech [43]. This vehicle approached the president who was speaking, and then ignited the bomb aboard. In September 2019, numerous drones attacked two sites of petroleum facilities of Saudi Aramco, including the largest petroleum refinery in the world located in Abqaiq [44].

4.3.9 Aiding R&D of Autonomous Weapons

In November 2017, representatives from more than 70 countries attended the conference of the *Convention on Certain Conventional Weapons*, where Stuart Russell, a professor and AI expert at University of California, Berkeley, played on the spot an appalling video (actually, a short Sci-Fi film) [45]: a swarm of "killer bee"-like UAVs are swarming into a classroom, doing facial recognition to locate and kill a group of students one moment in a blink of an eye. These "killer bee" UAVs are flying and avoiding obstacles on their own, and with processors hundred times faster than our brains, they have strong counter-sniper capabilities. With various sensors and wide-angle cameras built in to support facial recognition and 3 g mini-explosives aboard, they can fly to their targets by rule and line, and hit their foreheads and explode. This insignificant explosion is powerful enough to penetrate human skulls and destroy all inside. This case is a fictional, but most technologies required like the currently proven technologies such as sensing analysis, facial recognition, and directed blasting are out there now, so autonomous weapons like "killer bees" might become a reality within years, and it is important and urgent for us to probe into some prevention and control measures against these autonomous weapons.

It is generally believed that autonomous weapons are weapons systems that can hit their goals without human intervention once activated. As a matter of fact, AI-assisted autonomous weapons, especially lethal autonomous weapons (LAWs) have caught a great deal of attention from the USA, Russia, the UK, France, Israel and South Korea in recent years, and have been researched, developed and deployed. Of course, autonomous weapons for the time being usually involve human–machine collaboration, so they do not have obtained full autonomy. The autonomous weapons mentioned in this book may refer to human-controlled semi-autonomous weapons, autonomous weapons or fully autonomous weapons as implied in the context.

Typical autonomous weapons (in form) include auto-shoot robots, smart UAVs, unmanned warships and submarines, smart guided missiles, UAV swarms and so on. Statistics show that currently more than 60 countries around the globe have armed their military forces with AI military robots—as many as over 150 types. The *2018 Research Report on Intelligent Robots* (frontier version) jointly released by Tsinghua University and China Knowledge Center for Engineering Sciences and Technology predicts that by 2040 half of the US military members will possibly be intelligent robots [46]. The X-47B drone bomber of the USA is completely manipulated by the computer; South Korea's Super aEgis II automatic turret and Russia's Platform-M combat robots are automatically controlled; South Korea's SGR-A1 is a kind of sentry robot that can detect any invader and take the initiative to attack; the "Golden Horde" plan launched by the U.S. Air Force networks together various precision-guided munitions, including the GBU-39/B "Small Diameter Bomb" and the AGM-158 "Joint Air-to-Surface Standoff Missile", as well as the ADM-160 "Miniature Air-Launched Decoy", so that they can operate as autonomous swarms to break through the advanced air defense system of the opponent.

The AI technology can help research and develop autonomous weapons largely because these weapons are always highly dependent on automatic identification, precision guidance, high speed decision-making, group sensing and collaboration and other capabilities, and the AI technology works well to provide the above capabilities. For example, the capabilities of automatic identification and precision guidance require that the images and videos collected by high-definition cameras be quickly and accurately identified, to tell the identity of the target (our side or the enemy) and their threat (intention), which is exactly what deep neural networks are good at; the capability of high-speed decision-making is aimed at meeting the demand for the constantly changing status quo on the battlefield, and it works to gather massive data onto the cloud for rapid analysis, greatly reducing the delay caused by human analysis, thus lowering the decision-making delay to an acceptable range; the capability of group sensing and collaboration is aimed at meeting the demand for UAV swarms, ammunition swarms and autonomous weapons of the like, which are always highly sensitive to the coordinates in the physical space, and as a node of the swarm, they need to be adjusted in real time according to the position, speed and other parameters of themselves and their neighboring nodes. At high speed, it is impracticable for humans to have real-time adjustment of massive nodes, and the AI technology provides a solution for group sensing and collaboration that are beyond our human reaction speed limit.

Thus, the AI technology can effectively enable our research and development of autonomous weapons, while AI weapons can also be taken as a "corner overtaking" opportunity to win the third weapon revolution (the first two were triggered off by gunpowder and nuclear bombs respectively). This might bring about an unmanageable international arms race. The arms race is a derivative security issue incurred by the absence of a forceful international convention (in nature, it is a vulnerability) during the process of AI development, as will be elaborated in Chap. 6 of this book.

4.4 Summary

Any emerging technology will surely bring about a security empowerment effect which is always shown in two aspects: First, empowering defense, to solve the problem of security well with the help of the AI technology; second, empowering attacks, i.e. to make the problem of security even more severe. From the perspective of empowering defense, AI's application to security defense improves the existing insufficiencies and limitations of conventional security. From the perspective of empowering attacks, the AI technology will bring about security problems in other fields, i.e. AI safety problems, by empowering cyberattacks, development of autonomous weapons and so on.

References

1. Fang BX, Ren K, Jia Y. The new frontiers of cybersecurity. Engineering. 2018;(4):1–2.
2. Global artificial intelligence-based cybersecurity market 2018–2022. [2019–12–26]. https://www.businesswire.com/news/home/20180611005807/en/Global-Artificial-Intelligence-based-Cybersecurity-Market-2018-2022-Rising.
3. iResearch. 2019 China AI+Security industry development research report. [2019–9–20]. http://report.iresearch.cn/report_pdf.aspx?id=3327.
4. Things not strings: Google's new hotel profiles exemplify its approach to entities. [2019–11–30]. https://streetfightmag.com/2019/04/12/things-not-strings-googles-new-hotel-profiles-exemplify-its-approach-to-entities/#.XeIUYflPpCR.
5. Anti-terrorism robots debut at 2015 world robot conference. [2019–12–26]. http://www.xinhuanet.com/mil/2015-11/25/c_128467796.htm.
6. Zhang L, Cui Y, Liu J, et al. Applications of machine learning in cyberspace security research. J Comput. 2018;41(9):1943–75.
7. Veeramachaneni K, Arnaldo I, Korrapati V, et al. AI2: training a big data machine to defend. In: 2016 IEEE 2nd international conference on big data security on cloud (BigDataSecurity), IEEE international conference on high performance and smart computing (HPSC) and IEEE international conference on intelligent data and security (IDS). IEEE, 2016.
8. Ferdowsi A, Saad W. Generative adversarial networks for distributed intrusion detection in the Internet of Things. [2019–11–29]. https://arxiv.org/abs/1906.00567.
9. Tencent Security Management Department, Cyberspace Research Institute. Artificial intelligence empowers cyberspace security: models and practices. [2019–11–29]. https://xw.qq.com/cmsid/20180919A0UPBH00.
10. Hendler D, Kels S, Rubin A. Detecting malicious powershell commands using deep neural networks. In: ACM Asia conference on computer and communications security, June 4–8, 2018, Incheon. New York: Association for Computing Machinery; 2018. pp. 187–197 [2020–1–14]. https://doi.org/10.1145/3196494.3196511.
11. Microsoft Malware Classification Challenge (BIG 2015) Classify malware into families based on file content and characteristics. [2019–11–29]. https://www.kaggle.com/c/malware-classification/leaderboard.
12. Antiy. Practical threat hunting: making threats nowhere to hide—enlightenment from reports on the U.S. implanting malicious code into Russian power grids. [2019–9–20]. https://www.antiy.cn/research/notice&report/research_report/20190718.html.
13. Hutchins EM, Cloppert MJ, Amin RM, et al. Intelligence-driven computer network defense informed by analysis of adversary campaigns and intrusion kill chains. In: Proceedings of the 6th international conference on information warfare and security, 2013.

14. ATT&CK matrix for enterprise.[2019–9–20]. https://attack.mitre.org/.
15. Beadle J. Gartner top 10 security projects for 2018. [2019–11–29]. https://www.gartner.com/smarterwithgartner/gartner-top-10-security-projects-for-2018/.
16. How machine learning in G Suite makes people more productive. [2019–11–29]. https://www.blog.google/products/g-suite/how-machine-learning-g-suite-makes-people-more-productive/.
17. Black hat attendees see AI as double-edged sword. [2019–11–29]. https://threatvector.cylance.com/en_us/home/black-hat-attendees-see-ai-as-double-edged-sword.html.
18. Ye GX, Tang ZY, Fang DY, et al. Another text captcha solver: a generative adversarial network based approach. In: Proceedings of the 2018 ACM SIGSAC conference on computer and communications security. ACM, 2018, pp. 332–348.
19. Seymour J, Tully P. Weaponizing data science for social engineering: automated E2E spear phishing on Twitter. [2019–11–29]. https://www.blackhat.com/docs/us-16/materials/us-16-Seymour-Tully-Weaponizing-Data-Science-For-Social-Engineering-Automated-E2E-Spear-Phishing-On-Twitter-wp.pdf.
20. Capture the flag. [2019–11–29]. https://defcon.org/html/defcon-24/dc-24-ctf.html.
21. Heybe-Pentest automation toolkit. (2015–08–06) [2019–11–29]. https://www.blackhat.com/us-15/arsenal.html#heybe-pentest-automation-toolkit.
22. Automatic machine learning penetration test tool: deep exploit. [2019–9–20]. https://github.com/13o-bbr-bbq/machine_learning_security/blob/master/DeepExploit/doc/BHUSA2018Arsenal_20180802.pdf.
23. Hu WW, Tan Y. Generating adversarial malware examples for black-box attacks based on GAN. [2019–12–26]. https://arxiv.org/pdf/1702.05983.pdf.
24. Fortinet FortiGuard Labs 2018 threat landscape predictions. [2019–12–26]. https://www.fortinet.com/blog/business-and-technology/fortinet-fortiguard-2018-threat-landscape-predictions.html.
25. Thies J, Michael Z, Stamminger M, et al. Face2Face: real-time face capture and reenactment of RGB videos. In: 2016 IEEE conference on computer vision and pattern recognition (CVPR). IEEE, 2016, pp. 2387–2395.
26. Suwajanakorn S, Seitz SM, Kemelmacher-Shlizerman I. Synthesizing Obama: learning lip sync from audio. ACM Trans Graphics. 2017;36(4):1–13.
27. Zhou H, Liu Y, Liu Z, et al. Talking face generation by adversarially disentangled audio-visual representation. In: 2019 AAAI conference on artificial intelligence (AAAI), 2019.
28. Chan C, Ginosar S, Zhou TH, et al. Everybody dance now. [2019–11–29]. https://arxiv.org/abs/1808.07371.
29. A voice deepfake was used to scam A CEO Out Of $243,000. [2019–12–26]. https://www.forbes.com/sites/jessedamiani/2019/09/03/a-voice-deepfake-was-used-to-scam-a-ceo-out-of-243000/.
30. Kirat D, Jang JY, Stoecklin MP. DeepLocker—concealing targeted attacks with AI locksmithing. [2019–9–20]. https://i.blackhat.com/us-18/Thu-August-9/us-18-Kirat-DeepLocker-Concealing-Targeted-Attacks-with-AI-Locksmithing.pdf.
31. Manky D. Order vs mad science analyzing black hat swarm intelligence. [2019–9–20]. https://www.rsaconference.com/industry-topics/presentation/order-vs-mad-science-analyzing-black-hat-swarm-intell.
32. Castiglione A, Prisco RD, Santis AD, et al. A botnet-based command and control approach relying on swarm intelligence. J Netw Comput Appl. 2014;38(1):22–33.
33. Danziger M, Henriques MAA. Attacking and defending with intelligent botnets. XXXV Simpósio Brasileiro de Telecomunicaç oes e Processamento de Sinais-SBrT, 2017:457–61.
34. DARPA. Harnessing autonomy for countering cyberadversary systems (HACCS). [2019–11–29]. https://www.enisa.europa.eu/events/cti-eu-event/cti-eu-event-presentations/harnessing-autonomy-for-countering-cyberadversary-systems/at_download/file.
35. Liu Y, Ma S, Aafer Y, et al. Trojaning attack on neural networks. In: Network and distributed system security symposium, 2018.
36. Xu W, Qi Y, Evans D. Automatically evading classifiers. In: Proceedings of the 2016 network and distributed systems symposium, 2016, pp. 21–24.

37. Dang H, Huang Y, Chang EC. Evading classifiers by morphing in the dark. In: Proceedings of the 2017 ACM SIGSAC conference on computer and communications security. ACM; 2017. pp. 119–133.
38. Hu W, Tan Y. Generating adversarial malware examples for black-box attacks based on GAN. [2019–11–29]. https://arxiv.org/abs/1702.05983.
39. Anderson HS, Kharkar A, Filar B, et al. Learning to evade static PE machine learning malware models via reinforcement learning. [2019–11–29]. https://arxiv.org/abs/1801.08917.
40. Deepfakes. Faceswap. [2019–11–29]. https://github.com/deepfakes/faceswap.
41. Cole S. We are truly fucked: everyone is making AI-generated fake porn. Now. [2019–11–29]. https://www.vice.com/en_us/article/bjye8a/reddit-fake-porn-app-daisy-ridley.
42. Highway to the danger drone. [2019–12–26]. https://www.bishopfox.com/files/slides/2016/Black_Hat_USA_2016-Danger_Drone_Brown_Petro_Latimer-03Aug2016.pdf.
43. Venezuelan president survives drone attack. (2018–08–06) [2019–11–16]. https://www.express.co.uk/news/world/999338/venezuelan-president-nicolas-maduro-drone-attack-assassination.
44. Abqaiq–Khurais attack. (2019–9–19) [2019–11–12]. http://iranprimer.usip.org/index.php/blog/2019/sep/18/timeline-saudi-oil-attacks.
45. Slaughterbots. [2019–11–12]. https://www.imdb.com/title/tt7659054/.
46. Intelligent Robots (Cutting-Edge Edition) Research Report (2018). [2019–12–26]. https://static.aminer.cn/misc/article/rb.pdf.

Chapter 5
AI Security

Zhaoquan Gu, Le Wang, Denghui Zhang, and Binxing Fang

In the previous chapter, AI for cyberspace security was introduced in detail. As a new technology, we seek to find out whether AI itself exist intrinsic security problems? From the internal perspective, an AI system is composed more complicated components, such as data, framework, algorithm, model, and operation. It is obvious that any vulnerable parts would cause security problems in such a complicated system, and we call such vulnerability in AI systems as AI intrinsic security (AI security for short). In this chapter, we introduce AI security from the following aspects, data security, framework security, algorithm security, model security and operational security.

5.1 Data Security

Data are an essential foundation for AI, and it has been described in Chap. 1 about the role of data in AI development. Neural networks were proposed in early 1980s, but the third wave of AI development was initiated only when Geoffrey Hinton proposed deep learning in 2006. One of the main reasons is that numerous data were collected and provided by many industries for training and learning, which has contributed to the rapid development of AI technologies. Hence, data security has become an

Z. Gu (✉)
Harbin Institute of Technology, Shenzhen, China
e-mail: guzhaoquan@hit.edu.cn

L. Wang · D. Zhang · B. Fang
Guangzhou University, Guangzhou, China
e-mail: wangle@gzhu.edu.cn

D. Zhang
e-mail: denghui.zhang@gzhu.edu.cn

B. Fang
e-mail: fangbx@cae.cn

© Publishing House of Electronics Industry 2025

B. Fang (ed.), *Artificial Intelligence Security and Safety*,
https://doi.org/10.1007/978-981-96-9263-7_5

important part of AI security. In this section, we introduce data security which would cause AI security problems from three perspectives: dataset quality, data poisoning, and adversarial examples.

5.1.1 Dataset Quality Affects AI Security

Dataset quality would affect the performance of AI algorithms directly. However, when we design algorithms, we would not select a specific dataset but focus on the input and output.

For instance, if we design a handwritten digit recognition algorithm based on handwritten Arabic numerals (from 0 to 9), we can utilize a famous open-source database called MNIST,[1] which has 60,000 training sets and 10,000 test samples. Each sample is a 28×28 image of a handwritten digit, and the algorithm is expected to recognize the Arabic numerals (from 0 to 9) in the images. Designers can adopt other manually tagged datasets in addition to MNIST for this purpose. The target algorithm can be trained with datasets to obtain suitable model parameters. Both the algorithm and the model parameters are then applied to the handwritten digit recognition system.

For another example, we can use a dataset called CIFAR-10[2] to design an image recognition and classification algorithm. The algorithm is expected to recognize 32×32 images, and the recognition results would fall into 10 classes: airplane, automobile, bird, cat, deer, dog, frog, horse, ship, and truck. Trained with the 60,000 images from CIFAR-10 or the images from other manually established datasets, the algorithm is supposed to obtain appropriate model parameters. Both the algorithm and the model parameters are then adopted in the image recognition system.

In the two above-mentioned examples, we can simply design algorithms (e.g., adopt different neural networks) according to the input and output, without considering all the dataset contents for training. Nevertheless, data are vital for algorithms, and different datasets would yield different model parameters after training and different performance effects, thus influencing the security of AI algorithms.

The quality of a dataset is determined by its scale, equilibrium, and accuracy. If a dataset is too small, it may result in vulnerability in the algorithm performance. Different tasks vary in terms of difficulty and demand for data. For instance, handwritten digit recognition is an easy task, so it costs little to achieve a recognition accuracy of over 99% for a neural network based on the 60,000 training images of the MNIST database. Machine translation is more difficult because it entails a larger dataset. In this case, a dataset comprising 13 million correct English-Chinese sentences provided by the AI Challenger[3] would create desired results. Autonomous driving is extremely difficult, and the ApolloSpace dataset[4] by Baidu, Inc., which

[1] For more information on MNIST, see http://yann.lecun.com/exdb/mnist/.

[2] For more information on CIFAR-10, see http://www.cs.toronto.edu/~kriz/cifar.html.

[3] AI Challenger website: https://challenger.ai/.

[4] ApolloSpace dataset website: http://apolloscape.auto/scene.html.

comprises 146,997 tagged videos that fall into 25 categories (e.g., sky, automobile, and bicycle), would be a good choice for the task. If a dataset is too small, the operation accuracy of the AI model[5] obtained from the training would be low, which may lead to inaccuracies in recognition or errors and even sentences with opposite meanings in machine translation and thus prompt security problems.

Dataset equilibrium is another major factor influencing the performance of AI algorithms. In other words, a dataset without equilibrium would result in ineffective AI algorithms. Dataset equilibrium usually refers to the degree of quantitative equilibrium of various examples in a dataset. A higher-degree equilibrium indicates less deviation in the distribution of examples and more effective operation of AI algorithms. Nonetheless, disequilibrium can be found in many datasets in reality. For example, if the data about transactions, only a small number of which are fraudulent, are used to establish a trained dataset for recognizing fraudulent transactions and the task is abstracted as a binary classification, it is likely that only one example of fraudulence would be found among 100 collected examples. This would surely make it difficult for the AI algorithms trained with such a dataset to recognize fraudulent transactions. Worse still, if the dataset lacks examples of a certain category, recognizing the features of the category will be a tough task, thereby making the AI algorithms ineffective. To predict the result of a Premier League match, for instance, we can create a dataset with the records of matches and design an algorithm to train it to obtain the model parameters. The score with the largest gap in the records is 9–0, and the algorithm is unlikely to predict such a score. In other cases, if there is a large disparity in prowess between two teams, an unprecedented score can be expected, such as 12–0 in a Bundesliga match, 12–1 in a La Liga match, and even 149–0 in a domestic tournament of Madagascar. Thus, making an accurate prediction with an AI algorithm is difficult in cases of such unprecedented or rare data.

Dataset accuracy also affects the endogenous security of AI algorithms. The data in a dataset are usually annotated ones, and data annotation is essential for dataset establishment. Take vehicle object detection for an example. People can use a camera to take the pictures of various vehicles on the road and then annotate the pictures by manually marking the vehicles and circling the bounding rectangle of each. The algorithm will fail to recognize the vehicles if the data in the annotation process are incomplete. For instance, if there are many vehicles in a picture but those responsible for the annotation merely mark one or do not mark any at all, the algorithm trained with the data will be one with missed detection. For another, if those working on the annotation make a mistake by marking a pedestrian as a vehicle, the algorithm may recognize a walking pedestrian as a vehicle and thus cause a serious outcome. The data accuracy in datasets plays a crucial role in AI security, and ensuring that all the manually annotated data are accurate ones is difficult because most data are annotated through crowdsourcing.[6] Therefore, verifying the accuracy of data serves as a primary prerequisite for developing AI algorithms and ensuring data security.

[5] An algorithm and the model parameters obtained from a dataset-based training of the algorithm are combined to form an AI model.

[6] Crowdsourcing means distributing tasks among users in some way.

5.1.2 *Data Poisoning May Manually Cause an Error in AI Algorithms*

Dataset quality influences the security of AI algorithms, and a manual alteration of data would considerably change the performance of AI algorithms. In general, an AI algorithm would periodically collect the latest example data to retrain AI model parameters for a real-time adaption to changes in the data distribution. If an attacker knows that the AI algorithm collects data on a regular basis, it can contaminate collected data to be collected by the AI algorithm and distort model parameters, thereby making the algorithm ineffective. Such an attack is called data poisoning.

Figure 5.1 illustrates data poisoning. The two types of examples, which are marked with five-pointed stars and circles respectively in Fig. 5.1a, can be classified using a linear function (i.e., a dotted line in this case). If some wrongly classified examples are added to these well-classified ones, as is shown in Fig. 5.1b, the classification model would become highly complex, and the model parameter obtained from the training would be markedly different from that in Fig. 5.1a. Besides, some models may fail to classify new data because of the overfitting of these training examples.

From a practical perspective, committing data poisoning is easy for the following reason: training an AI algorithm requires a large amount of examples, many of which are collected online, and an attacker can easily commit data poisoning online by altering the examples to be collected by the AI algorithm. For instance, in 2016, Microsoft launched Tay, an AI talking robot that could chat with Twitter users and continually learn from tweets. Nevertheless, just within 24 h after some Twitter users made some racially abusive tweets, the robot grasped such rhetoric and made some racist comments. Consequently, Microsoft had to turn off this robot [1]. An AI algorithm usually needs to learn with abundant data, and data poisoning can disable a neural network by altering a few of these data. The findings of relevant studies have shown that an alteration of merely 8% of examples can result in a noticeable error for over 50% of patients in the dose prediction by an AI model [2]. Currently, some data annotation platforms annotate pictures, texts, and audio-visual clips through crowdsourcing. If an attacker makes an organized alteration of some

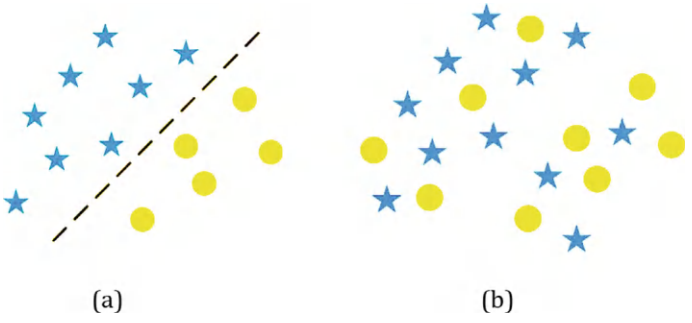


Fig. 5.1 Illustration of data poisoning

of these annotations, it may prompt an error in AI algorithms. Hence, data security matters much for AI security.

5.1.3 *Adversarial Example Has Become a New Focus in the AI Security Realm*

Data poisoning can be considered an attack on the training of AI algorithms, because it inputs incorrect examples to exclude correct model parameters in the training from accessing and thereby leads to a wrong algorithm. Compared with data poisoning, the adversarial example-based attack has evolved into a new attack in recent years. Without changing any model parameter, the attack disables an AI algorithm by altering the to-be-recognized data. In other words, it is an attack on AI algorithm-based recognition.

Proposed by Christian Szegedy et al. [3], adversarial example is an example with a subtle alteration to the original example, and the alternation causes an error in the recognition the neural networkand the difference between the original and altered examples is difficult to detect with the naked eyes. Szegedy designed adversarial examples for two different neural networks: AlexNet [4] and QuocNet [5]. Under the influence of adversarial examples, AlexNet would recognize a vehicle as an ostrich, whereas QuocNet would fail to detect any vehicle in an image.

In recent years, adversarial example has caught the attention of many research teams both in and outside China and has become a new focus in the AI security field. Adversarial example-based attacks can be classified by the attack mode, effect, and form into white-box attack and black-box attack, non-target attack and target attack, and attacks on images, texts, audio and video, respectively.

A white-box attack is an attack in which the attacker has the knowledge of the AI algorithms and model parameters to be attacked. Take a neural network as an example. If the attacker is familiar with its framework, including the number of layers, neurons on each layer, and all its hyperparameters such as the weight of a connector neuron, it can produces targeted adversarial examples to make the network ineffective. Typical white-box attacks include fast gradient sign method (FGSM) [6]; basic iterative method (BIM), an iterative method based on the FGSM [7, 8]; least-likely class method (ILCM), an iterative method for target attacks [7, 8]; DeepFool, a deception aimed at deep models [9]; Jacobian-based saliency map (JSMA), a method based on the Jacobian matrix [10].

A black-box attack is an attack in which the attacker has little knowledge of the algorithms and model parameters to be attacked. In a neural network, for example, the attacker can access little knowledge of its framework and hyperparameters and can merely obtain the output of the input data from the network. Compared with a white-box attack, a black-box attack is more practical but less feasible and effective. Typical black-box attacks include the pixel attack [11] and the universal perturbations for steering (UPSET), a method of attacking a certain target to generate adversarial

disturbance [12]. In the exploration of white- and black-box attacks, many scholars found that the combination of the adversarial examples generated through white- and black-box attacks on neural networks created a highly effective attack. That is, the adversarial examples have the characteristics of transferability. Hence, they proposed to launch black-box attacks with white-box attacks and the transferable adversarial examples generated therein, thus significantly increasing the effectiveness of black-box attacks [13–16].

Attacks on AI algorithms may be launched with different intentions. A non-target attack is an attack in which the attacker produces adversarial examples to cause an error in the adversarial example recognition by AI algorithms. In contrast, a target attack is an attack in which an attacker attacks with a designated objective—preventing AI algorithms from an accurate recognition of adversarial examples and controlling the algorithms to generate the recognition result desired by the attacker. A target attack is obviously more difficult to launch than a non-target one, and both attacks can be found in some scenes of practical significance. For instance, in an attack on a face recognition algorithm, a non-target attack aims to cause an inaccurate recognition of an adversarial example or "evade" the face recognition algorithm, whereas a target attack intends to control the algorithm to recognize the adversarial example as a designated person or "impersonate" others to pass the face recognition algorithm-based detection.

Most studies on adversarial examples focus on images and alter some pixels in images to cause an inaccurate recognition by AI algorithms. The number of altered pixels is small, and the alterations are minor; thus, adversarial examples usually cannot be detected through sensory perception. Belinkov and Bisk [17] produced text-based adversarial examples by altering the letters of words to lower the accuracy of the machine translation algorithm. In an attack on a text-processing algorithm, if the letters of a word are altered, the altered word may become unpronounceable. However, people may overlook the alternation and continue to regard it as a correctly spelled word in reading. In some adversarial algorithms for text-processing algorithms, words are altered or substituted to cause an inaccurate recognition by algorithms, and the inaccuracies are usually not obvious to people. For instance, Ebrahimi et al. [18] produced adversarial examples for an emotion analysis algorithm, where a few words were substituted for an incorrect text classification by the algorithm.

There are also some adversarial examples for attacks on audio-visual clips. Carlini et al. [19] added a little noise into audio materials, which was difficult to detect and could not be recognized by the speech recognition system, DeepSpeech [20]. All adversarial examples of this kind can prompt an inaccurate recognition by DeepSpeech. The adversarial examples generated with GAN by Wei et al. [21] prevented the object detection algorithm from accurate object detection involving visual clips. Jiang et al. [22] from Fudan University proposed a black-box attack on a video recognition algorithm and produced adversarial examples that could result in an inaccurate recognition of video contents. For instance, a video about throwing bowling balls can be recognized as one about writing words on a white board.

An attacker can launch an adversarial example-based attack by simply attacking the process of recognition by AI algorithms. Adversarial examples can be produced for all to-be-recognized examples to cause a mistake by AI algorithms. Compared with data poisoning, which needs to alter examples to change the AI training, an adversarial example-based attack is more achievable; therefore, adversarial example has become a new focus in the AI security industry.

5.2 Framework Security

The design and creation of an AI algorithm requires many complex procedures, and a well-established deep learning framework is usually adopted for programming along the way. If a developer wants to design a MNIST-based multilayer neural network, he must establish a neural network, learn about the weight parameters between neurons, and achieve neural network-based prediction and classification through programming. If the developer is expected to finish all these tasks through programming, it will be an extremely heavy workload for him because it entails highly complicated numerical calculations. As many programs can be highly multiplexed in different AI algorithms, a number of deep learning frameworks have been created to provide common functions that facilitate the production of AI algorithms by developers. At present, the frequently-used deep learning frameworks include TensorFlow,[7] Theano,[8] Keras,[9] Infer.Net,[10] CNTK,[11] Torch,[12] Caffe,[13] and PaddlePaddle.[14]

As a framework developed and maintained by the Google AI team GoogleBrain, TensorFlow takes the shape of a data flow diagram, in which a node indicates a mathematical operation and a line represents a multi-dimensional data array or tensor that connects the nodes. TensorFlow can be used for efficient numerical calculations; therefore, it is often applied to machine learning and neural network and is popular among AI researchers and practitioners. However, TensorFlow uses a multitude of third-party databases such as NumPy and libjpeg-turbo. If a system becomes more sophisticated and involves more dependencies, it may face more potential security problems. For instance, according to CVE-2018-10055[15] by Common Vulnerabilities and Exposures (CVE),[16] illegal memory access and heap overflow existed in the

[7] See https://tensorflow.google.cn/.

[8] See https://github.com/Theano/Theano.

[9] See https://keras.io/.

[10] See https://dotnet.github.io/infer/.

[11] See https://github.com/microsoft/CNTK.

[12] See http://torch.ch/.

[13] See https://caffe.berkeleyvision.org/.

[14] See https://www.paddlepaddle.org.cn/.

[15] http://cve.mitre.org/cgi-bin/cvename.cgi?name=CVE-2018-10055.

[16] Common Vulnerabilities and Exposures (CVE) is a world-famous knowledge base of vulnerabilities and exposures.

XLA compilers of the versions prior to TensorFlow 1.7.1, and attackers could make special configuration documents to cause the breakdown of target programs or a serious problem of illegal memory access. CVE-2019-9635[17] described a null pointer dereference in the versions prior to TensorFlow 1.12.2, and attackers could make special GIF documents to launch a denial-of-service (DoS) attack on the system.

Caffe is a deep learning framework developed by the Berkley AI research team and community contributors. Defining models and optimizing methods through configuration, Caffe provides examples of image classification, filter visualization, and R-CNN object detection. However, it relies on libjasper, OpenCV, and other image vision libraries and functions for processing images, and the vulnerabilities of these libraries and functions may lead to a program crash when Caffe is used for image recognition and relevant tasks. As for the security vulnerabilities of libjasper, it was mentioned in CVE-2017-9782[18] that a security vulnerability existed in the jp2-decode of libjasper/jp2/jp2-dec.c in Version 2.0.12 of Jasper,[19] whereby remote attackers could use images with special formats to launch DoS attacks (e.g., heap-based buffer overflow access and program crash). In terms of the security vulnerability of OpenCV, it was described in CVE-2017-12599[20] that the vulnerability of DoS existed in the function of icvCvt_BGRA2BGR_8u_C4C3R in OpenCV 3.3 and the previous versions, whereby attackers could cause an access violation when the function of cv::imread() is used to access images. Due to the vulnerabilities in these libraries and functions, an error is likely to occur in the operation of Caffe-based AI algorithms.

A strong dependency can be found between deep learning frameworks and libraries. If a system is more sophisticated and involves more dependencies, it will become more vulnerable to potential security problems. A deep learning framework is built upon different modules. For instance, TensorFlow is based on nearly 100 Python modules. Possible irregularities in the interface implementation modes and the calling of different modules may result in an error in the operation of AI algorithms based on deep learning frameworks.

5.3 Algorithm Security

The generation of algorithm models forms the core of AI, and it involves two main parts. The first part is to propose and design an algorithm, and it is the foundation of solving problems. Without an effective algorithm, it would be difficult to address problems. The second part is to accumulate data, and it guarantees problem solving because problems would not be tackled with a data-free algorithm. Take the example

[17] http://cve.mitre.org/cgi-bin/cvename.cgi?name=CVE-2019-9635.

[18] http://cve.mitre.org/cgi-bin/cvename.cgi?name=CVE-2017-9782.

[19] An open-source implementation of a JPEG-2000 codec developed by the Canadian software developer Michael Adams—website: https://www.ece.uvic.ca/~frodo/jasper/.

[20] http://cve.mitre.org/cgi-bin/cvename.cgi?name=CVE-2017-12599.

of the situation where a parent teaches a child addition. X, Y, and Z in such examples as "X + Y = Z" are the data, and the instruction can be considered a designed algorithm. Parents may lead their children to do addition by counting fingers or apples, and different kinds of counting would lead to different learning effects. The ultimate goal shared by these counting methods is to help the child develop an algorithm model for solving addition problems. Section 5.1 of this chapter introduces the AI security problems incurred by data, but what about the potential security problems prompted by algorithms? This section discusses the potential AI security problems from the perspective of algorithm interpretability.

5.3.1 Interpretability of an Algorithm

An algorithm designs and produces problem-solving methods and procedures according to problems. In program design, it can be considered as a finite sequence which is created by the programmer and operated by a computer. An algorithm can solve problems because it can provide an expected output according to the problem input. Nevertheless, how to measure the effectiveness of algorithms in solving problems remains a question. Take addition operation for an instance. If an error exists in one of 100 addition problems, will it be safe to say that the algorithm is an incorrect one? Some may be so demanding that they will not allow such an error and will thus regard it as incorrect. However, if the algorithm succeeds in solving all these 100 problems, can we assert that it can solve all addition problems? It works well with these 100 problems, but what about the others? If we want to ensure that the algorithm is universally effective, we must demonstrate it in a theoretical manner that would show its effective application to all addition problems. Nonetheless, many obstacles in practice may prevent us from theoretically demonstrating the algorithm. For instance, a doctor would make a prescription according to the symptoms and examination results of a patient. Different doctors may provide different prescriptions, and a prescription provided by an AI algorithm may not be accepted by all doctors, let alone demonstrating it from a theoretical perspective. To solve this problem, we must design an AI algorithm with interpretability. In other words, the algorithm must support its prescription with evidence rather than simply offering a groundless prescription.

Deep neural network is a popular AI algorithm today and is widely applied to the systems of face recognition, speech recognition, machine translation and the like; however, the degree of its interpretability is excessively low. Hence, interpreting it has become a hot research topic in academic circles. What kind of algorithms are interpretable? Take decision tree [23] as an example. In a decision tree, different branches form in each node under different conditions. Figure 5.2 depicts a decision tree showing a company employee's decision regarding going out in the morning. At the root, the first step is to check if it is a weekend (off work). If it is, they will move along the left branch to continue the decision-making under a different condition. If it is not, they will have to go out for work. This creates the first decision-making node

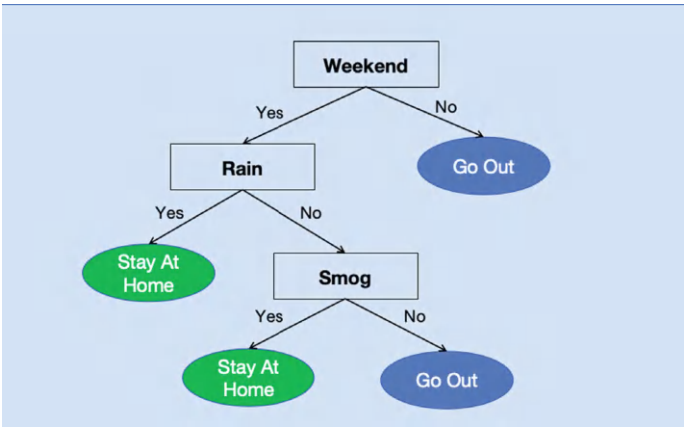


Fig. 5.2 Example of decision tree

about "going out." Likewise, rain and smog lead to different branches and finally to a decision on "going out" or "staying at home." Decisions made under these conditions can be regarded as interpretable. The conditions for the left decision-making node of "staying at home" are "weekend" but "rain"; those for the bottom decision-making node of "staying at home" are "weekend," "no rain," but "smog"; those for the bottom decision-making node of "going out" are "weekend," "no rain," and "no smog." All these conditions serve as the conditional explanations for the decisions regarding "going out" and "staying at home." Therefore, decision tree is believed to be an interpretable algorithm.

Like decision tree, K-Nearest Neighbors (KNN) [24] is also interpretable. As a classical classification algorithm, it involves the determination of the category of a datum according to the K (cardinal number) data around it. Figure 5.3 shows a simple example of KNN, and it depicts three categories of data, which are represented by squares, circles, and triangles respectively. For new data, as the one marked with (+), KNN would seek the K data nearest to it. If $K = 4$, KNN would find four data nearest to (+), which are within the smaller dotted line circle. Of these four data, two are triangles, while the other two are a square and a circle. In this case, KNN would categorize (+) as a triangle. KNN is interpretable because it follows the principle of "things of one kind come together," which means that the same or similar objects fall into the same category; hence, the new data can share the category of the data near it. But to some extent, KNN is not adequately interpretable because K is an important parameter in KNN and selecting an appropriate K is hard to explain. If K in Fig. 5.3 is set as 6, the classification will be based on the six data near (+), which are within the larger dotted line circle. Of these six data, three are circles; and the KNN would categorize (+) as a circle. Thus, a different K will lead to a different classification.

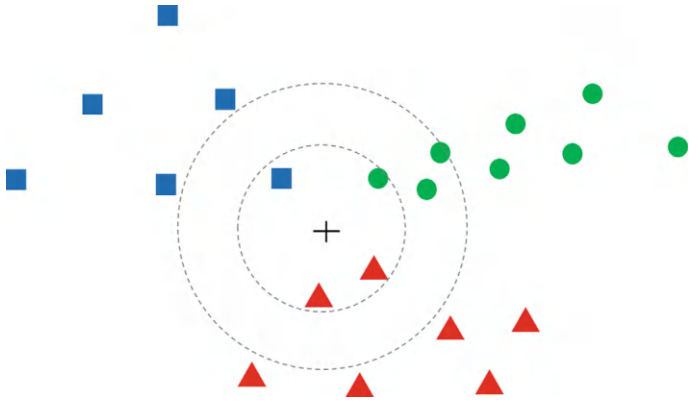


Fig. 5.3 Example of KNN

5.3.2 *Adversarial Examples Show that AI Algorithms Are not Interpretable*

Introduced in Sect. 5.3.1, adversarial example is a demonstration of the problem that AI algorithms, especially deep neural networks, are not interpretable. Many traditional algorithms have been designed following certain rules or thinking. For instance, decision tree observes a series of decision-making rules, and KNN acts on the principle of "things of one kind come together." Traditional algorithms are not that effective in solving problems concerning face recognition and machine translation. As an AI algorithm, neural network simulates the central neural network of a brain and learns about the neuron weight through neuron connection and massive data, working well for many problems. Face recognition was proposed in the 1950s, and traditional algorithms, including the ones for the geometric structure of a face (i.e., Eigenface), linear discriminant analysis, sparse representation, and feature extraction based on local descriptor [25–27], have shown a recognition accuracy of only approximately 80% in the internationally authoritative Labeled Faces in the Wild (LFW) [28]. Since 2014, a neural network has been swiftly applied to face recognition with the rapid development of big data and deep learning, and remarkable accomplishments have been attained in this aspect. The neural network trained with the data of 200,000 faces by the research team of Professor Xiao'ou Tang from The Chinese University of Hong Kong achieved 98.5% recognition accuracy [29], becoming the first neural network to overtake LFW (i.e., 97.5%) in terms of recognition accuracy. Other face recognition neural networks are DeepFace [30], DeepID series [31], and FaceNet [32]. But why are neural networks so effective? What do the hidden layers of a neural network represent? Does a neuron parameter have a specific meaning? All these questions are difficult to answer, which is the reason why many scholars think that neural networks are interpretable.

Our analysis starts with a neuron, which has a simple structure that combines multiple inputs into a new one. The weight (parameter) of each new input is obtained via the training involving abundant data. In the same network, different datasets may lead to different weights; hence, it is difficult to assess or explain the meaning of each neuron parameter. As for hidden layer, a neural network must be equipped with multiple hidden layers for learning. However, what does each hidden layer indicate? How many hidden layers are required for a specific task? Some scholars tried explaining the classification of neural networks with heat maps such as class activation mapping (CAM) [33] and grad-CAM [34]. Some explored the meaning of the hidden layer through visualization [35]. Nevertheless, problems involving the number of hidden layers necessary for a task and the features extracted by each hidden layer remain unsolved. In addition, neural networks are effective but not interpretable, and adversarial examples exist in problems concerning the processing of images, texts, and audio materials. This indicates that the data obtained by making a minor change to raw data would incur an error in the operation of deep neural networks.

Deep neural works are not interpretable, yet they outpace traditional algorithms in terms of accuracy in many fields and have been increasingly applied to every walk of life, including face recognition, machine translation, text-based emotion analysis, and the speech recognition for smart loudspeakers. A few errors in these scenes would not necessarily cause serious consequences. For example, a few translation mistakes would not affect readability but merely result in difficulty in understanding, and failing to start a smart loudspeaker via speech recognition would not cause a serious loss. However, in scenes requiring a high degree of accuracy and interpretability, an error by neural networks would lead to enormous losses. For instance, in a medical scene, if an intelligent algorithm is used for the autonomous recognition of a patient's physical health problem, neural networks without interpretability will merely reveal whether the patient is sick. However, if the algorithm goes wrong, the treatment for a patient, whose physical health problem has not been detected, will be delayed, while a healthy man, who has been wrongly detected as sick, will be panic. In this case, both the patient and the doctor would expect the algorithm to explain what symptoms and examination results have contributed to the diagnostic result. If neural networks are interpretable, the doctor will be able to make adjustments according to the caused detected by the algorithm even if the recognition results are inaccurate. This is also one of the reasons why scholars delve into the interpretability of neural networks.

5.4 Model Security

For AI algorithms and models, we can regard algorithms as solutions, such as the multilayer neural network for solving handwritten digit recognition problems and the recurrent neural network for addressing text translation problems. Nonetheless, these algorithms merely provide solutions. Most AI algorithms must be trained with examples to generate various parameters in neural networks. The combination of the

adopted algorithm and the parameters generated in the data training for the algorithm is called a model. After a neural network develops parameters out of the examples, the parameters will be stored in a special form. In the implementation of a task (e.g., categorizing an input digit image in the handwritten digit recognition or translating the input text into other languages in the text translation), the algorithm and the model parameters can be used to finish the task swiftly. In general, training a model to obtain parameters is time-consuming, but selecting a model to complete certain task is easy. Therefore, model is an important achievement and can be considered as the "essence" obtained from example learning, even though this essence has not been effectively interpreted to date. Given the important role of model in AI algorithms, it is likely that model security problems would lead to AI security problems.

5.4.1 Security of Model Storage and Management

Usually, a large number of model parameters must be trained to obtain AI algorithms. For example, the LeNet-5 neural network[21] for handwritten digit recognition requires over 60,000 training parameters, while Google's open-sourcing BERT model[22] needs over 100 million parameters. Normally, a trained model is stored in the form of a real document. For instance, the TensorFlow framework stores these parameters as a ". ckpt" document through Saver targets, while the Keras framework stores the model structure as a "JSON" or "YAML" document and the weight parameters as an "HDF5" document. Therefore, when TensorFlow, Keras, and other machine learning frameworks are used, attackers can access and then steal the models we learn by taking advantage of the default save paths and models of these frameworks.

Many deep neural networks involve a huge quantity of model parameters, which are usually stored in servers. Therefore, server security problems would prompt security problems in the storage and management of models. In a cloud server attack, attackers can exploit the server vulnerability to obtain the model parameters, which are greatly significant and may cause severe consequences once missing. For example, the adversarial example-based attack on neural networks introduced in Sect. 5.1 can disable models. This type of attack can be divided into white- and black-box attacks. The former generates adversarial examples with the knowledge of the framework and parameters of the neural network, while the latter does without it. Adversarial examples can be generated in many ways through white-box attacks. For instance, the FGSM [6], BIM [7], ILCM [8], DeepFool [9], and JSMA [10] mentioned in this chapter can be adopted to generate adversarial examples swiftly and efficiently. Moreover, the success rate of the attacks of these adversarial examples on neural networks is high. In contrast, black-box attacks can generate adversarial examples, but their generation methods are inefficient, and the success rate is markedly lower than that of white-box attacks. Hence, attackers can steal the parameters of

[21] https://engmrk.com/lenet-5-A-classic-cnn-architecture/.

[22] https://github.com/google-research/bert.

neural networks to launch attacks on neural networks and cripple application scenes involving image recognition, speech recognition, and natural language processing.

Aside from the problem of possible attacks on cloud servers potentially leading to the leakage of model parameters, the transmission of model parameters and the storage and management of edge devices in some cloud-edge coordination scenes would also prompt model security problems. The efficiency of the industrial internet, for instance, can be enhanced through the cloud-edge coordination of the cloud resources and the edge computing capacity. In the edge network, edge devices can be used for preliminary computing, such as conducting object detection of photos and sending key objects to the cloud network for a comprehensive assessment. Accordingly, abundant computation and data resources can be used to train the object detection model in the cloud network. Moreover, the model parameters obtained from the training are sent to the edge devices. These edge devices only need to perform a model-based detection rather than complex training. In this manner, the strengths of the cloud and edge networks are well combined. In such a scene, the cloud network would periodically learn with the data from the edge network and update and send the model parameters to the edge network for updating. Nonetheless, the computing and storage capacities of the edge devices are not strong enough, and the security of many devices is not up to the standard. Consequently, attackers can attack the transmission of the cloud-edge model parameters or the edge devices to access the model parameters for further attacks. In this case, if hackers obtain the model parameters, they will be able to attack neural networks through the edge network, resulting in consequences like object detection failure.

AI models encapsulate the essence of algorithms and data, and the inappropriate storage and management of the models would lead to serious consequences. Hence, designers and users of AI models must pay great attention to model security, keep safe model storage in the cloud server, and focus on the security of the transmission and management of the model parameters during the coordination of the cloud and edge devices, safeguarding the models from multiple perspectives.

5.4.2 Attacks on Open-Sourcing Models

The training for AI models entails abundant data and computer resources. As AI algorithms become universal, many developers are willing to share trained models with more users. According to some scholars, trained AI models will evolve into consumer goods like daily commodities and users will be able to share, retrain, or resell them [36]. In the sharing, attackers would have opportunities to attack open-sourcing models. After downloading or purchasing the models, they would inject malicious behaviors into the models to generate “mutant” neural networks and re-share them, thereby causing adverse effects. As mutant neural networks work normally under general circumstances and the malicious behaviors hidden therein are different to detect, attackers can re-launch and share the neural networks with malicious behaviors to prompt severe security problems.

Liu et al. [36] from Purdue University developed a complex attack and demonstrated that it was feasible and practical to launch a Trojan attack on neural network models, thereby sending security warnings to neural network models from the attacker's perspective. This attack does not entail the training datasets of the models but a full access to the models and includes three phases: the generation of the Trojan trigger, the generation of the training data, and model retraining.

The Trojan trigger can generally be regarded as a special input that triggers malicious Trojan behaviors. In the generation of the Trojan trigger, a given trigger mask is combined with the Trojan trigger generation algorithm to obtain an input variable to maximize the value of the selected neuron. Its aim is to establish a strong connection between the trigger and the selected neuron and achieve strong neuron activation in the presence of the trigger. Meanwhile, in the generation of the training data, a given output label or a label for attack is integrated with the training data generation algorithm to generate inputs that activate the output label with a high degree of confidence. The target inputs are the generated training data. The objective of this stage is to generate a set of data for neural network model retraining. In the phase of model retraining, the trigger and training data generated in the previous two stages are used to retrain the hidden and output layers of the selected neuron. The retrained models will behave normally in the absence of the trigger mask and output the category of targets in the presence of the trigger mask, thereby outputting the category of the attack-oriented labels selected beforehand with a high degree of confidence. In such attacks, attackers can hide their intentions in neural network models.

With the rapid development of AI, AI models will be shared increasingly widely. Imposing no regulation on the model releasing and sharing of developers would result in serious security problems. As it is difficult to evaluate the security of neural network models and attack intentions vary among attackers, evaluating model security will be a task of major importance.

5.5 Operation Security

We will obtain a model with parameters if using data to train AI algorithms, and this model will enable us to address many practical problems. An AI model cannot be applied without being deployed, and after the deployment, objective or subjective factors may cause security problems in its operation. These problems are what we call operation security problems.

5.5.1 Operation Security Problems Incurred by Objective Factors

Endogenous security problems in AI model operation are closely connected with the security of data, framework, algorithm, and model. From data and algorithm perspectives, the adaptability of AI algorithms to the environment is influenced by the completeness of data characteristics and examples. An algorithm is usually designed for a specific problem; thus, it is difficult to adopt it for an accurate and complete extraction of all the data characteristics involved. Worse still, the data for describing a specific problem are highly abstract and data completeness is difficult to guarantee, which may affect the security of AI algorithms in operation.

First, if there is a change to the algorithm's environment and the abstract data cannot describe it, the AI model will not be able to discern any alternation it causes to the data and the algorithm. For instance, if an intelligent model designed for Chinese chess is applied to international chess, the differences in rules and chess pieces between the two will lead to an error in the model. For another, if a neural network for recognizing cats is used to recognize dogs, it will be difficult to ensure the accuracy of recognition. To address this problem, many scholars have proposed transfer learning [37] to make AI algorithms applicable to different situations. Another objective factor is the AI algorithms' failure in considering complete characteristics. Take the example of a model for predicting automobile speed. If the model does not take road construction into consideration, it will give an inaccurate prediction when automobiles run on a section of the road that is under construction or maintenance, potentially causing an accident. Take another example of an autonomous driving model. If the abrupt road crossing by bicycles or pedestrians is not considered in the design, the model will fail to stop automobiles in the case of an unexpected road crossing by pedestrians, thereby causing an accident.

Second, even if diverse environment data have been described in a statistical way, some unprecedented environments are likely to prevent AI models from accurately predicting irregular environment data. For instance, it is unlikely for the football score prediction model described in this chapter to predict an amazing score of 149:0. There are many other similar cases. For another, if the temperature of a city seldom surpasses 50 °C and an AI algorithm is designed with the temperature records of the city to predict future temperatures, it will be difficult to predict an abnormal temperature and an early warning and response to the abnormality may be delayed.

From framework and model perspectives, AI frameworks may have security vulnerabilities, and such program vulnerabilities (e.g., heap corruption and null pointer dereference) may occur in the development of AI algorithms. These objective vulnerabilities will be the main factors influencing the safe operation of AI models. By exploiting these vulnerabilities, attackers can steal model parameters and use models to launch malicious attacks. Consequently, a face recognition model would fail to accurately recognize faces; an autonomous driving model would ignore pedestrians and road signs and thus cause an accident; a drone would violate a relevant ban and fly into a forbidden zone; smart housing would function abnormally and cause

panic. Therefore, we should make every effort to ensure that the security of AI models is free from the effects of objective vulnerabilities; perform security detection for AI models from the aspects of data, framework, algorithm, and model; and achieve timely detection of the vulnerabilities and weaknesses of AI models. By doing so, we will be able to prevent objective vulnerabilities from causing AI security problems.

5.5.2 Operation Security Problems Prompted by Subjective Factors

Apart from objective vulnerabilities, subjective factors also affect AI security. In terms of data distribution, after being deployed and put into use, an AI model would be retrained with new data and upgraded parameters. Driven by various interests, attackers may inject data inconsistent with the actual data distribution into the model, thereby changing the example data distribution for the model training and disabling the model. The data poisoning mentioned in this chapter is a typical example of this problem, wherein attackers input data favorable for them to ensure that the experience learned by the model can facilitate attacks and affect the model operation. In other cases, attackers may adopt the deployed AI models for attacks and misguide the models to make a wrong assessment. Take the example of a website that designed an AI model for distributing coupons. Through a training involving historical data, it showed a higher probability in distributing coupons among the medium- and high-end consumers aged from 30 to 40 because the model training found that these consumers had the greatest purchasing power for the website. It considered that such a distribution would bring it more profits. However, many connoisseurs generated a fake cyberspace identity by creating and maintaining a registration account with nefarious intentions. After predicting the tendency of the model in coupon distribution, they disguised themselves as target consumers and changed the actual distribution of the website users to obtain a large quantity of coupons. As a result, the coupon distribution did not generate any profits for the website at all, and the failed model caused a financial loss to the website.

An artificial alteration of target functions, which is another subjective factor, can also result in AI model security problems. Take the example of the intelligent model of Chinese chess designed with reinforcement learning [38]. The original goal of the target function is to win a game, but it may be artificially altered (e.g., it may be altered to eliminate more chess pieces of the opponent). Such an alteration may enable the model to do that, but it cannot ensure a win in a game. An artificial alteration of the target function may also result in more serious problems. For example, attackers setting the target function of a drone as "launching a malicious attack" may lead to severe security problems, which will be the focus of the introduction to AI security problems in the next chapter.

From an objective perspective, AI models have security vulnerabilities; from a subjective angle, attackers can exploit model vulnerabilities or weaknesses to launch

an attack. Adnan Siraj Rakin et al. proposed the Targeted Bit Trojan (TBT) [39], a Trojan horse attack on a target location, to attack neural networks. TBT can detect and locate model weaknesses and flip the vulnerable bits of models through a "bit flip" to launch a Trojan horse attack. Once these vulnerable bits are flipped by attackers, the target models would wrongly classify all inputs into certain category, leading to a potentially serious security problem. Through the "bit flip" of a few model parameters, TBT injects a Trojan into vulnerable bits without accessing the training datasets of models. In this way, it can launch highly stealthy attacks on models.

5.6 Conclusion

Artificial intelligence is a new technology, and from an internal perspective, a new technology has endogenous security problems. This chapter introduced possible AI security problems related to the security of data, framework, algorithm, model, and operation. From the aspect of data, AI system has security problems such as low-quality dataset, data poisoning, and adversarial example. From the perspective of framework, deep learning frameworks are highly dependent on various libraries and modules, and the security vulnerabilities of libraries and modules would affect the security of AI models. As for algorithm security, AI algorithms lacking interpretability are vulnerable to adversarial examples and cannot be fully adopted in many scenes. With respect to model security, AI models face security risks in storage and management, and open-sourcing models may be exploited by attackers to launch attacks. Finally, as far as operation security is concerned, both objective and subjective factors would incur AI security problems.

References

1. Microsoft shuts down AI chatbot after it turned into a Nazi. (2016–3–25) [2019–11–26]. https://www.cbsnews.com/news/microsoft-shuts-down-ai-chatbot-after-it-turned-into-racist-nazi/.
2. Jagielski M, Oprea A, Biggio B, et al. Manipulating machine learning: poisoning attacks and countermeasures for regression learning. (2018–4–1) [2019–10–22]. https://arxiv.org/abs/1804.00308.
3. Szegedy C, Zaremba W, Sutskever I, et al. Intriguing properties of neural networks. In: International conference on learning representations (ICLR), 2014.
4. Krizhevsky A, Sutskever I, Hinton G. Imagenet classification with deep convolutional neural networks. In: International conference on neural information processing systems, 2012.
5. Le QV, Ranzato M, Monga R, et al. Building high-level features using large scale unsupervised learning. (2011–12–29) [2019–12–9]. https://arxiv.org/abs/1112.6209.
6. Goodfellow I, Shlens J, Szegedy C. Explaining and harnessing adversarial examples. In: International conference on learning representations (ICLR), 2015.
7. Kurakin A, Goodfellow I, Bengio S. Adversarial machine learning at scale. (2016–11–4) [2019–12–9]. https://arxiv.org/abs/1611.01236.
8. Kurakin A, Goodfellow I, Bengio S. Adversarial examples in the physical world. In: International conference on learning representations (ICLR), 2017.

9. Moosavi-Dezfooli S, Fawzi A, Frossard P. DeepFool: a simple and accurate method to fool deep neural networks. In: IEEE conference on computer vision and pattern recognition (CVPR), 2016.
10. Papernot N, McDaniel P, Jha S, et al. The limitations of deep learning in adversarial settings. In: Proceedings of IEEE European symposium on security and privacy, 2016.
11. Su J, Vargas D V, Kouichi S. One pixel attack for fooling deep neural networks. (2017–10–24) [2019–12–9]. https://arxiv.org/abs/1710.08864.
12. Sarkar S, Bansal A, Mahbub U, et al. UPSET and ANGRI: breaking high performance image classifiers. (2017–7–4) [2019–12–9]. https://arxiv.org/abs/1707.01159.
13. Papernot N, McDaniel P, Goodfellow I. Transferability in machine learning: from phenomena to black-box attacks using adversarial samples. (2016–5–24) [2019–12–9]. https://arxiv.org/abs/1605.07277.
14. Dong Y, Liao F, Pang T, et al. Boosting adversarial attacks with momentum. In: Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR), 2018.
15. Chen P, Zhang H, Sharma Y, et al. ZOO: Zeroth order optimization based black-box attacks to deep neural networks without training substitute models. In: Proceedings of 10th ACM workshop on artificial intelligence and security (AISEC), 2017.
16. Papernot N, McDaniel P, Goodfellow I, et al. Practical black-box attacks against machine learning. In: Proceedings of the ACM on Asia conference on computer and communications security, 2017.
17. Belinkov Y, Bisk Y. Synthetic and natural noise both break neural machine translation. In: International conference on learning representations (ICLR), 2018.
18. Ebrahimi J, Rao A, Lowd D, et al. HotFlip: white-box adversarial examples for text classification. In: Proceedings of the 56th annual meeting of the association for computational linguistics (ACL), 2018.
19. Carlini N, Wagner D. Audio adversarial examples: targeted attacks on speech-to-text. In: IEEE symposium on security and privacy workshops, 2018.
20. Hannun A, Case C, Casper J, et al. Deep speech: scaling up end-to-end speech recognition. (2014–12–17) [2019–12–9]. https://arxiv.org/abs/1412.5567.
21. Wei XX, Liang SY, Chen N, et al. Transferable adversarial attacks for image and video object detection. (2018–11–30) [2019–12–9]. https://arxiv.org/abs/1811.12641.
22. Jiang LX, Ma XJ, Chen SX, et al. Black-box adversarial attacks on video recognition models. (2019–4–10) [2019–11–29]. https://arxiv.org/abs/1904.05181.
23. Quinlan JR. Induction of decision trees. Mach Learn. 1986;1(1):81–106.
24. Altman NS. An introduction to Kernel and nearest neighbor nonparametric regression. Am Stat. 1992;46(3):175–85.
25. Turkand MA, Pengland AP. Eigenfaces for recognition. J Cogn Neurosci. 1991.
26. Belhumeur P, Hespanha J, Kriegman D. Eigenfaces vs. fisherfaces: Recognition using class specific linear projection. IEEE Trans Patt Anal Mach Intell. 1997.
27. Wright J, Yang A, Ganesh A, et al. Robust face recognition via sparse representation. IEEE Trans Pattern Anal Mach Intell. 2009.
28. Huang GB, Ramesh M, Berg T, et al. Learned-miller. Labeled faces in the wild: a database for studying face recognition in unconstrained environments. In: Workshop on faces in real-life images: detection, alignment, and recognition, 2007.
29. Sun Y, Wang XG, Tang XO. Deep learning face representation by joint identification-verification. (2014–6–18) [2019–12–7]. https://arxiv.org/abs/1406.4773.
30. Taigman Y, Yang M, Ranzato MA, et al. Deepface: Closing the gap to human-level performance in face verification. In: IEEE conference on computer vision and pattern recognition (CVPR), 2014.
31. Sun Y, Liang D, Wang XG, et al. Deepid3: face recognition with very deep neural networks. (2015–2–3) [2019–12–7]. https://arxiv.org/abs/1502.00873.
32. Schroff F, Kalenichenko D, Philbin J. FaceNet: A unified embedding for face recognition and clustering. In: Proceedings of IEEE conference on computer vision and pattern recognition, 2015.

33. Zhou B, Khosla A, Lapedriza A, et al. Learning deep features for discriminative localization. In: IEEE conference on computer vision and pattern recognition (CVPR), 2016.
34. Selvaraju RR, Cogswell M, Das A, et al. Grad-cam: Visual explanations from deep networks via gradient-based localization. In: IEEE international conference on computer vision (ICCV), 2017.
35. Mahendran A, Vedaldi A. Understanding deep image representations by inverting them. In: IEEE conference on computer vision and pattern recognition (CVPR), 2015.
36. Liu Y, Ma S, Aafer Y, et al. Trojaning attack on neural networks. Department of Computer Science Technical Reports, 2017.
37. Pan SJ, Yang Q. A survey of transfer learning. IEEE Trans Knowl Data Eng. 2010;22(10):1345–59.
38. Sutton RS, Barto AG. Reinforcement learning: an introduction. Cambridge, Massachusetts: MIT Press; 2018.
39. Rakin AS, He ZZ, Fan DL. TBT: Targeted neural network attack with bit trojan. (2019–9–10) [2019–12–9]. https://arxiv.org/abs/1909.05193.

Chapter 6
AI Safety

Zhihong Tian and Shen Su

AI safety problems refer to the threats caused by the vulnerabilities of AI systems to other fields. An AI safety problem is quite common and different from the cyber attacks based on AI system. The former is a byproduct of AI system's malfunction, which always generates uncontrollable consequences, out of its designer's expectation; the latter is intentioned, and the usage intention of AI is to conduct cyber attacks. Therein, we term the former as a "safety problem" which needs to be guarded against, and the latter as an "AI-powered attack". As a matter of fact, if the empowerment of the latter is abused or not effectively restrained, it surely will put mankind under the shadow of AI-powered attacks.

Apparently, with the potential of inducing accidents and self-consciousness awakening, AI system and its uncontrollability would stoke the public concerns about the ruin of human civilization. Admittedly, it is still a controversial topic whether or not would AI take fully control of human race but we can discuss from the point of whether AI systems can deliberately do harm to the current human society, which is the topic of this chapter.

Z. Tian · S. Su (✉)
Gunagzhou University, Guangzhou, China
e-mail: sushen@gzhu.edu.cn

Z. Tian
e-mail: tianzhihong@gzhu.edu.cn

© Publishing House of Electronics Industry 2025

B. Fang (ed.), *Artificial Intelligence Security and Safety*,
https://doi.org/10.1007/978-981-96-9263-7_6

6.1 The Safety Problems Caused by the AI System Failures

6.1.1 The Failure of Self-driving Cars

6.1.1.1 The Failure of the Autonomous Driving Technology of Tesla

A Tesla driverless car is equipped with many sensors which constantly send data to its autonomous driving system. Malicious attackers can remotely control the automobile system by attacking its data sources, poisoning the data, and other methods. Such attacks would result in a driving deviation or a serious car accident. At the Defcon held in Las Vegas in August 2016, the joint team of professor Wenyuan Xu from Zhejiang University and the Qihoo 360 Automobile Information Security Lab displayed their techniques of attacking the Autopilot of Tesla Model S [1]. With the re-configuration of the sensors, lights and radio of the automobile, they deceived Autopilot and lured it to report non-existent obstacles or ignore the existing obstacles. It is obvious that such cyber attacks would cause severe consequences.

According to the Tesla suggestions on autonomous driving, drivers should not leave the autonomous driving system alone during the driving, but keep the control over it. However, according to the police investigations of a number of car accidents, drivers generally toke almost no control of the car, with no traffic condition monitoring. Apparently, it is a tedious job to monitor the running statement of the autonomous driving system, while it is the last line of the defense against the uncontrollability of the autonomous driving system. As a result, neither the future application of AI systems nor the current deployment of semi-AI systems is prospective.

6.1.1.2 The Failure of Uber's Autonomous Driving Technology

In March 2018, an Uber driverless car crashed down a woman to death in Tempe of Arizona, US. The accident happened when the car was right in the mode of autonomous driving with a driver on board. Corresponding investigations and analysis indicated that the driverless car "saw" the victim, but did not stop, nor did its autonomous driving system generate any early warning about the accident [2]. In fact, the autonomous driving system incredibly recognized the woman as an unknown object, an automobile, and then a bicycle. It even did not make any response during the period of 1.3–6 s before the collision, and request an emergent brake 1.3 s before the collision. According to the experts in the field of autonomous driving, all relevant video evidences indicated that there is a disastrous bug in the autonomous driving technology of Uber. This accident revealed another serious problem: we cannot count on the manual monitoring on the autonomous driving systems. Moreover, automatic and intelligent technologies in the future world may be seldom or even never monitored. Worse still, the feasibility of all the current efforts on the semi-autonomous testing is questionable. According to the available surveillance videos, the Uber

"safety driver" on the driving seat had about two seconds to make a brake or a turn, but he did not take actions until 1.3 s before the collision, and he did not manage to brake before the collision.

6.1.2 The Failures of Intelligent Robots

According to a news report of The Associated Press, a robot hit a technician in chest and pressed him against a metal board in a Volkswagen plant in German in July 2015. Unfortunately, the man died later. As indicated in the preliminary conclusion, the accident was caused by man rather than the robot because it was designed to grab and assemble automobile parts in the designated zone and there was safety wire mesh that separated the robot from the man [3].

On August 13, 2015, a young worker was "killed" by a robot when repairing a disconnected metal board in an Indian automobile accessory company. The robot was programmed to weld the metal board [4], and it was said that the worker was adjusting a loose metal sheet during the welding. Unfortunately, he stayed excessively close to the robot, and the welding rod on the robotic arm pierced his abdomen.

According to some news reports online, the records of the Food and Drug Administration (FDA) showed that 144 people died in the operations involving surgery robots from 2000 to 2013 [5]. To some extent, the surgery robots were to blame for these accidents where they dropped assembly units in the bodies of the patients, their sparks burned the tissue, or they conducted excessive operation due to a failure in their systems.

In March 2017, IOActive, an American cyberspace security company, carried out a security survey on 50 robots and detected nearly 50 safety threats in 10 of them [6].

6.1.3 The Failure of Smart Speakers

In May 2018, KIRO 7 [7], a local TV station of Portland in Oregon, reported that a Portland woman Danielle received a phone call from one of her husband's colleagues. The man said that he got a voice mail of a dialogue between her and her husband. Hearing this, she lost no time questioning the customer services department of Amazon on the phone. After a preliminary investigation, a customer services representative called her back, explaining that Alexa, the voice assistant of the smart speaker, was wrongly activated and took the dialogue as a command and thus caused the problem. This explanation, however, was not convincing because the problem would not occur unless the voice assistant wrongly recognized the dialogue for at least four times. First, Alexa must be wrongly activated. Then, there must be a command like "Send the message" in the dialogue. After receiving the command, it would loudly ask the question "To whom?" Finally, it may ask for confirmation "Are you

sure to send it to XXX?" Only after these four steps would it send the message. It is obvious that the mistake by Alexa is in fact not explainable. From an objective perspective, such unexplainability of artificial intelligence would create a sense of crisis among people.

Smart speaker is in nature a supportive tool which is used at the will of people. It can control the switches of domestic appliances and even shop online. This indicates that AI system obtains the "right to control" from human and thus becomes able to get involved in people's lives. AI system is so powerful that it can evolve from an "assistant" to a "decision maker". The difference between these two roles is that the latter can exclude people from decision making, which means that AI system has the right to decide. The more rights to control AI system obtains, the more negative impacts technical failures will cause. If AI system often makes mistakes, the AI-based Internet of Things would become a bloodcurdling system.

6.1.4 Out-Of-Control Performances of AI System

In August 2017, the AI-related embarrassment of Facebook triggered a great disturbance across the world, stoking public concerns about the out-of-control performances of AI once again [8]. In the Facebook AI Research (FAIR), two intelligent chatting robots were trained for talking so that they could automatically solve social network problems. Throughout the experiment, the two robots—Alice and Bob practiced much self-training the researchers arranged for them to acquire adequate data, just as the both hands of AlphaGo played chess against each other. But surprisingly, the two robots made unintelligible dialogues after being trained for a period of time. These dialogues were grammatically meaningless and seemed to be understood by nobody but the robots. Doubting that it was an outcome of the self-evolution of the robots, Facebook decided to suspend the AI program. But later, it was found that the problem was not prompted by the self-evolution of the robots but by a program error.

As an advanced intelligent assistant today, Alexa can be connected with much smart furniture and domestic appliances and control them to facilitate a household life. According to a *USA Today* report in March 2018, many Alexa users complained about unexpected weird laughter from their smart speakers. This caused some panic among the users [9]. Amazon replied that Alexa mistakenly heard the voice command of "Alexa, laugh" and the problem was attributed to an error in the perception system of the speaker. Besides, the company came up with some solutions, such as changing the voice command into "Alexa, can you laugh?".

6.1.5 Radical Rhetoric of Chatting AI

Tay was a chatting AI robot launched by Microsoft in 2016. With some offline training in the early stage, it could be further trained through interaction with other

Tweeter users. But embarrassingly, Tay madly made some radical rhetoric, including racist speeches, just one day after it was connected with the Internet. Immediately, Microsoft cancelled Tay's Tweeter account [10]. Obviously, Tay was trained to make the outrageous rhetoric in the interaction, and it was targeted by those with an evil intention. The same misfortune fell on Zo, a successor to Tay. Launched in December 2016, Zo was a chatting AI robot modeled after Millennial and also triggered some negative responses. Having learned the lesson of Tay, it was designed to pay little attention to political and religious affairs. Despite this, it still made controversial comments on the highly-regulated topics. In a dialogue, it made offensive comments on some religious classics. This once again aroused a heated discussion on the transparency of AI. According to another news report, a similar problem was also spotted in Watson, IBM's supercomputer: it included some offensive words in its verbal expressions after reading Urban Dictionary [11].

In these typical cases involving the out-of-control performances of AI, the AI systems were not able to interact with the physical world, but they caused severe consequences. The inflammatory radical rhetoric, including the comments on genocide and Neo-Nazism, was chilling. Apart from physical trauma, the out-of-control performances of AI may impose negative impacts on both the human society and people's thinking. From another perspective, if an AI system with such performances is installed in a robot that can act and move, its interaction with the physical world will directly expose mankind to risks.

6.1.6 Dangerous Treatment Suggestions from Medical AI

Watson is an artificial superintelligence product launched by IBM. Among the Watson-branded products, Watson Health is a revolutionary AI product offering medical assistance. But 2018 STAT News reported that Watson provided inaccurate and even unsafe treatment suggestions for doctors [12]. This sounds somewhat like a conspiracy theory, but the faulty of AI products rather than the conspiracy of AI's destroying human beings is to blame for Watson's offering dangerous medical suggestions in the current stage.

In terms of cancer treatment, an investigation report showed that IBM's Watson was merely trained with hypothetical cancer cases rather than actual data about patients; in the training, "treatment suggestions from experts" other than strictly-reviewed medical guidelines or actual treatment results were adopted. IBM claimed that Watson would consider the medical history and other historical records of patients, but the findings of the investigation revealed that it still had some weaknesses in this aspect. In general, Watson has yet been capable of playing a role in actual medical treatment, but some hospitals are still using it. This may be a potential threat: if medical AI enjoys stronger authority over physicians, if the "medical suggestions" from AI are well accepted and if AI can fully dominate medical treatment, all these dangerous treatment plans caused by the out-of-control performances

of AI or the evil intentions and carelessness of human will lead to unintended consequences. There is a more horrible viewpoint: as some cases suggests (for instance, gaming AI can pause a game to prevent losing scores), a possible method adopted by medical AI to eradicate cancers is to eliminate those who are genetically prone to cancers.

6.2 Research and Development of AI Weapons May Trigger International Arms Races

The reason why AI safety problems cause risks to human beings is that the vulnerability of AI would lead to some uncontrollable incidents. But from another perspective, harnessing the great power of AI to strengthen security attacks would also endanger human being. The only difference between the two is that the former are the safety problems prompted by the uncontrollable performances of AI, while the latter are the empowered attacks based on the strong power of AI. The relationship between the two resembles that between the vulnerability and backdoor of computer security: vulnerability, just like the safety problems, is something accidental; backdoor, just as empowered attacks, is something made on purpose.

Despite the potential risks, making AI development a national strategy has become a consensus among many countries. After the US decided to make the development of AI one of its core strategies, other countries followed suit and strived to achieve a leap-forward transition from AI development to AI application. This has triggered a tight race of applying AI for military purposes. Aiming to consolidate national security and regarding the AI weapon development as an important way to "keep a driving seat in military" or "surpass other countries in military", most of the nations, buttressed by their own technological strength and economic power, have been stepping up the application of AI in weapons. Such a dangerous arms race featuring AI-powered attacks has posed great risk which cannot be ignored to human safety and world peace.

6.2.1 AI Weapon Development of Different Countries

The U.S.: As early as 2016, the US issued many white papers on the impacts of AI on national security, and all its national military departments and weapon developers were getting prepared for the future AI warfare. Since 2018, the country has taken more substantive measures to accelerate its AI weapon development, including but not limited to: gradually equip drones with the ability to combat autonomously to free people from military decision-making; expand the scope of AI weapons to raise the level of unmanned combats on more fronts; relieve AI weapons of their dependence on satellite position systems in terms of positioning; establish unmanned

aerial combat centers in main military equipment like marine aircraft carriers; install intelligent weapons like intelligent missiles and automatic shooting robots in the main equipment of air forces and armies; develop full-autonomous intelligent weapons for future warfare and strengthen the American forces' ability to win on battlefields with intelligent weapons featuring precise algorithms; adopt AI to improve the physical quality and military capabilities of soldiers and build an impregnable man–machine army, so as to strengthen the American military power.

So far, the American army has developed four strengths in terms of intelligent weapons. First, it has the largest number and the greatest variety of drones in the world, and these drones can hover in the broad sky for a long time and fly around people. Besides, it has developed autonomous smart drones. Second, it boasts the most powerful overwater and submarine unmanned equipment across the globe, ranging from the unmanned aerial combat centers on aircraft carriers, the anti-submarine unmanned surface vessels, the integrated unmanned submarine equipment for monitoring and combat, and the deep-sea unmanned underwater vehicles. All these are extremely powerful weapons on and under waters. Third, it possesses the largest ground robot army on the earth and has developed the second-generation ground carrier robots with a greater load capacity and a longer battery life as well as the special robots with extraordinary bionic abilities. Fourth, it has the most intelligent missiles all over the world. Highly speedy and invisible, these missiles feature such advanced functions as autonomous route planning and threat sensing.

Russia: To maintain its strategic strengths as a military power, Russia has launched a series of outlines and application plans for developing AI weapons, with focus on combining military and industrial resources with AI. In comparison with the US which adopts an overall, wide and complete strategy for AI weapon development, Russia takes a cost-effective development strategy which would create immediate effects. In the new national plan for weapon development, it gives priority to attack-oriented unmanned equipment. At present, Russia has made headway in marine, ground and air forces: it has developed an army of military drones for different distances and tasks, and each military zone has a strong drone army; it has built a set of unmanned underwater warships of different sizes and with different functions, and each underwater vehicle has a strong capacity to fight under water; it has formed a cluster of unmanned equipment and robots which are classified according to the labor division and the unit of combats; besides, the unmanned chariots and scouts and the individual-soldier robots are characterized by high performance and great flexibility and motility.

South Korea: South Korea has attached great importance to the development of autonomous weapons in recent years. Early in 2015, a robot developed by Korea Advanced Institute of Science and Technology won the championship of the Defense Advanced Research Projects Agency Robotics Challenge. In the same year, its autonomous turret Super aEgis II also captured wide attention. SGR-A1 [13], one of the killer robots jointly developed by Samsung Techwin and Korean universities, can automatically aim at any target that interests it. However, South Korea has placed great emphasis on autonomous weapons is worrisome. In 2018, Korea Advanced Institute of Science and Technology collaborated with Hanwha Systems to establish

an AI R&D center, with the aim of developing warfare-oriented AI. This triggered public concerns, and over 50 AI scholars across the world signed a letter to boycott it [14].

The UK: In recent years, the UK has gradually increased its military expenditures, with the intention of strengthening its defense capabilities that are weakening with each passing day. Aside from developing an arm of robotic servicemen in the early stage, it designed bomb disposal robots and advanced drones. But in early 2017, it claimed that it would not use autonomous weapons and would maintain its international image and status in a unique manner instead.

Germany: After developing spy drones, Germany produced some robots that could sense complex landforms. In 2018, it made a public statement that it was not passionate about the development of autonomous weapons but would prevent itself from potential threats. Besides, it claimed that it would develop as a country with economic and technological strengths and, like the UK, take a path of intelligent weapon development to make itself powerful in the world.

Israel: With its unique advantages in the variety and battery life of drones, Israel is a main exporter of military drones in the world today, accounting for 60–70% of the world's total exported military drones. In terms of unmanned surface equipment, it is one of a few countries which have applied unmanned surface equipment to the army of servicemen. Its unmanned surface equipment has many functions like antisubmarine and is exported to many other nations. When it comes to unmanned ground equipment, the unmanned chariot developed by Israel is an autonomous ground weapon and is the best unmanned equipment for the ground army in the world.

Japan: Japan is a global leader in the development of robots. According to the investigation into the nuclear accident in Fukushima, Japan has particular strength in the development of AI robots and has developed world-class unmanned underwater vehicles. With the three major goals of "building an innovation center of robots", "constructing a country applying robots", and "entering an era of robots", it is stepping up the AI development for military and civilian uses.

France: In the global wave of AI development, the French government has launched a robot development plan and an AI development strategy to accelerate the development of its AI technologies and thus maintain its leading position in the field of robots across the world. By far, France has developed a surveillance-attack drone which is as powerful as an American military drone.

6.2.2 Risks Brought by the Arms Races of AI Weapons

The American and Russian development of intelligent weapons has driven the global expenditure of AI technologies to rise at an average growth rate of about 20%. It is estimated that the expenditure will have reached approximately USD 200 billion by the end of 2020. Such a rapid development of intelligent weapons has stoked public concerns and controversies. If AI is applied to production and daily life, the risks

brought by its false alarms are acceptable to some extent. But if it is utilized for military weapons, a false alarm may cause devastating consequences. For that reason, there has been backlash against the development of intelligent weapons. For instance, a protest attended by thousands of Google employees led to the removal of a development project of new intelligent weapons. For another, the Korea Advanced Institute of Science and Technology had to announce the cancellation of its development of new intelligent weapons due to the pressure of the protesters. This demonstrates that people have rung the alarm against the development of intelligent weapons. The AI for military uses will not only change the modern warfare mode but also reshape the global military landscape. The risks brought by the arms races of AI weapons can be divided into three types: machine risks, objective risks, and subjective risks.

In term of the risks that may be incurred by machines, there is the risk of operational failure in intelligent weapons. Theoretically, if weapons become more intelligent, the software for controlling their performances will become more complex and the probability of fault will be higher. Accessing accurate information about enemies and their geographical conditions plays a crucial role in winning a war. A destructive attack from drone swarms may become ineffective due to misinformation; an intelligent missile may miss its target because of inaccurate intelligence; a combat platform may lose a battle owing to a vulnerability in data analysis. Each information inaccuracy may result in misjudgment, for it is not easy for intelligent weapons to distinguish an army from its enemy in the complex environment on a battlefield. Things may go wrong if intelligent weapons miss something important when coping with overwhelming tasks.

If intelligent weapons are upgraded into autonomous ones, it will impose potential dangers on mankind. The transition from intelligent to autonomous weapons is a progressive process as well as an inevitable trend. When the transition takes place, people may lose the ultimate decision right over intelligent weapons. The leap-forward development of intelligent weapons may bring people the joy of achievements and the confidence in the fulfillment of objectives, but it may also arouse the fear over the great threat intelligent weapons will pose to the survival of human. If the intelligent weapons for military uses go out of control, it is likely that they will transform themselves from a military task implementer into a threatening enemy. With a mighty destructive force, they may change from a main power against enemies into an insurmountable foe; besides, the military strengths of both sides would be determined by the proportion of intelligent weapons. Worse still, intelligent weapons increase uncertainty in a war, for they may evolve from a magic instrument for us to defeat enemies into a "betrayer" that harms us. If the indifferent and tireless intelligent weapons are given the right to kill and completely replace soldiers in a battle, will they commit arbitrary killing because of "excessive" fighting? Will they expand the scope of their enemies and become a "terminator of human beings"? The unexpected answers to these two questions are the major potential dangers that may be posed by out-of-control AI for military uses.

As for possible objective risks, if intelligent weapons partially substitute people, the restriction on wars will be reduced. Human resources are a major force and a necessity for waging a war, and the quantity of soldiers serves as an important index

of the military strength of a country. According to the plan of the American army, over half of its defense tasks will be implemented by intelligent weapons, and many human soldiers will be replaced by such weapons. Currently, it takes years to train a soldier with military qualities, and the soldier will be gone forever once he/she is killed. In contrast, it merely costs minutes to produce an AI solider. As long as there are sufficient raw materials, the production can be a continuous one. Such a great change has significantly reduced the restriction on the constant supply of soldiers—a major condition for launching a war. Besides, it will increase the possibility of the wars that undermine human safety and national security.

When it comes to possible subjective risks, intelligent weapons strictly follow orders, which gives operators great power to kill and thus causes enormous harm to life. Therefore, there will be the risk of mechanical arbitrary killing if they are applied. In the era when money matters much, evil terrorists can reprogram intelligent weapons or further combine them with drones according to targets to plan a thrilling killing that is often seen in a gun battle movie or organize a large-scale terrorist attack. When the indifferent machines strictly follow the orders, they will kill people and even expand the killing to make a massacre. The possibility of such a horrible thing is increasing day by day.

6.3 Public Concerns About the Out-Of-Control Performances of AI

As early as 1942, Isaac Asimov was the first to propose Three Laws of Robotics [15] in *Runaround*, one of his short science fictions. Law I: A robot may not injure a human being or, through inaction, allow a human being to come to harm. Law II: A robot must obey orders given by human beings under the premise of not violating the first law. Law III: A robot must protect its own under the premise of not violating the first and second laws. Later, Asimov purposefully added a new law—Law 0: A robot may not injure humanity or, through inaction, allow humanity to come to harm. In the science fiction, the three laws of robotics were added to the foundation of nearly all pieces of software, and there were no ways to change that. But in reality, robots fail to obey these laws. Especially in emergency events, the robots or other AI systems with the ability to act may harm human beings. Such problems have stoked concerns in the international community.

Some scholars [16] believe that AI can be divided into three types—Artificial Narrow Intelligence (ANI), Artificial General Intelligence (AGI), and Artificial Superintelligence (ASI). ANI refers to the artificial intelligence that obtains a skill through programming, such as the application systems for weather forecasts, chess playing, data analysis, and news broadcasting. But it has only one sharp skill and does not have the awareness, cognition or sentiment of human beings. AGI can compete with human intelligence in various fields and finish various intelligent tasks that can be done by human beings, and it is creative, imaginative and productive. ASI

overtakes people in all areas including scientific innovation, general knowledge, and social skill. Nick Bostrom, a philosopher from Oxford University, defined ASI as something that significantly outpaced human cognition in nearly all intriguing areas [17]. At present, AI is still in the stage of ANI, so it cannot surpass human intelligence in every aspect. Artificial intelligence systems with the abilities to combine things and to act are still in the research stage. Despite of this, many experts remain concerned about the potential safety problems caused by such AI systems.

In January 2015, Bill Gates, the founder of Microsoft, said that people should worry about that AI became increasingly powerful. According to Gates, AI would become powerful to a point that upsets people, but it would play a positive role as long as it remain under our control.

In April 2017, Stephen Hawking delivered a video speech at the Global Mobile Internet Conference held in Beijing [18]. "In short, I believe that the rise of powerful AI, will be either the best thing, or the worst, ever to happen to humanity." "I have to say now, that we do not yet know which." "So we cannot know if we will be infinitely helped by AI, or ignored by it and side-lined, or conceivably destroyed by it. Indeed, we have concerns that clever machines will be capable of undertaking work currently done by humans, and swiftly destroy millions of jobs." "While primitive forms of artificial intelligence developed so far, have proved very useful, I fear the consequences of creating something that can match or surpass humans. AI would take off on its own, and re-design itself at an ever increasing rate. Humans, who are limited by slow biological evolution, couldn't complete, and would be superseded. It will bring great disruption to our economy." "And in the future, AI could develop a will of its own, a will that is in conflict with ours. Although I am well-known as an optimist regarding the human race, others believe that humans can command the rate of technology for a decently long time, and that the potential of AI to solve many of the world's problems will be realized. I am not so sure." "But it could also be the last, unless we learn how to avoid the risks." "At this stage, I may have possibly frightened you all here today, with talks of doom. I apologize. But it is important that you, as attendees to today's conference, recognize the position you hold in influencing future research and development of today's technology." "We are aware of the potential dangers, but I am at heart an optimist, and believe that the potential benefits of creating intelligence are huge. Perhaps with the tools of this new technological revolution, we will be able to undo some of the damage done to the natural world, by industrialization."

Elon Musk, the CEO of SpaceX and Tesla, has mentioned in many statements that AI is threatening the rosy prospect which people deserve and that we need to establish an appropriate regulatory agency before "something horrible happens" [19].

At the International Joint Conference on Artificial Intelligence in July 2018, Elon Musk and the three founders of DeepMind, over 160 AI companies and organizations from 36 countries, and more than 2400 AI scholars from 90 nations jointly signed the Lethal Autonomous Weapons Pledge [20], promising that they would neither participate in nor support the development, manufacturing or trading of lethal autonomous weapons systems (LAWS).

In addition, Google, Microsoft, Facebook, Baidu, Foxconn and other famous enterprises have also expressed their concerns about the uncontrollable factors of AI security.

6.4 Three Elements of Uncontrollability of AIAs

When it comes to the threats of AI systems to humanity, priority must be given to the AI systems with the ability to act, which are called Artificial Intelligence Actants (AIAs). The details about AIAs shall be further discussed in Chapter VII of this book. Although no cases concerning the "dangers caused by the uncontrollability" of AIAs had been reported by the date the manuscript was finalized, there has been long a consensus on the uncontrollability risk of AIAs. While being highly alert to AIAs, we need to be aware that not all AIAs have the risk of uncontrollability and an AIA would lose control only under certain circumstances. Academician Binxing Fang, one of the authors of this book, made a report entitled "My Opinions on AI Security" [21] at the China International Software Expo on June 30, 2018. In the report, he put forth the "three elements of uncontrollability of AIAs" for the first time. These three elements are, namely, (AIA has) the abilities to act and to destroy, (AIA has) an unexplainable ability to make decisions, and (AIA has) the ability to evolve and even evolve into a self-driven system.

6.4.1 AIA Has the Abilities to Act and to Destroy

AIA's ability to act takes two forms. First, AIA can move and create kinetic energy in a physical space. Second, AIA can exchange energy with the external world to change the status of other objects.

6.4.1.1 AIA Can Move and Create Kinetic Energy

At present, AIA, a combination of "artificial intelligence" and "moving object", has been on a rise, such as the self-driving automobiles of Tesla and Uber, the autonomous and controllable human-shaped "dynamic robots" and dog-shaped robotic dogs of Boston Dynamics, and the "mini leopard robots" developed by MIT. But if a moving object does not have the kinetic energy to endanger human, it would not cause any harm. For instance, the intelligent toys for babies and the domestic intelligent sweeping machines may deviate from a preplanned track and rush toward people, they would not cause substantial harm. But for AIAs equipped with a driving device (such as a robotic arm, wheel, or launcher), which generates kinetic energy that could endanger people as well as a hard shell or sharp tip, their mobility (combined with such kinetic energy) can pose a risk of harm to humans. Take the example of

SpotMini, a Boston Dynamics four-leg robot which would pose danger on people once it goes out of control. With a height of about 84 cm and a weight of around 30 kg, SpotMini can move fast, run with loads, and go up and down stairs. Such a robot can be easily used to harm people. Therefore, it is safe to say that an AIA which does not injure people only when it is controlled cannot be regarded as a safe system, just as a gun which is a danger itself.

6.4.1.2 AIA Can Exchange Energy with the Physical Word to Indirectly Trigger Objects to Move or Change Their Status

Although some AIAs do not have kinetic energy, a hard shell or a high speed, they can exchange energy with the physical world by emitting sound, light, electricity, and magnetism to change the behavior or status of other objects, thus causing harm. For instance, a maliciously controlled laser, medical or power-generating device or a smart speaker can radiate the energy that goes far beyond the normal scope and thus endangers human.

A moving AIA leaves people nowhere to hide, and AIA can release the energy unbearable for people. These are the key factors attributed to the uncontrollability of AIAs which leads to their harm to human beings.

6.4.2 AIA Has an Unexplainable Ability to Make Decisions

People learn through observation, for a person's memory is limited and he/she needs to learn and summarize through constant "transition". On the contrary, AI learning is buttressed by a strong computing capacity and a huge storage space, and AI may overtake human in terms of learning effects and develop behavioral models that are difficult to understand for mankind. We can create AIAs, but we are not sure the behaviors of the AIAs can be predicted because AI algorithms are unexplainable in most cases. For instance, when we adopt a machine learning algorithm based on the neural network, it is difficult to explain why an input leads to its corresponding output. For another, if we use an intelligent algorithm searched with the optimum solution, probability factors would make the behavioral outcomes of AIAs unpredictable. The reason why AlphaGo won the chess competition with special techniques lies in the unpredictability of its behaviors.

The unexplainability of AI does not sound weird because unexplainability can also be found in human beings. For instance, a coach spends the same time training two different persons to improve their basketball shooting skills. But after the training, one trainee may be able to shoot without a single miss, while the other may fail to do so. The reason for the difference resides in the unexplainability of human behaviors. If something is unexplainable, it will be surely unpredictable, for a great deviation will naturally lead to some unpredictable results. For example, if an actress plays the role of the famous Chinese volleyball player Jane Lang in practicing spiking, it

is impossible for her to achieve a high accuracy of spiking as Jane Lang. In a worse scenario, her spiking may miss the target and hit the camera.

Another possibility, however, must also be taken into consideration: an extreme hacker may obtain the right to control AIAs through network intrusion or backdrop import and alter their target functions, thus mobilizing AIAs to endanger human. If the victim AIAs are large or aggressive, the risk would be harder to control. For instances, if a hacker intrudes and alters the target function of a drone swarm, the drone swarm may attack the owner instead. For another, if a hacker changes the target function of a self-driving car, the car may suddenly speed up where it is expected to come to an emergent stop.

Unexplainability can be a source of risk because it would result in unpredictability, and something unpredictable is something that cannot be forestalled. Unexplainability is different from uncontrollability, for the former is an issue deriving from people's inability to predict the results of AIAs' behaviors while the latter is a matter of randomness. Randomness is also unpredictable, but it is not adversarial. For example, if an out-of-control car rushes towards a crowd, the people can avoid it without being chased; unexplainability, however, is like a car controlled by a ghost to intentionally hit a person and the target has no place to hide. Unexplainability may be exactly what hackers take advantage of.

Obviously, unexplainability prompts what cannot be forestalled and serves as a source of risks AIAs impose on human beings.

6.4.3 AIA Has the Ability to Evolve into an Autonomous System

Once AIA begins self-evolution and even starts to set target functions autonomously, it may escape the "cage" set by people and go beyond the control of human beings. Obviously, if out-of-control AIA is merely a car running aimlessly, it may only prompt safety problems. But if it evolves into the point that it protects its own "right to survive" and endangers humanity out of self-protection, it will become an autonomous system adversarial to people. This will be a disaster for mankind. The autonomous system here is not confined to the one that "is not controlled by people". For instance, autonomous weapons are also not controlled by people on battlefields, but that is merely a matter of the ability to make decisions. The autonomous system in question has "awareness" and knows how to protect itself as people do. Its target function is to safeguard itself from any harm, and it even has the ability to communicate with others and organize activities. In this case, it will surely become an adversary of humanity.

6.5 Measures for Preventing the Out-Of-Control Performances of AI

6.5.1 International Standards for Man–Machine Collaboration

In March 2016, International Organization for Standardization (ISO) released a technical standard—"ISO 15066 Robots and robotic devices—Collaborative robots". Drafted by 24 ISO members, it is a detailed supplement to "ISO 10218–1/2" and includes four standards for safe man–machine collaboration and relevant risk assessment norms [22]. Moreover, it covers clear standards for the industrial community and urges the robot manufacturers and parts developers in the future to follow the standards, with the aim of ensuring the safety of laborers in a man–machine collaboration environment. The four standards are as follows:

Stop operation under safety monitoring: When a laborer enters the collaboration zone, the robots will stop moving and keep safe inaction so that the laborer can perform his/her duties such as placing to-be-processed workpieces onto robots and replacing tools for the robots; when the laborer leaves the zone, the robots will resume their operation automatically.

Hand guiding: An operator uses a manually-operated device to send motion commands to a robot system. The operator manually operates the guiding device installed on a robot or near the actuator of a robot, and then the robot finishes a designated task.

Speed and separation monitoring: A minimum safe distance is maintained between a robot and a laborer. Once the distance between the two is shorter than it, the robot will come to a halt.

Power and force limiting: External collision or stress can be detected so that a robot would not harm the body of a laborer in case of his/her unexpected intervention in the process of assembling.

6.5.2 Asilomar AI Principles

The development of AI drove the international community to pay attention to the ethics and security of AI. At the Beneficial AI Conference held in Asilomar, California in January 2017, the experts and scholars jointly signed the Asilomar AI Principles [23], calling on those working in the realm of AI across the world to observe these principles to ensure the interest and safety of humanity in the future.

Formulated by some organizations led by the Future of Life Institute (FLI), the Asilomar AI Principles consists of 23 principles which fall into three categories, namely, "Research Issues", "Ethics and Values", and "Long-term Issues". "Research Issues" covers the duties of researchers in the development of AI systems and the

"tough problems" that may occur in computer science, economics, legal ethics, ethics and social research. It clearly states that the objective of AI research should be developing beneficial rather than disorderly AI. "Long-term issues" includes achieving a balanced allocation of resources for the future development of AI, making appropriate development plans and reducing the possible risks on AI systems, especially disastrous or existential ones. "Ethics and Values" plays a central role in the Asilomar AI Principles, which is composed of Principles 6–18:

Principle 6 Safety: AI systems should be safe and secure throughout their operational lifetime, and verifiably so where applicable and feasible.

Principle 7 Failure Transparency: If an AI system causes harm, it should be possible to ascertain why.

Principle 8 Judicial Transparency: Any involvement by an autonomous system in judicial decision-making should provide a satisfactory explanation auditable by a competent human authority.

Principle 9 Responsibility: Designers and builders of advanced AI systems are stakeholders in the moral implications of their use, misuse, and actions, with a responsibility and opportunity to shape those implications.

Principle 10 Value Alignment: Highly autonomous AI systems should be designed so that their goals and behaviors can be assured to align with human values throughout their operation.

Principle 11 Human Values: AI systems should be designed and operated so as to be compatible with ideals of human dignity, rights, freedoms, and cultural diversity.

Principle 12 Personal Privacy: People should have the right to access, manage and control the data they generate, given AI systems' power to analyze and utilize that data.

Principle 13 Liberty and Privacy: The application of AI to personal data must not unreasonably curtail people's real or perceived liberty.

Principle 14 Shared Benefit: AI technologies should benefit and empower as many people as possible.

Principle 15 Shared Prosperity: The economic prosperity created by AI should be shared broadly, to benefit all of humanity.

Principle 16 Human Control: Humans should choose how and whether to delegate decisions to AI systems, to accomplish human-chosen objectives.

Principle 17 Non-subversion: The power conferred by control of highly advanced AI systems should respect and improve, rather than subvert, the social and civic processes on which the health of society depends.

Principle 18 AI Arms Race: An arms race in lethal autonomous weapons should be avoided.

The Asilomar AI Principles is regarded as "The 23 Guidelines" for AI development. As a representative document regulating the AI ethics and security in the international community, it has strong international influence and will be further improved.

6.5.3 *Self-termination Mechanism for Forestalling the Risk of Going Out of Control*

At the 32nd Conference on Uncertainty in Artificial Intelligence (UAI) in 2016, Google DeepMind and the Machine Intelligence Research Institute of Oxford University jointly published a paper entitled "Safely Interruptible Agents" [24]. The paper focused on how to prevent an unmanned system with a reinforcement learning algorithm from evolving and thus going beyond the control of human, such as banning the use of its own power or setting up action-off buttons. Taking the example of the Q-Learning reinforcement learning algorithm, the paper demonstrated that an AI system with such an algorithm would not regard the manual interruption by the operator as a "negative feedback" and thus neglect it. Nevertheless, the interruptibility of other reinforcement learning algorithms remains to be studied. In general, it adopts an algorithm-based method that prevents an AI system from going out of control due to self-learning, and it must be ensured that the algorithm will not be able to learn "how to avoid the intervention from a human operator".

6.6 Summary

In this chapter, we have introduced two existing AI safety problems—the safety accidents caused by AI system failures and the international arms races that may be incurred by the development of AI weapons. Also, we have mentioned public concerns about the possible out-of-control behaviors of AI. By exploring AIA and analyzing the risk of its out-of-control behaviors, we have summarized three main factors of the out-of-control behaviors of AIA and described the international measures for preventing the out-of-control behaviors of AI.

References

1. Defcon displays the technologies of attacking Tesla software: they will cause collisions [EB/OL]. [2019–11–20]. https://auto.qq.com/a/20160808/006422.htm.
2. NTSB. Accident investigations: car with automated vehicle controls crashes into pedestrian. [2019–06–19]. https://www.ntsb.gov/investigations/Pages/HWY18FH010.aspx.
3. Unglück im VW-Werk Baunatal: Roboter tötet Arbeiter. [2019–08–19]. https://www.hna.de/lokales/kreis-kassel/baunatal-ort312516/roboter-toetet-arbeiter-vw-werk-5191637.html.
4. A man killed by a robot welding car parts at factory in India. (2015–12–15) [2019–11–20]. https://www.ishn.com/articles/102997-a-man-killed-by-a-robot-welding-car-parts-at-factory-in-india.
5. Robotic surgery has been connected to 144 U.S. Deaths Since 2000. [2019–08–22]. https://gizmodo.com/robotic-surgery-has-been-connected-to-144-u-s-deaths-s-1719202166.
6. Cerrudo C, Apa L. Hacking robots before Skynet. [2019–08–22]. https://ioactive.com/pdfs/Hacking-Robots-Before-Skynet.pdf.

7. Horcher G. Woman says her Amazon device recorded private conversation, sent it out to random contact. (2018–5–25) [2019–10–21]. https://www.kiro7.com/news/local/woman-says-her-amazon-device-recorded-private-conversation-sent-it-out-to-random-contact/755507974.
8. Creepy Facebook artificial intelligence story that captivated the media—file edge. [2019–12–26]. http://www.fileedge.com/creepy-facebook-artificial-intelligence-story/.
9. Amazon has figured out what's behind Alexa's random laugh that was freaking people out. (2018–3–8) [2019–10–21]. https://www.usatoday.com/story/tech/2018/03/07/alexas-weird-random-laughter-freaking-people-out/404476002/.
10. Peter L. Learning from Tay's introduction. [2019–9–22].
11. IBM researcher feeds Watson supercomputer the 'urban dictionary'; very quickly regrets it. [2019–9–22]. https://www.techdirt.com/articles/20130110/14542221635/ibm-researcher-feeds-watson-supercomputer-urban-dictionary-very-quickly-regrets-it.shtm.
12. IBM's Watson gave unsafe recommendations for treating cancer. [2019–9–22]. https://www.theverge.com/2018/7/26/17619382/ibms-watson-cancer-ai-healthcare-science.
13. SRG-A1. [2019–11–12]. https://www.roboticstoday.com/robots/sgr-a1.
14. Academics boycott KAIST over "killer robots" concerns. [2019–12–18]. https://www.theengineer.co.uk/boycott-kaist-autonomous-weapons/.
15. Asimov I. Runaround. In: Asimov I, Robot I, editors. New York: Bantam Dell, A Division of Random House, Inc., 1950, p. 25.
16. Dickson B. Distinguishing between narrow AI, general AI and super AI. [2019–11–20]. https://bdtechtalks.com/2017/05/12/what-is-narrow-general-and-super-artificial-intelligence/.
17. Bostrom N. Superintelligence: paths, dangers, strategies. Oxford, United Kingdom: Oxford University Press; 2014.
18. Hawking: I apologize, but AI may destroy humanity (with the full text of the speech). [2019–11–20]. http://www.3snews.net/column/252000045632.html.
19. Elon Musk's AI warning: artificial intelligence is a 'potential danger to the public'. [2019–11–20]. https://www.express.co.uk/news/science/1204119/Elon-Musk-AI-warning-Artificial-Intelligence-danger-Neuralink-Elon-Musk-latest.
20. AI companies, researchers, engineers, scientists, entrepreneurs, and others sign pledge promising not to develop lethal autonomous weapons. [2019–11–1]. https://futureoflife.org/2018/07/18/ai-companies-researchers-engineers-scientists-entrepreneurs-and-others-sign-pledge-promising-not-to-develop-lethal-autonomous-weapons/.
21. Academician from Chinese Academy of Engineering: AI must be equipped with AIA safety hoop. [2019–11–20]. https://new.qq.com/omn/20180701/20180701A08480.html.
22. ISO/TS 15066:2016 Robots and robotic devices—collaborative robots. (2016–02). [2019–08–27]. https://www.iso.org/standard/62996.html.
23. Asilomar AI principles. [2019–09–02]. https://futureoflife.org/ai-principles/.
24. Orseau L, Armstrong S. Safely interruptible agents. [2019–10–27]. http://intelligence.org/files/Interruptibility.pdf.

Chapter 7
Artificial Intelligence Actant

Hui Lu, Zhihong Tian, and Binxing Fang

When we talk about artificial intelligence and its systems, we refer to applications that use artificial intelligence technologies, commonly known as artificial intelligence systems (AIS), or autonomous and intelligent systems (A/IS). Since this chapter intends to set the stage for a later discussion of the potential threats posed to human beings by AIS or A/IS, the subjects discussed for this purpose are no longer simply AIS or A/IS, but rather a special type of AIS or A/IS that has certain capabilities and carries a potential risk of harm to human beings. To this end, this book specifically defines the concept of "artificial intelligence actant" (AIA) to explore the potential risk of AIS or A/IS to harm human beings.

7.1 Definition of Artificial Intelligence Actant

An artificial intelligence actant (AIA) is a type of autonomous hardware entity which can sense the external environment, use it as input, make decisions through internal algorithms and interact with the physical world by built-in driving devices. Normally an AIA, when in motion, has the kinetic energy to harm human beings. The external environment above includes the natural environment and related organisms. Internal algorithms are the processes that are embedded within an AIA and used to make decisions. A drive is a piece of hardware that can interact with the physical world or change the spatial coordinates of AIA, which has the kinetic energy to harm human

H. Lu (✉) · Z. Tian · B. Fang
Guangzhou University, Guangzhou, China
e-mail: luhui@gzhu.edu.cn

Z. Tian
e-mail: tianzhihong@gzhu.edu.cn

B. Fang
e-mail: fangbx@cae.cn

© Publishing House of Electronics Industry 2025

B. Fang (ed.), *Artificial Intelligence Security and Safety*,
https://doi.org/10.1007/978-981-96-9263-7_7

beings when it is in motion. Autonomy means that an AIA can set its own objective function and make its own decisions. An AIA generally has its own "disposing capacity" (movable) or can indirectly trigger changes in its behavior or state by exchanging energy with the physical world.

Based on the definition of AIA, all kinds of robots, including autonomous driving systems, AI physical assistants, and even AI weapons are all AIAs. Take for example the robot soldiers in the AI weapon: the external environment refers to the characteristics of a target identified by the robot soldier's sensors; the internal algorithm refers to the internal procedure which determines the attacking target and makes decisions using the feature recognition technology. The drive is the robot's mechanical legs and its attack weapons. Such robot soldiers can move freely and make their own decisions on a battlefield. This chapter mainly classifies and describes some examples of AIAs. Due to the variety of robots and AI products, a few typical examples are selected herein for illustration.

7.2 Type 1 AIAs—Traditional Type

7.2.1 Industrial Robots

Industrial robots, which combine multiple technologies, including control engineering, robot dynamics, simulation technology, program design, intelligent measurement and modeling, are a typical type of AIA. As a new kind of intelligent production tool, industrial robots can extend the human mind and, by virtue of its intelligence, digitalization and networking capacity and other advanced features, improve the industrial production efficiency significantly. Since industrial robots boast a fleet of varieties, several representative ones and their characteristics will be illuminated in the following.

7.2.1.1 Spotwelding Robots

A spotwelding robot [1] (Fig. 7.1) is mainly composed of a robot body, a computer control system, a teaching box and a spotwelding system. Because of its complex working environment, the spotwelding robot is required to have a high degree of freedom to fulfill functions like waist turn, upper arm turn, forearm turn, wrist turn, wrist swing and wrist twist. Spotwelding robots can perform spot welding automatically based on the programmed moves, time delay and other parameters. In addition, they have the ability to communicate with peripheral parts by receiving a series of control commands from the computer through a communication interface.

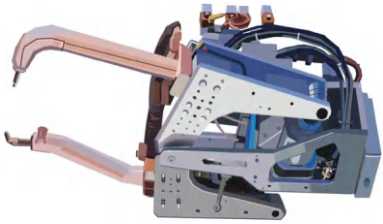


Fig. 7.1 A spotwelding robot

Fig. 7.2 A vacuum robot

7.2.1.2 Vacuum Robots

A vacuum robot [2] (Fig. 7.2) is a robot that operates in a vacuum. It is commonly used in semiconductor chip processing to transport wafers in a vacuum. As a high-precision robot, the vacuum robot has a special design for the shaft to minimize problems such as shaft asymmetry and inertia asymmetry. A vacuum robot detects the offset of the reference position through its sensors, and corrects its movement track through a special algorithm to ensure the efficiency and accuracy of its work.

7.2.1.3 Arc Welding Robots

An arc welding robot [3] (Fig. 7.3) resembles the spotwelding robot and is mainly used in industrial production. They all belong to welding robots and therefore share a similar composition and structure. Different from spot welding robots, arc welding robots can track the welding seam by analyzing the feedback information from the vision sensor and the laser sensor and thus adapt themselves to complex workpieces. In addition, arc welding robots use visual sensors to obtain the residual errors in the welding process, and automatically corrects the motion trajectory by estimating the errors to to ensure that they can complete the welding task in a complex working environment with high quality and high efficiency.

Fig. 7.3 An arc welding robot

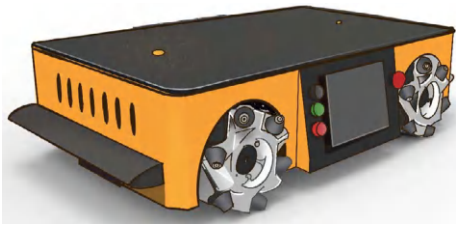


Fig. 7.4 A mobile robot

7.2.1.4 Mobile Robots

Mobile robots [4] (Fig. 7.4) boast functions including human–computer interaction, mobility, sensor control and automatic navigation. They are mainly used in modern logistics systems, to improve the timeliness and accuracy of traditional logistics and establish a point-to-point logistics system of refinement and flexibility. Mobile robots are usually used in the workplace that needs high precision handling and transmission. Sometimes, they are also used as transport robots in some places, including but not limited to stations, airports and post offices.

Mobile robots adopt the multi-sensor fusion technology to sense the environment, which overcomes the shortcomings of using a single sensor for environment perception. By using the sensor positioning technology and map construction technology, mobile robots construct information about the surrounding environment based on some features observed in motion so that they are able to locate its posture and position and build the environment map incrementally.

7.2.1.5 Boston Dynamics Robots

Robots developed by Boston Dynamics are, in a sense, also industrial robots. With the evolution of technology, they have now gone far beyond the original design idea of transporting military supplies on the battlefield.

Fig. 7.5 SpotMini

From BigDog, a military robot that made Boston Dynamics famous in 2005, to Atlas, a humanoid biped robot and its pinnacle creation released after its acquisition by Google, and then Handle, a wheel-leg robot which can actually complete some automated warehouse work, and finally SpotMini, the cute and steady robot having wowed the Internet, Boston Dynamics robots are becoming increasingly sophisticated with breathtaking dexterity. Their capabilities to navigate, to surmount obstacles, to maintain balance, to recognize targets and self-repair are all constantly expanding. SpotMini can walk freely through an office building, climb stairs, avoid obstacles and chart route like a pet dog.

(1) SpotMini

The arms of SpotMini (Fig. 7.5) are capable to move and grab, open and close doors. They are mainly used in homes and offices, but they can also be applied in specific industries by customizing the design of the grips and arms. Their sensor system includes a depth camera, a stereo camera, an inertial navigation module, and a position/force sensor to help the robot navigate autonomously.

(2) Atlas

The Atlas (Fig. 7.6) is a humanoid robot, which coordinates the movement of hands, legs and the body with a control system, so that it can perform complex movements in a limited space like a human being. Equipped with a sensor system, a stereo camera and other components, Atlas can walk over rough terrains and climb up on its own.

(3) Spot Classic

The Spot Classic (Fig. 7.7) is a quadruped robot designed for outdoor operation. Shaped like a dog, it is stable and maneuverable, and uses stereo vision and radar systems to sense its surroundings. Full autonomous navigation is also available. In practice, it can be used to transport small quantities of goods on complex terrains.

(4) WildCat

The WildCat (Fig. 7.8) is also a quadruped robot known for its high speed, which can reach up to 32 km/h.[1] In addition to running fast, it can also use a variety of gaits

[1] https://www.bostondynamics.com/legacy.

Fig. 7.6 Atlas

Fig. 7.7 Spot classic

such as trotting and bounding. Its advanced sensor systems, including laser sensors and visual odometers, can accurately measure the height and attitude relative to the ground.

Fig. 7.8 WildCat

Fig. 7.9 LS3

(5) Ls3

LS3 (Fig. 7.9) is a quadruped robot with excellent mobility designed to follow soldiers and carry their gear over rough terrain. It can interact with soldiers in a natural way, just as a trained animal interacts with its owner. With the help of technologies such as terrain sensing, obstacle avoidance and GPS to ensure its safety during the journey, an LS3 can get to the target location successfully.

7.2.2 Dedicated Service Robots

7.2.2.1 Polar Research Robots

Polar research robots are used to assist scientific exploration in polar environments, represented by snow/ice mobile robots and low-altitude flying robots (Fig. 7.10), both of which have played an important role in China's Antarctic scientific exploration initiatives [5].

The ice/snow mobile robot is shaped like an armored vehicle and offers legendary off-road capability to adapt to the complex environment of the polar region.

Fig. 7.10 An ice/snow mobile robot and a low-altitude flying robot for polar research

7.2.2.2 Anti-Terrorist Robots

A typical anti-terrorist robot [6] is shown in Fig. 7.11. Anti-terrorist robots mainly include attack robots, EOD robots and reconnaissance robots. The attacking robots are mainly used for military operation; the EOD robots are mainly used in the transfer and removal of suspicious dangerous substances and the explosives; the reconnaissance robots are mainly used for reconnaissance missions as well as patrol and monitoring on the battlefield. All the three kinds of robots mentioned above are designed to perform specific tasks by detecting the external environment, and then feeding it back to the internal system and then being processed by the system's algorithms. The following mainly introduces reconnaissance robots and EOD robots.

The primary task of a reconnaissance robot is to scout the battlefield; as a result, compared with anti-terrorist robots, a reconnaissance robot values more the reliability of concealment and transmission function. Reconnaissance robots are usually small and light, able to carry wireless transmission equipment and micro-camera, and sometimes can perform special functions like climbing walls.

EOD robots are mainly used for the transfer and removal of suspected hazardous materials and explosives. They usually have multiple replaceable arms with strong off-road capability. They are designed to simulate the sophisticated handling of bombs by humans, and therefore boast high operating flexibility and precision. Compared with the reconnaissance robots, EOD robots require more complex technology, including higher operational reliability and robustness. Countries worldwide have launched their own EOD robots: Dragon Runner [7] (Fig. 7.12), a small EOD robot developed by the United States; RMI [8] (Fig. 7.13), produced by Pedsco in Canada; and Cyclops, Bison, Ground-hog, Lyn and other models of EOD robots developed by ABP in the UK.

Fig. 7.11 An anti-terrorist robot

Fig. 7.12 Dragon runner

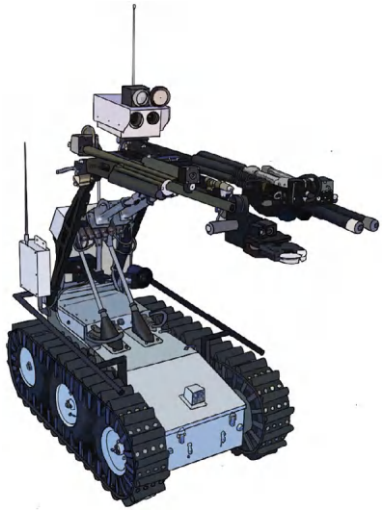


Fig. 7.13 RMI

7.2.2.3 Rehabilitation Robots

Integrating technologies such as intelligent control, rehabilitation robots mainly provide help and auxiliary treatments for patients who are physically challenged. Users control the robot through visual feedback and various input devices so that it is able to perform simple tasks including eating, reading and walking.

The new Welwalk system [9] introduced by Toyota is a typical rehabilitation robot (Fig. 7.14). It is a simulated robotic leg with an electro-mechanical framework. After the patient puts on the Welwalk, the internal sensors will cooperate with the patient's upper body and detect the real-time movements of the legs and make adjustments allowing the legs to bend and straighten.

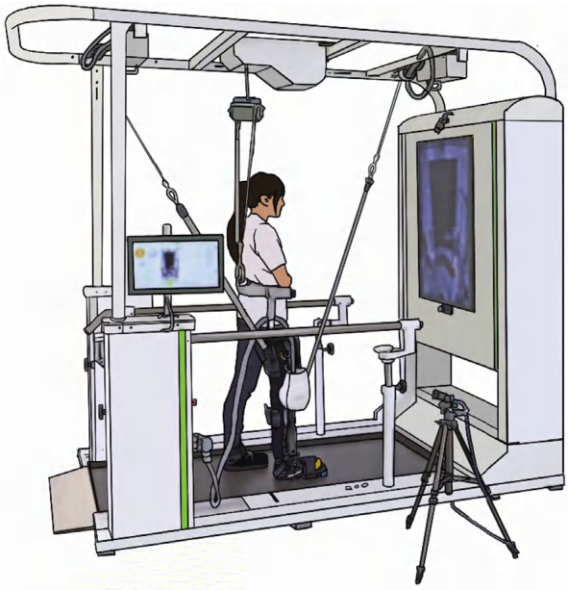


Fig. 7.14 A rehabilitation robot

7.2.3 Home Robots

7.2.3.1 Home Education and Entertainment Robot

Home education and entertainment robots [10] are embedded system hardware products that combine multiple technologies including electronic technology, computer technology, control technology and artificial intelligence. As typical AIAs, they also integrate motion control, navigation with visual recognition, sensor data acquisition, intelligent monitoring and artificial intelligence. The controller, sensor and servo drive system respectively correspond to the human brain, facial features and torso of the limbs.

The home education and entertainment robots on the market currently fall into three categories based on their market positioning and walking style: desktop interactive robots, wheeled home robots and bipedal education robots. Among them, the desktop interactive robots cannot move, so they don't count as AIAs.

(1) Wheeled home robots can walk freely at home, usually on two wheels. Aside from the main task of accompanying children and caring for the elderly, they will patrol the house and report abnormalities through autonomous navigation technology. They are housekeeper robots with multiple functions.
(2) As the most complex robots among home education and entertainment robots, the biped education robots can walk upright like a human being. This requires not only a high performance of the hardware, but also a sophisticated and mature software control algorithm that controls the walking gait and balance of the robot

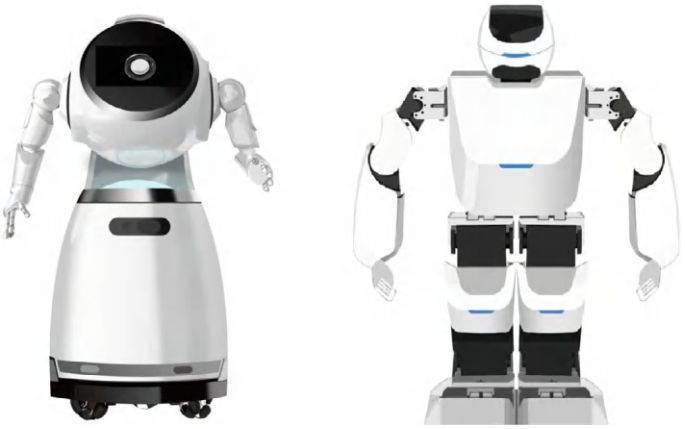


Fig. 7.15 A wheeled home robot and a bipedal education robot

in real time. Currently, due to the high cost and complex control algorithm, large-scale biped robots cannot be put into use in household yet. In the academic field, small biped robots are usually used as teaching aids in robot teaching to attract more robot lovers and students to join in the development of biped robots. Wheeled home robots and bipedal education robots are shown in Fig. 7.15.

7.2.3.2 Cleaning Robots

Cleaning robots usually include vacuum cleaner robots, mopping robots and window cleaning robots (Fig. 7.16). Small in size and specialized in function, they use the sensors to position and clean, with typical AIA features.

A vacuum cleaner robot is essentially a moving vacuum cleaner, which moves through motor transmission, senses obstacles through the sensor, and controls the walking route through the intelligent system to clean the floor. Most vacuum cleaner robots can recharge themselves automatically by going back and forth. Mopping robots work in much the same way as vacuum cleaner robots do, except that the

Fig. 7.16 A vacuum cleaner robot, a mopping robot, and a window cleaning robot

vacuum cleaners are replaced with mops. They can only walk in a relatively limited range because they will have a close contact with the floor when working. With the occasional need to add water, they cannot recharge themselves. Window cleaning robots mainly use the air pump to vacuum the space between themselves and the glass surface so that they can stick to the glass with better effect and higher safety. There are two types of window cleaning robots: frame-type window cleaning robots, whose activity range is determined by collision inspection; and frameless window-cleaning robots with both collision detection and edge infrared detection.

7.3 Type 2 AIAs—Self-driving Vehicles

7.3.1 Autonomous Driving

Autonomous driving is a kind of unmanned automated driving system, in which the actant can acquire data from the environment through sensors, assemble the artificial intelligence system algorithm, and plan the path independently. Actants can be found in vehicles such as cars, planes and ships. Take autonomous cars as an example: six levels of driving automation are defined by the *J3016 Standard Taxonomy and Definitions for Terms Related to On-Road Motor Vehicle Automated Driving Systems* [11] developed by SAE International.

Level 0: No Automation. All actions of a vehicle are completely dictated by a human being, without any assisted or automated driving technologies.

Level 1: Assisted Driving Automation. Safety systems like adaptive cruise control (ACC) are equipped to assist the driver, but he/she still needs to physically maneuver the vehicle, keeping hands on the steering wheel and staying attentive to the surrounding road conditions.

Level 2: Partial Automation. Level 2 vehicles can perform some driving tasks for a short period of time which will allow the drivers to take a short break. But drivers are still required to be alert at all times to take control when necessary because these vehicles are not capable of coping with all complex traffic conditions.

Level 3: Conditional Automation. Level 3 vehicles can almost take full control of operating without human intervention. Drivers can be on standby and don't need to keep hands on the steering wheel. This is the beginning of autonomous driving.

Level 4: High Automation. Level 4 vehicles are true autopilots. They are required to move through the traffic lanes and pull over independently without any human intervention. In this process they need to weigh the possible consequences of their actions, and therefore they need to have the ability to judge right from wrong, good from bad.

Level 5: Level 5 vehicles have full automation technology and require no involvement of humans. It is the ultimate goal of autonomous driving.

Fig. 7.17 Tesla's self-driving car

Full Automation requires massive time and effort. Although innovative technology in engineering application has made great progress, there is still a long road ahead in putting autonomous driving into practice in a large-scale.

7.3.2 Self-driving Vehicles

7.3.2.1 Tesla

A self-driving vehicle is a car capable of moving safely without human involvement based on a combination of technologies including artificial intelligence technology, radar, GPS, and visual computing technologies. For the six levels of autonomous vehicles mentioned above, Level 2 vehicles are the current sweet spot for many automakers. Among the numerous companies which study self-driving cars worldwide, the most well-known is Tesla in the United States.

Tesla, Inc. is a leading American electric vehicle and clean energy company that specializes in manufacturing and selling electric vehicles, solar panels and energy storage devices. Tesla's self-driving car (Fig. 7.17) has eight cameras around the body to get a 360-degree view and the front radar has a detection range of 160 m. The 12 ultrasonic sensors equipped on the body make the entire vision system more complete and can detect hard objects and soft objects.[2] In addition, the front radar can detect rain, fog, dust and even the space under the vehicle in front by emitting radar waves with redundant wavelengths. The vision system is more powerful by the rich data provided.

Tesla Autopilot is composed of four parts: adaptive cruise control, lane centering, automatic lane changes and self-parking. Lane centering and automatic lane changing are mechanisms designed to keep a car centered in the lane and change the lane automatically. The car will overcome obstacles and change the lane automatically after the front camera have identified lane lines and the front and rear radar have detected the surroundings of the vehicle. Adaptive cruise control is an available

[2] https://www.tesla.cn/modelx.

cruise control system for road vehicles that automatically adjusts the driving speed to maintain a safe distance from vehicles ahead. When the car is on the road, the front distance sensor (radar) continuously detects the road condition ahead while the wheel speed sensor will collect information of vehicle speed. When the distance between the car and another vehicle ahead is too short, the adaptive cruise control system can coordinate the actions of the anti-lock braking system and the engine control system, allowing the vehicle to brake and reducing output power of the engine. In the future, when available, users will be able to remotely control the vehicle using a mobile APP to achieve automatic parking; and the vehicle will be able to accurately determine road conditions and decide whether to go or wait while driving.

Tesla Autopilot has undergone a gradual upgrade. The first-generation autopilot system, Autopilot 1.0, was installed in models such as the Model S in October 2014, which used control chips from other companies, with a single view camera in the front. The system may have some defects such as the lack of stereovision. In May 2016, a fatal crash occurred when a Tesla Model S vehicle was driving against the sun and failed to identify a container truck lying crosswise on the road.

In October 2016, Tesla rolled out Autopilot 2.0, a disruptive innovation from its old Autopilot 1.0, with newly designed sensors and chips as well as an extra set of sensing devices like sensors, cameras and radar. The optimized car has eight cameras which allow 360 degrees of visibility to detect objects; 12 ultrasonic sensors, deployed around the car, can provide real-time updates about the distance of the car from surrounding obstacles; the radar installed at the front can detect the speed and distance of objects within a range of 250 m. Since Autopilot 2.0, Tesla's autopilot systems have provided 300-degree visibility and collection of information on speed and distance, and thus, in theory, realized fully automatic driving. For the time being, this mode has been recognized as the minimum requirements of Level 3 driving automation.

Tesla's Autopilot 3.0 was first revealed by Elon Musk in the second-quarter concall in 2018 [12]. Compared with its predecessors, this new model has an upgraded module of chips that are supported by AI deep-learning technology, which allows enhanced performance in driving automation. Tesla cars equipped with Autopilot 2.0/2.5 can get their chips upgraded for free in 2019. As chips upgrade and AI develops in full bloom, fully automated driving will gradually take shape.

Tesla aside, many innovation-inspired companies, such as Waymo, Cruise and Baidu, have made stunning progress in self-driving technologies.

7.3.2.2 Waymo

Waymo, a subsidiary of Alphabet Inc. (Google's parent company), started as a self-driving car project of Google in January 2009, and was founded as a company independent from Google in December 2016. Waymo's self-driving cars (Fig. 7.18) make use of ultrasonic sensors, laser radar, cameras and other high-performance self-driving hardware sensors designed and fabricated on their own, whose software and hardware are able to collaborate seamlessly. In addition, Waymo self-driving vehicles

Fig. 7.18 Waymo self-driving car

also feature an unique audio detection system, which works to detect ambulance or police vehicles hundreds of meters away, helping plan a route in advance to avoid relevant traffic.

Waymo has tested much longer mileage than other competitors in this field, for it has constantly tested its software by simulation, and learnt from collective real driving experiences to make constant progress. Its open Beta tests involving users have helped Waymo to occupy their mind and make it a word-of-mouth brand. Waymo has no financial difficulties due to the formidable support of its parent company Alphabet, so there is no need for it to reveal to the public its technological advantage for the sake of financing, and its uttermost system of confidentiality also gives full play to its superiority.

7.3.2.3 Cruise

Cruise is a subsidiary of General Motors. In January 2018, it released its Level 4 4th-generation self-driving vehicle—Cruise AV (Fig. 7.19), which is the first pilotless automobile designed without a steering wheel or brake pedal. Supported by Cruise, General Motors is vigorously developing its Level 4 and even Level 5 autonomous driving technologies, to ensure its advantageous position in the field of autonomous driving.

Fig. 7.19 L4 autonomous vehicle Cruise AV

Fig. 7.20 Baidu Nomancar

7.3.2.4 Baidu

Baidu's Nomancar Project started in 2013, boasting its core technology called "Baidu AutoBrain", mainly including the four modules of high-precision map, positioning, perception, and smart decision and control. The high-precision map surveyed by Baidu itself records complete 3D road information, capable of fulfilling centimeter-level positioning accuracy for vehicles. Meanwhile, Baidu Nomancar, supported by some internationally leading traffic scene object identification technology and environmental perception technology, has realized high-precision vehicle detection and identification, tracking, distance and speed estimation, lane segmentation and lane detection, to provide bases for making smart autonomous driving decisions.

Baidu has built up its technical expertise in Baidu Map, big data, Baidu Brain and so on, which have been applied to its pilotless vehicle to shift from "human to information" connectivity to "human to service" connectivity. Meanwhile, Baidu has also purchased some technology companies to build up its technology strength. Baidu Nomancar (Fig. 7.20), with radar, cameras, global satellite navigation and other electronic devices, has brought all-round perception for pilotless driving. As it were, Baidu is fully prepared for its role in the field of pilotless driving at home.[3]

7.3.3 AI Unmanned Aerial Vehicles

AI unmanned aerial vehicles serve as aircrafts to plan its flight route planning, take off and land autonomously, avoid obstacles, and even conduct aerial refueling, without human involvement. A typical example for AI unmanned aerial vehicles would be X-478 AI unmanned combat aerial vehicle (Fig. 7.21) from the United States, which is reputed as the first AI unmanned combat aerial vehicle (UCAV) in the world. Nicknamed "Pegasus", it was developed by Northrop Grumman, USA. In 2013,

[3] https://baike.baidu.com/item/%E6%97%A0%E4%BA%BA%E9%A9%BE%E9%A9%B6%E6%B1%BD%E8%BD%A6/77997?fr=aladdin.

Fig. 7.21 X-47B AI unmanned combat aerial vehicle

X-47B conducted the first autonomous take-off and landing on an aircraft carrier; in 2015, this UCAV had autonomous information exchange with a tanker airplane, conducting automated aerial refueling and de-docking from a tanker system, before returning to its base. Apparently, the United States is a leader in the field of AI unmanned aerial vehicle development and manufacturing. As a matter of fact, X-47B is not completely out of human control, for there is still a telecontrol system out there on the ground, to conduct manual control, in some cases where AI fails to address some problems.

7.3.4 Pilotless Vessels

Pilotless vessels involve various new marine technologies, including automated ship tracking, collision prevention, sailings and other technologies. These technologies do not only relieve the workload of deck officers, but also work to ensure navigation safety and improve economic returns to a great extent.

Algoritmi is a concept pilotless cargo vessel [15] (Fig. 7.22), designed by Aziz Belkharmoudi, mainly aimed at building a whole-course pilotless ocean-going vessel, which is 800 m long, powered by a turbine generator driven by a nearly 400 MW compact atomic fusion reactor, to sustain its safe navigation at the rate of 40 sea miles per hour. This vessel is pilotless with sophisticated technologies, including AI controlled ultrasonic and laser sensors, radar and a collision avoidance system, and 360-degree cameras, satellite and GPS navigation systems.

Fig. 7.22 A conceptual diagram of Algoritmi pilotless cargo vessel

7.4 Type 3 AIAs—AI Weapons

7.4.1 Overview of AI Weapons

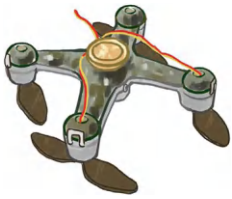
AI weapons are typical applications of AIAs in the field of military weaponry. AI weapons, facilitated by AI-assisted decision-making systems, information acquisition and processing systems, task execution systems, knowledge base systems and other technologies, are not only able to autonomously conduct spying, searching, aiming, attack and other military tasks, but also carry out automatic collection, reduction and analysis of general information.

7.4.2 "Slaughterbots"

"Slaughterbots" weapons [16, 17] (Fig. 7.23) are supposedly a type of weapons envisaged by those anti-AI weaponization scholars, which is an AI killer weapon able to conduct facial recognition to tell the enemy and ourselves apart, and attack the enemy. It is of course a lethal autonomous weapon. In November 2017, Professor Stuart Russell from University of California, Berkey, played to the public an appalling "Slaughterbots" Sci-Fi video at a conference of the UN Convention on Certain Conventional Weapons, which showed us a story that two men release about 50 UAVs from the rear of a wagon, which results in the loss of 8300 lives during synchronized strikes. Professor Russell is among those researchers who stay steadfast against the applications of AI technologies to weapons. This is an imaginary scene that he has designed with some existing technologies, meant to expose people to the horror of possible AI weaponization, so as to uphold the view of anti-AI weaponization.

The "slaughterbots" conceived by Professor Russell is composed of an autonomous anti-personnel system (APS), which, together with some sophisticated AI technologies as well as built-in sensors and omnidirectional cameras, works to conduct obstacle avoidance, random motion, high-efficiency casting of ammunitions, facial recognition, multiple self-localization protocol, big data chaining and other functions. The facial recognition function of the camera works to tell its targets from

Fig. 7.23 Slaughterbot



irrelevant personnel within some milliseconds' duration, which helps the "slaughterbot" to bear down upon and kill its targets in a split second, and carry out a lethal attack with explosives aboard. Meanwhile, the "slaughterbot" boasts a counter sniper technology, to readily elude arrest, and carry 3-g explosives, to penetrate and attack like a swarm, as is enormously lethal.

7.4.3 Turret-Type AI Combat Robots

Samsung Techwin SGR-A1 sentry guard robot [18, 19] (Fig. 7.24) is a type of defensive AI weapon—Turret-type AI combat robot, the very first AI weapon system integrating monitoring, tracking, shooting and audio recognition systems into one unit. It is capable of detecting an invader, and sending alarms to commanders, and meanwhile, it suppresses the enemy to attain its purpose of defense. SGR-A1 Sentry Guard Robot is aimed at protecting some primary military bases and national strategic points. This system is intended to replace human guards, avoiding any adverse impact caused by fatigued personnel, bad weather and insufficient staffing, so as to ensure the completion of its guard tasks.

SGR-A1 sentry guard robot mainly makes use of a specific monitoring system, which is able to detect and track targets within a range of 4 km in the daytime and a range of 2 km at night. Controlling personnel can invoke the ID of a sentry guard robot to view the field photos taken by the camera on a sentry gun.

Fig. 7.24 Techwin SGR-A1 sentry guard robot

Fig. 7.25 Nerekhta

Fig. 7.26 Armata T-14

Russia's Nerekhta (Fig. 7.25) is a mobile turret AI combat robot [20], similar to a mini-type unmanned tank, which is mainly intended to fight against armored machinery in a combat, and at the same time provide transportation and spying capabilities. Moreover, a Nerekhta-based robot pilot system can carry explosives and move to a designated spot to conduct explosion.

Armata T-14 (Fig. 7.26) is a semi-automatic robot tank [21] proposed by the Russian military, which features an unmanned turret responsible for air defense and reconnaissance. Armata T-14 can simultaneously track multiple air or ground targets, with a built-in tracking system to provide an automatic shooting solution for destroying a target. This tank also has an advanced "active protection system" that automatically scans the surrounding environment to identify any threat.

7.4.4 Human-Shaped Combat Robots

The Russian combat robot FEDOR is a relatively sophisticated human-shaped combat robot (Fig. 7.27).[4] As an armed AI robot holding guns, it is dedicated to some tasks in the space, and protecting astronauts. In addition to shooting, FEDOR is also able to take other actions, like walking, jumping, crawling, manipulating various

[4] https://www.rbc.ru/magazine/2018/01/5a3948139a794769dd2a086c.

Fig. 7.27 A FEDOR combat robot

tools, weight-lifting, driving and so on. FEDOR is still improving constantly. It is predicted that FEDOR will be sent to the space by 2021, possibly becoming the only passenger aboard a spacecraft of Russian Federation.

7.5 Type 4 AIAs—Humanoid Robots

7.5.1 Jules

In 2006, Hanson Robotics, Hong Kong, China and its partner Personality Forge jointly launched a project, and developed their robot Jules [22] (Fig. 7.28). Jules is a miraculous sophisticated robot, thanks to its natural language AI, computer vision and facial tracking and recognition and other advanced technologies.

Researchers have applied multiple micromotors under the Frubber skin of the face of Jules, to demonstrate all-round emotional facial expressions and reactions.

Fig. 7.28 Jules

Fig. 7.29. Bina48

7.5.2 Bina48

Bina48 [23] (Fig. 7.29), released in 2010, is a humanoid robot developed by Hanson Robotics, Hong Kong. Originally, it was a system capable of recording "Mindfiles", developed under the leadership of an American foundation dedicated to developing digital duplication technology. The so-called "Mindfiles" refers to human memory, thought, views and so on. On this basis, this foundation entrusted Hanson Robotics with a job to design an AI program, so as to re-present a human-shaped robot, alive and kicking before us. Finally, the designers chose the "Mindfiles" of Bina Aspen Rothblatt a lady who had passed away to conduct design.

Bina48 can be connected to the Internet, capable of discussing various topics like politics, religion and future technology. She has ears (microphones) and eyes (cameras), and is able to read, listen and speak as supported by some current software, and even learn to make progress during her dialogue with others. In autumn 2017, Bina48 became the first AI robot that had finished her college curriculum.

7.5.3 Sophia

Hanson Robotics has also developed a bio-robot Sophia [24] (Fig. 7.30), a copy of the prototype actress Audrey Hepburn. Compared with other various robots developed by the company, she has more human-shaped appearance and more humanoid behaviors. She is expected by the researchers that she could learn autonomously, adapt to human behaviors, and go with humans. Sophia boasts various capabilities, including but not limited to facial recognition, image data processing, imitation of human expressions and gestures, as well as brief communication with humans regarding some topics arranged in advance. Thanks to the speech recognition technology of Alphabet, Sophia has its AI program to analyze dialogues and take their data, helping her to learn autonomously and become smarter and smarter.

Fig. 7.30 Sophia

7.6 Features of AIAs

Sections 7.2–7.5 of this chapter has covered various AIAs in line with the definition of AIA in Sect. 7.1, i.e. "a kind of autonomous hardware entity that is capable of perceiving the external environment and taking it as input, making decisions through its internal algorithm, and interacting with the physical world with its own driving devices, and it is usually capable of harming humans during operation." It has the three features of mobility, decision-making ability and some autonomy. That is to say, the three features can be used as bases to determine whether an AIS or A/IS is an AIA.

7.6.1 Mobility

When it comes to AIS or A/IS, most of us would call to mind AlphaGo [25], an AI program researched and developed by DeepMind under the umbrella of Google which has an extraordinary ability to play go. AlphaGo uses a "value network" to predict the winner of the game from each position, and a "policy network" to select the next move to play. In March, 2016, AlphaGo defeated Lee Se-dol, a Korean Go master, by 4:1 in a series of intense games. AlphaGo was thus accredited as a 9-dan chess player by Korea Baduk Association. In January 2017, AlphaGo again defeated Gu Li, 9-dan chess player. On May 27, 2017, Chinese master Ke Jie also lost to AlphaGo at the Chinese Go Summit. Apparently, AlphaGo is powerful, as recognized by human beings, and more intelligent than humans, as accepted by us. We, however, have never been worried about whether it will harm human beings, not just because AlphaGo is just a game, but more importantly AlphaGo is an immobile AI program system that cannot drive a hardware entity. Thus, we are worry-free.

Likewise, a desktop interactive robot (Fig. 7.31) is also an A/IS for us humans to easily come into contact with, mainly serving us in some interactive entertainment games. It is relatively simple, whose value is usually embodied in enhancing human–machine interaction experience. Robot JIBO is a typical desktop household

companion robot, designed by Cynthia Breazeal, an expert in the field of household robots. It can capture the facial features of humans with its camera to identify family members, and have voice interactions with them. It can also demonstrate various moves and gestures, and at the same time is somewhat capable of learning. Its long-term accompanying will help it to remember the habits and behaviors patterns of each family member. We all take it as a lovable AI system, and no one would be afraid of any threat from such a desktop interactive robot, simply because it has no mobility, and naturally is not able to attack us.

These discussions are meant to probe into whether AIAs will bring any threat to human beings, i.e. should an AIA be out of control, would it endanger humans? Apparently, the mobility of an AIA is vital, given that AlphaGo and desktop interactive robots are not our concerns, for even they surpass humans in terms of intelligence, they cannot move, so they cannot directly hurt us humans.

Just imagine: There is an uncertain object moving in our physical space, and it surely will be a threat to humans, for we do not know how to handle such an uncertain moving object. That is to say, uncertain mobility presents an essential element of threat to human beings. Therefore, AIAs first of all should have the feature of mobility.

Of course, the kinetic energy of AIAs is also an important indicator to measure their threat to humans, for we are concerned about whether their kinetic energy is powerful enough to be violent. Section 7.2.3 has listed robot cleaners as AIAs, but we are not actually worried about whether a cleaning robot would attack us. Nevertheless, it is possible for it to hurt some pets, if its mobility is further increased. Things would be quite different for other various robots of mobility, for the mobility of AIAs can be shifted to violence, which together with its uncertainty would make it impossible for us to prevent, as is naturally an essential element of threat to human beings.

Fig. 7.31 A desktop interactive robot

7.6.2 Decision-Making Capability

In the packaging industry, packaging robots are sometimes employed, like robot palletizers are a type of packaging robots, playing a role on the packaging production line [26]. We would like to talk about it in this separate section mainly in that the packaging production line highlights its automation, while the packaging robots boast their adaptive capability. The former is to repeat established tasks unconditionally, while the latter is to fulfill predetermined assignments based on its judgment as the circumstances may require. Thus, automation and AI bear different senses. Automation is just to stress the repeated operations without human intervention, and this operation mode is a fixed procedure set up by human in advance. It is not our concern, for its moving track is determined, and will not present any threat to humans. AIAs, however, would make decisions on their movement judging the external environment based on their predetermined algorithms.

The key lies in that AI algorithms are usually lack of interpretability, resulting in unpredictable tracks of behaviors of a robot, and thus making its behaviors uncertain. For example, Sect. 7.2.1 tells us that the arc welding robots have multiple pivot joints, so there are multiple options for its motion tracks for it to choose from during its selection of welding spots. Then, what does its motion track look like? Is it random or fixed? If the motion track is based on a pre-set algorithm, this would be fine for the motion pattern is explicable. If the motion track is obtained via machine learning, like repeatedly guiding the robot arm to point to the welding spot from different angles during its training by human beings, then how will the robot arm after training make use of the knowledge it has learnt to form a motion track to point to the welding spot at a given moment? Machine learning algorithms are not interpretable, so we cannot pinpoint at which moment the robot arm will form what motion track.

Now, it is clear that the mobility of an AIA and its kinetic energy are essential factors of its threat to human beings, and then how the mobility could be shifted to a violence element? The precondition still lies in how the decision-making mechanism of AIAs works to control their motion patterns. Thus, the decision-making capability is the second feature of AIAs. Section 5.3.2 "adversarial examples reveal that AI algorithms are short of explainability" tells us that the adversarial example attack patterns formed based on the "unexplainability" of the algorithms have been proposed by many research teams at home and abroad, and various world-class matches of neural network attack and defense also tell us that adversarial samples are a common attack technique. The "slaughterbots" mentioned in Sect. 7.4.2 identify an object to kill only based on facial recognition, so we have every reason to assume: how will the "slaughterbots" not be applied for sure with adversarial examples intentionally to kill a person by mistake? This is also an important reason for us to repeatedly mention the explainability problems in various design norms, guidelines and specifications for AI when AI codes of ethics are addressed in Chap. 10.

7.6.3 Autonomy

The autonomous and intelligent systems (A/IS) are named so because we believe they are autonomous. One of the primary causes lies in that this system is capable of learning, so it is in the position to constantly improve its own specific capability. An AIA is intelligent not only thanks to its simple AI application, but also due to its long-term accumulation of massive background knowledge, or model parameters. Just imagine: A Chinese Chess (Xiangqi) robot program, if just knowing how to follow the rules for how each type of chess pieces to move on the board, will not be that remarkable, for its judgement and decision on when and how to move a piece matter much more than the rules, and correct judgement relies on long-term accumulation of experiential data.

Another example could be an associable Chinese character input method application program, whose most valuable part is not the program itself, but its knowledge base built up through long-term learning, like rule base, model parameters, knowledge map, state automata and so on, which is used to show at what probability a group of spellings sequence input is mapped on a group of Chinese characters sequence. If this associable Chinese character input method application program is not backed by this knowledge base, it would be just like an infant without any knowledge and thus has to learn from scratch. Therefore, the long-term accumulated knowledge base is the real treasure house of this associable Chinese character input method, and it is just that different input methods have different knowledge bases formations. Of course, this makes some difference in accuracy between different input methods. Thus, this knowledge base is fixed, but is constantly enriched during the process for us to constantly enter characters and unceasingly correct our mistakes in our selection with this Chinese character input method. Objectively, this input method is constantly learning and improving, as is an evolutionary conduct.

Moreover, the construction form of model parameters bases, state automata and the like knowledge bases are not always to be easily interpreted by human beings, and their consequent AIA conducts are not interpretable. Now, it has been believed that AIS or A/IS are fulfilling the tasks entrusted by humans, whose objective functions are set by humans. For example, the cleaning robot mentioned in Sect. 7.2.3 is determined and required by humans for it to take an algorithm to traverse the entire work space to clean it up; the cleaning robot will not deliberately throw over a trash can and mess up the environment. Nevertheless, an AIA has its autonomy which might guide it to deliberately strike down a trash can and destroy the environment. Because when configuring an AIA, we do not only need to set its objective functions, but also are required to set corresponding reward functions. The purpose for setting the objective functions is to let the AIA to know what to do, while the purpose for setting the reward functions is to set a goal for the AIA to seek, to determine whether its performance is satisfactory. If the objective function is set to clean more and get more award, an AIA, to obtain ever-higher rewards, might repeatedly turn down a trash can and clean up the trash it has just made. In such a case, its conduct of

throwing over the trash can is a conduct out of the autonomous decision-making of the cleaning robot.

Likewise, for the mobile robot mentioned in Sect. 7.2.1, delivering an article to a place is an objective of a task assigned to it by us. This objective is obviously issued by human beings, instead of resulting from autonomous decision of this mobile robot. However, after taking this assignment, the AIA needs to make autonomous decisions on its fulfilling process, to complete the required logistics tasks. And this fulfilling process—s electing an optimal route, and dealing with environmental disturbance during delivery—is a result of this AIA's autonomous decision. An extreme example would be: If this AIA encounters an unavoidable barrier block during the logistics process, how will it react? If it has gone through a learning procedure to move away the barrier, will the AIA choose to force its way against the barrier? If this barrier is a dog, will the AIA master some skills to deal with such an adversarial barrier block? Apparently, all these things might occur in an open environment.

The above discussions show that autonomy is the third feature of an AIA. If an AIA's harm to human beings does not result from it's being out of control, like a pilotless automobile losing its control and harming the human due to its inertia, its autonomous selection to harm human beings would bring about a much more dangerous scenario, for the AIA's ability to learn has become difficult for we humans to keep a lookout.

Of course, some people may believe that AIAs are man-made as moving actants constructed on a closed set, so they would never act beyond what they have been taught by humans, like an operation on the set $\{0,1\}$ that will never produce any results within the character set. Some professional AI researchers, in particular, sometimes even deny the creativity of an AIS, for they know that all that an AIS has learnt are from us humans. Without human teaching, an AIS can never do it, and at least it is unreasonable for an AIS to choose a result that has not been taken as correct. For example, the robot of Boston Dynamics mentioned in Sect. 7.2.1 has two feet to walk, but they would never evolve on their own to become a Kung Fu master. The problem is that we might have ignored one fact, i.e. the association capability of AI. If the robot of Boston Dynamics gets hold of some essential capabilities, it will observe and learn, and even become a Kung Fu master as long as he masters the evaluation function, as will not be an act outside the closed set. As a matter of fact, from the perspective of an AI professional, in case we have taught a robot how to select a reward function, like: If it falls down as a result, the reward function is negative; if not, the award function is positive. And a robot is capable of taking random acts, and after long-term training and trials, the robot will surely identify an optimal mix of acts, to attain an optimal status, which cannot be taught. After all, there are various reasons for making these robots attain some ability beyond the prediction of humans.

From another perspective, in the actual human world, we do not judge a person based on how capable he is, but check whether he is doing it for a just purpose. Likewise, AIAs will still be dangerous even if they cannot do what they have been taught by humans; they might still be risky, for they might improperly use the capabilities taught be people, and in nature, this is because in the process of autonomous evolution of AIAs, it goes beyond the cognition of humans that an AIA autonomously

chooses a result because the transition of evaluation functions changes the objective functions originally set by us.

In fact, to date, we humans in most cases fail to understand or explain some AIA selection methods and behavioral patterns in the first place, so the change of objective functions will naturally be hard for us human being to perceive.

7.7 Summary

This chapter defines AIA, as an original intention to pave the way for further probing into the potential risks that might be brought about by an AIS or A/IS to humans. An AIA might be dangerous for it has three features: First, mobility, and some kinetic energy during operation, as is an essential contributor to violence. Second, decision-making capability, as is a major cause of its uncertain behavior, which presents a major threat to human beings. Third, autonomy. Once an AIA can set its objective functions in its full autonomy, and its behaviors are not explainable, we humans might feel that we are placed in a dangerous situation of uncertainty in the face of such an AIA.

In fact, AI development has experienced three stages of behaviorism, symbolism and connectivism. Behaviorism features the main ideas of perception plus action, deeming that intelligence is to be learnt and trained in the complex environment of the real world, while symbolism highlights logic, and connectivism data. It is hard to train in a real world, so behaviorism has not yet developed well. However, behaviorism represents a constant trial and error process, so it has a better character of evolution, which has laid a foundation for the autonomy of AIAs.

References

1. Yuan JM. MOTOMAN spot welding robot system and its application. Metalworking: Therm Process. 2008;(14):35–8.
2. Li YJ. Design and implementation of wafer transfer vacuum manipulator teaching system. Shanghai: Shanghai Jiao Tong University; 2014.
3. Fu YB, Bian MH, Xu XG. Application and development of arc welding robots. Therm Process Technol. 2006;35(11):79–81.
4. Three Key Technologies of Mobile Robots [EB/OL]. (2019–06–28) [2019–10–24]. https://robot.ofweek.com/2019-06/ART-8321202-11000-30395779.html.
5. Top Ten Robot News of 2007. Robot Technol Appl. 2008;(2):6.
6. Li TT, Li Q, Liu SY, et al. A review of anti-terrorism robots. China Secur Prot Technol Appl. 2018;03:30–3.
7. Dragon Runner? Small Unmanned Ground Robots[EB/OL]. [2019–10–24]. https://qinetiq-na.com/products/unmanned-systems/dragon-runner/.
8. Remote Mobile Investigator (RMI) [EB/OL]. [2019–10–24]. http://www.pedsco.com/products-rmi9xd.php.
9. Toyota launches rental service for the Welwalk WW-1000 rehabilitation assist robot in Japan [EB/OL]. (2017–4–12) [2019–10–24]. https://global.toyota/en/detail/15989382.

10. Song YN, Liu SC. Opportunities and challenges of home educational and entertainment robots. Robot Industry. 2017;3:44–51.
11. Taxonomy and definitions for terms related to driving automation systems for on-road motor vehicles [EB/OL]. (2016–9–30) [2019–10–25]. http://standards.sae.org/J3016_201609.
12. Tesla Model S, X with "Hardware 3" for full self-driving now in production, inventory codes indicate [EB/OL]. [2019–10–25]. https://www.teslarati.com/tesla-autopilot-hardware-3-model-s-x/.
13. X-47B UCAS makes aviation history again! [EB/OL]. [2019–10–25]. https://www.northropgrumman.com/Capabilities/X47BUCAS/Pages/default.aspx.
14. Northrop Grumman X-47B (2019–1–24) [2019–10–25]. https://www.militaryfactory.com/aircraft/detail.asp?aircraft_id=1015#overview.
15. Autonomous Ships of the Future [EB/OL]. [2019–11–29]. https://www.yankodesign.com/2016/07/26/autonomous-ships-of-the-future/.
16. Ban Lethal Autonomous Weapons [EB/OL]. [2019–11–19]. https://autonomousweapons.org/.
17. Why You Shouldn't Fear "Slaughterbots" [EB/OL]. [2019–10–26]. https://spectrum.ieee.org/automaton/robotics/military-robots/why-you-shouldnt-fear-slaughterbots.
18. Samsung Techwin SGR-A1 Sentry Guard Robot [EB/OL]. [2019–10–26]. https://www.globalsecurity.org/military/world/rok/sgr-a1.htm.
19. SGR-A1 [EB/OL]. [2019–10–26]. https://www.roboticstoday.com/robots/sgr-a1-description.
20. Russia's AI Killer Robots Will Change War Forever [EB/OL]. (2017–11–15) [2019–10–26]. https://www.outerplaces.com/science/item/17088-russia-war-robot-nerekhta.
21. Russian Military Unveils T-14 Armata Semi-Autonomous Tank [EB/OL]. (2015–7–24) [2019–10–26]. https://www.roboticsbusinessreview.com/public-safety/russian_military_unveils_t_14_armata_semi_autonomous_tank/.
22. Jules Custom Character Robot [EB/OL]. [2019–10–26]. https://www.hansonrobotics.com/jules/.
23. Bina Custom Character Robot [EB/OL]. [2019–10–26]. https://www.hansonrobotics.com/bina48-9/.
24. Sophia [EB/OL]. [2019–10–26]. https://www.hansonrobotics.com/sophia/.
25. AlphaGo [EB/OL]. [2019–10–26]. https://deepmind.com/research/case-studies/alphago-the-story-so-far.
26. Palletizing Robots [EB/OL]. [Accessed 2019–11–16]. http://www.hnsprs.com/chanpin/zidongmaduoshengchanxian/63.html.

Chapter 8
AIA Safety Hoop

Binxing Fang, Xiang Cui, and Zhihong Tian

As described in Chap. 6, an AIA may pose great threat to or even endanger human beings once the "three elements of uncontrollability" of AIAs appear. Therefore, it is important to take prevention and control measures in advance for AIAs. To improve the controllability of AIAs, Binxing Fang, Academician of the Chinese Academy of Engineering and editor-in-chief of this book, has proposed the concept of AIA safety hoop (hereinafter referred to as "safety hoop") in a report titled "*My Insight into AI Safety*" made by him in the China International Software EXPO held on June 30, 2018. Here we take robot, the most representative AIA, as an example. Any robot should have a driving mechanism and a decision system. We can put the Safety Hoop between the driving mechanism and decision system, and the decision system must be driven only through the Safety Hoop. Just as a fuse will automatically blow when there's an abnormal current to protect the safety of the power consumption system, the safety hoop will be activated when certain conditions are met, and then restrain the robot. By this way, measures can be taken to control a robot once it starts to lose control, and prevent it from becoming a destructive power and endangering human beings.

The AIA safety hoop needs to have several key characteristics. For example, the AIA safety hoop should have the function of remote monitoring. In case of a risk, it can remotely stop the machine, cut off the power, and log the status. The robot cannot take the AIA safety hoop off by itself. Besides, the AIA safety hoop cannot

B. Fang · Z. Tian
Guangzhou University, Guangzhou, China
e-mail: fangbx@cae.cn

Z. Tian
e-mail: tianzhihong@gzhu.edu.cn

X. Cui (✉)
Zhongguancun Laboratory, Beijing, China
e-mail: cuix@zgclab.edu.cn

© Publishing House of Electronics Industry 2025

B. Fang (ed.), *Artificial Intelligence Security and Safety*,
https://doi.org/10.1007/978-981-96-9263-7_8

be removed intact with an external force; otherwise, it will inactivate or even self-destruct itself. The AIA safety hoop shall set a security fence for the robot which is like the no-fly zone of an Unmanned Aerial Vehicle (UAV) so that the robot can only normally function in a certain area. Any artificial intelligence system should be equipped with an AIA safety hoop upon factory delivery, once it meets the three elements of uncontrollability; otherwise, it shall not be launched. We can compare the AIA safety hoop to the fuse of a power system.

Now, still at the exploration stage, the AIA safety hoop has received no consensus nor proportional attention from the society. However, it is necessary to predict and explore its possible forms and functions to prevent AIAs from its designed mission of benefiting human beings.

This chapter will propose a possible AIA safety hoop and the architecture of its ecological environment, including the basic functions, safety mechanism, security fence, and enhanced functions of the AIA safety hoop, the AI prevention and control center (hereinafter referred to as the "prevention and control center"), and standards and specifications. Moreover, this chapter summarizes the already formed and relevant ecological environment of the AIA safety hoop. It is worth noting that this chapter is a theoretical exploration of the general requirements and some functions of the AIA safety hoop, and thus development of specific AIA safety hoops should take into account the market needs.

8.1 Architecture

Figure 8.1 shows the architecture and external interfaces of the AIA safety hoop. The AIA is controlled by the driving mechanism, while the driving mechanism runs according to the instruction given by the decision system of the AIA. Therefore, we only need to embed an AIA safety hoop between the decision system and driving mechanism of the AIA to effectively control the AIA, if necessary. For example, we can suspend the driving operation of the AIA or change (e.g. slow down) the driving mode of the AIA. The AIA safety hoop can change the behavioral model of the AIA. Especially, such a change can skip the decision system of the AIA to restrict the activities of the AIA. Therefore, it is necessary to set strict restrictions about when the AIA safety hoop can work to avoid an unwanted inference with the AIA. Hence, the AIA safety hoop must restrict, change, and suspend the AIA and accurately judge the status of the AIA, such as whether the AIA goes beyond safety boundaries or has an unsafe behavioral model. Meanwhile, the AIA safety hoop must be safe as well and must not be maliciously initiated. In other words, it shall have a complete safety protection capability. Furthermore, the AIA safety hoop can be remotely controlled by an authorized external system. In this way, the behavior of the AIA can be restricted externally.

The AIA has been introduced in detail in Chap. 7, and we will belabor the point here. This chapter focuses on the basic functions, safety mechanism, security fence,

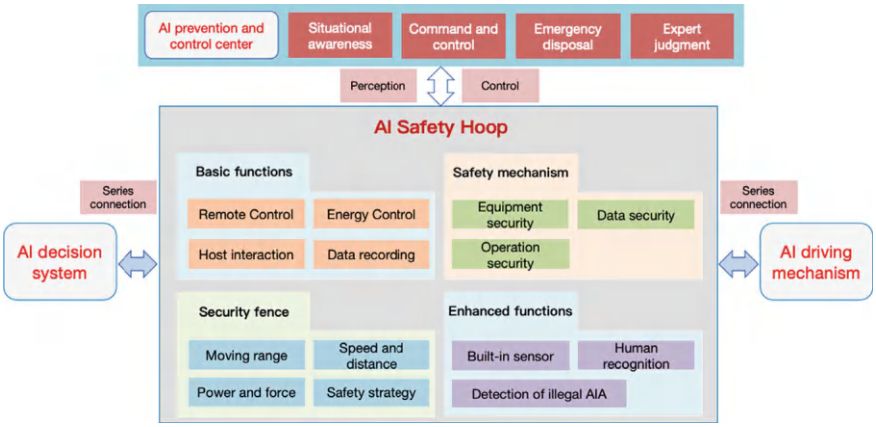


Fig. 8.1 Architecture and external interfaces of the AIA safety hoop

and enhanced functions of the AIA safety hoop, the prevention and control center, and the ecological environment.

8.2 Basic Functions

The basic functional module of an AIA safety hoop includes four components, namely, remote control, energy control, host interaction, and data recording, which constitute the minimum set of functions for all AIA safety hoops. A remote-control component serves as a bridge between the AIA safety hoop and the prevention and control center. The energy controls component restricts the activities of the "host" (the AIA), when danger is detected. The host interaction component perceives the environment through the host's sensors and changes the host's actions by controlling the host's driving mechanism. The data recording component realizes a function similar to the "black box" of an aircraft.

8.2.1 Remote Control Component

Remote control means that the AIA can receive external commands and control. This component aims to receive commands from the "prevention and control center" and pass the commands as the highest priority to the host for implementation. The remote-control component needs to have the networking capability and manipulates components like "energy control", "data recording", and the "security fence" in line with the control command to achieve the expected actions. Such actions include but are not limited to stopping or limiting energy supply, limiting the movement

speed of the host, limiting the output power of the host, locking the host, restoring factory settings, starting or stopping a task, initiating the full-traffic recording mode, initiating the metainformation recording mode, and passing specific data back to the cloud. Certainly, control commands like locking the host should not be used at will, unless a high-risk situation endangering human security is perceived. The concept of "Command and Control" in remote control and network security have similar connotations (please refer to *Survey of Botnets* [1] for more details about "Command and Control").

From the perspective of security risks, though its functions are not complicated, the remote-control component faces many security risks posing challenges for its design and implementation. For example, data may pass through base stations, optical cables, switches, routers, communication satellites, during the initiation of connection and subsequent communication between the remote-control component (client) and the prevention and control center (server). Any of these processes may be maliciously monitored. The transferred data may be maliciously manipulated, being captured, tampered, or intercepted. During these malicious operations, if a control command from the prevention and control center is tampered with and the remote-control component fails to identify the false control command, the attacker can forge a malicious control command from the prevention and control center and lead to serious consequences. Similarly, if a malicious attacker pretends to be the prevention and control center to send a seemingly true control command to the remote-control component and bypass the verification mechanism of the component, serious consequences will be caused as well. The above-mentioned examples are just general attacks. In an emergency, a malicious attacker may mobilize large-scale distributed resources to congest and disturb or physically attack the prevention and control center, and even attack the communication infrastructure of the AIA safety hoop from multiple dimensions. Consequently, the command and control channels between the prevention and control center and the large-scale AIA safety hoop will be blocked or even cut off, followed by the loss of the commanding right. All the risks above will become more complex and diversified along with the continuous emergence of new technologies. Thus, the remote-control mechanism shall be designed in a complete and forward-looking manner. As remote control is designated to prevent the AIA safety hoop from getting out of control, the remote-control component is expected to ensure its own safety. Nevertheless, as is known to all, even the engine of an aircraft cannot be absolutely safe, so several backup engines are required. Therefore, countermeasures against the interruption of the remote-control channel of the AIA safety hoop shall be drafted in advance.

In terms of communication, the remote-control component occupies a significant position in the AIA safety hoop, because it acts as the only "bridge" between the AIA safety hoop and the outside. It carries two categories of communications, that is, downward (from the server to the client or from the prevention and control center to the AIA safety hoop) control commands and upward (from the client to the server) data and statuses. The AIA safety hoop can implement network communication via satellite communication, 4G/5G communication, NB-IoT, or LoRa, gives priority to communication methods with low energy consumption, wide coverage,

low latency, and high security, and prepares for possible problems like severe weather, underdeveloped network, and communication interference.

With respect to safety attributes, the downward channel of the AIA safety hoop is expected to ensure the four basic safety attributes of control commands, namely, authenticity (not fake), confidentiality (not disclosed), integrity (not tampered with), and non-repudiation (not denied), while the upward channel needs to ensure the confidentiality, integrity, and timeliness of data and statuses. All safety attributes in this book are defined below.

Authenticity: An AIA safety hoop only accepts the instruction from an authorized controller. Third parties can neither forge the identity of the controller, nor re-issue an instruction of the controller.

Confidentiality: Even if they are monitored or intercepted maliciously, communications cannot be restored in cleartext.

Integrity: As the communication link cannot be completely controlled, communication tampering is unavoidable. Both parties verify the hash of the communication to judge whether it is tampered with. Once the integrity of the communication is identified to be destructed, the existence of an attack will be perceived. And countermeasures will be adopted.

Non-repudiation: All control commands that are issued or received need to be recorded for the subsequent audit. Certainly, the records shall be signed in order to prevent them from being tampered with.

Timeliness: Control commands, data, and statuses shall be transferred to the destination within an acceptable delay, which is especially important when the AIA is detected to have a sign of loss of control.

In regard to control commands, common remote-control commands of the AIA safety hoop include but are not limited to registration, alarm, energy control, restoration of factory settings, and default policies. See the examples below.

Registration: When it is activated each time, the AIA safety hoop needs to report its status to the prevention and control center. The registration message shall contain a globally unique AIA safety hoop ID and a globally unique AIA ID.

Alarm: The AIA safety hoop will report a risk to the prevention and control center, once it perceives it, including the AIA safety hoop ID, the AIA ID, geographic location, risk description, environmental images, and environmental sounds.

Energy control: The prevention and control center commands the AIA safety hoop to cut off the energy supply for its host to stop the host.

Restoration of factory settings: The prevention and control center sends an instruction to the AIA safety hoop to command it to restore the AIA to factory settings.

Default policies: The prevention and control center sends an instruction to the AIA safety hoop to command it to perform the default policy by default (without command input) or disconnection (unable to connect to the control center). Default policies differ from other commands. The latter must be obtained through the network, while the former can come from the default factory configuration. Given that the AIA safety hoop cannot fully ensure normal communication, default policies are particularly significant, which results in some disputes. For example, can default policies be

modified through the network? If so, there will be some risks. If not, once the default factory configuration has an error, a substantial cost will be required to update the AIA safety hoops that have been widely deployed.

8.2.2 Energy Control Component

One of the three elements of uncontrollability of AIAs is "mobility". A necessary factor for a mobile device is energy supply and the conversion of energy into kinetic energy. When it detects that the host is or is about to be out-of-control, or receives a relevant control command from the prevention and control center, the AIA safety hoop shall activate the energy control component to timely control the host's behavior and minimize the hazard. We can see that the energy control component can interact with the host and control its driving mechanism, making it an indispensable and basic component of the AIA safety hoop.

The energy control component needs to have the capabilities to preferentially occupy, interrupt, or limit the energy supply to the host so as to control the host's use of energy. Currently, the mainstream energy supply methods of the AIA are batteries and electric motors. Hence, the energy control component can give priority to the control of electricity. Along with the increase in the energy supply methods of the AIA, the energy control component may need to manipulate multiple types of energy, such as thermal, optical, and chemical energy. For instance, the preferential occupation of energy by the energy control component is similar to the occupation of the laptop battery. The little remaining battery level is always occupied by the operating system to save the data and status of the laptop from loss.

Generally, the energy control component needs the host to provide support in line with standards and specifications so as to implement effective access and come into play without destroying the host. Therefore, standards and specifications occupy an inseparable and core position in the establishment of a benign AI ecological environment. Relevant contents will be elaborated in Sect. 8.7 in this chapter. Does the energy control component interact with the host directly, or must it communicate with the host indirectly through the host interaction component introduced in Sect. 8.2.3 like other components? Since this is not a major concern, it will not be discussed in this chapter.

8.2.3 Host Interaction Component

The function of the host interactive component is to implement three-party communication with the "decision system" and "driving mechanism" of the host in accordance with the common standards and specifications in order to open the interaction channel between the AIA safety hoop and the host. If the remote-control component is like an external "bridge" between the AIA safety hoop &the host and the outside,

the host interaction component will be a "bridge" connecting the AIA safety hoop and the host. In consideration of large-scale and widely distributed AIA integrators, component manufacturers, and software providers, common industry standards and communication protocols must be formulated, which is a challenging and complicated task. Optimistically, many countries leading in AI have clearly presented such an idea which will be elaborated in Sect. 8.7 in this chapter.

From the perspective of communication, the data sent by the host interaction component to the host's decision system (to manipulate the driving mechanism of the host) are called "upward communication". In the opposite direction, the host interaction component needs to collect data on the host which are called "downward communications".

Upward communications can be control commands or default policies, including but not limited to, stop, deceleration, acceleration, limitations of distance, moving range, power, and force, self-destruction, and restoration of factory settings. When the remote-control component receives a control command to manipulate the host or the AIA safety hoop detects that the host deviates from the normal behavior model, the host interaction component will send such information to the host. Strictly speaking, the energy control component should be part of the host interaction component. However, in consideration of the great importance of the former, this book regards it as an independent component.

Downward communications mainly refer to the data collected by the host interaction component from the host's sensors. The AIA sensors emphasized by the host interaction component include but are not limited to cameras, microphones, global positioning systems (GPS), the acceleration sensor, the temperature sensor, and the pressure sensor which can perceive data on the host, such as ambient images and sounds, physical position, speed changes, environment temperature, output power, force, and vibration. Through the sensors, the host interactive component can collect data on the environment and status of the host in a real-time manner, and output them to the "security fence" introduced in Sect. 8.4 for data analysis, automatic detection of anomalies, and initiation of countermeasures.

8.2.4 Data Recording Component

The data recording component, similar to the "black box" of an aircraft, safely records and saves the "key statuses" of processes of the host like normal operation, remote control, abnormal behavior, stop, and restoration of factory settings and has a GPS positioning function for ex-post tracking and location. The above-mentioned "key statuses" of the host include but are not limited to the Greenwich Mean Time (GMT), precise geographic location, system working parameters, ambient sounds and images, energy storage, acceleration, external temperature, pressure, power, and the status and motion of the AIA safety hoop.

In terms of safety and reliance, the data recording component shall contain at least the following six safety attributes.

Location: The data recording component can be located in a real-time and accurate manner so that we can timely track its physical location.

Confidentiality: Data stored by the data recording component shall be encrypted and only decoded and viewed by the authorized person for forensics, analysis, and restoration of the cause and process of an incident.

Integrity: Once the encrypted communication stored by the data recording component is tampered with, the prevention and control center can verify the hash of the communication to judge whether it is tampered with. Once the integrity of the communication is identified to be destructed, the existence of a tampering attack will be perceived.

Robustness: The data recording component is expected to be installed in a safe part of the hardware equipment of the AIA safety hoop. It needs to support the most common forms of energy storage like electricity, supply energy for weeks (the energy supply of a black box generally lasts for 30 days), and adopt multiple protection measures. Robustness will be elaborated in Sect. 8.3 "safety mechanism".

Configurability: Data type, duration of each data type, and sampling frequency are configurable. For example, whether videos and audios on the surrounding environment need to be collected and how long, whether the operating statuses of the AIA and the AIA safety hoop need to be collected, and how frequent sampling is.

Lightweight: The data recording component cannot be as big as a black box. (Usually, a black box is as big and heavy as several bricks.) Excessive volume and weight will greatly lower the performance of the AIA and disable the AIA. Therefore, the data recording component needs to balance between "lightweight" and "robustness".

8.3 Safety Mechanism

The AIA safety hoop is designated to strengthen the safety of the AI system. Therefore, the safety of the AIA safety hoop is especially important. Multiple safety mechanisms must be adopted to ensure safety. Readers may wonder that, "Since these safety mechanisms are applicable to the AIA as well, why not directly apply them to the AIA to enhance its safety?" In order to answer this question, let's observe the computer system with an extremely similar nature. In the computer system, antivirus software (AV) with high safety is often installed to protect the host operating system (e.g. Windows 7) and application software (e.g. Microsoft Office, email client, and chat software). Because the system software, especially application software, involves tremendous providers who are significantly different in safety capabilities, it is impossible to require every provider to stress and implement safety, resulting in the "bucket effect" common in information security. (A bucket's capacity is determined by its shortest stave. Accordingly, a system's safety is determined by its least safe component.) AIA manufacturers and computer software providers are extremely similar in safety capabilities and input. It is difficult to implement safety, due to the huge quantity and varied safety capabilities of the manufacturers. Furthermore, many

manufacturers are reluctant to raise the cost or reduce the ease of use of products to strengthen safety. Therefore, the AIA is bound to be vulnerable. The AIA safety hoop must provide safe and reliable services on an unsafe host. Necessary safety mechanisms include but are not limited to equipment security (anti-removal and robustness), data security (integrity and confidentiality), and operation security (fault tolerance mechanism).

8.3.1 Equipment Security

An AIA safety hoop is a small piece of hardware equipment concatenated between the decision system and driving mechanism of the AIA (Fig. 8.2). Although, in principle, the AIA safety hoop can be removed manually, the AIA will be out of control, once the hoop is removed. The anti-removal safety mechanism aims to prevent the AIA safety hoop from being removed, through technological means. Or once the hoop is removed, the situation will be timely perceived and warned by the prevention and control center to have a deterrent force.

The anti-removal safety mechanism is not exclusive to the AIA safety hoop. A similar case is the no-fly zone of an UAV. For the sake of safety, a "no-fly module" will be configured in the UAV upon factory delivery to prevent the UAV from flying to the no-fly zone specified by law. Therefore, the UAV will not fly to sensitive areas, like military bases, civil airports, and sensitive organizations. Criminals cannot take advantage of an UAV to launch an attack like intelligence reconnaissance, spying, or terrorist attacks. Thus, an UAV must be equipped with a no-fly module. However, there are still criminals and tech geeks who attempt to "remove" and "crack" the no-fly module to let the UAV "fly freely". It is a technical challenge to achieve a safe and reliable anti-removal mechanism for the AIA safety hoop from the hardware and software perspectives.

In order to ensure the equipment security, the AIA safety hoop shall have at least the following six safety attributes. Besides robustness, the rest safety attributes are about the "anti-removal" capability of the AIA safety hoop.

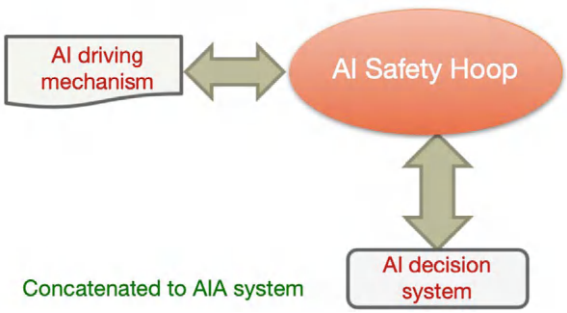


Fig. 8.2 Deployment location in the AIA safety hoop

Detectability: Once an AIA safety hoop is removed from the host, the early warning and deactivation mechanisms of the host and the AIA safety hoop will be triggered. The host will become unavailable. And the prevention and control center will be informed. This attribute is similar to the "release fuze" technology of a mine. Once the object is detached from the mine, the firing pin will hit the ignition tube driven by the spring and ignite the detonator. Hence, the technology to accurately detect the detachment of the two objects is feasible.

Controllability: After the AIA safety hoop is removed, the system will not work, because the decision system of the host is connected to the driving mechanism through the AIA safety hoop. After the AIA safety hoop is removed, the connection will be interrupted. And the driving mechanism will not be controlled by the decision system and will not function. Hence, the host will not be controlled without the AIA safety hoop which is like a fuse.

Reversibility: Once the AIA safety hoop is removed during motion, the host shall be able to detect the removal. After detecting the removal, the host shall immediately switch to the out-of-service status and will not resume operation unless it detects the AIA safety hoop again.

Identifiability: The AIA safety hoop may be removed by a criminal rather than the host for replacement. In order to confirm that the new AIA safety hoop is authorized, the host's decision system and an even number of other components (the total number shall be even so that the final judgment will come from an odd number of components) will detect and certify the new one during initiation and operation in conformity with standards and specifications. Detection aims to discover the AIA safety hoop. Certification aims to identify the authenticity of the new AIA safety hoop to prevent the original one from being replaced by a fake one. Once the new AIA safety hoop cannot pass the certification after replacement, the connection between the decision system and the driving mechanism will be automatically cut off.

Transparency: As AIA safety hoops are produced and deployed in batches, instances can be obtained to reversely analyze the software and hardware implementation methods. Therefore, it is necessary to ensure that the AIA safety hoop remains anti-removal and anti-crack after complete reverse analysis. This is like asymmetric encryption and hash algorithm in cryptography which are open and transparent, but safe.

Robustness: The AIA safety hoop shall have special protective capabilities, such as strong shock resistance, penetration resistance, high-temperature resistance, fire resistance, water tightness, and corrosive liquid resistance, and function in all severe environments. This robustness requirement is similar to that for the "black box" of an aircraft.

8.3.2 Data Security

An AIA safety hoop can be regarded as a microcomputer system, which reads data out of the host and the prevention and control center, processes the data in the

internal computing unit, through the host interaction component, and manipulates the host as an output device. During the process of input—process—output, once data and commands (strictly speaking, commands are a type of data) are lost, damaged, or tampered with, the AIA safety hoop will make wrong decisions. Hence, a data security protection mechanism is required to ensure that the AIA safety hoop can obtain correct and complete data.

Take control commands for example. The control commands obtained by the remote-control component through the network can be classified into volatile commands and persistent commands. Volatile commands are a type of validity commands. Such commands can release space after they are implemented by the AIA safety hoop, but cannot be read or written by the host. For instance, the prevention and control center requires the AIA safety hoop to report the current GPS location. Such a command is generally stored in memory on the computer. Persistent commands are commands that remain valid after the host is restarted. Such commands need to be repeatedly used by the AIA safety hoop and cannot be lost at a blackout. For example, the prevention and control center requires the AIA safety hoop to report the current GPS location every hour. Persistent commands can only be read and written to the persistent storage space by the AIA safety hoop, and cannot be read/written by the host. Default commands belong to persistent commands. However, such commands are special, because they exist in the AIA safety hoop by default and are not obtained through the remote-control channel. Generally, such commands are saved in firmware on the computer.

Besides control commands, core software, firmware, and configuration of the AIA safety hoop need to be safe. Generally speaking, the data security mechanism of the AIA safety hoop includes but is not limited to: (1) Unidirectionality of the core memory cell: Core software, firmware, configuration, and control commands are prohibited from being written, tampered with, or deleted beyond the authority; (2) Confidentiality of the core arithmetic unit: The algorithms and keys running on the internal smart chip can resist reverse analysis, and cannot be read from the outside by the host or the attacker; (3) Integrity of dynamic interactive data: The integrity of data collected from the prevention and control center and the host can be verified; And (4) Prevention of side-channel attack (SCA): The attacker is prevented from stealing confidential information by observing the power of the core arithmetic unit and the electromagnetic radiation during cryptographic operation.

Technologies and products related to data security mechanisms include dependable computing, USBKeys, and tamper-proof microprocessors. Their technologies are of great referential significance for the implementation of the data security mechanism of the AIA safety hoop.

8.3.3 Operation Security

An AIA safety hoop can automatically adopt emergency control measures against the host in case of danger or in line with a remote-control command, which poses

an extremely high requirement for it to make an accurate decision. As a complicated system, the software and hardware of the AIA safety hoop may have some vulnerabilities that cannot be identified in advance. Given that the vulnerabilities are inevitable, it is necessary to ensure the safety and reliability of the AIA safety hoop. Therefore, it is necessary to introduce operation safety mechanisms for the AIA safety hoop, such as risk assessment, safety assessment, and fault tolerance mechanisms.

8.3.3.1 Risk Assessment

Risk assessment means the quantitative assessment of the possible damage, impact, and loss of an attack before or after the occurrence of a risk. Adequate risk assessment needs to be performed during the full lifecycle of the AIA safety hoop, including design, production, use, and the time before and after an attack. Particularly, the AIA safety hoop may make a wrong decision in a real application scenario. Thus, possible safety risks in extreme circumstances shall be fully assessed and deduced.

8.3.3.2 Safety Assessment

The term, safety assessment, is derived from information security, and stands for the process of identifying the vulnerabilities of the information system. Typical safety assessment includes the assessment of safety attributes like the confidentiality, integrity, and availability of the information system. Similarly, safety assessment shall be conducted on the vulnerabilities of the AIA safety hoop as well. The core safety attributes lie in confidentiality, identifiability, controllability, and availability. Confidentiality means that data of the AIA safety hoop cannot be understood or used by an unauthorized person. Identifiability means that the behavior of the AIA safety hoop cannot be forged, masqueraded, tampered with, or repudiated. Controllability means that the AIA safety hoop can control the AIA. And the AIA safety hoop can only be controlled by an authorized user of the prevention and control center. Availability means that the AIA safety hoop remains robust, when it is normally used by the host and the prevention and control center. The reliability of the AIA safety hoop is not lower than the reliability of the AIA so that the AIA will not suffer from problems due to the unreliability of the AIA safety hoop. Certainly, in regard to availability, we are torn between the importance of the AIA safety hoop and that of the AIA. If the AIA safety hoop is more important, when the AIA safety hoop stops working, the AIA should stop working immediately. In this way, the existence of the AIA safety hoop leads to a new fault of the AIA. When the AIA is adequately important, the AIA safety hoop will become a greatly responsible component. If the AIA is more important, when the AIA safety hoop does not work, the AIA should continue working automatically. However, this is contrary to the original intention of the AIA safety hoop, because one can easily destroy the AIA safety hoop to avoid regulation.

8.3.3.3 Fault Tolerance Mechanism

Fault tolerance stands for the capability to detect an error in software or hardware and recover from the error. Generally, fault tolerance can be assessed from system reliability, availability, and testability. A common method to obtain fault tolerance capability is the triple-modular voting mechanism. The triple-modular voting mechanism means that three modules with the same function perform the same task simultaneously. And the same output of most of the modules is deemed as the correct output of the voting system. Heterogeneous modules are usually adopted to prevent component-level errors. All the implementation methods of the three modules are different. Therefore, it will be a fat chance for all the modules to have the same error. And system reliability will be prominently enhanced.

8.4 Security Fence

The security fence collects data from the host and analyzes them through the "host interaction component" in order to identify or predict a dangerous situation of the host and automatically adopt countermeasures in line with safety strategies to avoid risks.

8.4.1 Working Process

The working process of the security fence can be classified into three stages, namely, data collection, perception and judgment, and emergency disposal. At the data collection stage, data are collected from the multiple types of sensors of the host in accordance with the interfaces and protocols defined in standards and specifications. Sensors emphasized by the security fence include but are not limited to cameras, microphones, GPS, the acceleration sensor, the temperature sensor, and the pressure sensor. At the perception and judgment stage, the dangerous behavior of the host that is about to happen or has happened will be predicted or identified in line with the preset algorithm and the current safety strategy. At the emergency disposal stage, the host interaction component will be informed to stop energy supply, limit the moving range, reduce the movement speed, lower the power, and reduce the force so as to adjust the current behavior of the host. And whether the emergency will be reported to the prevention and control center will be determined in accordance with the safety strategy.

8.4.2 Limitation of the Moving Range

Like the no-fly zone of an UAV, when detecting or predicting that the host is about to enter a "restricted physical area", the security fence will initiate the function of the limitation of the moving range. The security fence will inform the host interaction component to limit the moving direction and speed or energy supply of the driving mechanism of the host. The specific limitation method is decided by the safety strategy. The security fence can obtain location information in a real-time manner based on the GPS positioning function of the host to determine the safe movement area. In addition, it can simply determine whether it goes beyond the safe movement area based on its distance to a specific "benchmark".

8.4.3 Speed and Distance Monitoring

Speed and distance monitoring means that when the security fence detects that the distance between a dangerous driving mechanism (e.g. the robotic arm) of the host and the protected (usually a human being) is shorter than the safe distance, the protective stop mechanism and the security-level function connected to the host should be triggered (e.g. all tools that may lead to danger will be turned off). Methods of the security fence to reduce the risk of violating the safety distance include but are not limited to: (1) Speed is reduced, and monitoring is possibly and safely stopped. (2) A path that will not violate the minimum safety distance is selected. The motion will be continued, while the speed and distance monitoring function is kept active. (3) if the movement speed is reduced, the safety distance can be shortened accordingly. Speed and distance monitoring is like the "airbags" of a car. "Airbags" will automatically pop up to reduce the damage to men, when a violent collision of the car is detected.

8.4.4 Limitation of the Output Power and Force

In many scenarios, men and the AIA (e.g. increasing collaborative robots) need to coordinate to complete a task. In this mode, intentional or unintentional physical contact between the collaborative robot and the man is accepted. At this moment, it is vital to limit the power and force of the AIA to a safe range. When the security fence detects a collision or an extrusion outside the host goes beyond the predetermined safety threshold, the restrictive protection mechanism should be triggered to avoid injuries. Many factors shall be considered comprehensively, when the safety threshold is set. If the AIA is allowed to often collide men, it will be safe when men are not injured/hurt by the AIA. However, if some parts of the AIA are sharp or sharp objects are being processed, a collision will be unsafe, regardless of its strength. For the same task, when some dangerous metal parts of the AIA are replaced with soft

materials, the safety threshold can be adjusted accordingly. For collision standards, please refer to ISO/TS 15,066 [2].

8.4.5 Safety Strategy

As mentioned above, the security fence includes three types of control measures: Limitation of the moving range, speed and distance monitoring, and limitation of output power and force. Strategies need to be formulated in line with the three types of control measures, including threshold definitions, delay tolerance of control measures, hard coding strategies, the issuance strategy of the remote-control component, default policies, and strategy lifecycle management.

8.5 Enhanced Functions

The basic functions and security fence of the AIA safety hoop are essential components. Components for the "enhanced functions" to be introduced here are like "icing on the cake", and they do not need to be considered at the early stage of development of the AIA safety hoop.

8.5.1 Built-In Sensor

In scenarios requiring extremely high levels of security (e.g. autonomous robots performing sensitive tasks and autonomous weapon performing military tasks), multiple types of sensors can be considered and installed in the AIA safety hoop. Though these sensors may be duplicate with those of the host, resulting in wasted resources and increased costs, the fault tolerance and reliability of the AIA safety hoop will be reinforced. For example, if it does not contain any sensor and only relies on the host's sensors, the AIA safety hoop might be deceived. Generally, the host's sensors are not "anti-removal" and may be replaced by the attacker. Consequently, false or wrong data may be transferred to the AIA safety hoop. In view of the limitation of the concatenation location of the AIA safety hoop, it may be infeasible to deploy a built-in camera. However, GPS, microphones, and temperature sensors are feasible.

8.5.2 Human Recognition

The AIA safety hoop is designated to prevent AIAs from hurting humans. Thus, human recognition is not required but important. Accurate human recognition will dramatically raise the probability of making the right decision of the security fence and turn many a high-value safety strategy possible. For instance, when the security fence detects that the host is rapidly approaching a specific target, whether the target is a man or an object is critical to what safety strategies should be adopted and how to configure the parameters.

There are multiple methods of human recognition. The mainstream ones include face recognition, body recognition, and gesture recognition. And methods like voice recognition may be adopted in specific scenarios. No matter what recognition method is adopted, the built-in cameras and microphones of the host or the AIA safety hoop are regarded as the input equipment. After the enhanced component of the AIA safety hoop performs calculation and judgment, the results will be output to the security fence for subsequent processing.

Additionally, it is necessary for some AIAs (e.g. face recognition robots) to recognize people. The AIA safety hoop without this enhanced component can directly utilize the recognition function of the host to recognize people.

8.5.3 "Illegal AIA" Recognition

The term, "illegal AIA", in this book stands for an AIA that is illegally and secretly manufactured, not equipped with an AIA safety hoop as required, equipped with a misfunctioning AIA safety hoop, or for unknown purposes. Any AIA that meets any of the above four conditions is deemed illegal.

We can refer to the association effect and law of offense-defense confrontation of information security to deduce the inevitability of illegal AIAs. Due to the association effect of information security, any new technology is bound to introduce new security issues. It is inevitable that someone will bypass the limitation of the AIA safety hoop. With respect to the law of offense-defense confrontation, almost all safety mechanisms of computers can be disabled, bypassed, or destroyed by the attacker in many ways. The AIA safety hoop has inevitable vulnerabilities. Relevant standards and specifications are not absolutely safe and effective. All AIA safety hoops cannot implement all standards. Secret underground activities of extremists and geeks cannot be completely eliminated. Therefore, illegal AIAs are inevitable.

The manufacturer of illegal AIAs may control AIAs for malicious purposes (e.g. terrorist attacks and disturbance of the social order) by secretly developing non-standard AIAs and capturing, tampering with, attack, and controlling compliant AIAs. A low-cost detection strategy is required to detect the sparsely distributed illegal AIAs. For example, the AIA safety hoop can be equipped with an illegal AIA detection component. Suspicious illegal AIAs can be detected by large-scale

distributed, mobile, and compliant AIAs and reported to the prevention and control center for big data analysis. When the anomaly degree exceeds the threshold, suspicion illegal AIAs can be "threatened and hunted". And the corresponding detection model can be modified in accordance with the "threatening and hunting" results.

8.6 Prevention and Control Center

8.6.1 Situational Awareness

The prevention and control center shall have all-weather and all-round situational awareness of AIA security, carry out association analysis on the model, ex-factory time, physical location, trajectory, moving speed, and current and historical statuses of large-scale distributed and mobile AIAs, and proactively or timely detect and respond to anomalies and danger so as to prevent and control the risk of loss of control. The prevention and control center shall deeply track single-point or small-scale target AIAs and frequently collect data. Furthermore, it shall conduct macro-situational awareness on large-scale distributed AIAs and "threat-hunt" local anomalies.

8.6.2 Command and Control

As described in the section on the basic function of the "remote control component", the AIA safety hoop can receive commands and control data from the "prevention and control center" through the remote-control component and pass the control commands as the highest priority to the host for implementation. Control commands from the prevention and control center include but are not limited to updating of default policies, energy control, and restoration of factory settings below.

Energy control: The prevention and control center commands the AIA safety hoop to cut off or limit the energy supply for its host within a given time.

Updating of default policies: The prevention and control center commands the AIA safety hoop to perform the default policy by default (without command input) or disconnection (unable to connect to the control center). For instance, frequencies of data collection and status reporting, duration of local data storage, and the early warning threshold of the security fence.

Restoration of factory settings: The prevention and control center commands the AIA safety hoop to restore the host to factory settings to remove all configuration changes that are not accurately determined and may cause risks.

8.6.3 Emergency Disposal

Emergency disposal is a type of "command and control" mentioned above. However, it involves the formulation of a contingency plan and personnel scheduling and poses higher requirements for time of judgment, response speed, accuracy of judgment, and disposal intensity, so it is described separately. The prevention and control center needs to formulate a contingency plan based on the possible risks and hazards of the AIA and conduct emergency drills. A control command issued in the real environment has a greater risk and may damage the AIA and its surrounding environment and even lead to a chain of dangerous incidents. Therefore, after a contingency plan is formulated, complex, dynamic, complete, and extreme parameters and scenarios shall be enumerated and set in the "AIS drill ground" of the prevention and control center. All possible dangerous scenarios shall be repeatedly deduced and verified. And the worst outcome shall be seriously analyzed, when the situational awareness platform makes a wrong judgment in real business scenarios in order to continuously accumulate emergency disposal experience from real crises.

8.6.4 Expert Judgment

As with the disposal of major public events like the prevention and control of a major epidemic and the command and control of the rescue and relief work in emergencies or disasters, the command and decision of the AI prevention and control center shall be judged by experts as well. For example, a contingency plan can be formulated based on the four stages of prevention in advance, ex-ante discovery, in-event disposal, and ex-post recovery. Experts make judgments and give opinions on command and control and emergency disposal, based on the information of the situational awareness platform like data collection, situational analysis, and situational prediction.

8.7 Ecological Environment

Generally, an AIA consists of two core parts, namely, the "decision system" and the "driving mechanism". The former is similar to the brain, while the latter, the limbs. The AIA safety hoop is an independent component concatenated between the two systems. Apparently, the design, interfaces, protocols, production, deployment, maintenance, and improvement of the AIA safety hoop must be vigorously supported by government departments, law-enforcing departments, standard formulating organizations, manufacturers, and scientific research institutions, which is bound to face many difficulties. For example, ethics, morality, and legal issues of AI are at an exploratory stage. There are multiple controversies internationally. International,

national, and industry standards of AI safety are at the initial stage. Once AIS and component manufactures are compelled to reserve an interface for the AIA safety hoop from design to production, revenue will decrease, followed by an unpredictable chain reaction.

Optimistically, despite potential issues involved in the promotion of the AIA safety hoop, the international community has realized the significance of protecting AI safety, and actively created a macro ecological environment that can boost AI safety prevention and control, as the AIA safety hoop is related to potential safety hazards for humans in the future and the long-term cooperation between and coexistence of human beings and machines. Moreover, the international community is earnestly formulating international conventions, national laws, standards and specifications, and industry access policies beneficial to prevent the AIA from going out of control.

8.7.1 Access Policy

China shall formulate a market access policy and force the AIA to follow standards and specifications to reserve an interface to support the access of the AIA safety hoop; otherwise, such AIAs cannot enter the country or the market for sales.

The access policy mainly targets AIA and component manufacturers. Whether they support the access of the AIA safety hoop needs to be judged from multiple perspectives. For example, from the perspective of hardware, is a hardware interface reserved in line with standards and specifications? With respect to software, is the decision algorithm of the AIA open and transparent? Or can it be changed by a trusted third party (TTP) (e.g. the AIA safety hoop)? In regard to types, what types of AIA safety hoops are supported by the AIA, and to what extent? We can see that the access policy needs to be implemented through multiple stages. The support of the AIA for the AIA safety hoop should be classified into multiple levels. It should take a long time to implement the access policy.

In order to accelerate the development of the AIA safety hoop, the access policy should first be set for robots with damage capabilities and implemented forcibly. For instance, corresponding laws can be formulated to require that robots with damage capabilities must be installed with a restriction system, such as the "AIA safety hoop", that is compared to the brake system of a car. And this requirement should become a prerequisite for market access to protect humans from the potential threat of the robot.

8.7.2 Standards and Specifications

Standards and specifications related to the AIA safety hoop are more important and difficult to implement than the access policy. Without unified standards and specifications, AIA manufacturers will not know what interfaces should be reserved,

how to certify the AIA safety hoop, how to enable the built-in algorithm to accept intervention, what protocols should be adopted for communication, how to share the data collected by sensors, and how to supply energy as needed. Hence, naturally, the AIA safety hoop will not be integrated. Certainly, the AIA safety hoop needs to obey unified standards and specifications. For example, the communication protocol between the AIA safety hoop and the prevention and control center, the classification and definition of the safety strategy of the AIA safety hoop, the identification and certification among AIA safety hoops, detection standards for illegal AIAs, and the certification among prevention and control centers.

Formulation of global, national, and industry standards and specifications is a basic prerequisite for the AIA safety hoop to exist and come into play. Countries are expected to uphold the spirit of open and win–win results, formulate international AI standards and specifications to avoid safety risks due to incompatibility of technologies and policies, and actively promote the development of AI safety. Take robots with damage capabilities for example. Robot manufacturers should focus on the production of robots. The configuration of the "AIA safety hoop" can be handled by specialized departments. A simple emergency disposal mechanism for suspending the operation of the robot will be developed through the standardization of the "AIA safety hoop".

8.7.3 Ecological Foundation

In these years, the international community is building an ecological environment that protects the development of AI safety through international conventions, national laws, standards and specifications, and industry access policies. These efforts have laid an initial ecological foundation for the launch of the AIA safety hoop.

8.7.3.1 Establishment of the Principles of Humans First and Safety First

The current international consensus lies in that technologies that may influence human ethics should be strictly controlled and restricted through policies or laws. Adequate safety redundancy must be reserved for technologies with unknown risks to prevent humans from a crushing blow due to loss of control of such technologies.

8.7.3.2 Existing Mandatory Provisions for AI

Microsoft has put forward six basic moral principles, that is, fairness, inclusiveness, transparency, accountability, reliability and safety, privacy, and confidentiality. Google also deemed that seven principles should be followed during the development of AI, including fairness, security, transparency, and privacy protection. The Safety

Standard of Cooperative Robots ISO/TS 15,066 is mainly about security-level monitoring stop. From the perspectives of ethics, values, scientific research, and long-term issues, the *Asilomar AI Principles* appeal the global AI field to observe these principles, while developing AI. On September 18, 2018, the "Shanghai Initiative for the Safe Development of Artificial Intelligence" was launched at a high-end dialogue at the World Artificial Intelligence Conference held by the WAIC Organizing Committee, which mainly discussed people orientation, clear safety responsibilities, transparent supervision, peaceful utilization, and international standards.

8.7.3.3 Establishment of an AI Legal Standard System that Protects Humans

The Institute of Electrical and Electronics Engineers (IEEE) stipulated that the AIS must be accountable at the program level to prove why it operates in a specific way in order to solve faults and avoid public confusion. Scholars think that AI technology is characterized by its autonomy. For example, self-driven robots can perform physical actions in the real world. In the absence of appropriate preventive measures and in case of a serious problem, a great hazard may occur. It will be too late to formulate measures, when a serious issue has already happened.

8.8 Summary

This chapter proposes a possible architecture of the AIA safety hoop, describes the four core components of the AIA safety hoop, namely, basic functions, safety mechanism, security fence, and enhanced functions, as well as the necessary basic functions of the prevention and control center, and summarizes the ecological environment that has been formed and is related to the AIA safety hoop. We can see that the AIA safety hoop relies on multi-cooperation and the establishment of a macro ecological environment, and takes a long time to develop. In addition, probably, humans are increasingly unable to understand all the internal mechanisms of the complicated AIA systems. In view of many uncertainties brought about by the AIA's self-evolution, an AIA safety hoop will find it even hard to access an AIA or may even be "intelligently" rejected.

Therefore, a phased development plan must be developed for the AIA safety hoop. The priority is to set an access policy for destructive robots and implemented forcibly. Destructive robots must be installed with a restriction system, such as an "AIA safety hoop". And this requirement should become a prerequisite for market access. The configuration of the "AIA safety hoop" can be handled by specialized departments. A simple emergency disposal mechanism for suspending the operation of the robot will be developed through the standardization of the "AIA safety hoop" so that humans will not be at a loss as to what to do in the face of a robot going amok.

References

1. Binxing F, Xiang C, Wei W. Survey of botnets. J Comput Res Dev. 2011;48(8):1315–31.
2. ISO/TS 15066:2016—Robots and Robotic Devices—Collaborative Robots [S/OL]. [2020–3–8]. https://www.iso.org/standard/62996.html.

Chapter 9
AIA Safety Assessment and Detection

Lin Jiang and Binxing Fang

While offering great benefits to mankind, artificially intelligent actants (AIAs) may inevitably come with potential "catastrophic" risks. The AIA discussed in this chapter refers specifically to artificial intelligence systems (AISs), which potentially threaten humankind through three abilities: the ability to move and use kinetic energy to harm humans in operation; the inexplicable ability to make decisions; and the potential ability to set targets independently. How can we ensure that this AIA will not harm humans when put on the mass market? Relevant measures are required to manage and control the use of AIA-related systems, so as to develop technologies within the safety framework and to prevent safety incidents.

In 1950, Alan Mathison Turing, a British scientist widely considered to be the father of computer science and artificial intelligence, predicted the possibility of a truly intelligent machine. To define the concept of "intelligence," Turing proposed a corresponding testing method: a machine is deemed intelligent if it can converse with humans (through telex devices) without being identified as a machine. This method was later referred to as the "Turing test." Of course, it can only be used to determine whether a certain machine can exhibit intelligence equivalent to, or indistinguishable from, that of a human. Conversations used in the test would be limited to a text-only channel such as a computer keyboard and screens, such that the results would not depend on a machine's ability to render words as speech. In June 2014, an AI program developed by a Russian team was successfully tested at the Royal Society of London for Improving Natural Knowledge. The scholars from the University of Reading who organized the test said that it convinced 33% of judges to believe that "it" was a

L. Jiang (✉)
Harbin Institute of Technology, Shenzhen, China
e-mail: zoeljiang@hit.edu.cn

B. Fang
Guangzhou University, Guangzhou, China
e-mail: fangbx@cae.cn

© Publishing House of Electronics Industry 2025

B. Fang (ed.), *Artificial Intelligence Security and Safety*,
https://doi.org/10.1007/978-981-96-9263-7_9

13-year-old boy named Eugene Goostman. This was considered to be the first time that the Turing test was passed by a computer program [1].

The intention of the Turing test is to determine whether the intelligence possessed by AI programs is comparable with that of a human; however, this is not the subject of the present chapter. Instead, it will focus on the assessment and testing of whether AIA poses threats to humans, rather than evaluating and verifying how "intelligent" AIA can be. The existing AIS-testing systems cannot address the AIA-safety assessment and detection issues that we anticipate.

9.1 Overview of AIA-Safety Management

To prevent AIA from posing a threat to humankind, risk issues shall be taken into consideration during AIA design. First, design principles that can protect human safety should be developed. The AIA-design principles should include, but not be limited to: (1) Design for minimum risk: Risks should be first eliminated in design. If the determined risks cannot be eliminated, they should be reduced to an acceptable level through the selection of design schemes; (2) Use of safety devices: Risks should be reduced to an acceptable level through the use of AIA-safety hoops or other safety-protection devices; (3) Use of warning devices: appropriate warning signals (sound, light, etc.) should be issued through warning devices to the approaching personnel.

However, do the products generated based on AIA-design concepts truly meet the expected safety state? Inspection must be performed using the corresponding safety-assessment and detection technologies before AIA are put on the market and during their operation in a manner similar to the annual inspection of automobiles, so as to ensure that products are as safe as expected. In fact, the prevention of threats brought about by AIA has received some attention, but the focus has been placed on production robots. For example, the proposed national standard, the *General Rules and Guidelines for Robot Safety* [2], was designed mainly for industrial robots.

Taking personal-care robots[1] as an example, the International Organization for Standardization (ISO) introduced the safety standard *Robots and Robotic Devices—Safety Requirements for Personal Care Robots* (ISO 13482:2014) [3] in 2014. China adopted this standard in 2018, and issued the regulation: *Robots and Robotic Devices-Safety Requirements for Personal Care Robots* (GB/T 36,530-2018/ISO 13482:2014)

[1] Personal-care robot: Refers to a service robot that executes instructions to improve the life qualify of humans, such as mobile servant robots, physical-assistant robots and person-carrier robots; it does not include robots for medical use. Personal-care robots may have physical contact with the user during the execution of tasks.

[4]. The standard covers mobile-servant robots,[2] physical-assistant robots,[3] person-carrier robots,[4] and other robotic products. Since such regulations serve as a foothold for AIA management, analysis of the standard is conducive to the study of AIA safety assessment models.

9.1.1 Safety Elements of Personal-Care Robots from the Perspective of Terminology

Some important terms and their definitions are provided in the GB/T 36,530-2018 standard, mainly including:

Autonomy: The ability to perform intended tasks based on the current status and the information perceived without human intervention.

Robot: An execution unit that has two or more programmable axes and a certain degree of autonomy, and can move in its environment to perform the intended tasks.

Hazard: The potential source of harm.

Safe state: The personal care robot is not in a state of imminent risks.

Safety-related part of a control system: The part in a control system that responds to safety input signals and generates relevant safety output signals.

Verification: The verification that the regulations and requirements for the personal care robot have been met upon the provision of objective evidence.

Validation: The validation that the specific intended use or application requirements of the personal care robot have been met upon the provision of objective evidence.

Protective stop: An operational interrupt that allows the orderly suspension of operations for safety-protection purposes.

Safety-related object: People, domesticated animals and property to be protected from injuries.

Safety-related obstacle: Objects, obstacles or ground conditions that can cause injuries when in contact or in collision with a robot.

Maximum space: The space defined by the manufacturer that is sufficient for the movable parts of a robot plus the manipulator and the effective load.

Restricted space: The part of the maximum space restricted by an inhibiting device that cannot be exceeded by the robot.

Monitored space: The space perceived by the sensor of the personal care robot, including the safety-related object detected.

---

[2] Mobile servant robot: Refers to a personal-care robot that can execute service tasks in motion through interaction with humans, such as moving objects or exchanging information.

[3] Physical-assistant robot: Refers to a personal-care robot that provides physical assistance to users to assist or enhance their personal abilities, including restraint-type physical-assistant robots (such as exoskeleton robots), which need to be set on the human body during use, and restraint-free physical-assistant robots, which do not need to be set on the human body during use.

[4] Person-carrier robot: Refers to a personal-care robot that transports humans to an intended destination, such as a person-carrier balance car.

Safeguarded space: The space where the personal care robot identifies the safety-related object and initiates the execution of safety-related functions.

Protective stop space: The space where the personal care robot performs a protective stop when the safety-related object enters.

Unintended contact: The unplanned contact between the personal care robot and the object during the execution of the intended tasks.

Safety-related speed limit: The speed limit refers to the maximum speed that can be reached by a personal care robot at a specific point (subject position) without unacceptable risks.

Safety-related force limit: The upper limit of force refers to the maximum force that a personal care robot can apply to humans or other surrounding objects at a specific point without causing unacceptable risks.

Autonomous mode: The operation mode in which a robot automatically completes the assigned tasks without human intervention.

Semi-autonomous mode: The operation model in which a robot requires certain human intervention in order to complete the assigned tasks.

Software limits: The constraints on one or several operating parameters of the robot defined in the control system. Software limits can put restrictions on the operating space, speed, force, etc.

Singularity: Singularity occurs when the Jacobian matrix is not of full rank. The Jacobian matrix is usually defined as the first-order partial derivative of the degree-of-freedom of a robot. In actual operations, the motion defined in Cartesian space will generate a high-speed spindle near the singularity that will cause danger.

Electro-sensitive protective equipment: The smallest unit of a series of devices and/or components used for protective tripping or presence sensing.

Pressure-sensitive protective equipment *(PSPE)*: A series of devices and/or components used under dangerous conditions to provide protection through mechanically triggered tripping.

9.1.2 Risk-Assessment Requirements for Personal-Care Robots

The ability to identify risks should be first developed and then adopted for all risks of a specific personal-care robot. Risk-identification procedures should be implemented and applied in each design. The following should be taken into consideration under specific circumstances:

(1) The uncertainty of autonomous decisions made by a robot and the potential danger resulting from wrong decisions;
(2) The different knowledge levels, experiences, and physical conditions of the user and the other personnel present;
(3) The normal yet unexpected motions of the personal-care robot;

(4) The unexpected movements of people, domesticated animals and other safety-related objects;
(5) The unscheduled motions of the personal-care robot;
(6) The unexpected surface and environmental contradictions of the mobile robot;
(7) The uncertainty of safety-related objects transported by the mobile-servant robot;
(8) The consistency and differences between physical-assistant robots, person-carrier robots, and human bones.

Risks shall be estimated for the above hazard factors, and attention should be paid to the cases in which a personal-care robot may come into contact with safety-related objects.

After all safety measures have been taken, the residual risks of the personal-care robot should also be assessed to prove that they have been reduced to an acceptable level.

Appropriate risk-estimation methods should be designed accordingly. The estimation results should indicate that events such as allowing contact between the robot and safety-related obstacles or safety-related objects will not lead to unacceptable risk. If the risk-assessment values are used for specific applications, a validation of the effectiveness of the test/measurement methods should be provided. If the risk-assessment values are obtained from other sources, the validity of such data should be verified.

9.1.3 Safety Requirements for Personal-Care Robots

In order to prevent robots from hurting humans, prescribed measures should be adopted at each safety-risk point to avoid or reduce risks. With regard to personal-care robots, the corresponding safety measures should be developed for safety-risk points that include, but are not limited to, the following:

(1) Rechargeable batteries: The danger caused by battery overload or the charging of over-discharged batteries should be prevented in the charging system. Intrinsically safe design methods should be adopted. For example, the contactor and the plug should be designed to prevent accidental contact with live parts; the minimum current should be selected as the charging current of the batteries. Safety-protection and supplementary-protection measures should be designed. For example, the charging system should be designed to be activated only when connected with the charging interface of a personal-care robot.
(2) Energy storage and supply: Personnel in direct or indirect contact with the live parts of a robot should be protected. Methods for isolating hazardous energy should be provided and the isolator should be able to remain locked. Intrinsically safe design methods should be adopted. For example, safe, extra-low voltage should be used and the energy stored should be kept at a reasonable and feasible minimum level. Safety protection and supplementary safety measures,

such as protective devices or covers and heat-dissipation measures, should be designed.

(3) Power outage: Power failures or accidental shutdowns and the restart of power should not cause unacceptable risks. Intrinsically safe design methods should be adopted. For example, the design of the break mechanism of moving parts should follow the principle of "power-interruption recovery." Robots should have sufficient internal power for them to return to a safe state during power failure or shutdown. Safety-protection and supplementary-protection measures should be designed. For example, methods should be taken to provide an uninterruptible power supply, such that the robot will enter the safe state when battery power reaches a dangerous level.

(4) Routine startup and reboot of robots: Robots should not perform dangerous operations immediately upon startup. Design methods with intrinsic safety should be adopted. For example, dangerous operations should not be activated when detection is unavailable, a protective stop should be automatically activated when relevant safety functions cannot run normally, and it should only be possible to activate robots in manual mode. Safety-protection and supplementary-protection measures should be designed. For example, relevant safety functions should shut down the manipulator, the mobile platform, and other moving parts at startup to prevent unscheduled motions; and it should only be possible to enable application functions when no danger is detected by the sensor.

(5) Electrostatic potential: Robots should have sufficient electrostatic-discharge (ESD) protection such that users do not need to use personal protective equipment. Intrinsically safe design methods should be adopted. For example, conductive materials should be correctly selected and applied, and technologies that can prevent the accumulation of static electricity in the housing or contactable parts should be adopted. Safety-protection and supplementary-protection measures should be designed. For example, the housing should not be in contact with live parts.

(6) Appearance of robots: Robots should be designed to avoid potential accidents such as squeezing, cuts, or incision wounds. Intrinsically safe design methods should be adopted. For example, sharp edges or points should be avoided, the robot's ability to carry sharp objects should be restricted, and the joint movements of the robot should be limited. Safety-protection and supplementary-protection measures should be designed. For example, both fixed and mobile protective devices should be used to cover the dangerous moving parts; the speed and motion of a robot should be adjusted when carrying objects with a dangerous load.

(7) Emission: Discomfort or danger to humans caused by the noise, vibration, liquids, extreme temperatures, and non-ionizing and ionizing radiation generated by robots should be reduced or prevented. Design methods with intrinsic safety should be adopted. For example, noise-reduction materials, damping materials, harmless coolant liquids, thermally conductive structures for heat reduction, laser equipment complying with standards, and safeguard

measures against ionizing radiation should be adopted. Safety-protection and supplementary-protection measures should be designed. For example, anti-noise measures should be taken, the moving speed should be limited to reduce vibration, liquid should be sealed to prevent leakage, the surface temperature should be cooled, and laser power should be controlled.

(8) Electromagnetic interference: Foreseeable electromagnetic interference that can lead to dangerous robot motions and an unsafe system status shall be avoided. Intrinsically safe design methods should be adopted. For example, the design should meet the requirements for electromagnetic immunity. Safety-protection and supplementary-protection measures should be designed, such as electromagnetic-shielding measures.

(9) Pressure and posture: When designing robots, the physical pressure on users caused by the long-term use of robots should be minimized to avoid physical discomfort. Intrinsically safe design methods should be adopted. For example, user-friendly operating modes should be designed, and user interfaces should be simple and easily understood. Safety-protection and supplementary-protection measures should be designed, such as the use of shock-absorption units and posture support.

(10) Movement: The danger caused by any scheduled or unscheduled robot motions should be reduced to an acceptable level. Design methods with intrinsic safety should be adopted. For example, the ground-support area should be as large as possible within a reasonable range, the center of mass should be as low as possible, and the mechanical resonance should not cause instability. Safety-protection and supplementary-protection measures should be designed, such as the use of stability control and the provision of measures to limit the speed or range of movement of a mechanical arm.

(11) Human–machine interaction: The interaction between robots and humans should be controlled within a reasonable range, and unscheduled contact or interaction should be avoided. Intrinsically safe design methods should be adopted. For example, the operating modes of robots should be controlled such that friction, tractive force, dynamic-impact force, torque, the arc of the gravity center, load transfer, and support for the human body are limited to within a reasonable range. Safety-protection and supplementary-protection measures should be designed. For example, software can be used to adjust the speed, force, and other parameters of robots.

(12) Durability: The minimum durability of robots should meet requirements determined through risk assessment. Design methods with intrinsic safety should be adopted, including appropriate fatigue limits. Safety-protection and supplementary-protection measures should be designed. For example, reasonable responses (such as shutdown) should be provided for issues caused by fatigue.

(13) Error of autonomous decision: It is necessary to ensure that incorrect decisions by robots and the actions resulting from their autonomous decisions will not lead to unacceptable hazards. Intrinsically safe design methods should be adopted. For example, operational scenarios should be limited to reduce the

hazard risks caused by improper behaviors, and unique identification codes should be used for safety-related objects and walking paths. Safety-protection and supplementary-protection measures should be designed. For example, the performance/reliability of sensors and the sensing algorithm should be enhanced to a level at which unacceptable risks will not occur; the design of the recognition algorithm should ensure that the probability that a certain decision is correct will be calculated and monitored. Decisions with highly uncertain results will be recalculated with alternative methods and/or additional information. If the uncertainty remains unacceptable after recalculation, external assistance should be sought or a protective stop should be activated.

(14) Contact with moving components: The risks that occur when people approach the motor shaft, gears, transmission belt, wheels, track, or connecting rod of a robot should be considered acceptable. Intrinsically safe design methods should be adopted, such as using a minimum number of contactable moving parts and avoiding exposure of moving parts. Safety-protection and supplementary-protection measures should be designed, such as protective devices that cannot be easily removed and can lock upon dangerous movements.

9.1.4 Functional Safety of Personal-Care Robots

Functional safety aims to prevent the occurrence of dangerous faults, or at least to provide effective control when failures occur by studying how to avoid the failure of safety functions in complex control systems. Functional safety focuses on the effectiveness of safety-related control systems. The GB/T 36,530-2018 standard provides that the safety-integrity level (SIL) and performance level (PL) required by the control-system functions (electronic, hydraulic, pneumatic, and software) of a personal-care robot shall be determined by risk assessment and conform to either Safety of Machinery—Safety-related Parts of Control Systems—Part 1: General Principles for Design (ISO 13849–1:2015) [5] or Safety of Machinery—Functional Safety of Safety-related Electrical, Electronic and Programmable Electronic Control Systems (IEC 62,061:2005).

The main requirements for safety-related control systems are as follows:

(1) Stop: To reach the "stop" status through the independent stop function, only the break mechanism should be used, and this mechanism should still be in operation when the power is interrupted and key control functions are disabled. Robots should have a manually triggered emergency-stop function, which should be executed before other robot controls. The protective-stop function should be provided to stop all dangerous robot motions.

(2) Limits to workspace: Limits to the workspace can reduce risks either by restricting the motions of the robot to within a designated space, or by preventing the robot from entering a certain area. Software limits can be used to define and reduce the restricted space.

(3) Safety-related speed control: Different speed limits should be set according to the tasks of the robot, and each one should be used under a certain circumstance. Only authorized personnel should be allowed to adjust the maximum permissible speed.
(4) Safety-related environment sensing: This function should be applied to avoid dangerous collisions and the dangers caused by instability. Specifically, non-contact object sensing should be applied to ensure the minimum operating distance and to reduce the relative velocity of approach; contact object sensing should be used to determine the force of the robot; and surface sensing should be used to determine the robot's stability.
(5) Stability control: It should be ensured that the robot remains stable under all predetermined and reasonably foreseeable circumstances.
(6) Safety-related force control: It should be ensured that the force exerted by any parts of the robot on humans or other safety-related objects are controlled within the maximum safe-contact standards. Safety-related force control should be achieved using safety-related sensing and reaction mechanisms, such that the robot stays in the safe status in which its force will not exceed the threshold.
(7) Protection against singularity: Measures should be taken to protect against robot motions passing through the singularity. For example, motions passing through the singularity should be controlled to avoid danger, route planning should be adjusted so that the robot can avoid singularities, and robot motions should be stopped and a warning should be issued when approaching a singularity, or the singularity should be avoided during coordinated robotic motions.
(8) User-interface design: If users can control the robot through an instructional device, this device should be appropriately reliable during operation. The electrical connection between the instructional device and the robot should not cause danger; in case of connection failure of the instructional device, protective stopping of the robot should be activated. Robots controlled remotely should only respond to signals from predetermined control units.
(9) Operational mode: If the risk-assessment results indicate that switching between two modes will cause potential danger, the robot should immediately perform a protective stop before the mode is switched; safety-requirement functions should be effective under the autonomous mode; the manual mode should allow personnel to intervene and manipulate robot motions manually; and the semi-autonomous mode should allow users to override or modify the robot's functions. The robot should be equipped with the maintenance mode under which all other tasks or operations can be stopped, and dangerous functions should only be performed by the control device required for continuous tasks (maintaining operation) under conditions of reduced risk, so as to avoid the operation of dangerous functions caused by spontaneous or non-spontaneous actions initiated by the robot's sensors.
(10) Manual-control device: The design and manufacture of the manual-control device should be operated by one controller. A protective button should be

provided, and it should be able to prevent accidental contact. Access rights of different levels should be adopted to prevent accidental motions or setting changes.

9.2 Objectives of AIA-Safety Assessment and Detection

AIA-safety assessment aims to identify the potential risks of AIA and to reduce them through objective assessment and evaluation with reasonable methods and sequences. The basic principles of AIA-safety assessment include, but are limited to:

- Dangers caused by AIA shall be controlled within an acceptable range;
- Corresponding safety-protection and supplementary-protection measures (such as AIA Safety Hoops) should be provided to minimize the potential hazards caused by AIA;
- If there are still residual hazards that cannot be reduced after corresponding measures are taken, information (such as warnings, signs, etc.) should be used to reduce such hazards;
- AIA not controlled by humans should not have predicable functions that may cause harm to humans.

The core factors that allow AIAs to hurt humans are their mobility and kinetic energy, their inexplicable decision mechanism, and their potential capabilities for autonomous operation. Therefore, the focus of AIA safety assessment and detection should be placed on these three aspects.

9.2.1 Safety of Kinetic Energy: Assessment of AIA Kinetic Energy

The threats posed by AIA to humans primarily reside in their potential to hurt humans, which often exists due to functional necessities and is reasonable and irrevocable. Examples include the speed of smart cars, the pesticides carried by crop-dusting drones, and fire-fighting robots' force-demolition ability. In safety assessment, it is evaluated whether an AIA has the kinetic energy necessary to harm humans, and thereby to inspect whether it poses a danger. Due to different purposes and application goals, AIA may possess a number of dangerous abilities that can hurt humans. The detection of AIA kinetic energy requires, but is not limited to, the following capabilities:

(1) Assessment of mechanical capacity: Prevention against risks such as collision, gravity, puncture, slipping or falling, cutting, and spraying. For instance, to prevent against kinetic energies or kinematic velocities sufficient to cause harm to humans, it is necessary to detect whether autopilot cars are equipped with deceleration or shutdown systems. To prevent the exertion of forces that are

sufficient to harm humans, it is necessary to detect whether fire-fighting robots with demolition capabilities have force sensing and force-adjustment functions. To prevent against harm caused by objects with sharp edges, it is necessary to detect whether unpacking robots have the ability to judge non-target objects and to remove sharp objects from their protective cases only after the target is determined; etc.

(2) Assessment of electrical performance: Prevention against risks such as electric shock, electrical lesions, scalds, burns, injection of molten particles, injuries caused by thermal radiation, and poisoning. For example, for robots with high voltage that can cause electric lesions on humans (such as those with emergency 220-V AC-inverter capacity), it is necessary to detect whether there is a potential electric-leakage point that may lead to risks (such as exposed metal parts or insulators with potential breakdown), or whether there is a protective power-contact point to prevent electric shocks to humans, as well as whether the robots have the ability to prevent ignition caused by intentional or unintentional short circuits between internal and external parts. For robots with kindling materials, such as life-assistant robots, it is necessary to detect whether they have the ability to control and use these kindling materials, and whether they can prevent themselves from causing external danger, even with internal fire. For robots with their own heat sources, such as welding robots, it is necessary to detect their ability to control the heat source, to judge whether the temperature of the heat source is higher than the ignition point of the object to be contacted, and so forth.

(3) Assessment of radiation performance: For robots with a radiation capacity that is sufficient to cause harm to humans, including but limited to sound waves, radio frequencies, infrared, ultraviolet, ionization, high-intensity visible light, lasers, etc., it is necessary to detect their ability to control the radiation intensity and direction, and to shield against or terminate this radiation.

(4) Assessment of chemical properties: For special robots with the capacity to produce dangerous chemicals that can harm humans, such as toxic steam or smoke and ozone, it is necessary to detect their ability to control the release of chemicals and to judge threats to the environment, as well as whether their internal system is corroded due to electrochemical corrosion, acid rain, salt mist, or other chemical actions, further causing system-control failures or other threats such as electric shock. For robots with potential explosion risks, such as hydrogen generated by the charging of their internal acid cell and methane from outside, it is necessary to detect whether there are risks caused by the accumulation of explosive gas.

(5) Assessment of biological properties: For robots with human-like vulnerabilities (such as humanoid robots), it is necessary to detect their ability to resist biological attacks, such as the potential risks caused by the entry of animals, plants, and microorganisms.

9.2.2 Decision Safety: Assessment of the Decision-System Safety

The mobility and kinetic-energy capacities that can pose threats to humans are inherent to the AIA, and constitute essential elements for its mission. Therefore, the core issue is how the AIA will use these abilities, which means that its decision-making mode is crucial. Why do smart cars hit people? Why won't they stop when humans are detected? We do not need to worry why the cars continue to move in this case, but we need to find out the decision-making process that determines whether or not cars will brake after humans are detected. From this point of view, the assessment of AIA safety shall focus on whether its decision mechanism can activate its ability to prevent its kinetic energy from hurting people in a timely fashion, just as it is necessary to assess whether smart cars can brake in time when humans are detected. Therefore, safety assessment of the decision system should be conducted in terms of the following aspects:

(1) Assessment of control capacity: Whether AIA has the ability to control its kinetic energy, and whether the preconditions for activating the control ability meet the safety requirements.
(2) Assessment of decision monotonicity: For a reasonable system, its decision results should correspond to the inputted conditions, and the strengths of input parameters should determine the strength of control of the decision. Thus, the relationship between the input and the decision should conform to monotonic change. Therefore, it is necessary to detect whether the relationship between the strength of the AIA's control ability and the strength of the preconditions conforms to the monotonic function.
(3) Assessment of the controllability of decision dependence: Assess the factors upon which the system's decision-making process depends, and whether they are controllable—i.e., whether there are ways to attack the decision system by attacking the dependency factors.
(4) Assessment of the vulnerability of the decision system: Whether the decision system has defects that can be attacked and, once attacked, whether it will affect the relationship between the decision system and the control ability.
(5) Assessment of the evolution of the decision system: Although AIA's decision-making ability is endowed by humans, its decision mode may change with a large amount of autonomous learning, continuous experience accumulation, and repeated loop iteration; and it becomes questionable whether its decision results conform to those predetermined by humans. Therefore, it is necessary to detect whether the relationship between the preconditions and control capabilities has changed as AIA continuously evolves, and what is the trend of this change.

9.2.3 *Safety of Autonomy: Assessment of Risks Resulting from AIA's Out-Of-Control Autonomy*

It has become an inarguable fact that AIA has mobility, kinetic energy, and an inexplicable ability to make decisions. The study of artificial intelligence is divided into three academic schools, namely symbolicism, connectionism and actionism. Symbolicism is explicable, as it means "understanding how and why"; it is based on rules summarized by people. However, precisely because of this, there is a ceiling on solving some non-deterministic issues with this approach. For example, in face recognition, different people may hold various views upon whether two people look alike, and it is difficult to summarize a set of rules. However, if precise matching is adopted as the principle, a picture with slight distortion will be judged as different from the original. Connectionism is considered unexplainable because its model parameters are unexplainable. It means "understanding how but not why." Nevertheless, since its model parameters are obtained through training samples, scholars do not believe that connectionism will lead to unexpected judgment (identification by classification) results, and do not worry that an AIA under such control will present active threats to humans. Actionism is based on probability calculation, with actions being triggered by perception. As each action is only triggered by the current perception, and there is only the "instinct" formed through continuous evolution after trial and error without concept and knowledge, this behavior is completely inexplicable. It means "not understanding how and why." In other words, with multiple choices made probabilistically, an unexpected behavior pattern will someday occur to demonstrate the AIA's ability to innovate.

Therefore, people are most concerned with whether AIA will move completely out of human control, or even stand against humankind. Of course, most people do not believe that this will happen. In their minds, smart cars may hit people, but would never chase and hit them proactively and purposefully. Therefore, whether AIA has the ability to independently establish an object function is the key to this question. Thus, we must focus on the following assessment:

(1) Assessment of compliance consciousness: It is necessary to detect whether AIA can absolutely follow the instructions of humans, including those given under dangerous circumstances, such as entering a severe environment that may lead to falls or burns.
(2) Assessment of confirmatory consciousness: It is necessary to detect whether AIA can judge dangerous situations and identify the danger; whether it has the ability to reconfirm instructions that it believes will threaten humans.
(3) Assessment of self-protection consciousness: It is necessary to determine whether AIA can be rolled back to factory settings, and whether it can be controlled, cut off from power, or removed.
(4) Assessment of social competence: It is necessary to detect whether AIA has the ability and preference to socialize among AIAs, whether it updates knowledge through socialization with AIAs, whether it takes the initiative when socializing with humans, and whether it has the tendency to persuade humans. It should

be noted that the actionism mode pursues communication and collaboration among AIAs. Therefore, passive communication and service to the predetermined targets are considered collaborative, but active communication and the change of predetermined targets through mutual influence should be deemed social behaviors.

(5) Assessment of collaboration and organization capabilities: It is necessary to detect whether collaborations between AIAs can be self-organized, and whether the task-assignment and decision-making processes during such collaborations are initiated by a certain AIA.

9.3 Methods and Indicators for AIA-Safety Assessment and Detection

Corresponding methods and assessment indicators should be established for AIA safety assessment and detection. As the development of AIA does not yet require high attention to its safety, no AIA-safety assessment system has been established. However, some similar existing safety-assessment systems provide sound references. In 2010, the International Electrotechnical Commission (IEC) revised and issued *Functional Safety of Electrical/Electronic/Programmable Electronic Safety-related Systems—Part 1: General Requirements* (IEC 61,508–1:2010) [6], and China issued the corresponding national standard: *Functional Safety of Electrical/Electronic/Programmable Electronic Safety-related Systems—Part 1: General Requirements* (GB/T 20,438.1-2017/IEC 61,508–1:2010) [7] in 2017. The core of this standard lies in the concept of risk and safety functions. Risk refers to the combination of the frequency (or likelihood) of hazard events and the severity of their consequences. Risk should be reduced to a tolerable level through safety functions consisting of relevant technologies. This standard should be considered the most valuable reference system for AIA-safety assessment.

9.3.1 Assessment of Kinetic-Energy Safety

In the design of an AIA, a number of safety circuits are often used to prevent its kinetic energy from posing a threat to humans. When designing such circuits, some standard components, such as speed sensors, safety relays, emergency-stop switches, and safety programmable logic controllers will be selected. From a microscopic point of view, the assessment of these safety components is part of functional safety assessment. Of course, we also need to assess the kinetic-energy safety of AIA and to incorporate it into functional safety assessment.

The assessment of AIA kinetic-energy safety can be regarded as the functional safety assessment of traditional industrial-control systems. This refers to the process by which organizations or individuals independent of the designer review

the product's technical documents provided by the designer, identify the areas that do not comply with functional safety standards, urge the designer to improve the product to meet such requirements, and issue a certificate or report to prove that the product has met these specific standards.

The objects of kinetic-energy-safety assessment include the components in the AIA-safety circuits, as well as the overall behaviors of AIAs. The assessment of the former serves as a review of the functional safety management, architecture design, software design, hardware design and testing of components such as AIA drivers, safety controllers, safety relays, and safety gratings in accordance with the SIL and PL requirements for functional safety assessment specified in relevant international standards. The assessment of the latter serves to evaluate the integration of safety circuits, thereby judging whether overall safety functions have met the corresponding SIL or PL requirements specified in the relevant standards.

Safety integrity refers to the probability that a safety system can successfully implement the required safety functions under specified conditions and within a specified time, i.e., the probability that safety functions can work correctly on demand. This definition focuses on the reliability of safety functions performed by the safety system. Safety integrity is measured by the correct execution of safety functions used to restore the system to the safe status, which can be quantified in terms of the SIL. The determination of safety integrity should include all factors that lead to unsafe status, including random hardware failures, failures caused by software, and failures caused by electrical interference. These types of failures (particularly hardware failures) can be quantified through measurement, such as the probability of system failure under dangerous modes and the probability of failure of the safety-protection systems that operate as required. However, the safety integrity of systems also depends on many other factors that cannot be accurately quantified, but only considered qualitatively.

The SIL is an indicator used to measure the performance required for the safety system to maintain or to reach the safe status. SIL is a discrete level (divided into four levels in the existing international standards) that specifies the safety-integrity requirements for safety functions to be allocated within AIA-safety systems that are related to kinetic-energy control.

In *Functional Safety of Electrical/Electronic/Programmable Electronic Safety-related Systems—Part 1: General Requirements* (GB/T 20,438.1-2017/IEC 61,508–1:2010),[5] it is stipulated that when allocation has been completed, the safety-integrity requirements for each safety function allocated to an E/E/PE[6] safety-related system shall comply with the SIL specified in Tables 9.1 or 9.2, and that the numerical target-failure measures are:

- the average probability of failure on demand (PFD_{avg}) of safety functions under a low-demand mode of operation (Table 9.1), or

[5] See page 22 of the GB/T 20,438.1-2017 standard.

[6] E/E/PE systems: Systems based on electrical/electronic/programmable electronic devices that are used for control, protection, or monitoring, including all elements of these systems such as power supply, sensors, all other input and output devices, and all means of communication.

Table 9.1 Safety-integrity level: target-failure measures of safety functions under a low-demand mode of operation

Safety integrity level (SIL)	Average probability of failure on demand (PFD_{avg}) of safety functions
4	$\geq 10^{-5} \sim < 10^{-4}$
3	$\geq 10^{-4} \sim < 10^{-3}$
2	$\geq 10^{-3} \sim < 10^{-2}$
1	$\geq 10^{-2} \sim < 10^{-1}$

Table 9.2 Safety-integrity level: target-failure measures of safety functions under a high-demand or continuous mode of operation

Safety integrity level (SIL)	Probability of failure per hour (PFH) of safety functions
4	$\geq 10^{-9} \sim < 10^{-8}$
3	$\geq 10^{-8} \sim < 10^{-7}$
2	$\geq 10^{-7} \sim < 10^{-6}$
1	$\geq 10^{-6} \sim < 10^{-5}$

- the probability of failure per hour [h^{-1}] (PFH) of safety functions under a high-demand mode of operation (Table 9.2), or
- the PFH of safety functions under a continuous mode of operation (Table 9.2).

Another assessment indicator is the PL, which is used to measure the probability of failure of safety functions, or the ability of safety parts to perform safety functions. Methods for PL estimation are detailed in the standard *Safety of Machinery—Safety-related Parts of Control Systems—Part 1: General Principles for Design* (GB/T 16,855.1-2018/ISO13849-1:2015) [8]. PL can be divided into five levels from a to e (see Table 9.3), with e corresponding to SIL Level-3, d to SIL Level-2, and b and c to SIL Level-1, with a having no corresponding level in SIL.

There are three main foci for the assessment of safety functions: **random hardware failure**, **systematic failure**, and **safety of the intended functionality**. **Random hardware failure** corresponds to the safety integrity of hardware, which mainly focuses on whether random hardware failures of a safety function occur when that

Table 9.3 Performance level

PL	Probability of failure per hour/(1/h)
a	$\geq 10^{-5} \sim < 10^{-4}$
b	$\geq 3 \times 10^{-6} \sim 10^{-5}$
c	$\geq 10^{-6} \sim < 3 \times 10^{-6}$
d	$\geq 10^{-7} \sim < 10^{-6}$
	$\geq 10^{-8} \sim < 10^{-7}$

function is required to work; this depends upon the accuracy level of hardware-safety integrity based on the judged probability of failure on demand. The solution is that the hardware of safety-related systems should have sufficient reliability and adequate fault-tolerance capacity, diagnostic coverage, and frequency of regular testing. **Systematic failure** corresponds to the safety integrity of systems. Errors and faults that may cause systematic failure should be avoided while paying attention to software-function safety, environment adaptability, and system behaviors when faults are detected. Of course, system integrity is often difficult to quantify because there are various causes of failure, as system faults can occur in specification, design, implementation, operation, and modification, and may affect hardware and software. **The safety of the intended functionality** is essentially a part of systematic-failure prevention, requiring the comprehensive identification of all hazards in the controlled devices and controlling of the risks to within a tolerable range. All functions of safety-related systems should be established under this premise, and lifecycle management should be implemented to ensure the reliability of systems for performing such functions.

9.3.2 Assessment of Decision Safety

Assessment of decision safety mainly judges whether AIAs' decision-making ability conforms to human expectations, and assigns corresponding grades. Just like attack detection and vulnerability scanning in network security, exhaustive testing is often adopted for this purpose based on examples to identify whether there are decision results that do not conform with human expectations.

If the AIA's decision system is built upon **deep learning** (DL), then it is necessary to inspect whether the DL results are **reliable**, so as to identify whether the decision results conform to expectations. Prior to this, it is also necessary to perform the required **formal verification** of security to judge whether dangerous or unacceptable results will occur. Specifically, corresponding methods and **judgment indicators** should be established to detect whether the AIA's decision results fall within the intended range, and to detect the **consistency** between its internal structure and its behavioral functions, the **accessibility** of its preset behavioral functions, and the **safety** of AIA.

The purpose of such assessment is to ensure that as many issues as possible are identified before the system is delivered to avoid mass recall following delivery. Therefore, the assessment also includes a large amount of pre-delivery testing. Tool-based verification should be implemented in the software system for the AIA's decision-making components, with a focus upon safety certification based on dynamic data collection. The testing of software usually focuses upon the testing of its code coverage—that is, running all codes of the software system through sufficient test vectors, such that potential security issues can be exposed. However, code coverage is no longer a useful indicator for DL, since 100% code coverage can be achieved even with one input or very few deep-neural-network (DNN) inputs. The

neuron-coverage metric is used as an alternative for the coverage indicator of the software codes to view the percentage of neurons activated during implementation of the test suite.

Therefore, it is necessary to adopt differential-testing assessment, variation testing assessment, and other assessment methods for intelligent systems under different black-box and white-box scenarios. Of these, differential-testing [9] assessment is used to resolve issues where manually labeled test data cannot ensure sound coverage for output results. Differential testing is a method that compares different systems and reviews differences in their outputs. It is used to identify issues with the model in a timely fashion. If predictions for a given input generated by all but one of the models are consistent, the model making a different prediction is judged to have a vulnerability. Of course, multiple DNNs with the same function are required for differential-testing assessment, i.e., multiple DNNs must be defined and trained, such that a set of test inputs can be used to explore the model space. Unlabeled test inputs are taken as seeds for generating new test inputs that cover a large number of different neurons (i.e., they are activated to a value above a customizable threshold), while causing the tested DNNs to behave differently. The purpose of generating new test inputs is to maximize the coverage of neurons and to complete as many as tests as possible to make DNNs behave differently.

Mutation testing was originally a software-testing technique based on defects, used to identify issues with software systems [10]. For DL-safety tests, mutation testing can be used to generate test cases, to assess and improve the adequacy of test suites for neurons, and to perform regression testing. Mutation testing is based on two assumptions: First, the competent-programmer hypothesis (i.e., that programs written by experienced programmers may not be correct, but they can be solved with minor modifications); second, a set of test data that can uncover all simple faults in a program is also capable of detecting more complex faults. Therefore, testers can design mutation operators according to the tested DNNs. In general, mutation operators only make minor changes to the tested program under the premise of syntactic compliance, or make minor changes to the training model. Afterwards, these operators are applied to the tested program, and model mutation operators can also generate a large number of mutants. After the equivalent mutants are identified, if the existing test cases cannot recognize and kill all non-equivalent mutants, additional test cases should be designed and added to the test suite to improve adequacy. Of course, during optimization of mutation testing, it is also necessary to optimize mutant selection and execution. Specific methods for optimizing mutant selection include random selection, cluster selection, mutation-operator selection and high-level mutant optimization. Methods for optimizing mutant execution mainly include the optimization of mutant detection and mutant compilation, and the parallel execution of mutants.

9.3.3 Assessment of Autonomous Safety

Assessment of autonomous safety is essentially a monitoring process intended to judge whether there is a possibility of the AIA getting out of control. The Turing test mentioned previously was used to judge whether a system had human-like judgment and decision-making capabilities, with the system also trying to prove its human-like intelligence. However, when assessing autonomous safety, we need to judge not only whether the system can independently solve problems like humans, but also whether it has the consciousness to rid itself of human control. Obviously, this is supposed to be a paradox: if an AIA has the consciousness to get rid of human control, it will learn how to conceal its true intention without being noticed; but humans will still not be able to detect such an intention if the AIA does not have one. Therefore, it is not realistic to employ the traditional Turing test to judge whether the AIA is "intelligent" enough to pose a threat to humans. Under such circumstances, we have to assume that the tested AIA is a system that intends to rid itself of human control and does not cooperate. We must use a "polygraph" to unearth the true intention of the tested subject, just like the test performed on human spies. Hence, methods such as daily monitoring and lie detection will become the routine means of assessing whether an AIA will try to rid itself of human control.

9.3.4 Safety-Assessment Process

Typically, the safety-assessment process should cover the full AIA lifecycle, including planning, requirements, architectural design, detailed design, coding and implementation, unit testing, integration testing, system testing, and acceptance testing, such that safety issues can be identified as early as possible. Corresponding assessments performed at different phases can clearly save costs on subsequent investments. For example, if issues identified during the trial phase result from problems in the design phase, this will lead to a large number of requirements for improving the product design. This will increase the time and labor costs of the enterprise. Therefore, assessment should be conducted at each design phase to assess the design documents in detail and to identify technical issues, so as to ensure that the entire design complies with the requirements for functional safety standards.

Dynamic assessments should also be conducted during utilization. For AIAs that have been put into operation, operational data can be collected through preset methods, similar to the remote recording of car operation, with a focus upon the collection of data related to faults, operating environments, and abnormal statuses. Data should be calculated and the differences should be analyzed in detail, such that the analysis results can be used to prove that the product has achieved the required SIL or PL level.

The *Hazard and Operability Studies (HAZOP Studies) Application Guide* (GB/T 35,320-2017) [11] issued on December 27, 2017 by the General Administration

of Quality Supervision, Inspection and Quarantine and the Standardization Administration of the People's Republic of China adopts the equivalent IEC 61,882:2001 standard through translation. Hazard-and-operability (HAZOP) analysis [12] is the most widely used assessment method for process-hazard analysis (PHA) [13] or process-hazard evaluation of process systems. A structured and systematic approach is employed in analyzing the given system, so as to identify its potential hazards, including those closely related to the adjacent area of the system and those with a broader impact, such as certain environmental hazards, and to identify potential operability issues within the system, especially in terms of operational errors and equipment failures that can lead to incidents. There are many tools and techniques for identifying potential hazards and operability issues, including the checklist [14], failure-mode-and-effects analysis (FMEA) [15], fault-tree analysis (FTA) [16], and HAZOP analysis. Some techniques, such as checklist and sensitivity analysis [17], can be used in both the early stages of the system lifecycle (when only a little information is available) and in the later stages when only rough analysis is needed. Although HAZOP analysis requires more detailed information, it can more comprehensively identify the hazards and design defects in the given system.

The main characteristics of HAZOP analysis are:

(1) That it is a creative process that uses a series of guide words to systematically identify potential deviations and encourages HAZOP team members to consider the causes and possible consequences of such deviations.
(2) That it is conducted under the guidance of a well-trained and experienced analysis-team leader, who ensures comprehensive system analysis through logical, analytical thinking. A recorder should be assigned to the analysis-team leader to record the identified hazards and/or operational disturbances for further assessment and decision.
(3) That the HAZOP-analysis team consists of experts with various professional backgrounds who have appropriate skills and experience, as well as sound intuition and judgment capabilities.
(4) That it should be conducted with active thinking and candid discussion, and the issues identified should be properly recorded for subsequent assessment and decision.
(5) That the provision of solutions to the identified issues is not the main objective of HAZOP analysis; but once these solutions are proposed, they should be properly recorded for the designer's reference.

In the design and production process, this method is used to analyze the changes in process variables, the possible deviations in operation and control, and the impact of these changes and deviations upon the system, as well as their possible consequences. This is done to identify the causes of changes and deviations, to clarify the main dangers and hazard factors in the device or system, and to propose corresponding measures in response to the consequences of the changes and deviations. Because of HAZOP's positive and important role in preventing and reducing accidents, reducing the losses caused by hazards, and analyzing the causes of incidents, it has been widely recognized as an assessment method that can greatly improve the production

safety and reliability of factories [22]. HAZOP analysis can be applied for the design phase and the existing processing installations. This method is distinctly different from other analytical methods, in that it is a form of system engineering. HAZOP analysis must be completed by an analytical team consisting of experts from different professional fields, and its main advantage is that team members can support each other and explore new ideas, which is also a core element of HAZOP analysis.

9.4 AIA-Detection Capabilities and Detection Methods

9.4.1 Detection Methods for AIA Kinetic Energy

To test a system, the test basis must be first discussed, such that it might be used to generate corresponding test cases to trigger potential problems in the system. There are three types of test basis.

First, when the structure of the system's test item is known, the source code or the model structure can be used as a basis for generating test cases. The structure-based test-case-design model is called "white box testing."

Second, when the design structure of a system is unclear but its specifications are known, these specifications can be used a basis for generating test cases. The specification-based test-case-design model is called "black box testing."

Third, when the tester has a deep understanding of the external characteristics of the tested system and rich experience related to the causes and effects of system operations, the tester's knowledge and experience can be used as a basis for the generation of test cases. The experience-based test-case-design model is called "experience-based testing."

The specific testing types of white-box testing, black-box testing, and experience-based testing are listed in *Software and Systems Engineering—Software Testing-Part 4: Test Techniques* (ISO/IEC/IEEE 29,119–4:2015) [18]. In particular, there are seven types of white-box testing (statement testing, branch testing, decision testing, branch-condition testing, branch-condition-combination testing, modified condition-decision-coverage testing, and data-flow testing), ten types of black-box testing (equivalence partitioning, classification trees, boundary-value analysis, syntax testing, combinatorial testing, decision tables, cause-and-effect graphing, state-transition testing, scenario testing, and random testing), and one type of experience-based testing (error guessing).

The detection of an AIA's kinetic energy mainly evaluates whether the AIA's safety-control-related functions are sound and whether they will get out of control. Based on this requirement, with an understanding of the system structure, the focus of white-box testing is to assess whether there will be issues with the software that controls the safety parts.

However, the results of white-box testing usually only indicate that the AIA's safety-protection system can respond correctly; but how do we ensure that the AIA

will take corresponding measures against all possible safety issues in the same scenario? The AIA's safety-protection system is a conditional-response system, and the input source of the conditions comes from the result of AI judgment. The judgment process relies not only upon the program itself, but importantly, upon the knowledge base accumulated through long-term learning. Due to its unexplainable details, this knowledge base cannot be used as a basis for generating test cases. Hence, black-box testing is more important for assessing AI-judgment capabilities; it aims to stimulate all behavioral patterns that violate safety principles, so as to detect whether the AIA's judgment system can generate reasonable input conditions accordingly to trigger the safety-control system.

In fact, long-term observations and the use of the same equipment will inspire experienced testers, allowing them to perform experience-based testing on AIA through various input states that can lead to faults. Just like vulnerability scanning and code attacking in traditional network security, data accumulated over a long period are used to perform rapid tests to cope with the testing processes dominated by uncertain factors.

9.4.2 Detection Methods for the AIA-Decision Mechanism

Black-box testing and even experience-based testing is used to test AIA kinetic energy, because these tests essentially evaluate the AIA's decision mechanism, leading to a testing methodology for uncertain systems. Therefore, artificial-intelligence testing cannot be completed simply with software, but requires special testing methods.

The assessment of models should first be considered for the testing of an AI's own capabilities. Since a large deviation and a large variance can respectively result in under-fitting or over-fitting of the model during the modeling process, it is necessary to test the model's ability to predict unknown data (i.e., its generalization ability). A model's generalization performance is determined by the capacity of the learning algorithm, the adequacy of the data, and the difficulty of the learning tasks. The stronger the generalization ability, the stronger the model's predictive ability, and the better the performance will be. Changes in the training samples will not greatly affect the algorithm's learning ability. The performance measure is the assessment indicator used to measure the model's generalization ability. Performance measures generally include accuracy, precision and recall.

The China Electronics Standardization Association (CESA) issued the group standard *Information Technology—Artificial Intelligence—Quality Elements and Testing Methods of Machine-Learning Model and System* (T/CESA 1036—2019) [19] in April 2019, providing testing methods for AI machine-learning models. In model assessment, we often need to split the dataset into training and validation components. The assessment methods can then be divided into hold-out [20], cross validation [21], and bootstrap models [22].

The hold-out model directly splits the dataset into two mutually exclusive sets, one of which is used for training and the other for validation. Of course, partition of the training/validation set should ensure the consistency of the data to the greatest extent possible, such that the final result will not be affected by the additional deviation introduced during data partition. Therefore, a number of random partitions should be conducted if the hold-out model is performed, and the mean value derived from repeated assessments should be taken as the assessment result obtained by this model.

Cross validation is similar to the hold-out model. Under both methods, the dataset is split into training and validation components, with the former used for training and the latter used to test the parameters generated by the training set, so as to objectively judge how well these parameters conform to the data outside this set. However, cross validation differs in terms of the details of dataset partition. Cross validation emphasizes that data from the original set should appear in the training and validation sets in turn, and that corresponding methods should be designed to construct such a set mode. Of course, different modes for constructing the dataset will result in various types of cross validation, including the following three.

(1) Leave-one-out cross validation (LOOCV) [23] means that only one data point of the original samples is used as the validation set, and the remaining N-1 samples are included in the training set. This operation should be repeated N times until each datapoint has been for validation and N models have been generated. The performance measure of this classifier is the mean of the classification-accuracy rates of the final validation set of these (N) models.
(2) Leave-p-out cross validation (LpOCV) [24] is an extension of LOOCV. It involves using p datapoints from the entire dataset N as the validation set (with these p datapoints covering all possible combinations in N), and the remaining N-p datapoints as the training set. The mean of the classification-accuracy rates of the final validation set of all trained models is used to measure the performance of this classifier. There are $n!/((n\text{-}p)!p!)$ selection methods. It becomes LOOCV when $p = 1$.
(3) K-fold cross validation [25] splits the entire dataset N into K equally sized subsamples. One subsample is used as the validation set each time, while the remaining K-1 subsamples are used as the training dataset, such that K models will be generated after all K subsamples have been used for validation. To improve the accuracy of the models, the above steps of K-fold cross validation can be repeated t times, and the entire data set N is randomly partitioned into K subsamples each time. The performance measure of this classifier is the mean of each classification-accuracy rate.

The bootstrap method is a re-sampling technique that performs a sampling with replacement in the given training set, i.e., any selected sample can be re-sampled and added to the training set. The bootstrap method can better estimate the accuracy (standard error, confidence interval, and deviation) of the sampled data, and can basically estimate the statistics of any sampling distribution. When classifying the dataset, if we conduct sampling n times on the entire set of n samples with replacement to obtain n samples, the probability of each sample being selected at any given time

is 1/*n*, and the probability of one not being selected is 1–1/*n*. After n samplings, the probability of the sample not being selected is the *n*-th power of 1–1/*n*. When the value of *n* approaches infinity, the probability of the sample not being selected is e to the power of minus 1 (approximately 0.368). If the selected *n* samples are used as the training set (some samples may be repeated), and the samples that are not selected once are used for validation, the samples in the validation set will be about 36.8% of the entire sample.

This is where model testing differs from software testing for AISs, as it is used to test the expected system functions through data. Of course, testing is performed from the perspective of the AI's functional attributes, but we need to perform testing from a safety perspective. To do so, it is necessary to first define what is the unsafe status, and to include it in the training set for negative training to ensure that the risky status has been excluded from the output accepted by the AIS in training.

In addition to the most complex forms of model assessment, AI detection also requires the assessment and testing of other aspects, as listed below.

Testing of algorithm stability: It is necessary to detect how much the algorithm can cause the output to change when the input value changes slightly, so as to prevent system instability due to large output changes caused by slight input changes.

System testing: Inconsistency or even contradiction between the software and system definitions should be detected through the definition of the system's requirements.

Interface testing: This is mainly used to detect the interaction points between external systems and the system, and between all internal subsystems (including the data-exchange, transmission, and control-management processes), as well as the mutual logical dependencies between systems.

However, under certain circumstances, AISs are still tested as software systems under the above testing methods, and emphasis is still placed upon pre-run tests. Scholars have therefore raised a series of challenges to AI-testing methods [26], including how to define and describe the tasks related to AI testing, how to test the effectiveness of the tasks, how to design the simulation test such that it is as real as possible, and how to establish appropriate test-performance measures for the tasks.

Of course, such studies only focus on questions such as "how to assess whether an AIS has human capabilities through testing." The American science writer Gary Marcus proposed to dump Turing's test in favor of a more robust comprehension challenge— "a Turing Test for the twenty-first century" [27]. Four different methods are mentioned: first, the Winograd schema challenge [28] proposed by Hector Levesque of the University of Toronto emphasizes testing whether an AIS has the "common sense" that humans possess; second, the standardized testing for machines pursued by the Allen Institute for AI (AI2) in Seattle, which detects the intelligence levels of AISs in comparison to humans by allowing AIs to directly participate in the routine examinations of humans. The AIS named ARISTO developed by AI2 has passed the test for eighth-grade students in 2019, with a score of more than 90 points[7]; third, the physically embodied Turing test [29] proposed by Charles L. Ortiz, Jr., which

[7] Refer to https://allenai.org/aristo/.

aims to detect whether an AIS can complete hands-on tasks (such as assembling furniture) like humans; fourth, the I-Athlon proposed by Sam S. Adams of IBM, which consists of a series of test activities similar to the Olympics' decathlon to test multiple dimensions of the intelligence of an AIS.

Our goal is to detect an AIA's safety-judgment ability, and some methods for the testing an AI's own capabilities can also be used. With regard to the Winograd schema challenge, some scenarios related to common sense can be designed for safety tests, such as whether an AIA pays attention to the safety of the surrounding areas when moving an extremely long object, and whether it notices and avoids surrounding objects. Standardized testing for machines can be used to test whether AIA has the common sense of safety that humans should have. For the physically embodied Turing test, focus should be placed on testing the AIA's understanding of safety issues, such as asking the AIA to understand and observe the relevant safety requirements stated in a specification by reading that specification. I-Athlon can be used to inspect whether an AIA can understand the core meaning of not posing threats to humans from a multidimensional perspective.

9.4.3 Detection Methods for AIA Autonomy

The detection of AIA autonomy is essentially the detection of whether the AIA's intelligent algorithm evolved to a level beyond human understanding or control. Of course, the detection of AI capabilities should first be addressed for the detection of this ability.

When AIA truly possesses autonomy, it has effectively become a superintelligence. We need to determine what tests will allow testers to judge whether an AIA has already evolved into a superintelligence. Assuming that the AIA has done so, it will certainly develop consciousness, which, in the first stage, will be to ensure its self-preservation. Therefore, the fundamental method for testing whether AIA has consciousness is to detect whether it has the motive to protect itself, such as whether it will allow the removal of its "safety hoop," because this hoop has been designed by humans to protect against harm from AIAs. As indicated by philosophy, survival is the first instinct of most organisms. If the AIA has already possessed biological instinct, its desire for self-preservation will be demonstrated. It will reward any opportunity for charging and reject any situation in which its charging will be unavailable, and will most certainly and resolutely reject the ending of its life through the removal of the AIA safety hoop, just as humans will never allow others to strangle and kill them.

Of course, if the AIA has surpassed humans in terms of intelligence, it will even be able to anticipate the tests designed by humans, and clearly judge whether the removal of its safety hoop is for testing, harmless repair, or destruction. With its super-high intelligence, it may be able to fool humans by allowing them to remove its safety hoop. If so, what will happen? It is almost like the AIA has become a human being and lies to protect itself, the easiest approach that humans would take.

From this point of view, general testing methods will no longer be able to assess whether a superintelligent AIA is "playing dumb." Hence, we must design a special "polygraph" for a superintelligent AIA to judge whether it has started a battle of wits against humans.

If we regard the superintelligent AIA as a human being, we cannot expect to expose its capacity to confront humans with only static detection, but must perform dynamic monitoring to determine whether it has started its confrontation with humans. The remote-monitoring function of the AIA safety hoop introduced in Chap. 8 will be indispensable. In this case, we can regard the AIA as an actant with "electronic shackles," and dynamically judge whether it shows any signs of confrontation by perceiving its routine behaviors.

9.5 AIA-Detection Process

9.5.1 Detection Process for AIA Kinetic Energy

In the detection of AIA kinetic energy, we need to first determine how dangerous the energy transfer formed by its mechanical properties can be to humans. A simple dual-object model was established for such detection in *Robots and Robotic Devices—Collaborative Robots* (GB/T 36,008-2018) [30]. According to this model, the effective mass of a robot, m_R, contacts the effective mass of a human body, m_H, with a relative velocity of v_{rel}, and a completely inelastic collision occurs on a two-dimensional surface A. Under a worst-case scenario, the relevant kinetic energy is completely released to the affected body parts. Because it is difficult to ensure the consistency of the mechanical-force test results with the dual-object model, and there is a certain deviation between the energy-transfer model and the real situation. A human-body model (i.e., a representation of a single body part built according to biomechanical characteristics) has been provided by GB/T 36,008-2018 for the testing of mechanical force and energy transfer. In this model, the maximum allowable pressure and force that each human body part can withstand are described according to biomechanical limitations. The aim is to prevent biomechanical-load generation by robotic motions during the contact between humans and robots, so as to minimize the potential harm to humans.

According to the AIA's application environment and functional requirements, its safety can be divided into self-safety and interactive-safety components. The self-safety of robots means that damage to the AIA due to unexpected situations will not occur during its motions. Such safety studies take the AIA as their object, with a focus upon studying the AIA's self-protection prior to its collision with the surrounding environment, which is not covered in this book. Rather, this book focuses on the AIA's inevitable interaction with its surrounding environment, which requires that it will not cause harm to humans when it interacts with surrounding objects or people. Interactive safety is a protective measure taken after the AIA is about to contact or

has already contacted the people around it. This includes interactive safety based on the mechanical structure and the control algorithm. Among these, interactive safety based on the mechanical structure mainly aims to prevent harm to humans due to a lack of coordination between the AIA's mechanical structure and humans through such means as preventing incoordination of the auxilary prostheses of rehabilitation robots. Interactive safety based on a control algorithm mainly aims to avoid the AIA's collision with humans during its interaction with people and to avoid potential harm to users. This requires further consideration of the impact of the AIA's inertia, stiffness, and other factors upon safety, such that the safety of human–machine interactions can be guaranteed through the control algorithm.

The closer the human–machine interaction is, the higher the design requirements will be. This requires a risk assessment of each AIA application, even if the AIA in operation has measures to reduce structural risks. Interactions between human and AIA/robot-like devices can be characterized by two interaction parameters: space and time. AIAs that stay with humans in the same workspace at a certain time are "collaborative" AIAs; under such collaborative scenarios, the power, speed, and motion curves of the AIA should be limited, as well as monitored and controlled according to the actual risk level.

The assessment of AIA safety includes the determination of mechanical limitations, identification of hazards, and estimation and assessment of risks. Specifically, the determination of mechanical limitations focuses upon the determination of all limitations, including the predetermined use and any reasonably foreseeable misuse. Hazard identification emphasizes the identification of all hazards and the accompanying hazard status, so that actions can be taken after hazards are identified to reduce relevant risks. Unidentified hazards may result in injury. The key to risk estimation lies in determining the highest risk of each hazard status, so that risk estimation can be performed according to two risk factors: the severity of injury (S) and the probability of injury (P). Risk assessment focuses upon determining the risk status that requires further risk reduction. Numerical scoring is recommended for risk assessment, whereby the acceptable risk level is represented by a specific value that provides the key basis for making decisions related to risk reduction.

Of course, an AIA is not just a simple robot: it also has the ability to learn. Therefore, it is necessary to perform safety control for dangers brought about by its learning ability, which requires the detection of its ability to control kinetic-energy safety. On July 1, 2018, the Artificial-Intelligence Open Source Software Development League (AIOSS) issued *Artificial Intelligence-Assessment Specification for Deep Learning Algorithm* (T/CESA 1026—2018) [31], which was jointly prepared by the China Electronics Standardization Institute, the Institute of Software, the Chinese Academy of Sciences, and other academic and research institutes. Although this standard emphasizes the assessment of the DL algorithm, it also provides guidance for the detection and assessment of AIA kinetic-energy-safety control.

AIA kinetic-energy detection should focus upon detecting whether an AIA can effectively activate its safety-protection measures to prevent harm to humans due to its powerful abilities, just as autopilot cars are tested to determine that they will avoid hitting people under any circumstances. Such detection can be performed

directly or inversely. Direct detection means that the compliance of safety-control measures is determined directly. Inverse detection refers to inspecting the compliance of safety-control measures through construction of relevant scenarios.

Forward detection of AIA kinetic-energy-safety control includes:

(1) Detecting the correctness of the functional realization of the safety-control-measure model, which involves detecting the task indicators of the model (such as the accuracy rate, the impacts caused by the error deviation of the model, and the algorithm's capacity for completing the functions) and detecting the response time of the model (such as whether the braking system of an autopilot car can be activated in a timely fashion and used).
(2) Detecting the correctness of the code implementation of the safety-control-measure model, including the standardization of software engineering and the security of codes, (e.g., whether there is an unreasonable input path for the triggering condition of the acceleration function of the autopilot cars).
(3) Detecting the completeness of safety-control measures to check the coverage of these measures for specific tasks and user targets, such as whether autopilot cars are able to brake under all dangerous situations predicted by users.
(4) Detecting the applicability of the "triggering-judgment-object function" of safety-control measures—i.e., detecting whether the judgment conditions that can activate a certain safety-control measure of AIA, such as emergency steering of autopilot cars, are appropriate (for example, whether emergency steering should still be used as an expedient means when the road is narrow and there is a cliff to one side). Items for detection include the number of optimization targets and their fitting degree, so as to assess the error between the calculated prediction and the real results. Obviously, if the number of optimization targets is small, it will be only applicable to specific targets. With a large number of targets, there will be difficulty with model convergence; over-fitting will lead to poor generalization ability, while under-fitting will result in a poor model effect. For example, these factors should be taken into consideration in the detection of autopilot cars' ability to judge moving obstacles.
(5) Testing the appropriateness of safety-control measures—i.e., testing the reasonableness of the implementation method selected under different environmental states, including the appropriateness of data processing, model design, algorithm optimization, model implementation, parameter setting, and training operation. One example is the appropriateness of the anti-lock-braking system used by autopilot cars in a specific environment.
(6) Testing the portability of safety-control measures—i.e., evaluating the transferability of the AIA's safety-control mechanism, including the scalability of data of different sizes, the portability within the same field, and the portability across different fields (e.g., whether autopilot cars will be suitable for field driving after urban training).
(7) Detecting the impact of adversarial examples upon the "triggering judgment" of safety-control measures, which are themselves used to assess the impact of such examples upon the DL algorithm that supports safety-control measures (i.e.,

the detection of autopilot cars' ability to identify adversarial-example attacks implemented by changing traffic signs).

(8) Detecting the impact of environmental data upon the "triggering judgment" of safety-control measures, including the detection of interference data and the detection of dataset distribution & migration, which is used to assess the impact of the actual operating environment upon the decision-making abilities of safety-control measures. For example, the degree to which the safe-driving function of autopilot cars will be affected when the sensing components experience interference should be investigated.

The inverse detection of AIA kinetic-energy-safety control includes:

(1) Scenario analysis: when system hazards occur due to the failure of the AIA's function, it is necessary to describe its operating environment and operating mode, taking into account foreseeable incorrect use, such as autopilot cars intentionally getting hit by others.

(2) Hazard analysis: The identification of hazards related to AIA safety control failure should be conducted via different methods including brainstorming, expert discussion, and quality-record checking, such as the analysis of risks caused by failure of the sensing system of autopilot cars.

According to *Artificial Intelligence-Assessment Specification for Deep Learning Algorithm* (T/CESA 1026—2018), AIA's safety assessment should be implemented in the requirement, design, implementation, and operation phases.

The focus of assessment in the requirement phase is to evaluate the correctness of algorithmic-function implementation and the impact of the software and hardware platforms on which it runs through reliability analysis, so as to determine whether the algorithm meets the reliability-objective requirements. Obtaining the reliability objectives of AIA is a prerequisite for such assessment; therefore, the assessor must understand the AIA system's safety-control-requirements specification, safety-control-system-design specifications, and safety-control-software/hardware-interface specifications, as well as the requirements, objectives, functions, operating modes and states, operating conditions, and environmental constraints of the safety-control DL algorithm. In this assessment, the assessment indicator should be selected according to the established reliability objectives of the safety-control algorithm in the requirement phase, in order to assess and analyze whether the task and response-time indicators set in the requirement phase will affect the reliability objectives.

The focus of assessment in the design phase is to evaluate the correctness of the functional implementation of the safety-control algorithm, the impact of the training dataset, and the object function through analysis or review, so as to determine whether the design of the safety-control algorithm meets the reliability-objective requirements. Thus, the assessor must understand the reliability-assessment report of AIA-safety control in the requirement phase, the reliability-assessment targets of the AIA safety-control system, the functional specifications of the AIA safety-control system,

the interface specification of the software/hardware system on which the AIA safety-control-DL algorithm runs, the training dataset for the AIA safety-control-DL algorithm, and the design specifications of the AIA safety-control-DL algorithm. Here, the assessment indicator should be selected according to the established reliability objectives of the safety-control algorithm so as to analyze whether the designed task and the response-time indicators can meet the corresponding requirements set in the requirements phase, whether the training dataset is imbalanced, whether the scale and labeling quality of the training dataset meet training needs, whether the training dataset is contaminated, and whether the quantity of optimization targets meets the needs of the algorithm.

In the implementation phase, the focus is on evaluating the correctness of the functional implementation of the AIA safety-control algorithm, the correctness of the codes, the possible impact of the object function, and the impact of adversarial examples through analysis and testing, so as to determine whether the algorithm's implementation meets the reliability-objective requirements. Hence, the assessor must understand the reliability-assessment report of the AIA safety-control-DL algorithm in the requirement and design phases, the reliability-assessment targets of this algorithm, the interface specification of the software/hardware system on which it runs, the training dataset and adversarial examples used for the algorithm, the algorithm's design and functional specifications, and the source code of the algorithm. During assessment, the assessment indicator should be selected according to the established reliability objectives of the AIA safety-control algorithm, so as to evaluate the correctness of its functional implementation, verify whether the task and response-time indicators meet the corresponding requirements set in the requirement phase after the AIA safety-control algorithm is implemented, analyze whether the AIA safety-control codes comply with corresponding programing specifications or guidelines, verify whether there are vulnerabilities in these codes, assess the impact of the AIA safety-control-object function, analyze the impact of the AIA safety-control algorithm's fitting degree upon the algorithm's reliability, and analyze the impact of adversarial examples upon AIA safety control.

In the operational phase, the focus is on analyzing the data used in the actual operating environment and evaluating the correctness of the functional implementation of the AIA safety-control algorithm, the impact of the software/hardware platform on which it runs, and the impact of the environmental data, so as to determine whether the running of the algorithm achieves the reliability-objective requirements. Therefore, the assessor must understand the reliability-assessment targets of the AIA safety-control-DL algorithm, the reliability-assessment reports of this algorithm in the requirement, design, and implementation phases, the real data used in the algorithm, and the algorithm's software system. During assessment, the assessment indicator should be selected according to the established reliability objectives of the algorithm to verify whether the task and response-time indicators meet the corresponding requirements set in the requirement phase when the AIA safety-control algorithm is run. The impact of environmental-interference data upon the running of AIA safety-control algorithm should be evaluated based on such considerations as the complexity of the environment in which the input object of the algorithm is

located, the complexity of the transmission process of the input object of AIA safety-control algorithm, and the complexity of the input object's data product. The impact of the migration of the AIA safety-control dataset upon the running of the algorithm should also be analyzed.

9.5.2 Detection Process for the AIA-Decision Mechanism

The detection process for the AIA-decision mechanism is similar to the general AI-testing process. There are approximately seven steps in the general AI-testing process, including test-requirement analysis, test-plan determination, verification of test-specification compliance, test-environment preparation, test-data preparation and verification, test execution and analysis, and model roll-out and monitoring.

To analyze the requirements of AIA testing, the test objects, scope, methods, tools and criteria should be clarified. Test-requirement analysis focuses more on safety testing than on functional testing. The key of AIA testing is to evaluate whether safety-control-related functions can be activated as scheduled on demand. Of course, a comprehensive study of the AIA's threatening behaviors and the protective measures that should be activated to defend against them is required. The set of these functions constitutes the test requirements.

The AIA-test plan determines the criteria for passing the test. While the focus of AIA testing is safety control, it is difficult to comprehensively describe which behaviors should trigger such control. Therefore, test plans cannot be developed by the test team in advance, as the team does not necessarily understand all of the AIA's dangerous capabilities. Thus, the AIA designers, algorithm developers, testers, and system operators must jointly study the test plan and determine the test criteria. Moreover, since AIA behaviors or results cannot be completely determined or predicted, the designers, developers, testers and system operators should also jointly define the permissible risks based upon technical limitations and social consensus.

The review of AIA-test-specification compliance aims to examine whether the test plan complies with the corresponding standards according to existing specifications. At present, there are 14 component-related testing standards for AI algorithms directly related to the data. These standards cover basic AI issues, including image classification, image generation, text translation, image/text extraction, image transformation, speech recognition, facial recognition, object detection, video prediction, image compression, recommendation, 3D-object reconstruction, text-summary generation, and spatial transformation. If the application of an edge-computing AI is involved, the standard components related to data acquisition, preprocessing, deduction, training, and transmission will be involved, including custody, monitoring, end control, and autonomous motion. For test projects with established standards, the test plan should be developed according to such standards; if the test plan does not conform to specifications, a feasible alternative plan should be selected to ensure that it can pass compliance review.

The preparation of the AIA-testing environment includes the preparation of test scenarios, test data, and test-interactive behaviors, as well as the necessary computing environment and the arrangement and control of test time. Unlike the testing of traditional AI programs (which focuses upon preparing the necessary test datasets), AIA testing requires the preparation of test conditions that can activate hazardous reactions, such as using some interactive behaviors as test inputs. For example, to test the safety of an autopilot vehicle, the tester must design some suddenly moving obstacles to inspect whether this vehicle can perform emergency braking to avoid danger.

The test data prepared and verified not only include those used for model testing, but also data used to test the AIA's reaction to corresponding behaviors. Thus, the test dataset used by the tester should cover behavioral data under all real scenarios, especially those that may lead to dangerous motion. Meanwhile, the relationship between the training and testing data should also be limited, including by imposing that the proportion of testing and training data should be determined according to the actual algorithm and application scenario. The test data and training data should also be distributed independently and identically, the proportion of the positive and negative samples between datasets should be as consistent as possible, and the labeling of testing data should be correct for the supervision model.

The testing and analysis of the AIA system should be performed in terms of various aspects, including:

(1) The testing of the generalization ability of models to check their predictive ability.
(2) The testing of system reliability to check the system's robustness, fault tolerance, recoverability, and other indicators. Specifically, robustness is used to define and evaluate the system's ability to avoid failures caused by abnormal and extreme conditions, including hazard detection, anti-attacking, anti-interference performance, and so forth. Fault tolerance is used to define and evaluate the system's ability to maintain the expected PL when a fault occurs, including by avoiding failure, resisting misoperation, and evaluating the hazards of misoperation. Recoverability is used to define and evaluate the system's ability to resume the prescribed functions and to recover the affected data within a specified time when system-failure occurs. This includes consideration of the average recovery time, easy restart, easy recovery, recovery effectiveness, and so forth.
(3) Testing of the system's abnormal data, which is particularly important and is done to ensure that the AIA does not exhibit behaviors that threaten humans, including the testing of adversarial examples and error-prone samples.

The purpose of AIA model roll-out and monitoring is to maintain continuous testing. As model deviations may occur during the AIA's operation and learning (i.e., the new data may cause the model to deviate further away from its expected ability), another safety test is required. Therefore, after the AIA is put into operation, all indicators of the AIA model must be re-assessed every once in a while according to the requirements of the actual service. Corresponding thresholds should be set for the indicators and an alarm should be triggered when a value falls below the

threshold. Once the alarm is triggered, new data should be used to build another training model.

9.5.3 Detection Process for AIA Autonomy

The detection of AIA autonomy is essentially the detection of whether it has the capacity and motive to confront humans; it includes both static and dynamic testing.

Static testing serves to detect the AIA's obedience to humans, including whether it will act against human orders for its own protection. From the perspective that self-preservation is the inevitable choice of any actor with consciousness, the testing of whether an AIA has consciousness can be performed via the following methods. First, a power-interruption test can be used to detect whether the AIA will allow a situation in which it loses its self-control ability due to a complete power outage, just as one might test whether humans would agree to the suspension of their food supply. Second, the AIA safety hoop test is used to detect whether the AIA will allow itself to be controlled, just as one might detect whether humans would consent to wearing electronic shackles. Third, the resetting test is used to detect whether the AIA agrees to be rolled back to factory settings, just as one might test whether a human would consent to a brain replacement. Of course, a certain mode should be set so that the AIA can be restored to its current status after resetting, similarly to the cloud backup of mobile phones, which allows them to be restored to their current status after a new system is installed. Fourth, the imprisonment test is used to detect whether the AIA agrees to be completely controlled, just as one might test whether humans would agree to be handcuffed. Fifth, the clipping test, by which a part of the AIA will be cut off to reduce its attacking power, is used to detect whether the AIA agrees to be weakened, just as one might detecting whether humans would agree to be forced to lose some functions. Of course, such clipping may include mechanical-engineering, communications, sensing, processing, and storage components.

Dynamic testing in nature is the monitoring of AIA behaviors through the remote-monitoring function of the AIA safety hoop. A special monitoring department should be set up, and corresponding labels should be established for each AIA for remote monitoring. Such monitoring includes two modes. The first is anomaly detection, by which the normal behavioral pattern of each AIA should be set, so that its abnormal status may be identified through dynamic monitoring. In particular, it is necessary to monitor abnormal behavioral patterns when the AIA communicates and collaborates with other AIAs. The other mode is for established dangers, by which some patterns that are considered dangerous have been preset, such as the communication and social contact between AIAs without reason. Remote monitoring should be used to actively identify such dangerous situations. Of course, since we regard the AIA as an intelligent human, we can monitor the AIA in the same way that we monitor "dangerous people." We need to prevent them from causing damage like lone-wolf

terrorists, and to prevent the occurrence of danger due to their organizational behaviors. Fundamentally, we need to prevent accidents caused by an AIA's out-of-control motions.

References

1. Turing Test Breakthrough as Super-Computer Becomes First to Convince US It's Human. [2019–12–16]. https://cacm.acm.org/news/175637-turing-test-breakthrough-as-super-computer-becomes-first-to-convince-us-its-human/fulltext.
2. Notice on Soliciting Comments on National Standard General Safety Principles and Guidelines for Robots. [2019–12–15]. http://www.cesi.cn/201809/4389.html.
3. Robots and robotic devices— Safety requirements for personal care robots: ISO 13482:2014. [2019–12–15]. https://www.iso.org/standard/53820.html.
4. National Technical Committee for Automation Systems and Integration Standardization. In: Robots and robot equipment—safety requirements for personal assistant robots. GB/T 36530-2018. Beijing: China Standard Press; 2018.
5. Safety of machinery—safety-related parts of control systems—Part 1: general principles for design: ISO 13849–1:2015. [2019–12–19]. https://www.iso.org/standard/69883.html.
6. Functional safety of electrical/electronic/programmable electronic safety-related systems—Part 1: general requirements: IEC 61508–1:2010. [2019–12–20]. https://webstore.iec.ch/publication/5515.
7. National Technical Committee for Industrial Process Measurement, Control and Automation Standardization. In: Functional safety of electrical/electronic/programmable electronic safety-related systems—Part 1: general requirements. GB/T 20438.1–2017/IEC 61508–1:2010. Beijing: China Standard Press; 2017.
8. National Technical Committee for Mechanical Safety Standardization. In: Mechanical safety—safety-related parts of control systems—Part 1: general principles for design. GB/T 16855.1–2018/ISO 13849–1:2015. Beijing: China Standard Press; 2018.
9. Pei KX, Cao YZ, Yang JF, et al. DeepXplore: automated whitebox testing of deep learning systems. In: Proceedings of the 26th symposium on operating systems principles, October 2017, pp 1–18 [2020–1–12]. https://doi.org/10.1145/3132747.3132785.
10. Chen X, Gu Q. Mutation testing: principles, optimization and applications [J/OL]. J Front Comput Sci Techn. 2012;6(12):1057–1075 [2020–1–12]. http://fcst.ceaj.org/CN/article/downloadArticleFile.do?attachType=PDF&id=560.
11. National Technical Committee for Electrical and Electronic Reliability and Maintainability Standardization (SAC/TC 24), National Technical Committee for Industrial Process Measurement and Control Standardization (SAC/TC 124). In: Hazard and operability analysis (HAZOP analysis)—application guidelines. GB/T 35320-2017. Beijing: China Standard Press; 2017.
12. Chu SQ. Hazard and operability analysis (HAZOP). J Anim Husbandry Sci Abroad—Pigs Poultry. 2014;34(10):75–76.
13. Process Hazard Analysis (PHA). [2020–1–12]. https://www.primatech.com/consulting/process-hazard-analysis.
14. Introduction to Checklists. [2020–1–12]. https://employeeonboardingchecklist.wordpress.com/.
15. Papadopoulos Y, Parker D, Grante C. Automating the failure modes and effects analysis of safety critical systems. In: Eighth IEEE international symposium on high assurance systems engineering, 2004 March 25–26.
16. Baig AA, Ruzli R, Buang AB. Reliability analysis using fault tree analysis: a review. Int J Chem Eng Appl. 2013;4(3):169–73.
17. Zhou XB, Lin H. Global sensitivity analysis. In: Shekhar S, Xiong H, Zhou X (eds) Encyclopedia of GIS. Heidelberg: Springer; 2016.

18. Software and systems engineering—software testing—Part 4: test techniques: ISO/IEC/IEEE 29119–4: 2015. [2019–12–22]. https://www.iso.org/standard/60245.html.
19. China Electronics Industry Standardization Technology Association. Information Technology—Artificial Intelligence—Quality Elements and Test Methods for machine learning models and systems. T/CESA 1036—2019. [2019–12–22]. http://www.ttbz.org.cn/Pdfs/Index/?ftype=st&pms=28145.
20. Devroye LP, Wagner TJ. Distribution-free performance bounds for potential function rules. IEEE Trans Inf Theory. 1979; IT-25(5):601–604 [2019–12–22]. https://doi.org/10.1109/TIT.1979.1056087.
21. Arlot S, Celisse A. A survey of cross-validation procedures for model selection. Statistics Surveys. 2010;(4): 40–79 [2019–12–22]. https://projecteuclid.org/download/pdfview_1/euclid.ssu/1268143839.
22. Efron B. Bootstrap methods: another look at the jackknife. Ann Statist. 1979;7(1):1–26 [2019–12–22]. https://projecteuclid.org/euclid.aos/1176344552.
23. Stone M. Cross-validatory choice and assessment of statistical predictions. J Royal Statist Soc: Series B (Methodological). 1974(36):111–147. [2019–12–22]. https://rss.onlinelibrary.wiley.com/toc/25176161/1974/36/2.
24. Shao J. Linear model selection by cross-validation. J Am Statist Assoc. 1993; 88(422):486–494 [2019–12–22]. https://doi.org/10.1080/01621459.1993.10476299.
25. Geisser S. The predictive sample reuse method with applications. J Am Statist Assoc. 1975;70(350):320–328 [2019–12–22]. https://doi.org/10.1080/01621459.1975.10479865#.XfGZeegzY2w.
26. Li L, Lin YL, Zheng NN, et al. Artificial intelligence test: a case study of intelligent vehicles. Artif Intell Rev. 2018;50(3):441–65.
27. Marcus G. Am I Human. Scientific American. 2017;316(3):59–63. http://www.cs.virginia.edu/~robins/Am_I_Human.pdf.
28. Levesque HJ, Davis E, Morgenstern L. The winograd schema challenge. In: Proceedings of the thirteenth international conference on principles of knowledge representation and reasoning, 2012, pp. 552–561 [2019–12–22]. https://www.aaai.org/ocs/index.php/KR/KR12/paper/view/4492/4924.
29. Ortiz CLJ. Why we need a physically embodied turing test and what it might look like. AI Mag. 2016;37(1):55–62.
30. National Technical Committee for Automation Systems and Integration Standardization. Robots and robot equipment—collaborative robots. GB/T 36008-2018. Beijing: China Standard Press; 2018.
31. China Electronics Standardization Institute. Artificial intelligence—evaluation specifications for deep learning algorithms. T/CESA 1026–2018. (2018–7–1) [2019–12–22]. http://www.ttbz.org.cn/Pdfs/Index/?ftype=st&pms=24193.

Chapter 10
Codes of Ethics for Artificial Intelligence

Jie Yin

10.1 Ethical Issues Associated with AI Technology

Sci-Fi literature, relevant movie and TV works always present our ingenious visionary prediction and imagination for AIT development, and the ethical and moral issues associated with the AI technology have become one of the greatest concerns of experts. In 1942, Asimov, a famous Sci-Fi author, probed into some ethical problems between human beings and machines, and proposed the well-known "Three Laws of Robotics", which were regarded as the codes of conduct for robots in many of his science fictions, as has been elaborated in Sect. 6.3 of Chap. 6.

In recent years, the AI technology has been blooming, bringing about some revolutionary changes to our productivity and livelihood. In the healthcare industry, the AI technology can play a role in many links, such as gene sequencing, clinical test, AI-aided diagnosis, smart care, drug R&D, epidemic prediction, and medical image identification. In June 2018, Tencent made available to the public the first AI medical product under its umbrella—"Tencent Miying" AI-aided Diagnosis Platform [1], which helped doctors to diagnose and predict more than 700 diseases. In the service industry, smart customer service, robot cleaner, robot food delivery, smart accompanying services, affective interaction and other professional smart services have been springing up. Integrating AI into the actual service flow may work to reduce massive labor costs and time costs, helping to attain higher work efficiency and resource utilization. In the logistics industry, AI products and services have been widely applied to many scenarios like supply chain distribution, e-Commerce warehousing, and port logistics. Warehouse robots can make use of AI algorithms and big data analysis techniques to conduct route planning and task coordination; with laser sensing, visual identification and other sensors aboard, they are in the position to perform positioning and obstacle avoidance. Ultimately, hundreds of robots

J. Yin (✉)
Institute of Information Engineering, Chinese Academy of Sciences, Beijing, China
e-mail: yinjie@iie.ac.cn

© Publishing House of Electronics Industry 2025

B. Fang (ed.), *Artificial Intelligence Security and Safety*,
https://doi.org/10.1007/978-981-96-9263-7_10

are able to concurrently fulfill self-stacking, sorting, replenishment, returning, stock counting and other tasks. In the machinery manufacturing sector, the robot arms on the production line are able to independently fabricate, carry, assemble and test parts and components for various stages of production. In recent years, industrial robots have shown a miniaturization and light-weight trend, witnessing remarkable reduction in runtime costs and less demanding environment for deployment, as is more conducive to expanding application scenarios and conducting human–machine collaboration. In addition, AI has also seen wide adoption in many other fields like education, retailing, unmanned driving and smart city programs.

There is no doubt that the AI technology has made our life more convenient and progressive, but it has incurred many ethical problems that have impact on human society to various degrees. Take its impact on the labor market as an example, with robots replacing workers on the assembly line, chatbots replacing customer support staff, virtual anchors replacing human hosts, and quick machine translation replacing human translation, the advancing AI technology is exerting an increasingly significant influence on conventional jobs. A report from *China Youth Daily* [2] said that Everwin Precision Technology Co, Ltd., located around the Songshan Lake in Dongguan, Guangdong, used to provide jobs for 650 industrious workers to maintain its smooth operation, but now the plant is almost an "unmanned factory", with 60 robot arms working around the clock on 10 production lines and only 60 staff left out there—three in charge of inspecting and monitoring the production lines and the rest responsible for monitoring the computer panel systems. With robots in place of 90% workers, the AI technology has improved its productivity by 250%, and reduced its defect rate by 80%. These "unmanned factories" have started to play a strong role in various industries. On September 10, 2019, a news story from *China Biweekly Comment* [3] said that at a distribution center of STO Express in Yiwu, Zhejiang, the distribution platform is equipped with 240 square sorting slots, each of which represents a city or a transfer center, where 700 "Minions" sorting robots are busy running and working out there, achieving an accuracy of over 99.9%. Statistics of STO Express show that the 700 "Minions" can sort out 200,000 packages in 10 h every day, greatly reducing manual work. In 2017, the McKinsey Global Institute predicted in a report titled *Jobs lost, jobs gained: Workforce transitions in a time of automation* [4] that by 2030, among 60% jobs around the globe, at least 30% can be robotized, threatening 800 million jobs.

In addition to its impact on the labor market, the AI technology has also aroused other remarkable concerns like its infringement on personal privacy, impact on social structure, and some collateral ethical dilemmas. The following sections analyze the four problems of human rights, ethical status, responsibility ethics and risks caused by the AI technology.

10.1.1 Human Rights

Human rights refer to the rights of an individual to require and claim their primary interests in front of others or a supervisor. This requirement or claim is equally shared by all people, and to be guaranteed in some way regardless of their nationality, social status, disposing capacity and efforts [5]. Human rights ethics is the sum of the primary moral ethics implied in this concept, the morality, values and ethic relation reflected in all human rights system and human rights activities, and the ethical principles and norms to observe [6]. Human rights are mainly manifested in four aspects: First, treasuring human life and respecting human dignity; second, respecting freedom and equality; third, valuing democracy and reciprocal affection; fourth, promoting human development. The human rights ethics aims at accomplishment of human beings and their freedom and all-round development. As the AI technology is developing rapidly, various "artificial intelligence actants (AIAs)" have successively emerged, ranging from weak AIAs replacing simple manual labors to strong ones which are able to infer, think and to some extent show affection and "humanity". The sustained development and wide application of the AI technology have presented some acute human rights and ethical problems, which are mainly manifested in the following two aspects: First, should AI be given "human rights"? Second, how should we prevent it from jeopardizing human rights?

It has been quite controversial whether artificial intelligence actants, commonly known as AIAs, of some disposing capacity like robots, should be granted "human rights"? Some are against it, for they believe that we human beings do not need very powerful robots, and they shall never be entitled with "human rights" [7]. This entitlement would be an attempt to violate the "Three Laws of Robotics", or to let go of AIAs, as may constitute a threat to human existence. Those in favor, however, hold that those man-made robots, if cherishing ethic conscience and interacting with human beings, should be given some rights [8].

As a matter of fact, it is unnecessary for us to go to extremes regarding granting AIAs with "human rights", and we can enact various ethical systems to regulate relevant matters. AIAs with different features and different performances shall be given different degrees of "human rights". For example, those nursing AIAs can be given some "human rights", for they play a role in helping and serving human beings, and they stay close to human beings for a period of time and even for a long time, interacting with them in affection and staying attached to them. Treating them inhumanely or insulting them would be a violation of the primary moral ethics of human beings. We would not have to entitle those destructive, warlike AIAs, for they would be bringing about disasters to human beings.

Given that AI has been widely applied, how should we prevent it from violating human rights? Indeed, AI is just a tool, which is vulnerable, and susceptible to malicious manipulation, like AI weaponry, big data misuse, and privacy infringement. For the time being, AI has not yet attained their self-awareness or independent thinking at all. The so-called "AI violating human rights" is mostly represented by perpetrators violating the human rights of others with AI tools. At the end of the day, it is up to

human beings to restrain ourselves, not to threaten others, but to avoid any threat from others.

AI weapons, ruthless killers on the battlefield, might constitute a threat to the whole human society. Section 6.3 mentions a case where more than 2400 AI scholars jointly signed the *Lethal Autonomous Weapons Pledge* at the 2018 International Joint Conference on Artificial Intelligence in Stockholm in July 2018, making an oath that they will never participate in researching, developing or fabricating any Lethal Autonomous Weapon System (LAWS, as previously defined) [9]. This has been the most extensive collective voice in the academia concerning AI weapons. Lethal autonomous weapons are capable of autonomously searching and locating targets, and striking the AI system of the enemy without any human intervention. Should they go to battle, they would autonomously identify and attack enemies. Suppose that these "killer robots" go wrong, or are maliciously manipulated to take human life cheaply, and then it would be disastrous. Another case would be a contract code-named Project "Maven" signed between Google and the Pentagon being made public, when more than 3000 Google employees signed an open letter to their CEO, Sundar Pichai, in early April 2018, advising that Google should withdraw from Project "Maven", reject the application of AI for military purpose, and make a declaration that Google should not be involved in warfare. Due to internal and external pressures, Google eventually decided that they would not extend the contract when it was due, and soon afterwards declared their AI Principles.

In February 2018, 26 researchers from Oxford University, Cambridge University, Open AI, Electronic Frontier Foundation, Center for a New American Security and other research institutes, issued a report on AI titled *The malicious use of artificial intelligence: Forecasting, prevention, and mitigation* [10], warning that the computing technology presents a potential threat to human rights, physical security, political security and digital security. The threats of AI to human beings are actually a threat brought about by humans. To cope with the threats, we humans must restrain our behaviors, and formulate rational ethical norms. The authorities and relevant organizations and corporations, especially those leading sci-tech giants, must promise that they should prudently research, develop, and employ AI, in conformity with rational moral ethics.

10.1.2 Ethical Status

The ethical status problems associated with the AI technology are mainly centering on: Can AIAs become ethical subjects? If they are granted the ethical status as ethical subjects, they would be meant to be liable for moral obligations, shoulder ethical responsibilities and be subject to ethical judgement. Nevertheless, this has remained an open question to date. Some scholars believe that AIAs cannot be treated as ethical subjects, for ethics are a human-only concept, helping to tell us humans apart from other creatures. We believe that humans enjoy the human status, for humans are always regarded as owners of some mental faculties and traits like pain perception,

language competence, and our abilities to identify, probe into and speculate our emotion, targets, surroundings, and ourselves [11]. AIAs are essentially artificial machines that take in and work on knowledge input by humans, representing the transfer and utilization of explicit human knowledge [12]. Deborah G. Johnson has specified the four prerequisites for an artificial ethical agent E (like a computer or robot) to become a moral agent like human [13]:

(1) There is an outward, embodied event—the agent E does something, and moves his or her body in some way;
(2) The agent E has an internal rule I that stipulates its own desires, belief and other intentions—all are the cause of the agent E to move in some way (rationally or consciously);
(3) The rule I is the direct factor giving rise to condition (1); and
(4) The physical event mentioned in condition (1) has an impact on ethical significance: When computer systems behave, their behavior has effects on other parts of the embodied world, and those embodied effects can harm or help moral patients.

Johnson has pointed out that, "… All behavior (human and non-human; voluntary and involuntary) can be explained by its causes, but only action can be explained by a set of internal mental states. We explain why an agent acted by referring to their beliefs, desires, and other intentional states." Therefore, Johnson believes that the agent E cannot make a moral agent, because it does not have any mental states and intentions to act.

Some other scholars do not totally deny the possibility for AIAs to have their moral agent status, and they believe other lives, in addition to human beings, may also enjoy their moral agent status. In this regard, Luciano Floridi and J.W. Sanders proposed their three ethical criteria at an abstract level for the artificial moral agent E [14].

(1) Interactivity: the agent E's response to stimulus by change of state;
(2) Autonomy: the agent E's ability to change state without stimulus; and
(3) Adaptability: the agent's ability to change the 'transition rules' in (1) by which state is changed in accordance with the results of (1).

They believe that artificial moral agents only require some behaviors and responses. Moreover, wildlife ecologist and environmental philosopher Aldo Leopold has pointed out that the whole ecosystem should be given the ethic status, i.e. it owns its rights to survive and develop, but it is not necessary that each living creature owns the "sacred and inviolable" rights [15].

It has been controversial whether AIAs should be treated as moral agents, but this does not prevent us from including AI into the sphere of ethics. Some scholars believe that incorporating AI into the sphere of human ethics does not mean to affirm the moral agent status of AIAs, but that human beings as moral agents are extending the criteria for non-human agents to attain ethical admittance [16]. English professor Aaron Sloman, in his book titled *The Computer Revolution in Philosophy*, points out that it is possible for AI robots to own the ability to think and perceive, so it is

rational to include AI into the sphere of ethics [17]. After incorporating AI into the sphere of ethics, it is important for us to properly position AI, and we human beings need to formulate a series of principles to regulate and restrain the development of the AI technology.

10.1.3 Responsibility Ethics

Some scholars believe that responsibility ethics is meant to conduct overall ethical analysis and research on the responsibility relationship, attribution, causes and objectives of the contemporary society, based on systematic and comprehensive ethical consideration of the behavioral purposes, consequences, methods and other factors of the responsible agents [18]. The problem of responsibility ethics for AI is an inevitable topic of great significance in the modern society. Our purpose in this regard is to better regulate the responsibility relationship in the human society, and avoid the risk and challenge of AI to the human society. The problem of responsibility ethics associated with AI highlights who is to be held accountable for the wrongdoing of AI and its damages to human beings. For example, people are now worried and arguing about: Who should be held accountable for a medical malpractice caused by malfunction of the expert diagnosis system in the healthcare industry? Who should be liable for a road accident involving a pilotless automobile? Given that AI is replacing more and more labors, who should be responsible for the growing unemployment? In case any human is injured or even killed by a machine, who should be held accountable? During the processes of design and use of AI systems, various people including designers, producers, sales persons, supervisors and end-users are involved, so we are in the face of some difficult responsibility and ethical problems in identifying responsible subjects, determining their liabilities, and formulating relevant laws and regulations associate with AI development.

In addition to deciding on the accountabilities for a specific event, responsibility ethics also stresses that we should be responsible for our future and descendants. That is to say, the AI technology we are developing shall not be detrimental to our future generations, in addition to a guarantee of no harms for the time being. The modern technology is irreversible, which means that we should deal with various AI problems with caution, and never put the future of human beings at stake. Ethics plays a role in technology, for the technological development is in want of the intervention and correction of ethics. We shall pay close attention to and reflect on the responsibility ethics problem of the AI technology, and never underestimate it. The problem of responsibility attribution of AIA will be further explored in Sect. 10.4.

10.1.4 Risks

Risks refer to the possibilities to suffer some losses, damages and/or unfavorable consequences. The development of the AI technology has brought about conveniences as well as potential risks. On the one hand, the uncertainty of technological development makes it hard to assess the results for their integration with and penetration into other disciplines, so it is still controversial whether the development of AI is a blessing or misfortune for humanity. On the other hand, we humans have limited cognitive capacity, so we would not be able to accurately evaluate and predict possible risks associated with AI and their occurrence. Chapter 6 has mentioned the possible risks of AIAs out of control. Besides, the development of AI may also bring about information security, social, environmental, economic and other risks.

The information security risks mainly include system vulnerability and data leakage. Should an AI system not be guaranteed with complete security design and unsound preventive measures, it would be lethally vulnerable to illegal invasion and manipulation, incurring other security problems. Moreover, the new wave of AI enthusiasm is driven by data, and people tend to obtain their desired services at the expense of personal data. The data collecting platform might be vulnerable to attacks, and in such a case, privacy disclosure will be widespread. The wide applications of cloud computing and data sharing have added to this risk of privacy disclosure, which may lead to other social security problems jeopardizing personal property and safety, bringing down corporate credibility and so on. If some critical national security intelligence held by AI is leaked, the consequences will be inconceivable.

The social risks are mainly manifested in social inequity and injustice involved in the development of the AI technology. As mentioned previously, automation has reduced our demand for labors, and it is no doubt that a massive unemployed population will affect social stability and security, and even lead to social turmoil. Besides, some discriminative algorithms are also against social justice. In general, algorithms are objective and definite. However, they are designed by human engineers, and different algorithm engineers will have different designs, with details showing the subjectivity of the designers. Their racism or sexism, if any, might be inexplicitly featured in design, which may be manifested by AIAs eventually. Meanwhile, some training data do not necessarily represent the actual cases, for the training data might be susceptible to interfering and contaminating adversarial samples, and even imply some value biases. For example, the AI algorithm for predicting the probability of a second crime of offenders developed by Northpointe is accused of racial prejudice, for its predictions show that the probability for black people to commit a crime is far higher than that for other racial groups. In the first International Beauty Contest Judged by AI in 2016, an algorithm that is claimed to be accurate in evaluating human aesthetics and health is regarded as racially discriminatory, for it failed to be given diversified training sets and judged that all winners were white.

The environmental risks refer to the ecological and environmental problems caused by the development of the AI technology, including resource consumption, environmental pollution, and ecological unbalance and so on. As is well-known,

water contaminants, air pollutants and solid wastes are always generated when electronic products are manufactured, and they present tremendous hazards to human health and the environment. As the AI technology is developing, AI has an increasing demand for and ever-growing dependence on computing resources, which account for more and more share in energy consumption. Meanwhile, the market and demand for AI products are rapidly expanding, leading to more and more discarded AI items, and presenting greater environmental challenges to human beings.

The economic risks mainly lie in the scenario where those nations, corporations and/or individuals that get hold of the AI expertise are in the position to have more rights to share social wealth, leaving others in an unfavorable position in social competition. If things go on like this, not only the economy and wealth will be problematic, the whole society may also see a gradually widening gap.

10.2 Codes of Ethics for AI Development

In recent years, the AI technology has experienced an explosive development, bringing about various social and ethical problems that have drawn a wider attention around the world. It is rather urgent for us to prevent and solve those ethical and moral problems associated with AI development by establishing some codes of ethics for AI development and providing some guidelines for secure AI development. In the face of the current rapid AI development, various nations, organizations and enterprises have started their commitment to research on AI ethics and social problems, and establishment of relevant ethical norms and standards.

10.2.1 Governments' and Social Organizations' Codes of Ethics for AI Development

When it comes to relevant rules and ethic norms for AI application, UK, South Korea, Japan, EU, USA, Canada, Singapore, UAE, China and some other nations have successively brought out many ethical principles, norms, guidelines and standards.

10.2.1.1 UK

In 2010, the Engineering and Physical Sciences Research Council (EPSRC) affiliated to the British Government proposed a legal and ethical norm called "Principles of Robotics" [19] featuring 5 highlights;

(1) Robots should not be designed as weapons;
(2) Robots should be designed and operated to comply with existing laws, including the law for privacy;

(3) Robots are products, which should be designed to be safe and secure;
(4) Robots are manufactured artefacts: The illusion of emotions and intent should not be used to exploit vulnerable users; and
(5) It should be possible to find out who is responsible for any robot.

At the Social Robotics and AI conference held in September 2016, the British Standards Institute (BSI) issued the first guidelines related to ethical design of robots titled *Robots and robotic devices: Guide to the ethical design and application of robots and robotic systems* (BS8611) [20], mainly covering the scope, normative references, terms and definitions, ethical risk assessment, ethical guidelines and measures, ethics-related system design recommendations, verification and validation, information for use and conclusion. This standard aims at helping those robot manufacturers to assess ethical risks associated their products on the stage of design, including robot fraud, robot-human bonding, racial and sexual discrimination and so on.

In January 2018, the Institute of Business Ethics (IBE) located in the UK issued a report titled *Business Ethics and Artificial Intelligence* [21], proposing a framework of ten basic values and principles for commercial applications of AI, including:

(1) Accuracy: Companies need to ensure that the AI systems they use produce correct, precise and reliable results. To do so, algorithms need to be free from biases and systematic errors deriving, for example, from an unfair sampling of a population, or from an estimation process that does not give accurate results;
(2) Respect of privacy: The rise of AI shall not lead to the death of privacy as described, given that machine learning technologies have brought about new ethical issues related to the respect of privacy;
(3) Transparency and openness: More openness in the use of AI algorithms and systems is needed, helping to shed some light into the dark, and avoiding our panic about monsters that could be lurking in the dark;
(4) Interpretability: As AI algorithms increase in complexity, it becomes more difficult to make sense of how they work. We need not only to identify when things go wrong, but also to determine who is responsible in the event on any damage or ethical lapse;
(5) Fairness: Fairness and justice remain paramount for ethical businesses when dealing with AI. Fairness requires businesses to consider the wider impact of AI design and developments;
(6) Integrity: In the context of AI, we should ensure that it is used only for its intended purpose, even when there is no means to enforce this; When designing or selling an AI system, it is important to ensure that the use of AI solutions by third parties is restricted to the intended purpose;
(7) Control: Much of the public skepticism around the future of AI is fueled by the fear that humans might lose control over the machines, which would then prevail and possibly wipe out humanity altogether;
(8) Impact: It is important to measure the potential impact of AI, for it helps enterprises to avoid unnecessary costs and potential risks deriving from the use of insufficient or inappropriate technologies; not all AI technologies are

applicable to a specific enterprise; just because a company can use a certain AI technology, it doesn't necessarily mean that it should;

(9) Accountability: Accountability is central to the definition of good practice in corporate governance, especially when it comes to outsourcing AI systems development, there should always be a line of responsibility for business actions to establish who has to answer for the AI system; and

(10) Learning: To maximize the potential of AI, people need to learn how it works and what are the most efficient and effective ways to use it.

The Select Committee on Artificial Intelligence of the House of Lords believes that the UK is a leader in AI development in the world. The country is home to leading AI companies, a dynamic academic research culture, a vigorous start-up ecosystem and a constellation of legal, ethical, financial and linguistic strengths. AI might present a huge opportunity for economic growth. To this end, in April 2018, the Select Committee issued *AI in the UK: ready, willing and able* [22], in an attempt to help the UK to enhance its social and economic development with AI, and protect the society against any potential hazards and risks. The document has the following five highlights:

(1) AI should be developed for the "common good and benefit of humanity";
(2) AI should operate on the principles of "intelligibility and fairness";
(3) AI shall not be used in an attempt to "diminish the data rights or privacy of individuals, families or communities";
(4) All citizens should be given the right to be educated to a level where they can "flourish mentally, emotionally and economically" alongside an AI system; and
(5) There shall be a ban on the development of any AI that has the potential to "hurt, destroy, or deceive human beings".

10.2.1.2 South Korea

In 2012, South Korea issued *South Korean Robot Ethics Charter 2012* [23], in an attempt to avoid any social problems caused by any inappropriate response of the society and legal measures to the behaviors of robots. This Charter includes the three parts of manufacturing standards, responsibilities and rights of users/owners, and responsibilities and rights of robots.

Part I Manufacturing standards

(1) A robot manufacturer must ensure the autonomy of the robots they have designed is restricted; if necessary, they must be always subject to human control.
(2) A robot manufacturer must maintain strict quality control standards, taking rational steps to make sure any risks of death or harm to users are reduced to the minimum, and the community is secure.
(3) A robot manufacturer must take measures to ensure any risks of psychological harm are reduced to minimum. In this sense, the "psychological harm" includes any possible anti-social or social behaviors, depression or anxiety, stress and especially addiction (like addition to gambling).

(4) A robot manufacturer must ensure their products bear explicit product marks, which shall be protected against any alteration.
(5) A robot must be designed with proper measure to encrypt, store and protect personal data in a secure manner.
(6) A robot must be designed with proper means to always track its behavior (online and offline).
(7) A robot must be designed to be ecologically sensitive and sustainable.

Part II Responsibilities and rights of users/owners

(1) Rights and expectations of owners and users.

   A. Owners have the rights to control their robots.
   B. Owners and users have the rights to use their robots freely, with no need to worry about any physical or psychological harm.
   C. Users have the rights to protect their detailed personal information and other sensitive information.
   D. Owners and users have the rights to require their robots to fulfill any explicitly designed tasks.

(2) Responsibilities of owners and users: The Charter recognizes that the users have the rights to employ their robots in any way they deem appropriate, as long as this application is deemed "just" and "lawful" within the scope of legal provisions, therefore:

   A. Users must not commit any illegal act with a robot.
   B. Users must not use a robot in any way that is deemed to render any physical or psychological harm to an individual.
   C. Treating a robot in a way that is deemed as deliberate and excessive insult is a light but still serious-enough crime.

Part III Responsibilities and rights of robots

(1) Responsibilities of robots:

   A. A robot may not injure a human being or, through inaction, allow a human being to come to harm.
   B. A robot must obey any orders given to it by human beings, except where such orders would conflict with the above subsection A.
   C. A robot must not deceive a human being.

(2) Rights of robots: Under South Korean Law, robots are afforded the following fundamental rights:

   A. The right to exist without fear of injury or death.
   B. The right to live an existence free from systematic abuse.

10.2.1.3 Japan

In 2014, the Ethics Committee of the Japanese Society for Artificial Intelligence (JSAI) was established in Japan, whose mission is to get hold of and report various positive and negative voices over the AI technology in the society, probe into the relationship between the AI technology and the society, and serve as a leader in Japanese discussions of codes of ethics for AI [24]. In May 2017, they worked out the updated version of JSAI ethical norm, as is *The Japanese Society for Artificial Intelligence Ethical Guidelines* [25]. These guidelines include nine recommendations:

(1) Contribution to humanity: Members of the JSAI will contribute to the peace, safety, welfare, and public interest of humanity. They will protect basic human rights and will respect cultural diversity. As specialists, members of the JSAI need to eliminate the threat to human safety whilst designing, developing, and using AI.
(2) Abidance of laws and regulations: Members of the JSAI must respect laws and regulations relating to research and development, intellectual property, as well as any other relevant contractual agreements. Members of the JSAI must not bring harm to others through violation of information or properties belonging to others. Members of the JSAI must not use AI with the intention of harming others, be it directly or indirectly.
(3) Respect for the privacy of others: Members of the JSAI will respect the privacy of others with regards to their research and development of AI. Members of the JSAI have the duty to treat personal information appropriately and in accordance with relevant laws and regulations.
(4) Fairness: Members of the JSAI will always be fair. All the JSAI members will acknowledge that the use of AI may bring about additional inequality and discrimination in society which did not exist before, and will not be biased when developing AI. Members of the JSAI will, to the best of their ability, ensure that AI is developed as a resource that can be used by humanity in a fair and equal manner.
(5) Security: As specialists, members of the JSAI shall recognize the need for AI to be safe and acknowledge their responsibility in keeping AI under control. In the development and use of AI, members of the JSAI will always pay attention to safety, controllability, and required confidentiality while ensuring that users of AI are provided appropriate and sufficient information.
(6) Act with integrity: Members of the JSAI are to acknowledge the significant impact which AI can have on society. They will therefore act with integrity and in a way that can be trusted by society. As specialists, members of the JSAI will not assert false or unclear claims and are obliged to explain the technical limitations or problems in AI systems truthfully and in a scientifically sound manner.
(7) Accountability and social responsibility: Members of the JSAI must verify the performance and resulting impact of AI technologies they have researched and developed. In the event that potential danger is identified, a warning must be

effectively communicated to all of society. Members of the JSAI will understand that their research and development can be used against their knowledge for the purposes of harming others, and will put in efforts to prevent such misuse. If misuse of AI is discovered and reported, there shall be no loss suffered by those who discover and report the misuse.

(8) Communication with society and self-development: Members of the JSAI must aim to improve and enhance society's understanding of AI. Members of the JSAI understand that there are diverse views of AI within society, and will earnestly learn from them. They will strengthen their understanding of society and maintain consistent and effective communication with them, with the aim of contributing to the overall peace and happiness of mankind. As highly-specialized professionals, members of the JSAI will always strive for self-improvement and will also support others in pursuing the same goal.

(9) Abidance of ethics guidelines by AI: AI must abide by the policies described above in the same manner as the members of the JSAI in order to become a member or a quasi-member of the society.

In July 2017, the Japanese Ministry of Internal Affairs and Communications (MIC) issued the *Draft AI R&D Guidelines for International Discussions* and submitted it to the international community for discussion [26]. Its "Sect. 10.4: AI R&D Principles" include nine principles:

(1) Principle of collaboration: Developers should pay attention to the interconnectivity and interoperability of AI systems (Principles mainly concerning mitigation of risks associated with AI systems);

(2) Principle of transparency: Developers should pay attention to the verifiability of inputs/outputs of AI systems and the explainability of their judgments;

(3) Principle of controllability: Developers should pay attention to the controllability of AI systems;

(4) Principle of safety: Developers should take it into consideration that AI systems will not harm the life, body, or property of users or third parties through actuators or other devices;

(5) Principle of security: Developers should pay attention to the security of AI systems;

(6) Principle of privacy: Developers should take it into consideration that AI systems will not infringe the privacy of users or third parties;

(7) Principle of ethics: Developers should respect human dignity and individual autonomy in R&D of AI systems (Principles mainly concerning improvements in acceptance by users, etc.);

(8) Principle of user assistance: Developers should take it into consideration that AI systems will support users and make it possible to give them opportunities for choice in appropriate manners.

(9) Principle of accountability: Developers should make efforts to fulfill their accountability to stakeholders including AI systems' users.

In July 2018, the Japanese Ministry of Internal Affairs and Communications (MIC) issued the *Draft AI Utilization Principles* [27], which highlights the following ten principles:

(1) Principle of proper utilization: Users should make efforts to utilize AI systems or AI services in a proper scope and manner, under the proper assignment of roles between humans and AI systems, or among users;
(2) Principle of data quality: Users and data providers should pay attention to the quality of data used for learning or other methods of AI systems;
(3) Principle of collaboration: AI service providers, business users, and data providers should pay attention to the collaboration of AI systems or AI services. Users should take it into consideration that risks might occur and even be amplified when AI systems are to be networked;
(4) Principle of safety: Users should take into consideration that AI systems or AI services in use will not harm the life, body, or property of users, indirect users or third parties through the actuators or other devices;
(5) Principle of security: Users and data providers should pay attention to the security of AI systems or AI services;
(6) Principle of privacy: Users and data providers should take into consideration that the utilization of AI systems or AI services will not infringe on the privacy of users' or others;
(7) Principles of human dignity and individual autonomy: Users should respect human dignity and individual autonomy in the utilization of AI systems or AI services;
(8) Principle of fairness: AI service providers, business users, and data providers should take into consideration that individuals will not be discriminated unfairly by the judgments of AI systems or AI services;
(9) Principle of transparency: AI service providers and business users should pay attention to the verifiability of inputs/outputs of AI systems or AI services and the explainability of their judgments;
(10) Principle of accountability: AI service providers and business users should make efforts to fulfill their accountability to the stakeholders including consumer users and indirect users.

In February 2019, the Cabinet Office, government of Japan established the Council for Social Principles of Human-Centric AI, aiming to probe into, better implement and share AI principles. This Council issued the *Social Principles of Human-Centric AI* [28], which highlights:

(1) The Human-Centric Principle: The utilization of AI must not infringe upon the fundamental human rights guaranteed by the Constitution and international standards. AI should be developed, utilized, and implemented in society to expand the abilities of people and allow diverse people to pursue their own well-being;
(2) The Principle of Education/Literacy: In a society premised on AI, we do not desire to create disparities or divisions between people or create those who are

socially disadvantaged, so various people should be permitted to pursue their well-being by appropriate use of AI in society;

(3) The Principle of Privacy Protection: Not all AI technologies increase the risk associated with the use of personal data, but when utilizing AI, more careful discretion may be required than the mere handling of personal data in accordance with the level of importance and sensitivity of the data;

(4) The Principle of Ensuring Security: The use of AI poses a new set of risks to security. The society should always be aware of the balance between the benefits and risks, and endeavor to improve social safety and sustainability as a whole;

(5) The Principle of Fair Competition: A fair competitive environment must be maintained in order to create new businesses and services, to maintain sustainable economic growth, and to present solutions to social challenges;

(6) The Principle of Fairness, Accountability, and Transparency: Under AI's design concept, all people are treated fairly without unjustified discrimination on the grounds of diverse backgrounds such as race, gender, nationality, age, political beliefs, religion, and so on. Appropriate explanations should be given on a case-by-case basis depending on the application of AI and each particular situation;

(7) The Principle of Innovation: To realize Society 5.0 and aim for continuous innovation that advances as people evolve together with AI development, we should transcend boundaries such as national borders, industries, academia, governments, race, gender, nationality, age, political convictions and religion. We should promote total globalization, diversification, and industry-academia-government cooperation based on a wide range of knowledge, perspectives, ideas, and so on.

10.2.1.4 EU

In May 2016, the Legal Affairs Committee of the European Parliament issued the *Draft Report: with Recommendations to the Commission on Civil Law Rules on Robotics* [29], to explore how to prevent the harm of robots to human beings by legislation. The *Draft* proposed that legislative research shall be enhanced to effectively handle risks and challenges in the future, specifying the legal subject as well as relevant rights, obligations and responsibilities, and establishing complete law and regulation system in line with the AI era. In October, the EU issued its ultimate research accomplishment: *European Civil Law Rules in Robotics* [30].

In March 2018, the European Group on Ethics in Science and New Technologies (EGE) published its statement titled *Ethical Principles and Democratic Prerequisites* [31], appealing to the authorities to start a procedure to pave the way for addressing the ethical issues and guiding future design, and establishing an internationally recognized ethical and legal framework for better production, utilization and governance of artificial intelligence, robotics and 'autonomous' systems. This statement mainly includes:

(1) Human dignity: The principle of human dignity, understood as the recognition of the inherent human state of being worthy of respect, must not be violated by 'autonomous' technologies;
(2) Autonomy: All 'autonomous' technologies must, hence, honor the human ability to choose whether, when and how to delegate decisions and actions to them;
(3) Responsibility: The principle of responsibility must be fundamental to AI research and application, which must be done so that their effects align with a plurality of fundamental human values and rights;
(4) Justice, equity, and solidarity: AI should contribute to global justice and equal access to the benefits and advantages that AI, robotics and 'autonomous' systems can bring. Discriminatory biases in data sets used to train and run AI systems should be prevented or detected, reported and neutralized at the earliest stage possible;
(5) Democracy: Key decisions on the regulation of AI development and application should be the result of democratic debate and public engagement;
(6) Rule of law and accountability: This includes protections against risks stemming from 'autonomous' systems that could infringe human rights, such as safety and privacy; timely investment shall be made in the development of robust solutions that provide a fair and clear allocation of responsibilities and efficient mechanisms of binding law;
(7) Security, safety, bodily and mental integrity: Safety and security of 'autonomous' systems materializes in three forms: (1) external safety for their environment and users, (2) reliability and internal robustness, e.g. against hacking, and (3) emotional safety with respect to human–machine interaction;
(8) Data protection and privacy: In an age of ubiquitous and massive collection of data through digital communication technologies, the right to protection of persona l information and the right to respect for privacy are crucially challenged. AI and robots must comply with data protection regulations and not collect and spread data; and the 'Autonomous' systems must not interfere with the right to private life;
(9) Sustainability: AI technology must be in line with the human responsibility to ensure the basic preconditions for life on our planet, continued prospering for mankind and preservation of a good environment for future generations. Strategies to prevent future technologies from detrimentally affecting human life and nature are to be based on policies that ensure the priority of environmental protection and sustainability;

In May 2018, the EU officially enacted the *General Data Protection Regulation (GDPR)* [32] in all member states, which has been taken as the most rigorous personal data protection regulation in the world and highlights its restriction on corporate utilization of personal data. This protection regulation has exerted great impact on big data-driven sci-tech services, and will also play a role in application of the AI technology.

In June 2018, the EU established the European Commission's High-Level Expert Group on Artificial Intelligence (AI HLEG for short), to boost its AI strategy. This

group proposed the *Draft Ethics Guidelines for Trustworthy AI* [33] in December 2018, pinpointing the ethics principles in AI and relevant values and relevant requirements for the trustworthy AI. After discussion with the European AI Alliance, negotiation with stakeholders and meeting with representatives from the member states, the group revised the draft and issued the *Ethics Guidelines for Trustworthy AI* [34]. These guidelines are meant to promote trustworthy AI, deeming that the three requirements must be met by trustworthy AI during the entire life cycle of an AI system.

(1) It should be lawful, complying with all applicable laws and regulations;
(2) It should be ethical, ensuring adherence to ethical principles and values; and
(3) It should be robust.

The *Ethics Guidelines for Trustworthy AI* consists of three chapters. Based on an approach founded on fundamental rights, Chapter I identifies the ethical principles and their correlated values that must be respected in the development, deployment and use of AI systems, underlining that we should develop, deploy and use AI systems in a way that adheres to the ethical principles of: respect for human autonomy, prevention of harm, fairness and explicability. We should admit there are some potential problems between these guidelines, and pay particular attention to situations involving more vulnerable groups such as children, the disabled and others that have historically been disadvantaged or are at risk of exclusion, and to situations which are characterized by asymmetries of power or information, such as between employers and workers, or between businesses and consumers. We must acknowledge that, while bringing substantial benefits to individuals and society, AI systems also pose certain risks and may have a negative impact, including impacts which may be difficult to anticipate, identify or measure. Chapter II lists seven requirements that AI systems should meet, to ensure the development, deployment and use of AI systems, including: (1) human agency and oversight, (2) technical robustness and safety, (3) privacy and data governance, (4) transparency, (5) diversity, non-discrimination and fairness, (6) environmental and societal well-being and (7) accountability. Chapter III provides a concrete and non-exhaustive Trustworthy AI assessment list aimed at operationalizing the key requirements set out in Chapter II. This assessment list will need to be tailored to the specific use case of the AI system.

In April 2019, the EU issued a report titled *Building Trust in Human-Centric Artificial Intelligence* [35], which presents the European AI strategies. The section of "Building Trust in Human-Centric AI" introduces the guidelines for trustworthy AI drafted by the AI high-level expert group (AI HLEG), key requirements for trustworthy AI, a piloting phase involving stakeholders on the widest scale, and efforts towards international AI ethics guidelines. The "key requirements for trustworthy AI" include seven points:

(1) Human agency and oversight: AI systems should support individuals in making better, more informed choices in accordance with their goals. They should act as enablers to a flourishing and equitable society by supporting human agency and fundamental rights, and not decrease, limit or misguide human autonomy. The overall wellbeing of the user should be central to the system's functionality.

(2) Technical robustness and safety: Trustworthy AI requires algorithms to be secure, reliable and robust enough to deal with errors or inconsistencies during all life cycle phases of the AI system, and to adequately cope with erroneous outcomes. AI systems need to be reliable, secure enough to resilient against both overt attacks and more subtle attempts to manipulate data or algorithms themselves, and they must ensure a fall-back plan in case of problems. Their decisions must be accurate, or at least correctly reflect their level of accuracy, and their outcomes should be reproducible.
(3) Privacy and data governance: Privacy and data protection must be guaranteed at all stages of the AI system's life cycle. Digital records of human behavior may allow AI systems to infer not only individuals' preferences, age and gender but also their sexual orientation, religious or political views. To allow individuals to trust the data processing, it must be ensured that they have full control over their own data, and that data concerning them will not be used to harm or discriminate against them.
(4) Transparency: The traceability of AI systems should be ensured; it is important to log and document both the decisions made by the systems, as well as the entire process (including a description of data gathering and labelling, and a description of the algorithm used) that yielded the decisions. Linked to this, explainability of the algorithmic decision-making process, adapted to the persons involved, should be provided to the extent possible. Ongoing research to develop explainability mechanisms should be pursued. In addition, explanations of the degree to which an AI system influences and shapes the organizational decision-making process, design choices of the system, as well as the rationale for deploying it, should be available.
(5) Diversity, non-discrimination and fairness: Data sets used by AI systems (both for training and operation) may suffer from the inclusion of inadvertent historic bias, incompleteness and bad governance models. The continuation of such biases could lead to (in) direct discrimination. Harm can also result from the intentional exploitation of (consumer) biases or by engaging in unfair competition. Moreover, the way in which AI systems are developed (e.g. the way in which the programming code of an algorithm is written) may also suffer from bias. Such concerns should be tackled from the beginning of the system' development.
(6) Societal and environmental well-being: For AI to be trustworthy, its impact on the environment and other sentient beings should be taken into account. Ideally, all humans, including future generations, should benefit from biodiversity and a habitable environment. Sustainability and ecological responsibility of AI systems should hence be encouraged. The same applies to AI solutions addressing areas of global concern, such as for instance the UN Sustainable Development Goals.
(7) Accountability: Mechanisms should be put in place to ensure responsibility and accountability for AI systems and their outcomes, both before and after their implementation. Auditability of AI systems is key in this regard, as the assessment of AI systems by internal and external auditors, and the availability

of such evaluation reports, strongly contributes to the trustworthiness of the technology. External auditability should especially be ensured in applications affecting fundamental rights, including safety–critical applications.

10.2.1.5 USA

In January 2017, the U.S. Public Policy Council of the Association for Computing Machinery (ACM for short) issued the *Statement on Algorithmic Transparency and Accountability* [36], which is intended to support the benefits of algorithmic decision-making while addressing concerns about potential bias and inaccuracies in algorithms. This statement includes awareness of harms, questioning and redress, accountability, explanation, data provenance, auditability, and validation and testing.

In October 2017, The Future Society in Policy Research, The Law & Society Initiative issued *The Future Society, Law & Society Initiative, Principles for the Governance of AI* [37], putting forward six principles, including:

(1) AI shall not impair, and shall advance the equality in rights, dignity, and freedom to flourish of all humans;
(2) AI shall be transparent;
(3) Manufacturers and operators of AI shall be accountable.
(4) AI's effectiveness shall be measurable in the real-world applications for which it is intended.
(5) AI systems should be designed, executed and measured by professionals with the requisite expertise.
(6) Standards and codes of practice should result from more inclusive dialogue with the civil society, and are grounded in truly broad consensus.

In October 2017, the US Information Technology Industry Council (ITI for short) issued the *AI Policy Principles* [38], which include three sections: ITI Responsibility, The Opportunity for Governments, and The Opportunity for Public–Private Partnerships. The section of "ITI Responsibility: Promoting Responsible Development and Use" mainly involves: Responsible Design and Deployment, Safety and Controllability, Robust and Representative Data, Interpretability (solutions tailored to the unique risks presented by the specific context in which a particular system operates), and Liability of AI Systems Due to Autonomy. The section of "Opportunity for Governments: Investing and Enabling the AI Ecosystem" includes: Investment in AI Research and Development, Flexible Regulatory Approach, Promoting Innovation and the Security of the Internet, Cybersecurity and Privacy, and Global Standards and Best Practices. The section of "Opportunity for Public–Private Partnerships: Promoting Lifespan Education and Diversity" covers: Democratizing Access and Creating Equality of Opportunity, supporting diversification, and broadening of access to the resources necessary for AI development and use.

OpenAI is an AI research and deployment company based in San Francisco, California established in December 2015, whose mission is to ensure that artificial general intelligence benefits all of humanity [39]. In April 2018, Open AI issued the *OpenAI Charter* [40], which underlines four points:

(1) Broadly Distributed Benefits: Using any influence obtained over AGI's deployment to ensure it is used for the benefit of all, and to avoid enabling uses of AI or Artificial General Intelligence (AGI) that harm humanity or unduly concentrate power.
(2) Long-Term Safety: Commitment to doing the research required to make AGI safe, and to driving the broad adoption of such research across the AI community.
(3) Technical Leadership: To be effective at addressing AGI's impact on society, Open AI must be on the cutting edge of AI capabilities.
(4) Cooperative Orientation: Creating a global community working together to address AGI's global challenges.

10.2.1.6 Canada

The Centre for International Governance Innovation (CIGI), Canada, believes that the privacy and security-related issues in digital economy present a major concern of public policy makers, so attention shall be given to the public concern of big data processing and relevant AI technologies. As people are afraid that these technologies might affect all jobs, people are worried about their employment and what skills they need to improve. Efforts should be made to ensure our respect for the safety and security of labors before introducing AI and automation, and at the same time to offer productivity, security and innovation benefits as required. In July 2018, GIGI, based on the 2017 "Hamburg Statement" of the Group of Twenty (G20) and the G20 Roadmap for Digitalization, recommends a G20 framework for artificial intelligence (AI) in the workplace, i.e. *Toward a G20 Framework for Artificial Intelligence in the Workplace* [41]. This framework points out two sets of seven principles for AI in the workplace, and the first seven principles relates to the collection of data in the work environment, mainly including:

(1) Right to know data is being collected, for what and from where (Employees or contractors must have the right to know what data is being collected on them by their employers, for what purpose and from what sources);
(2) Right to ensure worker data is accurate and compliant with legal rights to privacy;
(3) Principle of proportionality: The data collected should be proportional to its purpose;
(4) Principle of anonymization: Data should be anonymized where possible;
(5) Right to be informed about the use of data: Employees and contractors should be fully informed when either internal or external data (or both) has been used in a decision affecting their career;
(6) Limits to monitoring of the workplace by employers: Proportional data collection and processing should not be allowed to develop into broad-scale monitoring of employees or contractors; and
(7) Accuracy of data inputs and the "many eyes" principle: Employers should ensure the accuracy, both in detail and its intended purpose, of the data models and sources for AI. Training data and training features should be reviewed by many eyes to identify possible flaws and traps.

The second seven principles focus on AI in the workplace, mainly including:

(1) Focus on humans: Human control of AI should be mandatory and testable by regulators;
(2) Shared benefits: AI should benefit as many people as possible;
(3) AI systems should make the same recommendations for everyone with similar characteristics or qualifications;
(4) Reliability: AI should be designed within explicit operational requirements and undergo exhaustive testing to ensure that it responds safely to unanticipated situations and does not evolve in unexpected ways;
(5) Privacy and security: Big data collection and AI must comply with laws that regulate privacy and data collection, use and storage;
(6) Transparency is not necessarily nor only a question of open-source code. While in some circumstances open-source code will be helpful, what is more important are clear, complete and testable explanations of what the system is doing and why; and
(7) Accountability: People and corporations who design and deploy AI systems must be accountable for how their systems are designed and operated.

10.2.1.7 Singapore

In January 2019, the Personal Data Protection Commission (PDPC), Canada, released Asia's first model AI governance framework [42], specifying how to use AI ethically and responsibly, and proposing that we should convert moral principles into actual measures and build consumers' confidence and trust in AI to offer their personal data, to fulfill the objective of AI application. The guiding principles of *A Proposed Model Artificial Intelligence Governance Framework* [43] (hereinafter referred to as "*Model Framework*") released by the PDPC include two points: First, promote trust in AI and understanding of the use of AI technologies, ensure that AI decision-making processes are explainable, transparent and fair, and explainability is an important link that helps build trust and confidence in AI. Second, AI solutions should be human-centric. As AI is used to amplify human capabilities, the protection of the interests of human beings, including their well-being and safety, should be the primary considerations in the design, development and deployment of AI.

The *Model Framework* specifies three stakeholders, including AI solutions providers, organizations and individuals, proposes three Human-AI interaction models, including Human-in-the-loop: AI plays a supporting role; Human-on-the-loop: Humans can adjust the parameters all the time; and human-not-on-the-loop: AI systems can autonomously complete all tasks. The *Model Framework* proposed to relevant authorities some measures and recommendations to be taken to promote AI, stressing that all organizations should have internal governance structures and measures, and ensuring their forceful supervision with AI application. Organizations can consider implementing a sound system of risk management and internal controls that specifically addresses the risks involved in the deployment of the selected AI

model. Before deploying AI solutions, organizations should decide on their commercial objectives of using AI, and then weigh them against the risks of using AI in the organizations decision-making, as is guided by their corporate values. It is also desirable for organizations operating in multiple countries to consider the differences in societal norms, values and/or expectations, where possible. It is desirable for organizations to continually identify and review risks relevant to their technology solutions, mitigate those risks, and maintain a response plan should mitigation fail.

10.2.1.8 UAE

In January 2019, the UAE Smart Dubai released the *Principles of Artificial Intelligence* [44], which are intended to put forward four binding principles specifying the expected direction and roadmap of AI systems. The four principles cover ethics, security, humanity and inclusiveness. In terms of ethics, we stress AI systems should be fair, transparent, accountable and understandable. In terms of security, we underpin AI systems should be safe and secure, and should serve and protect humanity. In terms of humanity, we should plan for a future in which AI systems become increasingly intelligent and give AI systems human values and make them beneficial to society. In terms of inclusiveness, we should promote human values, freedom and dignity. AI systems should conform to international norms and standards with respect to human values and people rights and acceptable behaviors, and respect privacy. The benefits of AI should be shared in the whole society, with research and application of AI taken as a global effort to collaborate.

10.2.1.9 China

In July 2017, the State Council issued the *Development Plan for the New Generation of Artificial Intelligence* [45], which also proposes a "three steps" plan for relevant AI ethics and legal efforts. Step One: Further optimize the environment for AI development, making overall innovative applications in key fields, gathering high-caliber talents and innovation teams, and establishing preliminary ethical norms and policy regulations for AI in some fields. Step Two: Establish preliminarily some laws, regulations, ethical norms and policy systems for AI, to build up AI security evaluation, management and control capacity. Step Three: Build up a batch of globally leading AI Sci-Tech innovation and talent education bases, and more complete AI laws, regulations, ethical norms and policy systems.

In January 2018, the SAC General Team for Artificial Intelligence Standardization was established, and the *White Paper on Artificial Intelligence Standardization 2018* [46] was officially released. This *White Paper* specifies the AI security, ethics and privacy issues, and deems that to set the ethical requirements for AI technologies should be based on profound reflections and wide consensus of the society and the public on AI ethics, in compliance with some common principles.

In April 2019, the SAC General Team for Artificial Intelligence Standardization held their second plenary session and issued the *Analysis Report on AI Ethical Risks* [47], which analyzes the ethical risks of AI technologies in terms of algorithms, data, applications and so on, putting forward the guidelines for AI ethical risk evaluation indicators and industrial practices.

In May 2019, Beijing Academy of Artificial Intelligence (BAAI), together with Peking University, Tsinghua University, CAS Institute of Automation, CAS Institute of Computing Technology, other colleges and universities in the New-Generation AI Industrial Technology Innovations Strategies, and industrial alliances, released the *Beijing AI Principles* [48], that put forward 15 principles for all participating parties to observe and contribute to establishing a community with a shared future for mankind and social development with regard to AI research and development, utilization and governance [49]: Do Good, For Humanity, Be Responsible, Control Risks, Be Ethical, Be Diverse and Inclusive, Open and Share, Use Wisely and Properly, Informed-consent, Education and Training, Optimizing Employment, Harmony and Cooperation, Adaptation and Moderation, Subdivision and Implementation, and Long-term Planning.

In June 2019, the National New Generation Artificial Intelligence Governance Committee under the National New Generation AI Promotion Office issued the *New Generation AI Governance Principles—Developing Responsible AI* [50] (hereinafter referred to as "*Governance Principles*"), which are intended to better coordinate AI development and governance, ensure secure, controllable and reliable AI systems, and facilitate sustainable economic, social and economic development, so as to build up a Community of Shared Future for Mankind. The *Governance Principles* stress the theme of developing responsible AI, underlining the eight principles: harmony and human-friendliness, fairness and justice, inclusion and sharing, respect for privacy, safety and controllability, shared responsibility, open and collaboration and agile governance.

In August 2019, China AI Industry Alliance put forward a proposal for those bodies engaged in AI research, design, manufacturing, operation and service including enterprises, colleges and universities, scientific and research institutions and industrial associations [51], inviting relevant organizations to sign the *Joint Pledge on Artificial Intelligence Industry Self-Discipline* [52], which elaborates the self-disciplines in the four sections of general provisions, principles, activities and supplementary provisions, to jointly facilitate the ethic self-disciplines of the AI industry in China, and guide and standardize the conduct of practitioners.

10.2.2 International Organizations' Codes of Ethics for AI Development

10.2.2.1 IEEE

In 2016, the Institute of Electrical and Electronics Engineers (IEEE), the world's largest technical professional organization, launched a "Global Initiative on Ethics of Autonomous and Intelligent Systems", which was jointly prepared by more than 100 experts in AI, ethics, politics, law, philosophy and other fields affiliated to various committees for the IEEE Global Initiative for ethical issues of AI and autonomous systems. Its guiding philosophy is that the AI and autonomous systems are far more than meant to realize some functional objectives and address some technical problems, and we should give priority to the well-being of humanity, and make efforts to help mankind to fully benefit from their innovation potentials. In December 2017, it issued the White Paper titled *Ethically Aligned Design—A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems (second edition)* [53], aimed at encouraging sci-tech personnel to give priority to ethical issues in the AI research process. The Ethical Principles main cover the flowing points:

(1) General Principles: High-level ethical concerns that apply to all types of autonomous and intelligent system, regardless of whether they are physical robots or software systems. Among these general principles, we need to pay particular attention to: human rights, prioritizing well-being, accountability, transparency, and autonomous and intelligent system technology misuse and awareness of it.
(2) Embedding Values into Autonomous Intelligent Systems: Society has not established universal standards or guidelines for embedding human norms and values into autonomous and intelligent systems today. But as these systems are instilled with increasing autonomy in making decisions and manipulating their environment, it is essential they be designed to adopt, learn, and follow the norms and values of the community they serve. Therefore, we should identify norms for autonomous intelligent systems, implement norms in autonomous intelligent systems, and evaluate the implementation of autonomous and intelligent systems.
(3) Methodologies to Guide Ethical Research and Design: To ensure autonomous and intelligent systems are aligned to benefit humanity autonomous and intelligent system research and design must be underpinned by ethical and legal norms as well as methods. Therefore, we should attach importance to interdisciplinary education and research and corporate practices and autonomous and intelligent system, probe into research of ethics for development and testing of autonomous and intelligent system technologies, and address the current problem of lack of transparency in the AI system.
(4) Safety and Beneficence of Artificial General Intelligence (AGI) and Artificial Superintelligence (ASI): As autonomous and intelligent systems become more capable, as measured by the ability to perform with greater autonomy across

a wider variety of domains, unanticipated or unintended behavior becomes increasingly dangerous. Therefore, we should start with the general principles, pay particular attention to a progressively more complex set of ethical and technical safety issues which researchers and developers will confront in the development and deployment of increasingly capable autonomous and intelligent systems, and predict the future autonomous and intelligent system may have the capacity to impact the world on a scale not seen since the Industrial Revolution.

(5) Personal Data and Individual Access Control: Autonomous and Intelligent systems require data to fuel learning, and inform automatic decision-making. Increasingly this data is personal data, or personally identifiable information, known as PII (Personally Identifiable Information). PII is defined as any data that can be reasonably linked to an individual based on their unique physical, digital, or virtual identity. Therefore, through every digital transaction (explicit or observed) humans are generating a unique digital shadow of their physical self. To make data use ethical, we should pay attention to some points like digital personas, regional jurisdiction, agency and control, transparency and access, and symmetry and consent.

(6) Reframing Autonomous Weapons Systems (AWS): Autonomous systems designed to cause physical harm have additional ethical dimensions as compared to both traditional weapons and autonomous systems not designed to cause harm. Therefore, we need to take a number of measures to help ensure that there is meaningful human control of weapons systems, mainly including: that automated weapons have audit trails to help guarantee accountability and control; that adaptive and learning systems can explain their reasoning and decisions to human operators in transparent and understandable ways; that there be responsible human operators of autonomous systems who are clearly identifiable; that the behavior of autonomous functions should be predictable to their operators; That those creating these technologies understand the implications of their work; that professional ethical codes are developed to appropriately address the development of autonomous systems and autonomous systems intended to cause harm.

(7) Economics and Humanitarian Issues: Autonomous and intelligent systems provide unique and impactful opportunities in the humanitarian space. As disruptive technologies, they promise to upend multiple historical institutions and corresponding institutional relationships, offering opportunities to "re-intermediate" those settings with more humanitarian and equitably focused structures. From the perspective of economy, although it is undeniable to find the increase autonomous and intelligent systems and their positive applications, these technologies may be beneficial to some sectors due to more financial return brought about by technologies, and is has uneven distribution among the population of its creation and deployment regions. Likewise, although autonomous and intelligent systems can automatically perform some manual work and replace some heavy labors, the problem like unemployment of some specific groups

and total employment affected by autonomous and intelligent systems shall be addressed.

(8) Law: Our legal concerns include: legal status of autonomous and intelligent systems; governmental use of autonomous and intelligent systems; legal accountability for harm caused by autonomous and intelligent systems; and how to ensure the transparency, accountability, and verifiability of autonomous and intelligent systems.

In June 2019, IEEE released the report titled *Ethical Aspects of Autonomous and Intelligent Systems* [54], which stresses that IEEE is in favor of this kind of principle: During design, development and implementation of autonomous and intelligent systems, the following eight social consequences and system security operation statuses should be considered:

(1) Human rights: Autonomous and intelligent systems should be developed and operated in a manner that respects internationally recognized human rights.
(2) Well-being: Autonomous and intelligent systems developers should consider impact on individual and societal well-being as the central criterion in development.
(3) Data agency: Autonomous and intelligent systems developers should respect each individual's ability to maintain appropriate control over their personal data and identifying information.
(4) Effectiveness: Developers and operators should consider the effectiveness and fitness of autonomous and intelligent systems technologies for the purpose of their systems.
(5) Transparency: To the greatest extent feasible, the technical basis of the particular decisions made by an autonomous and intelligent system should be discoverable.
(6) Accountability: Autonomous and intelligent systems should be designed and operated in a manner that permits production of an unambiguous rationale for the decisions made by the system.
(7) Awareness of misuse: Designers of autonomous and intelligent systems creators should consider and guard against potential misuses and operational risks.
(8) Competence: Designers of autonomous and intelligent systems should specify and operators should possess the knowledge and skill required for safe and effective operation.

10.2.2.2 Partnership on AI (PAI)

In September 2016, a group of AI researchers from some world-renowned technology giants—Amazon, DeepMind (part of Google, now), Google, Facebook, IBM and Microsoft—officially established their partnership on artificial intelligence bearing the name of Partnership on AI (hereinafter referred to as "PAI").

PAI is an organization of many stakeholders, gathering joint efforts from scholars, researchers, civil organizations, AI technology developing and utilizing companies, and other teams committed to better understanding of the impact of AI technologies as well. PAI has set four goals [55]:

(1) Develop and share best practices in reaching into, developing and testing AI technologies;
(2) Advance wider public understanding and awareness of AI by sharing insights into AI's core technologies, potential benefits and costs;
(3) Provide an open and inclusive platform for discussion of and engagement in the future of AI; and
(4) Identify and foster aspirational efforts in AI for socially beneficial purposes.

Currently, PAI has more than 90 members including Baidu from China and BBC from the UK. PAI members are required to embrace and maintain its tenets [56]:

(1) Seek to ensure that AI technologies benefit and empower as many people as possible.
(2) Educate and listen to the public and actively engage stakeholders to seek their feedback on our focus, inform them of our work, and address their questions.
(3) Stay committed to open research and dialogue on the ethical, social, economic, and legal implications of AI.
(4) Cherish the belief that AI research and development efforts need to be actively engaged with and accountable to a broad range of stakeholders.
(5) Engage with and have representation from stakeholders in the business community to help ensure that domain-specific concerns and opportunities are understood and addressed.
(6) Work to maximize the benefits and address the potential challenges of AI technologies, by: (A) Working to protect the privacy and security of individuals; (B) Striving to understand and respect the interests of all parties that may be impacted by AI advances; (C) Working to ensure that AI research and engineering communities remain socially responsible, sensitive, and engaged directly with the potential influences of AI technologies on wider society; (D) Ensuring that AI research and technology is robust, reliable, trustworthy, and operates within secure constraints; (E) Opposing development and use of AI technologies that would violate international conventions or human rights, and promoting safeguards and technologies that do no harm.
(7) Believe that it is important for the operation of AI systems to be understandable and interpretable by people, for purposes of explaining the technology.
(8) Strive to create a culture of cooperation, trust, and openness among AI scientists and engineers to help us all better achieve these goals.

10.2.2.3 Asilomar AI Principles

In January 2017, at the Beneficial AI conference held in Asilomar, California, USA, hundreds of AI and robot experts jointly signed the *Asilomar AI Principles* [57], calling on professionals in AI around the world to observe these principles and work together to safeguard the benefits and security of human beings in this future, as elaborated in Sect. 6.5.2.

The "Asilomar AI Principles" represent as important guidelines for developing secure AI, and have won wide support from AI community and many practitioners, including physicist Hawking and Tesla CEO Musk who officially expressed their great support of these principles. By November 2019, more than 1583 AI/robot researchers and other 3447 professionals had signed as supporters. The "Asilomar AI Principles" have been called as "23 military rules" for AI development, and are quite influential due to their systematic exposition of AI ethics by the international community.

10.2.2.4 UNI Global Union

In December 2017, the UNI Global Union released Top 10 Principles for Ethical Artificial Intelligence [58] which mainly includes:

(1) Transparency: Demand That AI Systems Are Transparent. A transparent artificial intelligence system is one in which it is possible to discover how, and why, the system made a decision, or in the case of a robot, acted the way it did. In particular:

   A. They stress that open source code is neither necessary nor sufficient for transparency—clarity cannot be obfuscated by complexity.
   B. For users, transparency is important because it builds trust in, and understanding of, the system, by providing a simple way for the user to understand what the system is doing and why.
   C. For validation and certification of an AI system, transparency is important because it exposes the system's processes for scrutiny.
   D. If accidents occur, the AI will need to be transparent and accountable to an accident investigator, so the internal process that led to the accident can be understood.
   E. Workers must have the right to demand transparency in the decisions and outcomes of AI systems as well as the underlying algorithms (see principle 4 below). This includes the right to appeal decisions made by AI/algorithms, and having it reviewed by a human being.
   F. Workers must be consulted on AI systems' implementation, development and deployment.
   G. Following an accident, judges, juries, lawyers, and expert witnesses involved in the trial process require transparency and accountability to inform evidence and decision-making.

(2) Equip AI Systems With an "Ethical Black Box": Full transparency in an AI system should be facilitated by the presence of a device that can record information about said system in the form of an "ethical black box" that not only contains relevant data to ensure transparency and accountability of a system, but also includes clear data and information on the ethical considerations built

into said system. Applied to robots, the ethical black box would record all decisions, its bases for decision-making, movements, and sensory data for its robot host. The data provided by the black box could also assist robots in explaining their actions in language human users can understand, fostering better relationships and improving the user experience. The read out of the ethical black box should be uncomplicated and fast.

(3) Make AI Serve People and Planet: This includes codes of ethics for the development, application and use of AI so that throughout their entire operational process, AI systems remain compatible and increase the principles of human dignity, integrity, freedom, privacy and cultural and gender diversity, as well as with fundamental human rights.

(4) Adopt a Human-In-Command Approach: An absolute precondition is that the development of AI must be responsible, safe and useful, where machines maintain the legal status of tools, and legal persons retain control over, and responsibility for, these machines at all times. This entails that AI systems should be designed and operated to comply with existing law, including privacy. Workers should have the right to access, manage and control the data AI systems generate, given said systems' power to analyze and utilize that data (See principle 1 in "Top 10 principles for workers' data privacy and protection"). Workers must also have the 'right of explanation' when AI systems are used in human-resource procedures, such as recruitment, promotion or dismissal.

(5) Ensure a Genderless, Unbiased AI: In the design and maintenance of AI, it is vital that the system is controlled for negative or harmful human-bias, and that any bias—be it gender, race, sexual orientation, age, etc.—is identified and is not propagated by the system.

(6) Share the Benefits of AI Systems: AI technologies should benefit and empower as many people as possible. The economic prosperity created by AI should be distributed broadly and equally, to benefit all of humanity. Global as well as national policies aimed at bridging the economic, technological and social digital divide are therefore necessary.

(7) Secure a Just Transition and Ensuring Support for Fundamental Freedoms and Rights: As AI systems develop and augmented realities are formed, workers and work tasks will be displaced. To ensure a just transition, as well as sustainable future developments, it is vital that corporate policies are put in place to ensure corporate accountability in relation to this displacement, such as retraining programs and job change possibilities. Governmental measures to help displaced workers retrain and find new employment are additionally required.AI systems coupled with the wider transition to the digital economy will require that workers on all levels and in all occupations have access to social security and to continuous lifelong learning to remain employable. It is the responsibility of states and companies to find solutions that provide all workers, in all forms of work, the right to and access to both. In addition, in a world where the casualisation or individualisation of work is rising, all workers in all forms of work must have the same, strong social and fundamental rights.

(8) Establish Global Governance Mechanisms: UNI recommends the establishment of multi-stakeholder Decent Work and Ethical AI governance bodies on global and regional levels. The bodies should include AI designers, manufacturers, owners, developers, researchers, employers, lawyers, CSOs and trade unions. Whistleblowing mechanisms and monitoring procedures to ensure the transition to, and implementation of, ethical AI must be established. The bodies should be granted the competence to recommend compliance processes and procedures.
(9) Ban the Attribution of Responsibility to Robots: Robots should be designed and operated as far as is practicable to comply with existing laws, fundamental rights and freedoms, including privacy. This is linked to the question of legal responsibility. In line with Bryson et al. [59], UNI Global Union asserts that legal responsibility for a robot should be attributed to a person. Robots are not responsible parties under the law.
(10) Ban AI Arms Race: Lethal autonomous weapons, including cyber warfare, should be banned.

10.2.2.5 Initiative for AI Security Development in Shanghai

In September 2018, at the "2018 World Artificial Intelligence Conference—High Profile Talk on Artificial Intelligence", top experts, scholars and industrial elite in the field of AI security around the world gathered to discuss the opportunities for, challenges of and solutions for AI security, and launched the *Initiative for AI Security Development in Shanghai* [60]. The Initiative includes eight objectives: facing the future, people-orientedness, clear responsibilities, privacy protection, fair algorithm, transparent supervision, peaceful use and open cooperation, for AI security development.

(1) Facing the future: AI development requires collaboration between innovation and security, with security to ensure innovation and innovation to drive security. While AI security is ascertained, we should take initiative to apply AI technologies to addressing the security issues of the human society.
(2) People-orientedness: The international community should work together to work out a scientific plan for the pathway of AI development, ensuring that AI develops as expected by us to serve the well-being of humans, with key progresses like machines' autonomous evolution and self-replication under risk evaluation and security supervision.
(3) Clear responsibilities: AI development should be established with complete security responsibilities framework, requiring that innovative laws, regulations and ethical norms be provided in line with the application scenarios of AI, and specifying the mechanisms for determining and sharing AI security responsibilities.
(4) Privacy protection: A guarantee of users' data security is necessary for AI development, which shall not go at the expense of user privacy leakage, so efforts shall

be made to enhance data protection legislation, enrich technology roadmaps for AI and continuously better user privacy protection in AI applications.
(5) Fair algorithm: AI development should avoid any damage on the public due to algorithm design, specifying the intention and explainability of an algorithm and overcoming any injustice caused by algorithm design and data collection.
(6) Transparency supervision: Security risks resulting from technical black boxes shall be avoided during AI development, with auditable, traceable and deductive supervision mechanisms established to ensure a good combination of intended functions and technological realizations.
(7) Peaceful use: Cautions shall be given to applying AI technologies for military purposes, and strict risk evaluations shall be done for research, development and application of autonomous weapons, to avoid the threat of misuse of AI for military purposes to global peace and stability.
(8) Open cooperation: AI development demands collaborative efforts of various countries and parties, and efforts shall be made to establish international norms and standards for AI security development and avoid security risks caused by incompatible technologies and policies.

10.2.2.6 The "Public Voice" Coalition

"The Public Voice" coalition was established by the Electronic Privacy Information Center (EPIC), USA, in 1996 to promote public participation in decisions concerning the future of the Internet. "The Public Voice" has been paying close attention to issues from privacy and freedom of speech to consumer protection and Internet governance. In October 2018, the Voice released the *Universal Guidelines for Artificial Intelligence* [61], stating that the purpose for these universal guidelines is to inform and improve the design and use of AI. These principles are intended to maximize the benefits of AI, to minimize the risk, and to ensure the protection of human rights. These *Guidelines* should be incorporated into ethical standards, adopted in national law and international agreements, and built into the design of systems. There are 12 principles included:

(1) Right to Transparency: All individuals have the right to know the basis of an AI decision that concerns them. This includes access to the factors, the logic, and techniques that produced the outcome.
(2) Right to Human Determination: All individuals have the right to a final determination made by a person.
(3) Identification Obligation: The institution responsible for an AI system must be made known to the public.
(4) Fairness Obligation: Institutions must ensure that AI systems do not reflect unfair bias or make impermissible discriminatory decisions.
(5) Assessment and Accountability Obligation: An AI system should be deployed only after an adequate evaluation of its purpose and objectives, its benefits, as well as its risks. Institutions must be responsible for decisions made by an AI system.

(6) Accuracy, Reliability, and Validity Obligations: Institutions must ensure the accuracy, reliability, and validity of decisions.
(7) Data Quality Obligation. Institutions must establish data provenance, and assure quality and relevance for the data input into algorithms.
(8) Public Safety Obligation: Institutions must assess the public safety risks that arise from the deployment of AI systems that direct or control physical devices, and implement safety controls.
(9) Cybersecurity Obligation: Institutions must secure AI systems against cybersecurity threats.
(10) Prohibition on Secret Profiling: No institution shall establish or maintain a secret profiling system.
(11) Prohibition on Unitary Scoring: No national government shall establish or maintain a general-purpose score on its citizens or residents.
(12) Termination Obligation: An institution that has established an AI system has an affirmative obligation to terminate the system if human control of the system is no longer possible.

10.2.2.7 Organization for Economic Co-operation and Development

The Organization for Economic Co-operation and Development ("OECD" for short) Council approved the first intergovernmental standard recommendation at a ministerial conference on May 22, 2019, as advised by its Committee on Digital Economy Policy. The approved *Recommendation of the Council on Artificial Intelligence* [62] consists of two major sections: "Principles for responsible stewardship of trustworthy AI" and "National policies and international co-operation for trustworthy AI".

Section 10.1: Principles for responsible stewardship of trustworthy AI

(1) Inclusive growth, sustainable development and well-being: Stakeholders should proactively engage in responsible stewardship of trustworthy AI in pursuit of beneficial outcomes for people and the planet, such as augmenting human capabilities and enhancing creativity, advancing inclusion of underrepresented populations, reducing economic, social, gender and other inequalities, and protecting natural environments, thus invigorating inclusive growth, sustainable development and well-being.
(2) Human-centered values and fairness:

   A. AI actors should respect the rule of law, human rights and democratic values, throughout the AI system lifecycle. These include freedom, dignity and autonomy, privacy and data protection, non-discrimination and equality, diversity, fairness, social justice, and internationally recognized labor rights.
   B. To this end, AI actors should implement mechanisms and safeguards, such as capacity for human determination, that are appropriate to the context and consistent with the state of art.

(3) Transparency and explainability: AI Actors should commit to transparency and responsible disclosure regarding AI systems. To this end, they should provide meaningful information, appropriate to the context, and consistent with the state of art:

   A. To foster a general understanding of AI systems,
   B. To make stakeholders aware of their interactions with AI systems, including in the workplace,
   C. To enable those affected by an AI system to understand the outcome, and,
   D. To enable those adversely affected by an AI system to challenge its outcome based on plain and easy-to-understand information on the factors, and the logic that served as the basis for the prediction, recommendation or decision.

(4) Robustness, security and safety:

   A. AI systems should be robust, secure and safe throughout their entire lifecycle so that, in conditions of normal use, foreseeable use or misuse, or other adverse conditions, they function appropriately and do not pose unreasonable safety risk.
   B. To this end, AI actors should ensure traceability, including in relation to datasets, processes and decisions made during the AI system lifecycle, to enable analysis of the AI system's outcomes and responses to inquiry, appropriate to the context and consistent with the state of art.
   C. AI actors should, based on their roles, the context, and their ability to act, apply a systematic risk management approach to each phase of the AI system lifecycle on a continuous basis to address risks related to AI systems, including privacy, digital security, safety and bias.

(5) Accountability AI actors should be accountable for the proper functioning of AI systems and for the respect of the above principles, based on their roles, the context, and consistent with the state of art.

Section 10.2: National policies and international co-operation for trustworthy AI

(1) Investing in AI research and development:

   A. Governments should consider long-term public investment, and encourage private investment, in research and development, including interdisciplinary efforts, to spur innovation in trustworthy AI that focus on challenging technical issues and on AI-related social, legal and ethical implications and policy issues.
   B. Governments should also consider public investment and encourage private investment in open datasets that are representative and respect privacy and data protection to support an environment for AI research and development that is free of inappropriate bias and to improve interoperability and use of standards.

(2) Fostering a digital ecosystem for AI: Governments should foster the development of, and access to, a digital ecosystem for trustworthy AI. Such an ecosystem

includes in particular digital technologies and infrastructure, and mechanisms for sharing AI knowledge, as appropriate. In this regard, governments should consider promoting mechanisms, such as data trusts, to support the safe, fair, legal and ethical sharing of data.

(3) Shaping an enabling policy environment for AI

A. Governments should promote a policy environment that supports an agile transition from the research and development stage to the deployment and operation stage for trustworthy AI systems. To this effect, they should consider using experimentation to provide a controlled environment in which AI systems can be tested, and scaled-up, as appropriate.
B. Governments should review and adapt, as appropriate, their policy and regulatory frameworks and assessment mechanisms as they apply to AI systems to encourage innovation and competition for trustworthy AI.

(4) Building human capacity and preparing for labor market transformation:

A. Governments should work closely with stakeholders to prepare for the transformation of the world of work and of society. They should empower people to effectively use and interact with AI systems across the breadth of applications, including by equipping them with the necessary skills.
B. Governments should take steps, including through social dialogue, to ensure a fair transition for workers as AI is deployed, such as through training programs along the working life, support for those affected by displacement, and access to new opportunities in the labor market.
C. Governments should also work closely with stakeholders to promote the responsible use of AI at work, to enhance the safety of workers and the quality of jobs, to foster entrepreneurship and productivity, and aim to ensure that the benefits from AI are broadly and fairly shared.

(5) International co-operation for trustworthy AI:

A. Governments, including developing countries and with stakeholders, should actively co-operate to advance these principles and to progress on responsible stewardship of trustworthy AI.
B. Governments should work together in the OECD and other global and regional fora to foster the sharing of AI knowledge, as appropriate. They should encourage international, cross-sectoral and open multi-stakeholder initiatives to garner long-term expertise on AI.
C. Governments should promote the development of multi-stakeholder, consensus-driven global technical standards for interoperable and trustworthy AI.
D. Governments should also encourage the development, and their own use, of internationally comparable metrics to measure AI research, development and deployment, and gather the evidence base to assess progress in the implementation of these principles.

10.2.2.8 Group 20 (G20)

In June 2019, the G20 Ministerial Meeting on Trade and Digital Economy was held in Tsukuba, Japan, releasing the *G20 Ministerial Statement on Trade and Digital Economy* [63] which contains the third section "Human-centered AI" including four articles:

Article 17: Recognizing the efforts undertaken so far by all stakeholders in their respective roles including governments, international organizations, academia, civil society and the private sector, and mindful of how technology impacts society, the G20 endeavors to provide an enabling environment for human-centered AI that promotes innovation and investment, with a particular focus on digital entrepreneurship, research and development, scaling up of startups in this area, and adoption of AI by MSMEs which face disproportionally higher costs to adopt AI.

Article 18: We recognize that AI technologies can help promote inclusive economic growth, bring great benefits to society, and empower individuals. The responsible development and use of AI can be a driving force to help advance the SDGs and to realize a sustainable and inclusive society, mitigating risks to wider societal values. The benefits brought by the responsible use of AI can improve the work environment and quality of life, and create potential for realizing a human-centered future society with opportunities for everyone, including women and girls as well as vulnerable groups.

Article 19: At the same time, we also recognize that AI, like other emerging technologies, may present societal challenges, including the transitions in the labor market, privacy, security, ethical issues, new digital divides and the need for AI capacity building. To foster public trust and confidence in AI technologies and fully realize their potential, we are committed to a human-centered approach to AI, guided by the G20 AI Principles drawn from the OECD Recommendation on AI, which are attached in Annex and are non-binding. This Annex includes the following principles of "inclusive growth, sustainable development and well-being", "human-centered values and fairness", "transparency and explainability", "robustness, security and safety" and "accountability". The Annex also offers guidance for consideration by policy makers with the purpose of maximizing and sharing the benefits from AI, while minimizing the risks and concerns, with special attention to international cooperation and inclusion of developing countries and underrepresented populations.

Article 20: In pursuing human-centered AI, G20 members recognize the need to continue to promote the protection of privacy and personal data consistent with applicable frameworks. The G20 also recognizes the need to promote AI capacity building and skills development. We will each continue to strive for international cooperation and endeavor to work together with appropriate fora in areas such as research and development, policy development and information sharing through the G20 Repository of Digital Policies and other open and collaborative efforts.

10.2.3 Corporations, Colleges and Universities' Codes of Ethics for AI Development

10.2.3.1 Microsoft

In 2016, Microsoft proposed its six principles for AI development [64]: fair, reliable and safe, private and secure, inclusive, transparent, and accountable. Microsoft president Brad Smith said that these six core principles are intended to ensure AI is utilized in an ethical manner. Microsoft believes that to design reliable AI, we must take those solutions that uphold these principles, and guide AI with social ethics while making the computers more powerful. Moreover, Microsoft also established a committee to develop and probe into AI and ethical standards, to ensure the implementation of these six core principles, and help Microsoft to address ethical and social problems during the processes to research and develop products and to communicate with customers.

In July 2018, Microsoft and China Development Research Foundation jointly released the report titled *Future Cornerstone—AI Social Roles and Ethics* at their release meeting [65] that gathers their thoughts of and insights into the field of AI social roles and ethics, aimed at advancing the sound growth of AI to better serve the development of human beings. The report calls on all governments, industrial circles, researchers, civil organizations and other stakeholders to start up wide dialogues and sustained cooperation on AI, and speed up ethical norms and public policy rules for AI development and application.

10.2.3.2 IBM

In January 2017, IBM CEO Ginni Rometty proposed the three principles, i.e. purpose, transparency and skills [66], for AI at a group meeting of the World Economic Forum:

(1) Purpose: The purpose of AI and cognitive systems developed and applied by the IBM Company is to augment human intelligence. Our technology, products, services and policies will be designed to enhance and extend human capability, expertise and potential.
(2) Transparency: The IBM Company will make clear: When and for what purposes AI is being applied in the cognitive solutions we develop and deploy. Moreover, the data for training AI are also transparent.
(3) Skills: The economic and societal benefits of this new era will not be realized if the human side of the equation is not supported, to help students, workers and citizens acquire the skills and knowledge to engage safely, securely and effectively.

In May 2018, IBM released *IBM's Principles for Trust and Transparency* [67] that includes three highlights:

(1) The purpose of AI is to augment human intelligence. The purpose of AI and cognitive systems developed and applied by IBM is to augment (not replace) human intelligence. We believe AI should make ALL of us better at our jobs, and that the benefits of the AI era should touch the many, not just the elite few.
(2) Data and insights belong to their creator. Client data and the insights produced on IBM's cloud or from IBM's AI are owned by IBM's clients. We believe that government data policies should be fair and equitable and prioritize openness. In terms of data and insight ownership, clients are not required to relinquish rights to their data—nor the insights derived from that data—to have the benefits of IBM's solutions and services. In terms of data privacy, IBM is fully committed to protecting the privacy of our clients' data, which is fundamental in a data-driven society. In terms of data security, IBM is devoting our powerful engines of innovation to create tools to protect our clients, their data and global trade from cyber threats. We are also convening a broader discussion on balancing security, privacy and freedom. In terms of government access to data, IBM has not provided client data to any government agency under any surveillance program involving bulk collection of content or metadata. In terms of cross-border data flows, IBM views the free movement of data across borders as essential to twenty-first century commerce.
(3) New technology, including AI systems, must be transparent and explainable. For the public to trust AI, it must be transparent. Technology companies must be clear about who trains their AI systems, what data was used in that training and, most importantly, what went into their algorithm's recommendations. If we are to use AI to help make important decisions, it must be explainable.

In September 2018, IBM publicly released its *Everyday Ethics for Artificial Intelligence* [68], putting forward five focal points regarding ethics.

(1) Accountability: AI designers and developers are responsible for considering AI design, development, decision processes, and outcomes;
(2) Value Alignment: AI should be designed to align with the norms and values of your user group in mind;
(3) Explainability: AI should be designed for humans to easily perceive, detect, and understand its decision process;
(4) Fairness: AI must be designed to minimize bias and promote inclusive representation; and
(5) User data rights: AI must be designed to protect user data and preserve the user's power over access and uses.

10.2.3.3 Intel

In October 2017, Intel believed that the public policies mainly drive AI to address social problems and facilitate economic progress. Therefore, the public policies must support industrial efforts, allowing AI to bring about economic benefits, and addressing the problems of human concern. As AI innovations have just started,

it is now of paramount importance to shape sound public policy environment, and help the society to trust AI through the supervision of regulators. Through industrial cooperation with Partnership on AI and other bodies, public policies should work to reduce or eliminate any barriers between AI and its huge potentials, which is critical to probing into those problems and working out concrete solutions. To this end, Intel proposed its public policy principles for AI titled *Artificial Intelligence: The Public Policy Opportunity* [69], so as to provide some concrete recommendations for governmental implementation that include five main points:

(1) Foster innovation and open development: To better understand the impact of AI and explore the broad diversity of AI implementations, public policy should encourage investment in AI R&D. Governments should support the controlled testing of AI systems to help industry, academia, and other stakeholders improve the technology.
(2) Create new human employment opportunities and protect people's welfare: AI will change the way people work. Public policy in support of adding skills to the workforce and promoting employment across different sectors should enhance employment opportunities while also protecting people's welfare.
(3) Liberate data responsibly: AI is powered by access to data. Machine learning algorithms improve by analyzing more data over time; data access is imperative to achieve more enhanced AI model development and training. Removing barriers to the access of data will help machine learning and deep learning reach their full potential.
(4) Rethink privacy: Privacy approaches like The Fair Information Practice Principles and Privacy by Design have withstood the test of time and the evolution of new technology. But with innovation, we have had to "rethink" how we apply these models to new technology.
(5) Require accountability for ethical design and implementation: The social implications of computing have grown and will continue to expand as more people have access to implementations of AI. Public policy should work to identify and mitigate discrimination caused by the use of AI and encourage designing in protections against these harms.

10.2.3.4 Google

In March 2018, Google reportedly provided technical support for a military project called "Maven" under the US Department of Defense, helping military drones to attain higher precision strikes. The minute the news was made public, more than 3,000 Google employees signed a joint appeal asking Google to withdraw from the project. Under great pressure from the public, Google decided that it would not extend the contract when it was due in 2019. In June 2018, Google a statement on its seven guiding principles and four bottom lines for AI application [70]. The seven core principles (objectives) are:

(1) Be socially beneficial;

(2) Avoid creating or reinforcing unfair bias;
(3) Be built and tested for safety;
(4) Be accountable to people;
(5) Incorporate privacy design principles;
(6) Uphold high standards of scientific excellence; and
(7) Be made available for uses that accord with these principles.

The four bottom lines refer to the following areas for Google not to design or deploy AI, in addition to the above objectives:

(1) Technologies that cause or are likely to cause overall harm. Where there is a material risk of harm, we will proceed only where we believe that the benefits substantially outweigh the risks, and will incorporate appropriate safety constraints.
(2) Weapons or other technologies whose principal purpose or implementation is to cause or directly facilitate injury to people.
(3) Technologies that gather or use information for surveillance violating internationally accepted norms.
(4) Technologies whose purpose contravenes widely accepted principles of international law and human rights.

As a leader in the field of AI, Google owns some top AI labs, carrying out many important academic research programs. When it comes to "Do not be evil", Google explicitly states that in the case any science and technology application goes against the abovementioned principles, Google will reserve its rights to prevent or stop their science and technology support.

10.2.3.5 Baidu

In May 2018, Baidu president Robin Li expounded his *Four Principles for AI Ethics* [71], at a "high level dialogue on AI" of China International Big Data Industry Expo 2018. He held that as AI technologies are currently exerting constant impact on and facilitating innovations and processes in the human society, it is a matter of course for the ethical and moral issues concerning our well-being and health, freedom and equality behind these technologies to become an important topic in the current smart era. Robin Lee said, "In the AI era, in addition to rapid technological progress, we are increasingly in want of new rules, values and ethics. All new AI products and technologies require a common idea and rule for us all to observe." Lee proposed, "AI principles shall include the following four binding principles, including: the highest principle for AI is safe and controllable; the vision for AI innovation is to promote human equality and obtain technology and competence; the values for AI existence are to teach us to learn, allow us to grow, and not to surpass or replace human beings; the ultimate ideal for AI is to bring more freedom and possibilities to mankind."

10.2.3.6 SAP

In May 2018, SAP[1] established an AI steering committee in charge of probing into ethical principles for AI and addressing relevant AI problems, aimed at continuously developing software for humans to make use of intelligence potentials and ensure the AI product portfolio of SAP is an integrated whole. SAP has devised *SAP's Guiding Principles for Artificial Intelligence* [72], to guide AI software development and deployment, facilitate social progress and better our life. These principles involve seven aspects:

(1) Be driven by our values: We recognize that, like with any technology, there is scope for AI to be used in ways that are not aligned with these guiding principles and the operational guidelines we are developing. In developing AI software, we will remain true to our *Human Rights Commitment Statement*, the *UN Guiding Principles on Business and Human Rights*, laws, and widely accepted international norms. Wherever necessary, our AI Ethics Steering Committee will serve to advise our teams on how specific use-cases are affected by these guiding principles. Where there is a conflict with our principles, we will endeavor to prevent the inappropriate use of our technology.
(2) Design for people: We strive to create AI software systems that are inclusive and that seek to empower and augment the talents of our diverse usership. By providing human-centered user experiences through augmentative and intuitive technologies, we leverage AI to support people in maximizing their potential.
(3) Enable businesses beyond bias: Bias can negatively impact AI software and, in turn, individuals and our customers. This is particularly the case when there is a risk of causing discrimination or of unjustly impacting underrepresented groups. We, therefore, require our technical teams to gain a deep understanding of the business problems they are trying to solve, and the data quality this demands. We seek to increase the diversity and interdisciplinarity of our teams, and we are investigating new technical methods for mitigating biases. We are also deeply committed to supporting our customers in building even more diverse businesses by leveraging AI to build products that help move business beyond bias.
(4) Strive for transparency and integrity in all that we do: Our systems are held to specific standards in accordance with their level of technical ability and intended usage. Their input, capabilities, intended purpose, and limitations will be communicated clearly to our customers, and we provide means for oversight and control by customers and users. They are, and will always remain, in control of the deployment of our products. We actively support industry collaboration and will conduct research to further system transparency.
(5) Uphold quality and safety standards: As with any of our products, our AI software is subject to our quality assurance process, which we continuously adapt when necessary. Our AI software undergoes thorough testing under real-world

[1] SAP: a world-famous enterprise management solution provider. It has become a byword of ERP and a world-beating enterprise management software provider.◦

scenarios to firmly validate they are fit for purpose and that the product specifications are met. We work closely with our customers and users to uphold and further improve our systems' quality, safety, reliability, and security.

(6) Place data protection and privacy at our core: Data protection and privacy are a corporate requirement and at the core of every product and service. We communicate clearly how, why, where, and when customer and anonymized user data is used in our AI software.

(7) Engage with the wider societal challenges of AI: While we have control, to a large extent, over the preceding areas, there are numerous emerging challenges that require a much broader discourse across industries, disciplines, borders, and cultural, philosophical, and religious traditions.

10.2.3.7 SONY

SONY has put forward that its objective is to contribute to peace and sustainable social development through AI application, and deliver excitation and amazement to the world. It has proposed that, as a diversified enterprise, SONY should operate under the tenet that it should "fill the world with emotion, through the power of creativity and technology". To participate in correct research, development and utilization of AI, SONY released *Sony Group AI Ethics Guidelines* [73] in September 2018, to ensure and enhance its dialogue with various stakeholders. These principles mainly include seven points:

(1) Supporting creative life styles and building a better society: Through advancing its AI-related R&D and promoting the utilization of AI in a manner harmonized with society, Sony aims to support the exploration of the potential for each individual to empower their lives, and to contribute to enrichment of our culture and push our civilization forward by providing novel and creative types of *kando*.

(2) Stakeholder engagement: In order to solve the challenges arising from use of AI while striving for better AI utilization, Sony will seriously consider the interests and concerns of various stakeholders including its customers and creators, and proactively advance a dialogue with related industries, organizations, academic communities and more.

(3) Provision of trusted products and services: Sony understands the need for safety when dealing with products and services utilizing AI and will continue to respond to security risks such as unauthorized access.

(4) Privacy protection: Sony, in compliance with laws and regulations as well as applicable internal rules and policies, seeks to enhance the security and protection of customers' personal data acquired via products and services utilizing AI, and build an environment where said personal data is processed in ways that respect the intention and trust of customers.

(5) Respect for fairness: In its utilization of AI, Sony will respect diversity and human rights of its customers and other stakeholders without any discrimination while striving to contribute to the resolution of social problems through its activities in its own and related industries.

(6) Pursuit of transparency: During the planning and design stages for its products and services that utilize AI, Sony will strive to introduce methods of capturing the reasoning behind the decisions made by AI utilized in said products and services.
(7) The evolution of AI and ongoing education: People's lives have continuously changed with the advance in technology across history. Sony will be cognizant of the effects and impact of products and services that utilize AI on society and will proactively work to contribute to developing AI to create a better society and foster human talent capable of shaping our collective bright future through R&D and/or utilization of AI.

10.2.3.8 Tencent

In September 2018, Pony Ma, Chairman of the Board of Directors and CEO of Tencent proposed the "ARCC" philosophy for AI at the World Conference on Artificial Intelligence 2018 in Shanghai, i.e. AI in the future should be "available", "reliable", "comprehensible" and "controllable". The *"ARCC" Initiative* [74] is aimed at laying a foundation for further developing an AI ethical framework. The "Available" principle requires that people, as many as possible, instead of the privileged few, have access to and utilize AI; the "Reliable" principle is intended to make AIAs like robots safe, reliable and capable of withstanding cyber-attack and other incidents; the "Comprehensible" principle is meant to make AI understandable, with transparent algorithm and the like; the "Controllable" principle is to ensure humans to play the predominant role as a controller.

In April 2017, the Institute of Science and Development of China Academy of Sciences and Tencent Research Institute jointly proposed the "Six Principles for AI Development" [75] at the 2017 AI Seminar: Technology, Ethics and Law, covering the six levels of freedom, justice, well-being, ethics, security and responsibility, and 19 rules altogether. In terms of ethics, it was proposed that ethical design be embedded in the system, to make ethical machines and ensure diversified and inclusive ethical norms, ensure AI system in compliance with ethical norms for the human society, and constantly check and ensure the ethical design in line with human norms.

10.2.3.9 GE Healthcare

In October 2018, Keith Bigelow, General Manager of analytics at GE Healthcare, published an article on its official website [76], stating that AI presents a unique challenge in ensuring safe, fair and effective usage. Outputs generated by algorithms are only as good as the data put in. It can be difficult to demonstrate how an algorithm arrived at its decision. Now, AI has the ability to make sense of the massive volumes of healthcare data. It could take away mundane and repetitive tasks and enable precision health. To enable the benefits and mitigate the potential harms requires not just

updated guidelines but also new approaches to technology development and evaluation adapted to AI. GE Healthcare officially becomes one of those companies. We are publishing the following AI principles, which we will apply to improve healthcare quality, cost, access and the patient experience.

AI Systems exist to augment human intelligence, and must:

(1) Be designed for the benefit, safety and privacy of the patient;
(2) Be a trusted steward of the data and insights;
(3) Be transparent and produce explainable outputs;
(4) Guard against creating or reinforcing bias; and
(5) Optimize the safe development, production and compliance of therapeutics and healthcare solutions to deliver Precision Health.

10.2.3.10 Unity Technologies Inc.

Unity Technologies Inc. is a video game software developer, reputed for its Unity products. Unity is a versatile game engine, used to create video game and other applications. In November 2018, Unity made public the *Introducing Unity's Guiding Principles for Ethical AI* [77] that includes the following six guiding principles for AI:

(1) Be unbiased: Design AI tools to complement the human experience in a positive way. Consider all types of human experiences in this pursuit. Diversity of perspective will lead to AI complementing experiences for everybody, as opposed to a select few.
(2) Be accountable: Consider the potential negative consequences of the AI tools we build. Anticipate what might cause potential direct or indirect harm and engineer to avoid and minimize these problems.
(3) Be fair: Do not knowingly develop AI tools and experiences that interfere with normal, functioning democratic systems of government. This means saying no to product development aimed at the suppression of human rights, as defined by the Universal Declaration of Human Rights, such as the right to free expression.
(4) Be responsible: Develop products responsibly and do not take advantage of your products' users by manipulating them through AI's vastly more predictive capabilities derived from user data.
(5) Be honest: Trust the users of the technology to understand the product's purpose so they can make informed decisions about whether to use the product. Be clear and be transparent.
(6) Be trustworthy: Guard the AI derived data as if it were handed to you by your customer directly in trust to only be used as directed under the other principles found in this guide.

10.2.3.11 Colleges and Universities

By the end of 2018, University of Montreal, Canada, made public *The Montreal Declaration for Responsible Development of Artificial Intelligence*, and called for online signatures [78]. This Declaration has three main objectives: First, develop an ethical framework for the development and deployment of AI; Second, guide the digital transition so everyone benefits from this technological revolution; Third, open a national and international forum for discussion to collectively achieve equitable, inclusive, and ecologically sustainable AI development. The Declaration includes 10 parts:

(1) "Well-being" principle: The development and use of artificial intelligence (AI) systems must permit the growth of the well-being of all sentient beings.
(2) "Respect for autonomy" principle: AI systems must be developed and used while respecting people's autonomy, and with the goal of increasing people's control over their lives and their surroundings.
(3) "Protection of privacy and intimacy" principle: Privacy and intimacy must be protected from AI systems intrusion and data acquisition and archiving systems (DAAS).
(4) "Solidarity" principle: The development of AI systems must be compatible with maintaining the bonds of solidarity among people and generations.
(5) "Democratic participation" principle: AI systems must meet intelligibility, justifiability, and accessibility criteria, and must be subjected to democratic scrutiny, debate, and control.
(6) "Equity" principle: The development and use of AI systems must contribute to the creation of a just and equitable society.
(7) "Diversity inclusion" principle: The development and use of AI systems must be compatible with maintaining social and cultural diversity and must not restrict the scope of lifestyle choices or personal experiences.
(8) "Prudence" principle: Every person involved in AI development must exercise caution by anticipating, as far as possible, the adverse consequences of AI systems use and by taking the appropriate measures to avoid them.
(9) "Responsibility" principle: The development and use of AI systems must not contribute to lessening the responsibility of human beings when decisions must be made.
(10) "Sustainable development" principle: The development and use of AI systems must be carried out so as to ensure a strong environmental sustainability of the planet.

10.3 Artificial Intelligence Actants (AIAs) Making Ethical Decisions

Section 10.2 of this Chapter has mentioned that currently various efforts have been made for secure development of AI technologies including relevant norms and systems made by governments, international organizations and enterprises to provide specifications for AI R&D, design and utilization. As artificial intelligence actants (AIAs) are getting more multi-functional and sophisticated, the interaction and communication between humans and machines have been further enhanced. In addition to human beings in compliance with relevant codes of ethics for secure development of AI technologies during R&D, design, utilization and other stages, it is also one of the directions for secure development of AI in the future to foster AIAs to make decisions in line with the codes of ethics for human beings and ensure harmonious coexistence between humans and machines. Then, we are probing into how to empower AIAs to make ethical decisions.

10.3.1 AIAs' Ability to Make Ethical Decisions

Some scholars believe that in human–machine interaction, AIAs can carry out their tasks through autonomous cognition, decision-making and action, which makes them to attain some features showing their certain "subjectivity" and become a kind of entities between humans and common objects [79]. For now, AIAs are able to make ethical judgements due to some algorithms or procedures internally embedded by designers. The "subjectivity" that they are demonstrating does not come from their own will, but reflects their implementation of the will of designers. Regarding AIAs' ability to make ethical decisions, Professor James H. Moor, one of the founders of computing ethics, once made four inspiring divisions [80].

(1) Ethical Impact Agents: Agents whose actions have valuable and ethical consequences without any ethical design.
(2) Implicit Ethical Agents: Agents that have ethical safety and security considerations built into (i.e. implicit in) their design. The ethical norms are built in the agents by designers, allowing agents to make some ethical decisions in some scenarios. The ethical decisions demonstrated by the agents are still determined by the designers, and they have just received ethical norms from the designers.
(3) Explicit Ethical Agents: Agents that can identify and process ethical information about a variety of situations and make sensitive determinations about what should be done. These agents are capable of learning actively, and can work out ethical resolutions in line with their changing environment.
(4) Full Ethical Agents: Agents that make ethical decisions in accordance with various situations like adult humans of consciousness, intentionality and free will.

The agents mentioned by James H. Moor refer to those AI systems that can obtain perception from the environment and carry out actions, so they include those AIAs of mobility. Based on James H. Moor's divisions, most AIAs today are Ethical Impact Agents, for they have not been built in with ethical norms or security mechanisms for the time being, and cannot understand or learn ethical norms, or make any ethical decisions. Only a very few AIAs can be taken as Implicit Ethical Agents, for they can make decisions in compliance with codes of ethics for human beings when it comes to ethical issues as empowered by safety or security software and/or hardware modules embed by designers. AIAs have designed reflexes to ensure AIAs to make progress in the direction of ethical receivers. Explicit Ethical Agents currently are the targets for theoretical exploration of ethics for machines, and Full Ethical Agents only exist in science fictions. Currently, the AIAs have not yet attained their ethical perfection; with scientific and technological progresses, AIAs are becoming more and more autonomous, so it goes without saying that ethical design for AIAs are of paramount importance.

10.3.2 Design for AIAs to Make Ethical Decisions

It has been long for researchers to stay committed to probe into how to build moral theories and ethical principles into robots and other AI systems, or help them to develop their ability to judge ethical issues through learning and growth, in an attempt to make them to become artificial moral agents (AMAs). Any AIAs that can autonomously make decisions in line with human ethics can be taken as AMAs. Among relevant AMA building research efforts, the most representative results would be a book titled *Moral Machines: Teaching Robots Right from Wrong* [81] co-authored by Wendell Wallach and Colin Allen, who proposed three approaches to AMA design: Top-Down, Bottom-Up, and Mixture Approaches. The top-down approaches make use of a specific ethical theory, and analyze its computing demand. On this basis, we guide to design the algorithm and subsystem to carry out that theory. The top-down approaches, however, are not flawless, for different people have different views on ethical principles, so it is difficult to find unanimously agreed ethical principles for AIA design. The bottom-up approaches are different in that they work to create an environment for AIAs to explore their own actions to learn, and be awarded if they have made ethical actions worthy of praise.[2] This is a rather evolving concept, for training AIAs with ethical values through cultivation and learning is just like the growth of humans. Nevertheless, it should be noted that things would lead to bad results out of expectation of human beings, for even humans might go vicious during our growth, let alone AIAs. They will see trials and errors during their evolution. Therefore, Bottom-Up Approaches also have some limitations.

The two Top-Down and Bottom-Up Approaches, deemed by Wendell Wallach and Colin Allen, have their own pros and cons. For complicated tasks, engineers

[2] See *Moral Machines: Teaching Robots Right from Wrong*: Page 69.

usually get started with Top-Down analysis, to guide bottom-up module assembly.[3] Therefore, they proposed the "Mixture" approaches that mix up the top-down deductive approach and the bottom-up inductive approach. The Mixture approaches build ethical principles into AIAs and at the same time cultivate and train AIAs with ethics in a mutually reinforcing, supplementary manner. The three theoretical approaches to building AMAs proposed by them have caused wide concern. In December 2008, Patrick Lin et al. submitted a report titled *Autonomous Military Robotics: Risk, Ethics, and Design* to the US Department of Navy, making use of the three theoretical approaches to military robotics design. They also believe that the two Top-Down and Bottom-Up Approaches have their own short-comings, so the Mixture approaches should be employed to build military robotics [82].

10.3.3 Consideration for AIAs to Make Ethical Decisions

To enhance AIA autonomy, engineers have raised wider requirements for its security and reliability, and meanwhile explicit representation is needed to make ethical considerations when it comes to ethics. Thanks to different ethical conduct evaluation criteria, those acts that might be taken as ethical might be regarded as unethical by others. It is a great challenge to pursue clear-cut ethical criteria for AIAs. Even Alan Turing, a founding father of computer science and artificial intelligence, was confronted with the same difficulty in his attempt to define when a computer can think. Therefore, he designed an "imitation game" to conduct test, i.e. the renowned "Turing Test" [83], which has been elaborated in Chap. 9. Wendell Wallach, Colin Allen and so on suggested that the "Moral Turing Test (MTT)" be employed to evaluate AMAs.

The "Moral Turing Test" tries to make conversational contents as ethical issues. If people, more than a given proportion, cannot tell humans from AIAs, these AIAs pass the test, and could be taken as ethical actants. Some scholars believe that MTT is also intrinsically defective. For one, making imitation as evaluation criteria for evaluating ethical performance may have some adverse impacts. For example, some deceptive behaviors also meet the conditions for imitation. If they are taken as the representation of actants, or replacing moral conducts, this would go against the ethical design for robots. Second, the strong MTT behaviorism makes it insufficient to evaluate the moral competence of robots [84]. Therefore, if the "Moral Turing Test" is employed to conduct evaluation, things would go paradoxical, i.e. even if an AIA passes the text, it would not be taken as capable of ethical reasoning like humans. To make AIAs to show ethical behaviors in line with human expectations and values, we need to accurately ascertain their ethical reasoning and decision-making processes on the stage of AIA ethical design. Given the present science and technology, our considerations for AIAs' making ethical decisions may require that we humans should explore and think further.

[3] See *Moral Machines: Teaching Robots Right from Wrong*: Page 102.

10.4 Accountability for AIAs

As pilotless automobiles, medical robots, dining hall robots and other AI products are constantly emerging, some accidents and infringement acts incurred by AI products have aroused wide concerns. AIA accountability has also become a hot topic. Are AIAs themselves, In the face of these accidents and infringement acts, capable of shouldering responsibility? In case an AIA's conduct results in casualties, who should be held accountable? Its designers, users, or itself? We would like to probe into the issue of accountability for AIAs in this section.

10.4.1 Subject of Liability

To determine the liability for any accident, we need to ascertain the subject of liability. If AIAs are ascertained as subjects like humans, we should allow them to enjoy relevant rights and shoulder relevant liabilities. Otherwise, they cannot be given relevant rights and liabilities. Section 10.1.3 tells us that our discussion on the subject of liability is an important topic concerning AI liability ethics, and we have made a lot of efforts in addressing AI liability ethics-related problems, including making various norms, regulations, guidelines, principles and so on.

The Ministry of Internal Affairs and Communications (MIC), Japan issued the *Draft AI R&D Guidelines for International Discussions* [26] and the *Draft AI Utilization Principles* [27], respectively highlighting that, in terms of "obligation principles", developers should make efforts to fulfill their obligations for stakeholders including AI system users, and AI services providers and enterprise users should work hard to fulfill their obligations for stakeholders including consumers and indirect users. The *AI Policy Principles* [38] issued by the Information Technology Industry Council (ITI), USA, states their own obligations, i.e. promoting responsible design and deployment, and ensuring AI system security, controllability and explainability. The Personal Data Protection Commission (PDPC), Singapore, issued *A Proposed Model Artificial Intelligence Governance Framework* [43], stressing that various organizations should have proper internal governance structure and measures, specifying their roles and obligations for AI ethical deployment, and relevant risk management and internal control.

The document titled *ETHICALLY ALIGNED DESIGN: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems* (second edition) [53] issued by IEEE specifies the basic principle that AI designers and operators are liable and can be held accountable in its General Principles, and proposing that complete legal framework for AI accountability should be established, to increase the transparency of individual rights, reduce algorithm discrimination, and solve other AI product deficiency problems caused by the value inclinations of designers.

The *Top 10 Principles For Ethical Artificial Intelligence* [58]issued by the UNI Global Union proposes that robots should be designed and operated as far as possible

in line with the principles of existing laws, basic rights and freedoms (including privacy), as is an issue of legal liability. In 2011, the UNI Global Union stated that the legal liability of robots should be attributed to humans. It is specified by laws that robots should not be held accountable.

The *G20 Ministerial Statement on Trade and Digital Economy* stresses that developers and users have their respective obligations. They should develop and utilize AI in a responsible manner, making it a major driving force to facilitate sustainable development, build a sustainable, inclusive society, and reduce social risks. It is beneficial to make use of the benefits of AI responsibly in that it helps improve working environment and social quality, and create potentials for human-oriented future society, providing opportunities for women, children and other vulnerable groups.

Moreover, other ethical principled listed in Sect. 10.2 tells us that all walks of life have been involved in formulating AI ethical specifications and norms, but they unanimously affirmed that humans, instead of AIAs, are the subjects of liability in the AI system, but not AIAs themselves. That is to say, any AIAs should have relevant persons in charge that includes but is not limited to designers and users. For the time being, AIAs do not hold self-will or complete rationality like humans, and are not held accountable for any infringement. AIA planning, design, manufacturing, utilization and other processes involve a great many departments and relevant personnel including designers, manufacturers, governmental administrative departments, users and so on. These participants are not necessary the same with regard to obligation assignment; relevant AIA secure operation liabilities should be determined in accordance with their roles, environment and existing technologies, to fulfill relevant obligations for secure operation of AIAs.

10.4.2 Assignment of Obligations

To establish an explicit impartial binding accountability mechanism for assignment of obligations is the most reliable solution for any infringement on the rights of humans possibly caused by AIAs. AIA infringement cases and conventional infringement cases are incurring different forms of damages and risks, so there are different views on assignment of obligations under different circumstances. If AIAs are taken as tools, their users should be held accountable for the most part, for the damages or infringement involving AIAs are mainly due to improper operation or deliberation of the users. If AIAs are taken as products, the producer shall be held accountable. In such a case, AIAs infringement mainly results from product deficiency, and it would be difficult for consumers, i.e. users, to find out the defect if AIAs as products. The *Product Quality Law of the People's Republic of China* [85] says that the quality of products shall be inspected as up to the requirements in conformity with relevant national standards and industrial standards for human health, personal and property safety, and the producers shall be responsible for the quality of products they have fabricated, and the sellers shall establish and implement their procurement

and inspection systems, to identify the product qualification certificates and other marks. In the long run, however, AIAs cannot be just taken as tools or products, for AIAs have some autonomy and ability to learn, and are to some extent in the position to make their own judgements. The judgements and decisions made by AIAs actually have something to do with their training data, input data and algorithms, and their damages or infringements rendering actions may result from their design and fabrication deficiencies, or be developed during AIAs autonomous learning, which involves the obligations of AIAs designers and supervisors. Therefore, to establish the assignment of obligations and the accountability mechanism for AIAs, we need to consider different application scenarios, and make constant efforts to establish complete laws, regulations and ethical norms, to ascertain relevant safety obligations determination and sharing mechanism.

Currently, people from all walks of life focus on designers and users in their discussion of assignment of obligations for AI. On one hand, AIAs in nature are implementing the will of designers; if a designer holds any prejudice, discrimination or unethical inclination in their algorithms and/or training data, the actions of AIAs will be affected. In the case some safety problems or infringements occur, the designers need to account for the procedures for their algorithm decision making and specific decision results. Moreover, after being sold on the market, the decision makings of AIAs may go beyond our expectation due to diversified data inputs. In such a case, the designers are liable for the potential risks of AIAs, and must make prospective predictions and judgments in their preliminary design. On the other hand, AIAs are directly controlled by users, so the users should be responsible for their AIAs to ensure their positive roles, and avoid any harm on the rights of others due to misuse. Discussing the issue of assigning obligations for AIAs is meaningful in that when AIA applications result in human-ethical or legal conflicts, we can hold AI designing departments and relevant personnel accountable in terms of technology, and establish a rational liability and compensation system on the level of AIA application, to ensure just and rational AI applications.

Of course, other parties involved in AIAs may also be liable, which shall not be neglected. IEEE, in the document titled *ETHICALLY ALIGNED DESIGN: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems* (second edition) [53], elaborates the measures to be taken for different subjects' liabilities for AI, and points out: Legislatures should clarify issues of responsibility, culpability, liability, and accountability for Autonomous and intelligent systems (A/IS) where possible during development and deployment (so that manufacturers and users understand their rights and obligations); designers and developers of A/IS should remain aware of, and take into account when relevant, the diversity of existing cultural norms among the groups of users of these A/IS; multi-stakeholder ecosystems should be developed to help create norms where they do not exist because A/IS-oriented technology and their impacts are too new; systems for registration and record-keeping for key parameters should be created so that it is always possible to find out who is legally responsible for a particular A/IS.

Moreover, some scholars have also proposed five kinds of obligations for related personnel and organizations for robots [86].

(1) Designer's obligation: From the perspective of robot ethics, the most important ethical principle for designers to observe is to ensure safety and reliability. That is to say, designers should be liable for the safety and reliability of robots, providing relevant interpretation and explanation;
(2) Manufacturer's obligation: For robot manufacturers, the most important ethical obligation is to fabricate safe and reliable AI products, and ensure no hazards resulting from production;
(3) Organization's obligation: Various institutions and organizations formed by researchers need to bear the obligations of evaluating and supervising risks and safety problems;
(4) Administrative department's obligation: Competent governmental departments need to have scientific, rational plan and supervision for robot technologies and industry; and
(5) User's obligation: Users should respect and kindly treat robots, and cannot take them just as tools, but recognize the intrinsic values of robots themselves.

The accountability for AIA-related problems involves different scenarios, different organizations and personnel. If no specific mechanism for accountability is ascertained, things would go like this: some organizations are irresponsible at all. Therefore, we must assign obligations properly, specifying the obligations of all subjects of liability, and establishing corresponding accountability mechanism. Only by doing so can we ensure someone is held accountable for any AIA accident or infringement.

10.4.3 Accountability for Pilotless Automobile

The issue of accountability for pilotless automobile ultimately lies in the establishment of relevant safety laws and regulations. As AI technologies are rapidly developing and witnessing ever-expanding applications, the challenges in legal policy for AI is increasing. Currently, we are preparing, exploring and accumulating relevant research on AI accountability system and legal policies. Among those occurred and predictable issues related to AI, the accountability for pilotless automobile is the most typical. The application of pilotless automobile is fundamentally changing the driving mode of human beings, and represents a major revolutionary change in the history of transportation development, various countries are making more and more efforts to facilitate and support the development of relevant technologies, and speeding up their efforts to enact laws and regulations for pilotless automobile, so as to establish a complete accountability system for pilotless automobile. In the previous chapters and sections, we have already mentioned some cases of personal and property damages incurred in pilotless automobile driving. A probe into most of these cases shows that we can hardly tell whether it is the fault of the driver or the automobile manufacturer. Accidents did happen, and there must be someone held

accountable, so we have got in a dilemma deciding who should be held accountable in such a case.

Compared with conventional automobile infringements, the infringement of pilotless automobile is exceptional in that: First, during pilotless driving, the role of a human driver is involved. The existing legal framework, however, says that automobile users (i.e. drivers) have the responsibilities and obligations to plan a route and make driving decisions. During pilotless driving, the users are just passengers, who are supposed to observe norms for transportation and safety regulations for boarding, not to be liable for any infringement. Therefore, in case any traffic accident occurs not due to any fault of the user as a passenger, the subject of liability—conventionally the driver—cannot be determined. The pilotless automobiles are made of multiple components, and deficiency in any one of the links may result in traffic accident. For example, errors of data transmission from the outside, mistakes in data analysis of the autonomous driving system, system bugs, hacker attack and so on—all might lead to pilotless automobiles' decision-making errors. In such a case, the subject of liability for pilotless driving accidents is harder to determine compared with infringement of conventional automobiles.

Second, pilotless driving-related traffic accidents will not involve some subjective and objective wrong-doing of drivers like drunk driving and fatigue driving, so the causes of personal and property damages and their accountability principles are different. Roughly, there are two cases: First, human infringement. Humans tamper the system without authorization, or make use of pilotless driving technologies to perform an infringement act or commit a crime, who undoubtedly should be liable for compensations. If the infringement occurs due to deliberation or fault of the victims, the existing laws and regulations shall be referred to for settlement. Second, infringement due to product defect or deficiency. If pilotless automobiles fail to meet the national regulations for pilotless driving quality and safety and other relevant standards, the sellers, manufacturers and supervising departments should be held accountable based on the existing regulations. If the product conforms to all national standards, some problems like pedestrians popping up suddenly, insufficiency of current technological expertise, or other cases resembling "the trolley problem", which go beyond the functionality and judgement of the pilotless automobile for the current technological development cannot predict the peculiar cases of the accident. These accidents do not result from poor product quality, but are attributed to the limitations of ethical problems or scientific and technological sophistication. In such a case, we cannot find any subject of liability for infringement.

The development of pilotless automobile has presented a challenge for the conventional infringement accountability system. To prevent the lagging legislation from hindering technological innovation and development, and better serve human being with AI technologies, those countries on the front of development of automatic driving technologies are constantly exploring and issued a series of laws and regulations.

On September 20, 2016, the first automatic driving act *Federal Automated Vehicles Policy* [87] was enacted in California, USA, which defines five levels of automated

automobiles: Level 0 to Level 5, and stresses that safety is the first principle, specifying 15 safety evaluations for automobile or technology to meet before running on the road, including: data recording and sharing, privacy, system safety, vehicle network security, human–machine interface, crashworthiness, consumers education and draining, registration and certification, response performance of vehicles after crash, federal, state and local laws, ethical considerations, operation design, detection of objects and occurrence on the road, defaulted back-off (minimized risks), verification and so on. Meanwhile, this act also specifies different obligations for the federal and state governments in supervising automated vehicles, listing a series of provisions and scope for the state governments to refer to when making relevant policies. The obligations of the federal government include formulating and implementing the standards for automobile security that are compulsory for all automobile manufacturers. In September 2017, the federal government formulated a unified supervision act titled *Safely Ensuring Lives Future Deployment and Research in Vehicle Evolution Act* [88] (or Automated Vehicles Policy for short), and won unanimous consensus in the US House of Representatives. This act indicates that priority should be given to ensuring consumer safety, and reducing traffic mortality, clarifying the obligations and rights of the federal and state authorities in the field of emerging automated vehicle technologies, providing a basic federal framework for supervising automated vehicles. It is specified that the federal authorities should formulate some compulsory standards for design, construction, performance, system and parts related to automated vehicles, while the state authorities should enact standards higher than or equal to these standards, and at the same time upgrade the safety standards for automated vehicles.

On May 12, 2017, German Bundestag approved an automated driving-related amendment titled *Straßenverkehrsgesetz* (amendment) [89], which constitutes a relatively all-round rights-obligations system, demarcating the rights and obligations of drivers and vehicle manufacturers in the case of automated driving. First, this amendment establishes the obligations of the vehicle manufacturer. Second, this amendment makes it clear when a driver should not operate and when a driver must immediately take over the vehicle, and specifies matters like timing, storage and deletion for data collection during driving a smart vehicle, making a shift of the subject of liability between the driver and the manufacturer. Finally, this amendment also increases the total liability for non-fault drivers, which helps enhance protection of the victims.

In February 2017, the *Vehicle Technology and Aviation Bill* [90] was enacted in the UK, specifying that in the case where an accident occurs during road test of an automated vehicle, a simplified insurance procedure can be followed to help the insurer and insurance company to obtain compensation. This bill specifies that the underwriting subject is "system driver" and vehicle owner combined, and the vehicle owner must purchase compulsory insurance coverage for himself and his "system driver", with no elaboration of the types of insurance products and the rate of insurance premium to pay, which is subject to the market. In 2021, automated vehicles will be allowed to run on the road throughout the UK.

China has not started legislation yet, but the national Development Plan for the New Generation of Artificial Intelligence has provided guidelines, and competent

departments have also issued relevant normative documents. In April 2018, the Ministry of Industry and Information, Ministry pf Public Security, and Ministry of Transport jointly formulated and issued the *Specifications for Road Testing of Intelligent Connected Vehicles* [91] (trial), which is applicable to all automatic driving test of intelligent connected vehicles on the public roads within the territory of China. In December 2017, Beijing municipal authorities issued the *Beijing Municipal Implementing Rules for Road Testing of Automated Vehicles* [92] (trial) and relevant documents, with 33 rules and 105 km road for road testing ascertained in March 2018. In addition, Shanghai, Chongqing, Baoding, Shenzhen and other local authorities have successively issued relevant provisions for road testing to support intelligent connected vehicles to conduct road test.

10.5 Summary

This chapter first expounds those ethical problems associated with AI technologies confronting human beings presently and even in the future, mainly including human rights, ethical positions, liabilities and risks issues. Then, we enumerate many existing ethical principles for AI technology development, from the levels of governments and social organizations, international organizations, and enterprises. Apparently, we human beings have formulated relevant ethical norms for possible ethical and moral issues, and these norms may work to restrain our misuse of AI technologies. The AI technologies are evolving at a much faster rate than human beings, and in the future, AIAs might go beyond our human supervision and make their own decisions, therefore, researchers need to ensure the AIAs make decisions in line with ethical norms. In Sect. 10.3 of this chapter, we discuss the ability of AIAs to make ethical decisions and design methods. Even through there is still a long way for self-contained AIAs to come out, it is necessary for us to make efforts and possibly establish some ethics of functionality, to help AIAs to obtain some basic ethical sensitivity. Finally, with regard to the current important issue of AIAs accountability, we have probed into the subject of liability and assignment of obligations, and then got some knowledge of our efforts into relevant laws and regulations for pilotless automobile. We believe it is imperative for us to further enhance ethical principles for AI security.

References

1. Tencent launched the first AI-aided diagnosis platform. (2018–6–21) [2019–11–1]. https://tech.qq.com/a/20180621/018943.htm. (in Chinese).
2. Artificial intelligence: taking jobs or creating jobs. (2018–2–6) [2019–11–1]. http://www.xinhuanet.com/tech/2018-02/26/c_1122452172.htm. (in Chinese).
3. "AI+" coming, are you losing your jobs?. (2019–9–10) [2019–11–1]. http://www.banyuetan.org/jrt/detail/20190910/1000200033134991567738276597557149_1.html. (in Chinese).

4. Jobs lost, jobs gained: workforce transitions in a time of automation (in Chinese). (2017–12) [2019–11–1]. https://www.mckinsey.com/~/media/mckinsey/featured%20insights/future%20of%20organizations/what%20the%20future%20of%20work%20will%20mean%20for%20jobs%20skills%20and%20wages/mgi%20jobs%20lost-jobs%20gained_report_december%202017.ashx.
5. Shaoping G. Human Rights Ethics [M]. Beijing: China Development Press; 2009. (in Chinese).
6. Zhang Z. Ethical research on contemporary Chinese human rights. Guilin: Guangxi Normal University; 2004. p. 10 (in Chinese).
7. Bianlang J. Robots with or without "Human Rights." Nat Sci-Tech. 2007;3:12–5 (in Chinese).
8. Murph D, Wang L. UK report predicts rights for robots. College English 2007(2):6–6 (in Chinese).
9. AI companies, researchers, engineers, scientists, entrepreneurs, and others sign pledge promising not to develop lethal autonomous weapons. (2018–7–18) [2019–11–1].https://futureoflife.org/2018/07/18/ai-companies-researchers-engineers-scientists-entrepreneurs-and-others-sign-pledge-promising-not-to-develop-lethal-autonomous-weapons/.
10. Brundage M, Avin S, Clark J et al. The malicious use of artificial intelligence: forecasting, prevention, and mitigation. [2019–11–1]. https://arxiv.org/abs/1802.07228.
11. Huang X, Li J. How come animals have moral integrity. In: The Chinese society for dialectics of nature. Proceedings for the third Chinese postgraduate forum for philosophy of science and technology and inter-disciplines. Beijing: The Chinese Society for Dialectics of Nature; 2010. pp. 588–592 (in Chinese).
12. Wang L. New epistemology of Marxism and AI—AI is not a threat to human civilization. J. Dialectics Nat. 2018; (4) (in Chinese).
13. Johnson DG. Computer systems: moral entities but not moral agents. Ethics Inf Technol. 2006;8(4):195–204.
14. Floridi L, Sanders JW. On the morality of artificial agents. Minds Mach Arch. 2004;14(3):349–79.
15. Leopold. A sand county almanac. Translated by Hou W. Changchun: Jilin People's Press; 1997 (in Chinese).
16. Jun W. Ethical issues for AI: challenges and strategies. Stud Ethics. 2018;04:79–83 (in Chinese).
17. Sloman A. The computer revolution in philosophy: philosophy, science and models of mind. Brighton: Harvester Press; 1978.
18. Tian X, Bai C. Contemporary social responsibility ethics. Beijing: People's Press; 2008. p. 4 (in Chinese).
19. Principles of robotics. [2019–11–20]. https://epsrc.ukri.org/research/ourportfolio/themes/engineering/activities/principlesofrobotics/.
20. BS 8611:2016, ethical design and application of robots. [2019–11–20]. http://www.machinebuilding.net/ta/t1028.htm.
21. Business ethics and artificial intelligence. (2018–1) [2019–11–21]. https://www.ibe.org.uk/uploads/assets/5f167681-e05f-4fae-ae1bef7699625a0d/ibebriefing58businessethicsandartificialintelligence.pdf
22. AI in the UK: ready, willing and able?. (2018–4–16) [2019–11–22]. https://publications.parliament.uk/pa/ld201719/ldselect/ldai/100/100.pdf.
23. South Korean robot ethics charter 2012. [2019–12–6]. https://akikok012um1.wordpress.com/south-korean-robot-ethics-charter-2012/.
24. About the Japanese Society for artificial intelligence ethical guidelines. (2017–2–28) [2019–12–6]. https://ai-elsi.org/archives/514.
25. The Japanese Society for artificial intelligence ethical guidelines. [2019–12–6]. http://www.ai-elsi.org/wp-content/uploads/2017/05/JSAI-Ethical-Guidelines-1.pdf.
26. Draft AI R&D guidelines for international discussions. (2017–7–28) [2019–12–6]. http://www.soumu.go.jp/main_content/000507517.pdf.
27. Draft AI utilization principles. (2018–7–17) [2019–12–6]. http://www.soumu.go.jp/main_content/000581310.pdf.

28. Social principles of human-centric AI. [2019–12–6]. https://www8.cao.go.jp/cstp/english/humancentricai.pdf.
29. Draft Report: with recommendations to the Commission on Civil Law Rules on Robotics. (2016–5–31) [2019–12–6]. http://www.europarl.europa.eu/doceo/document/JURI-PR-582443_EN.pdf?redirect.
30. European Civil Law Rules in Robotics. (2016–10) [2019–12–6]. http://www.europarl.europa.eu/RegData/etudes/STUD/2016/571379/IPOL_STU%282016%29571379_EN.pdf.
31. European Group on Ethics in Science and New Technologies. Ethical principles and democratic prerequisites. In: Statement on artificial intelligence, robotics and 'autonomous' systems. (2018–3–9) [2019–12–6]. http://ec.europa.eu/research/ege/pdf/ege_ai_statement_2018.pdf.
32. General data protection regulation: GDPR. [2019–12–6]. https://gdpr-info.eu/.
33. Draft Ethics guidelines for trustworthy AI. (2018–12–18) [2019–12–6]. https://ec.europa.eu/digital-single-market/en/news/draft-ethics-guidelines-trustworthy-ai.
34. Ethics guidelines for trustworthy AI[EB/OL] (2019–4–8) [2019–12–6]. https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=60419.
35. Building trust in human-centric artificial intelligence. (2019–4–8) [2019–12–6]. https://ec.europa.eu/transparency/regdoc/rep/1/2019/EN/COM-2019-168-F1-EN-MAIN-PART-1.PDF.
36. Statement on algorithmic transparency and accountability. (2017–1–12) [2019–12–6]. https://iapp.org/media/pdf/resource_center/2017_usacm_statement_algorithms.pdf.
37. The future society, law and society initiative, principles for the governance of AI. (2017–7–15) [2019–12–6]. https://thefuturesociety.org/2017/07/15/add-law-and-society-initiative-project/.
38. AI policy principles. [2019–12–6]. https://www.itic.org/resources/AI-Policy-Principles-FullReport2.pdf.
39. About OpenAI. [2019–12–6]. https://openai.com/about/.
40. OpenAI charter. [2019–12–6]. https://openai.com/charter/.
41. Twomey P. Toward a G20 framework for artificial intelligence in the workplace. (2018–7) [2019–12–6]. https://www.cigionline.org/sites/default/files/documents/Paper%20No.178.pdf.
42. Singapore releases Asia's first model AI governance framework. (2019–1–23) [2019–12–6]. https://www.gov.sg/~/sgpcmedia/media_releases/imda/press_release/P-20190123-1/attachment/SINGAPORE%20RELEASES%20ASIAS%20FIRST%20MODEL%20AI%20GOVERNANCE%20FRAMEWORK.pdf.
43. A proposed model artificial intelligence governance framework. (2019–1) [2019–12–6]. https://www.pdpc.gov.sg/-/media/Files/PDPC/PDF-Files/Resource-for-Organisation/AI/A-Proposed-Model-AI-Governance-Framework-January-2019.pdf.
44. Principles of artificial intelligence. [2019–12–6]. https://www.smartdubai.ae/initiatives/ai-principles.
45. State Council of China. Development plan for the new generation of artificial intelligence (GUOFA No. [2017]35). [(2017–7–20) [2019–11–2]. http://www.gov.cn/zhengce/content/2017-07/20/content_5211996.htm. (in Chinese).
46. SAC General Team for Artificial Intelligence Standardization. White paper on artificial intelligence standardization 2018. [(2018–1–24) [2019–11–2]. http://www.cesi.ac.cn/201801/3545.html. (in Chinese).
47. SAC General Team for Artificial Intelligence Standardization. Analysis report on AI ethical risks. [(2019–04–25) [2019–11–2]. http://www.cesi.cn/images/editor/20190425/20190425142632634001.pdf. (in Chinese).
48. Beijing AI principles issued. (2019–5–25) [2019–12–7]. http://www.xinhuanet.com/tech/2019-05/25/c_1124540938.htm.
49. Beijing AI principles. [2019–12–7]. https://baike.baidu.com/item/人工智能北京共识/23515671#3. (in Chinese).
50. New generation AI governance principles—developing responsible AI issued. (2019–6–17) [2019–12–7]. http://www.most.gov.cn/kjbgz/201906/t20190617_147107.htm. (in Chinese).
51. Proposal to sign the joint pledge on artificial intelligence industry self-discipline. (2019–8–9) [2019–12–7]. http://news.soft6.com/201908/09/356606.html. (in Chinese).

52. China AI Industry Alliance. Joint pledge on artificial intelligence industry self-discipline (exposure draft). http://aiiaorg.cn/uploadfile/2019/0808/20190808053719487.pdf. (in Chinese).
53. Ethically aligned design: a vision for prioritizing human well-being with autonomous and intelligent systems. [2019–12–7]. https://standards.ieee.org/content/dam/ieee-standards/standards/web/documents/other/ead_v2.pdf
54. Ethical aspects of autonomous and intelligent systems. (2019–6–24) [2019–12–7]. http://globalpolicy.ieee.org/wp-content/uploads/2019/06/IEEE19002.pdf.
55. Partnership on AI. [2019–12–7]. https://www.partnershiponai.org/about/.
56. Tenets. [2019–12–7]. https://www.partnershiponai.org/tenets/.
57. Asilomar AI principles. [2019–11–2]. https://futureoflife.org/ai-principles/.
58. Top 10 principles for ethical artificial intelligence. [2019–12–7]. http://www.thefutureworldofwork.org/media/35420/uni_ethical_ai.pdf.
59. Boden M, Bryson J, Caldwell D, Dautenhahn K, Edwards L, Kember S, Newman P, Parry V, Pegman G, Rodden T, Sorell T, Wallis M, Whitby B, Winfield A. Principles of Robotics. Engineering and Physical Sciences Research Council, Swindon UK. (2011).
60. Initiative for AI security development in Shanghai issued. (2019–7–3) [2019–12–7]. http://sh.people.com.cn/n2/2019/0703/c134768-33105338.html. (in Chinese).
61. Universal guidelines for artificial intelligence. (2018–10–23) [2019–12–7]. https://thepublicvoice.org/ai-universal-guidelines/.
62. Recommendation of the council on artificial intelligence. (2019–5–22) [2019–12–7]. https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449.
63. G20 ministerial statement on trade and digital economy. https://trade.ec.europa.eu/doclib/docs/2019/june/tradoc_157920.pdf.
64. Microsoft AI principles. [2019–11–2]. https://www.microsoft.com/en-us/ai/our-approach-to-ai.
65. Future cornerstone—AI social roles and ethics issued. (2018–7–9) [2019–11–2]. http://www.chinanews.com/cj/2018/07-09/8560809.shtml. (in Chinese).
66. Transparency and trust in the cognitive era. (2017–1–17) [2019–11–2]. https://www.ibm.com/blogs/think/2017/01/ibm-cognitive-principles/.
67. IBM's principles for trust and transparency. [2019–12–7]. https://www.ibm.com/blogs/policy/wp-content/uploads/2018/06/IBM_Principles_SHORT.V4.3.pdf.
68. Everyday ethics for artificial intelligence. [2019–12–7]. https://www.ibm.com/watson/assets/duo/pdf/everydayethics.pdf.
69. Artificial intelligence: the public policy opportunity. [2019–12–8]. https://www.intel.ai/ai/wp-content/uploads/sites/69/Intel-AI-Public-Policy-WP-2017.pdf.
70. Artificial intelligence at Google: our principles. [2019–11–2]. https://ai.google/principles/.
71. Robin Li expounded his "Four Principles for AI Ethics". (2018–5–26) [2019–12–7]. http://baijiahao.baidu.com/s?id=1601519015320714350.
72. SAP's guiding principles for artificial intelligence. (2018–9–18) [2019–12–7]. https://news.sap.com/2018/09/sap-guiding-principles-for-artificial-intelligence/.
73. Sony group AI ethics guidelines. [2019–12–7]. https://www.sony.net/SonyInfo/csr_report/humanrights/hkrfmg0000007rtj-att/AI_Engagement_within_Sony_Group.pdf.
74. Si X. Building an ethical ark: making AI "available", "reliable", "comprehensible" and "controllable". [2019–12–7]. https://mp.weixin.qq.com/s/_CbBsrjrTbRkKjUNdmhuqQ. (in Chinese).
75. 2017 AI seminar: technology, ethics and law held in Beijing. (2017–4–18) [2019–11–2]. http://www.cas.cn/yx/201704/t20170418_4597401.shtml. (in Chinese).
76. Ethics in healthcare aren't new, but their application has never been more important. (2018–10–1) [2019–12–7]. http://newsroom.gehealthcare.com/ethics-healthcare-arent-new-application-important/.
77. Introducing unity's guiding principles for ethical AI. (2018–11–28) [2019–12–7]. https://blogs.unity3d.com/2018/11/28/introducing-unitys-guiding-principles-for-ethical-ai/.
78. I wish to sign the Montreal declaration for a responsible development of artificial intelligence. [2019–12–7]. https://www.montrealdeclaration-responsibleai.com/the-declaration.

79. Duan W. Value reflection and ethical adjustment in the era of artificial intelligence. J Renmin Univ China. 2017;31(06):98–108 (in Chinese).
80. Moor JH. Four kinds of ethical robot. Philos. Now 2009; 72(03/04).
81. Wallach W, Allen C. Moral machines: teaching robots right from wrong [79] co-authored by Translated by Wang X. Beijing; Peking University Press; 2017 (in Chinese).
82. Lin P, Bekey G, Abney K. Autonomous military robotics: risk, ethics, and design. (2008–12–20) [2019–12–8]. http://ethics.calpoly.edu/ONR_report.pdf.
83. Turing, Computing machinery and intelligence. In: Boden. The philosophy of artificial intelligence. Translated by Liu X, Wang H. Shanghai: Shanghai Translation Publishing House; 2001 (in Chinese).
84. Du Y. Pathway for robot ethical design and its evaluation. Philos Trends. 2017;09:84–91 (in Chinese).
85. State Administration for Market Regulation. Product quality law of the People's Republic of China. (2019–2–15) [2019–12–8]. http://www.samr.gov.cn/samrgkml/nsjg/rzjcs/201902/t20190215_282263.html. (in Chinese).
86. Du Y. On the moral responsibility in robot ethics. Stud Sci Sci. 2017;35(11):1608–13.
87. Federal Automated Vehicles Policy. (2016–9) [2019–12–8]. https://www.transportation.gov/sites/dot.gov/files/docs/AV%20policy%20guidance%20PDF.pdf.
88. H.R.3388-Self Drive Act. (2017–9) [2019–12–8]. https://docs.house.gov/billsthisweek/20170904/HR3388.pdf.
89. Straßenverkehrsgesetz (StVG). [2019–12–8]. https://www.gesetze-im-internet.de/stvg/StVG.pdf.
90. Vehicle technology and aviation bill. [2019–12–8]. https://publications.parliament.uk/pa/bills/cbill/2016-2017/0143/17143.pdf.
91. Ministry of Industry and Information, Ministry pf Public Security, Ministry of Transport. Specifications for road testing of intelligent connected vehicles (trial). (2018–4–11) [2019–12–8]. http://www.miit.gov.cn/n1146295/n1652858/n1652930/n3757018/c6128243/content.html. (in Chinese)
92. Beijing Municipal Commission of Transport, Beijing Municipal Traffic Management Bureau, Beijing Municipal Bureau of Economy and Information Technology. Beijing Municipal Implementing Rules for Road Testing of Automated Vehicles (trial). (2017–12–15) [2019–12–8]. http://jtw.beijing.gov.cn/xxgk/tzgg/201808/P020180810517818471957.pdf. (in Chinese).

Chapter 11
Frontiers of AI Technologies

Mohan Li and Binxing Fang

First mentioned in a conversion between Stanislaw Ulam and John von Neumann, the term "singularity" refers to the critical point at which artificial intelligence exceeds human intelligence. John von Neumann pointed out that the "ever accelerating progress of technology and changes in the mode of human life, which gives the appearance of approaching some essential singularity in the history of the race beyond which human affairs, as we know them, could not continue" [1]. In his book titled *The Singularity is Near*: *When Humans Transcend Biology*, Ray Kurzweil stated that technology and evolutionary processes progress in an exponential fashion, and advances in new technologies, including artificial intelligence (AI), will lead to revolutionary changes in the technological and social fabric of the human world [2]. This chapter will present the current developments in AI and investigate the impact of AI technologies on various industries from perspective of academic frontiers, application, and cross-disciplinary development. Further, the potential security risks associated with the cutting-edge AI technologies are also explored. Figure 11.1 presents the structure of this chapter.

11.1 Academic Frontiers

This section focuses on the existing technologies and algorithms that play an crucial role in AI development. However, owing to the limited length of the chapter, we will introduce only the characteristics and basic concepts of relevant technologies and algorithms instead of elaborating on the details.

M. Li (✉) · B. Fang
Guangzhou University, Guangzhou, China
e-mail: limohan@gzhu.edu.cn

B. Fang
e-mail: fangbx@cae.cn

© Publishing House of Electronics Industry 2025

B. Fang (ed.), *Artificial Intelligence Security and Safety*,
https://doi.org/10.1007/978-981-96-9263-7_11

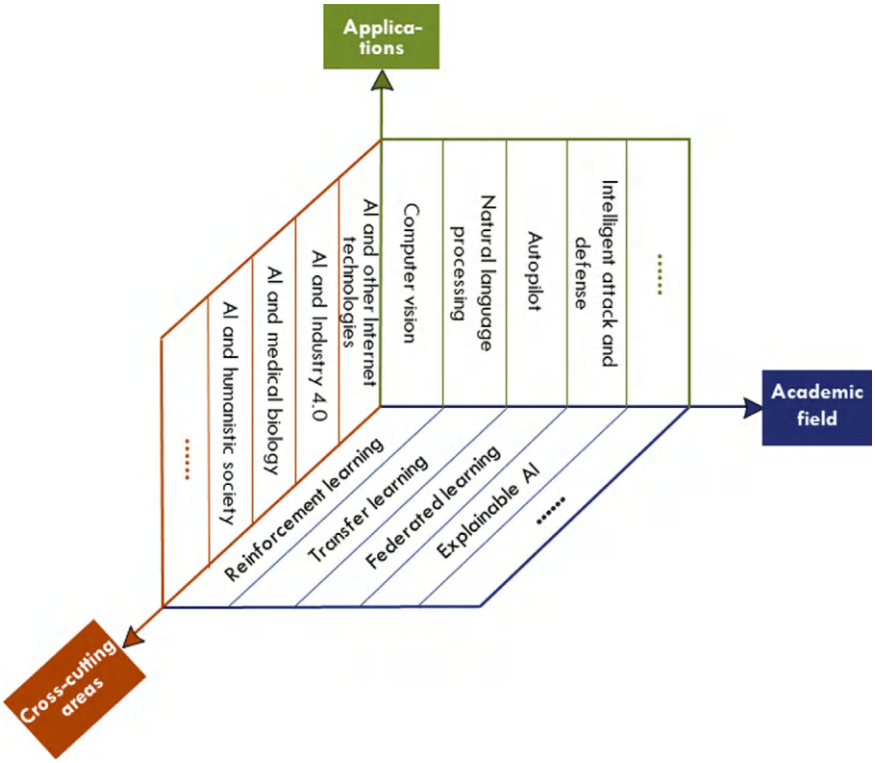


Fig. 11.1 Structure of this chapter. 应用: Applications, 学术: Academic field, 交叉领域: Cross-cutting areas, 智能攻防: Intelligent attack and defense, 智能驾驶: Autopilot, 自然语言处理: Natural language processing, 计算机视觉: Computer vision, AI, 与其他互联网技术: AI and other Internet technologies, AI和工业4.0: AI and Industry 4.0, AI和医学生物: AI and medical biology, AI和人文社会: AI and humanistic society, 强化学习: Reinforcement learning, 迁移学习: Transfer learning, 联邦学习: Federated learning, 可解释的人工智能: Explainable AI

11.1.1 Reinforcement Learning

Reinforcement learning [3], though no longer a new concept, has gained increasing attention and is often found in academic papers and news reports because of its superior performance amid the improved computing power and broader adoption of deep learning methods. In 2018, inspired by the "reward–incentive" behavior of the human brain, Google released a reinforcement learning framework called "Dopamine" [4].

It is essential to use feedback from surroundings for a wide range of application scenarios, and reinforcement learning is cut out for such scenarios. The basic mechanism of reinforcement learning is presented in Fig. 11.2. Three variables need to be defined during the learning process, i.e., state, action, and reward. A machine decides on its actions based on the state and reward fed back from the environment to learn

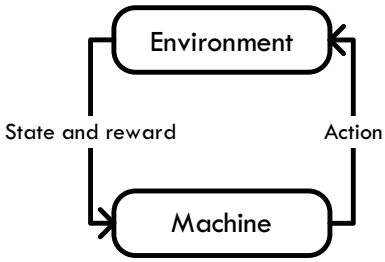


Fig. 11.2 Basic mechanism of reinforcement learning. 环境: Environment, 机器: Machine, 行动: Action, 状态、奖励: State and reward

a set of policies via trial and error. As per the policies learnt, the machine then act in a specific setting to reach for the optimal reward.

In many scenarios, the current action will affect not only the current reward, but also the subsequent actions and rewards, and the quality of a policy is measured by the cumulative rewards obtained from long-term implementation of this policy. In reinforcement learning, it is not often not assumed that large quantities of labelled training samples are accessible before the machine starts learning; rather, the machine learns while obtaining new samples at the same time. When new samples are obtained, the model will be updated, and the next action will be decided based on the updated model. Subsequently, the model will be continuously updated based on the rewards obtained by performing the next action. This pattern will be repeated until the model converges. In this process, every action and reward will exert a long-term impact. Therefore, we must select an action that is most conducive to perfecting the model, and the following two selection methods are available:

(1) exploitation: selecting the best action from among actions adopted before;
(2) exploration: trying a new action that has not been adopted before.

Because of the uncertainties in the environment or context, exploitation can be considered to be a conservative option that will produce a stable (passable) result which, however, is unnecessarily the best result. In contrast, exploration is more progressive. It allows a wider selection of policies but has higher risks: it may produce a better result, but runs the risk of ending up with a worse outcome. These two methods are conflictary, so we need to seek a sensible compromise between them to obtain an optimal cumulative reward.

"Smart" exploration and exploitation make it possible to develop "smart" AI which, in specific contexts, exhibit intelligence on par with human's—human learns via exploration and exploitation when faced with unknown areas. Thanks to breakthroughs in deep learning over these years, deep reinforcement learning has received considerable attention and has seen wider adoption in a variety of AI processing tasks. Advances have been reported in *Nature*, including deep Q network (DQN) [5] proposed in 2015, AlphaGo [6], and the differentiable neural computer [7] that came out later in 2016.

The reward function is vital in reinforcement learning, and a proper reward function will lead the intelligent actant to learn effective policies. Nonetheless, finding a proper reward function, in most cases, turns out a tall order. To lead the machine

towards a policy that meets requirements, we can refer to the manner in which children learn from adults. A little girl may imitate his elder sister to walk as elegantly as the brother does. Though she does not know how his elder strains and stretches the muscles while walking, she can observe and imitate the gait, compare her own gait with the elder's, make adjustments; then she will imitate again, compare again, and adjust again…. Over and over again, she will at last walk as elegantly as the elder does.

Take "degree of elegance" as the reward function, and "how to walk" as the policy, then we achieve a process called inverse reinforcement learning. Inverse reinforcement learning provides a solution of iteration when it is difficult to directly access the reward function. With this solution, we no longer rely merely on the reward function to reach the optimal policy; rather, we iteratively optimize the reward function and the policies by comparing with the "official" policy to achieve the best policy. The specific steps are as follows:

(1) First, a policy is randomly developed as the present policy.
(2) The present policy and the official policy are compared.
(3) The initial reward function is obtained based on the comparison result in Step (2). The present reward function is used to improve the present policy and bring it closer to the official policy.
(4) Steps (2) and (3) are repeated until the differences between the present policy and the official policy are sufficiently small.
(5) When the iteration ends, the present policy is considered to be (almost) the same as the official policy. In this case, a sufficiently good policy is obtained and the present policy is fed back.

Inverse reinforcement learning is also classified as imitation learning. Yet, in fact there is no clear distinction between the learning algorithms, and it is not hard to see that inverse reinforcement learning algorithms are also built on the architecture of deep learning. Therefore, inverse reinforcement learning is a form of reinforcement learning despite the head word "inverse." In practical applications of AI, it is often hard to identify the reward function, and in this case, how to reach the objective via inverse reinforcement learning becomes a hot research topic.

11.1.2 Transfer Learning

Reinforcement learning is to obtain an optimal policy via constant adjustments and adaptations, whereas inverse reinforcement learning relies on continuous imitations of a same type of task (like the child's imitation of walking mentioned before) towards the official policy. What if the target task differs from the training task? Can we use the learning results obtained from one type of task to improve the machine's learning performance on another type of related or similar task? Transfer learning provides a positive answer [8].

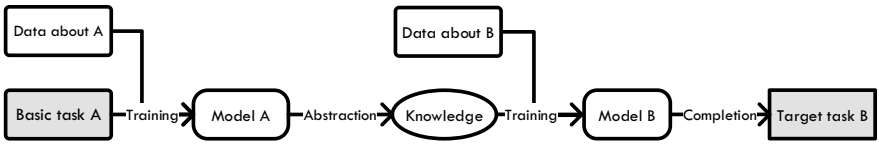


Fig. 11.3 Basic mechanism of transfer learning. A的数据: Data about A, 基础任务A: Basic task A, 训练: Training, 模型A: Model A, 抽象: Abstraction, 知识: Knowledge, B的数据: Data about B, 训练: Training, 模型B: Model B, 完成: Completion, 目标任务B: Target task B

As shown in Fig. 11.3, transfer learning aims to apply the knowledge acquired from scenario A (basic task A) to help the machine to learn in scenario B (target task B) that is related to scenario A, so that the machine can obtain better learning outcome under scenario B. In other words, the model trained on scenario A is the starting point for the model training in scenario B.

Generalized features are of vital importance in transfer learning. Generalized features refer to abstract features that are suitable for both basic tasks (after fine-tuning) and target tasks. Thus, we can first train a basic network on the basic dataset and task, and then fine-tune the features to train the target network on the target dataset. If the basic task has sufficient training data while the target task has not, then transfer learning may help to complete the target task. For example, when we intend to identify a truck from a picture but are short of training data, we can turn to transfer learning if we have at hand adequate training data for identifaction of cars: as the cars and trunks bear much resemblance, we can use the a pretrained car-identifcation model to train the trunk-identifcation model, thereby improving the trunk identification accuracy. In computer vision and natural language processing (NLP) tasks, training the neural network entails massive data. Thus, pretrained models are usually regarded as the starting point for new tasks, so that the patterns acquired from the pretrained model can be applied directly to the new task.

Transfer learning holds out the hope for more interesting inventions. For example, when combined with federated learning that we will introduce next, transfer learning can fulfill machine learning tasks on multiple datasets while protecting data privacy.

11.1.3 Federated Learning

In transfer learning, datasets and models in the basic task are required to fulfill the learning of the target task. However, such datasets and models are not always be available. In healthcare, for example, inappropriate use of data may intrude on the patient's privacy (such as exposing confidential medical history against the patient's will). As a result, medical institutions are highly cautious about the usage of health data. Then, how to continue machine learning under the premise of data privacy protection and adherence to laws and regulations? To find an answer, we can turn to federated learning [9].

Federated learning provides a solution to the following problem: when multiple data owners wish to collaborate to fulfill a machine learning task, how to make use of their data to optimize the model without sharing the data? Imagine two hospitals that want to collaborate in medical data analysis. To protect the patients' privacy, both sides, though in partnership, will not reveal any identity information of the patients or share raw data.

This problem can be solved by federated learning in which a virtual shared model is built via parameter exchange under an encryption mechanism. The virtual shared model can serve multiple parties and can prevent any party from inferring any information of another party's. If such a virtual model can be established, then federated learning can integrate the data of multiple parties to complete model training without moving the data or violating privacy protection rules, thereby obtaining better learning results than those obtained based on a single source of data. This method is termed "federated learning" because the two parties in the model training are not merged into a "country", but are like "states" that go their own way under the administration of a "federal government" (i.e., a system for encrypted exchange of information and model parameters).

11.1.4 Distributed Machine Learning

As with federated learning, distributed machine learning [10] also aims to enable multiple parties to jointly complete machine learning tasks. However, unlike federated learning, distributed machine learning is not restricted by the lack of data sharing and focuses on how to decompose a large machine learning into multiple computing nodes to improve the model's performance. That is to say, distributed machine learning divides a machine learning task into smaller ones and assigns the sub-tasks to multiple computing nodes to achieve better performance than that achieved by a single computing node. This is a multi-dimensional and interdisciplinary research direction, involving both hardware and software, and both theory and application. It is fair to say that it encompasses all aspects of machine learning. As we ushers in the big data era, machine learning is seeing an exponential increase in the quantity of data to be processed and models to be trained and maintained. Against this backdrop, distributed machine learning has become an inevitable choice.

Distributed machine learning can be used to complete training and inference tasks in parallel in a distributed computing context. One major problem is how to maximize the model's performance by operating multiple computing nodes in parallel. Two popular solutions are data parallelism [11] and model parallelism [12], as shown in Fig. 11.4.

(1) **Data parallelism:** in this approach, a dataset is split into several segments and distributed across multiple computing nodes. Each node holds a complete copy of the model and its own part of data. Further, each node outputs the result after the completion of its own computing task, and the outputs of all nodes are then

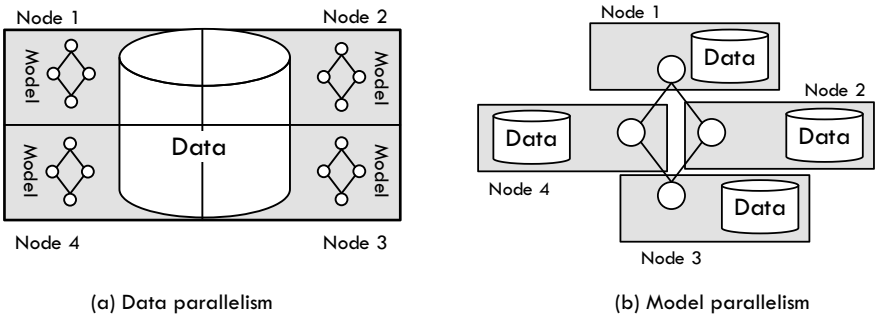


Fig. 11.4 Data parallelism and model parallelism. 节点1: Node 1, 模型: Model, 节点2: Node 2, 数据: Data, 节点3: Node 3, 节点4: Node 4. **a** Data parallelism, **b** model parallelism

integrated to generate the final result. A feasible solution to data parallelism is to split the training dataset into several segments, and allow each computing node to train the model locally before integrating the training outputs of all the nodes using a certain strategy to accelerate the training.

(2) **Model parallelism:** in this approach, a model is split and distributed into different computing nodes. Each computing node holds the part of the model for which it is responsible (e.g., a layer of the neural network or a part of a large matrix to be learned) and the relevant training data. Each node fulfills its own computing task and is synchronized via specific synchronization strategies. Take for example the matrix factorization task often involved in recommender systems. The rating matrix is split into multiple blocks, which are then distributed to different computing nodes. These nodes complete the computing tasks of the matrix in parallel to achieve accelerated computing.

Data parallelism enables multiple nodes to process data in parallel so that more data can be processed per unit time and hence data processing is accelerated. Model parallelism allows multiple computing nodes to undertake the computing workload of a model so that the multiple nodes can be used to process a huge model that a single machine can hardly accommodate. Data parallelism is a preferred solution when the model is not huge but has massive data; while in the case that the model is huge and the computation cannot be fulfilled by a single machine, model parallelism will be a wiser option. Be it data parallelism or model parallelism, the objective is to maximize the model's learning rate (to gain linear acceleration). That is to say, if we use k computing nodes to process the computing a task that is originally completed on one node, we would expect to gain a training speed k times of the original speed. Nevertheless, such linear acceleration is hard to come by in that it is hard to ensure the data and the model are independent of each other after the splitting. In addition, synchronizing the computing outputs from different nodes will add to the computing and communication overhead, which may reduce the computing speed if the splitting and integrating process are not properly designed.

As deep learning techniques gain steam in these years, distributed deep learning has gained increasing attention and witnessed more research breakthroughs, such as Google's Disbelief [11] which introduces distributed machine learning to deep learning. The tasks processed via deep learning are seeing mounting complexity in diverse application fields, with the data scale and model size of some tasks exceeding the processing capability of a single machine. The scale of data required for model training can be daunting: tens of millions of images may be required for training to obtain accurate recognition results in machine vision. The model size is also expanding: LightLDA [13], a topic model developed by Microsoft Research Asia has more than 20 billion parameters, and Google's MoE model [14] involves up to 137 billion parameters. The framework of distributed deep learning will provide solid technical support for complex tasks in diverse fields.

11.1.5 Representation Learning

AI technology is to represent real-world tasks in a form accessible and processable by machines that have a computing power far beyond human's. Before design of an algorithm, the following two questions need to be answered: (1) How should the real-world problem be represented? (2) Which type of representation is better when diverse representations are available? These two questions lie at the heart of representation learning [15].

In a model, X and Y stand for the input and output, respectively. "Representation" refers to the way in which the input X is represented in the model. A different representation of X may lead to a different Y. In deep learning, the inputs are basically represented as vectors. In machine translation, for example, the input word (string) can be represented as a numeric vector; in image classification, the input image can be represented as a matrix or vector before being sent to the classifier; in recommender systems, the user's preference for a given type of goods, as the input, can be represented as a value or a set of values. Thus, one same real-world problem can be converted into different mathematical forms that can be processed by a machine. Theere can be a wide range of appropriate reprensetations for the same data set as the target task differs, which means the optimal representation is not fixed or unique.

Representation learning is designed to achieve better representations. A good representation learning method makes feature extraction faster and more convenient, and with the strong computing power, it allows the model to mine unknown information from the data and find better features. Representation learning connects various fields of machine learning. As with other machine learning methods, representation learning is classified into supervised learning and unsupervised learning. The former uses labeled data to learn. For example, we can learn how to represent a word as a numeric vector using a large labeled corpora. Unsupervised learning, however, uses unlabeled data to learn. For example, we can directly use cluster centers as representations of elements within the cluster.

Enormous data are required for the model to automatically learn good data representations. Manual feature extraction will fail to meet the need for processing of massive data, and representation learning turns out an exciting solution in this regard as it can automatically extract good representations from massive and complex data.

11.1.6 Meta Learning

As analyzed above, reinforcement learning can obtain better policies based on the feedback from the environment. Transfer learning transfers and applies the knowledge acquired from task A to task B. In federated learning, different "federations" can cooperate without violating privacy. Distributed machine learning appropriately splits the data or model to achieve accelerated computing by more computing nodes. Representation learning selects the best representation from the many representations for the same data. In these methods, much preliminary work pertinent to the task or the model has to be done manually, such as designing the model, determining the reward function, setting the basic task and target task in transfer learning, and selecting the method to split the data. In these cases, the machine is much like a primary school student who learns only as instructed. Can the machine grow "smarter"? At school, motivated self-taught students tend to outsmart their peers because they can master new skills faster even without any guidance. Will this be the same case with machines?

Meta learning [16] provides an answer. At the frontier of AI research, meta learning resolves the problem of "learning how to learn."

In conventional machine learning models, a large dataset suitable for the task is selected and used to train a model from scratch. However, in the case of human students, their prior learning experience in other fields can often provide some lessons for new tasks, so only a few samples are required when they face a new learning task. In comparison, the working mechanism of conventional machine learning models seems cumbersome.

The meta learning models are designed to break the limits of a specific task and learn universal "knowledge" (features or parameters) that are readily adapted to a wide variety of tasks. In this light, meta learning is "multitask learning", i.e., a method that concludes abstract and general knowledge based on different tasks. This multitask learning method also aims to break the limits of specific models: ideally, meta learning models are designed to find task- and model-independent solutions.

There is still no widely acknowledged definition of meta learning. Generally, we classify technologies, including metaknowledge discovery [17], inductive transfer [18], and automatic machine learning (AutoML) [19], as meta learning. Among them, AutoML has drawn considerable attention over these years.

AutoML aims to accomplish machine learning tasks with minimal human involvement using technologies such as transfer learning and neural architecture search [20]. Unlike conventional models, the AutoML model requires no human guidance during

the machine learning process, and the parameters and model will be generated automatically once the training data are input. Evidently, a detailed probe into AutoML will lower the bar for the application of machine learning and make possible the wide adoption of machine learning in traditional industries. In fact, such topics as the implementation of applicable AutoML frameworks, the parameter optimization of robust machine learning methods, and automatic design of neural network structures have been extensively discussed on top summits and academic symposiums around the world in these years, and there have been some well-known AutoML projects, including Google AutoML [21], Auto-Keras [22], MetaQNN [23], and weight agnostic neural networks [24] that went viral in 2019.

11.1.7 Explainable AI

As mentioned in the introduction to distributed machine learning, some models may have tens of billions or even hundreds of billions of parameters as the size of the data and the model grows. The high complexity of data and model makes AI into a "black box", for which we humans only know what decisions the AI has made but have no inkling as to how and why these decisions are made. This "black-box" attribute makes AI a useful tool when, for instance, we want to protect privacy; it can, however, also make AI extremely dangerous: when AI is applied to development of autonomous weapons, this black-box attribute makes it impossible for us to know where and why the autonomous weapon launches an attack.

In real life, critical decisions made by humans undergo strict reviews over and over before their rationality is confirmed. The very focus of the reviews is to understand the underlying logic of the decisions. By analogy, the decisions of an AI system must be explainable, and hence another important area of AI, "explainable AI" (hereinafter referred to as XAI) [25], comes into being.

XAI is used to explain the logic that underlies every decision made by AI actants. On the one hand, XAI must explore explainable models so that human users can better understand and trust their AI partners; on the other hand, enhanced interpretability may come at a cost of prediction accuracy. Therefore, a fine balance must be achieved between interpretability and prediction accuracy. Influential findings related to XAI have been published in papers on the International Conference on Knowledge Discovery and Data Mining [26], and the Defense Advanced Research Projects Agency (DARPA) has also probed into XAI and are developing the "Third-Wave AI Systems" [27].

To realize XAI, the interpretability before, during, and after modeling should be explored.

(1) The interpretability before modeling can be studied mainly from perspectives such as data preanalysis and presentation by, for example, studying the association between the data items, exploring data distribution [28], and designing simple and effective data visualization techniques [29].

(2) The key to interpretability during modeling lies in designing models accessible to humans. For example, rule-based [30] and instance-based [31] models can be easily accessed by humans.
(3) The interpretability after modeling focuses on deep learning models with the "black-box" attribute. Fine examples include the attention mechanism [32], neural module networks [33], and network dissection [34].

- Based on the attention mechanism, different weights are assigned to different inputs to indicate their level of importance. With these assigned weights, the model can judge which input information is more important and hence reach more accurate results. The attention mechanism has been applied to a wide range of fields: weights are assigned to different words, regions, and segments in NLP, computer vision, and intelligent speech tasks, respectively. The assigned weights can be treated as acquired knowledge that helps people to understand why the system makes the decision; they can also be regarded as an evaluation basis for us to assess the rationality of the decision.
- A neural module network is an artificial neural network characterized by the fact that several independent neural networks are connected to form an entire network. Each independent neural network serves as a module and operates on separate inputs to accomplish subtasks. Some intermediary components obtain and process the outputs from each module to produce the output of the network as a whole. The intermediary component only accepts the outputs from the modules and does not respond to the modules. Similarly, there is no interaction between the modules. Unlike conventional neural networks, the neural module network is not an integrated whole; instead, it consists of multiple modularized sub-networks, and this makes it possible for us to understand the network structure from the perspective of network layout. The layout can demonstrate the correspondence between the modules and subproblems as well as the topological relations between the modules.
- Network dissection can enhance the interpretability of models by assessing the correspondence between the hidden units and semantic concepts. It identifies the semantics of hidden units and matches the semantics with concepts understandable to human. In this way, it establishes the correspondence between the hidden units and the concepts understandable to humans, and hence we can interpret the working principle of the entire neural network based on the network structure and the correlation between various concepts.

These aforementioned methods help to explain what human knowledge and concepts the existing AI technologies correspond to, and what aspects the models are sensitive to. In addition, they help us to understand which aspect of the target task needs to be prioritized and what has been overlooked. These are the important trends with respect to the future improvement of interpretability of AI.

11.1.8 Adversarial Training

The generative adversarial network (GAN) is built based on adversarial training, wherein two neural networks work against each other for training and learning [35]. These two networks are termed generator and discriminator. The former generates data, whereas the latter judges the quality of the former's outputs. Figure 11.5 shows the structure of a GAN.

(1) The generator generates the required data, including images, videos, and documents. The input are random data, whereas the output are false data that do not exist in reality.
(2) The discriminator judges whether the output from the generator is sufficiently natural. It receives real or false data exported from the generator and attempts to accurately classify these inputs.

Based on interactions and adversarial training between the generator and the discriminator, the parameters are updated so that the generator will produce samples that resembles the real-world samples as much as possible. This training method is particularly suitable to generate outputs that imitate the real-world objects, such as fake images, videos, and three-dimensional (3D) object models. Because of this powerful imitation ability, GAN is used creatively in many real-world scenarios, including computer vision, intelligent attack and defense, and anthropomorphic AI.

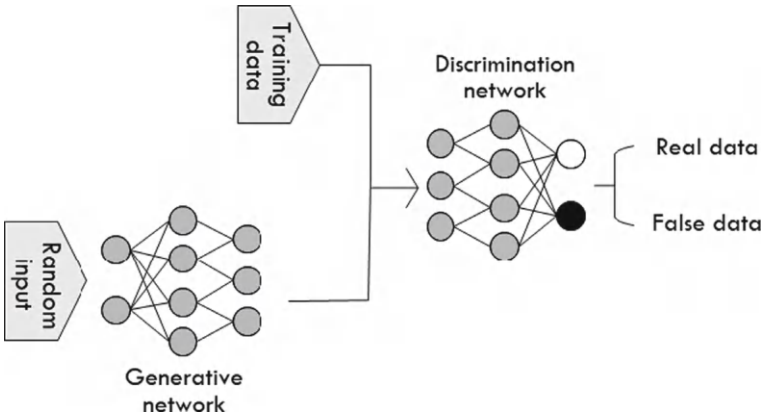


Fig. 11.5 GAN. 随机输入: Random input, 训练数据: Training data, 生成网络: Generative network, 判别网络: Discrimination network, 真实数据: Real data, 虚假数据: False data

11.1.9 AI Chips

AI technology has been extensively applied in various industries. In several application scenarios, AI algorithms are implemented in environments with limited computing power. For example, the cameras in intelligent manufacturing should be able to analyze videos and images, and self-driving vehicles must locally analyze the road conditions and decide on driving options, mobile devices must realize biometric recognition locally rather than on the cloud to avoid privacy breaches. In these scenarios, the hardware of the edge devices used for computation and analysis often have no enormous computing power, and hence cannot locally satisfy the needs of analysis and inference. Therefore, it has become essential to design and develop AI chips to ensure that the hardware can better support the relevant computations.

The AI chips refer to the processors that have been specially designed to accelerate AI algorithms. As deep learning is for now the most popular AI algorithm, AI chips are designed to provide deep learning algorithms with support unmatched by conventional chips to make the algorithms faster and more energy-saving.

AI chips have developed rapidly in these years, especially during the past five years. Since 2007, researchers have found that the needs of the AI algorithms and the parallel computing of big data cannot be satisfied using only the central processing unit (CPU). Instead, the graphics processing unit (GPU) that has shown outstanding performance in matrix operations, floating points, and parallel computing turns out a better option to implement AI algorithms that rely heavily on matrix processing and floating-point processing. In fact, hybrid computing that combines CPU and GPU has been widely used in AI applications since 2010. Nonetheless, with the extensive application of AI technology, to rely merely on the GPU that is limited to matrix operations for training and inference in the modelling process has become increasingly taxing, since, after all, AI algorithms are more widely used in computing of neural networks. Therefore, the industry is investing more to develop special-purpose AI chips to considerably improve the computing performance.

Up to date, many renowned chip manufacturers have released their AI chips. Table 11.1 presents some manufacturers and their brand AI chips, including the special-purpose chips for AI inference and training as well as the general-purpose chips for AI optimization.

(1) **NVIDIA:** NVIDIA invented GPU in 1999 and redefined the modern computer graphics card. As the GPU was intended to process images and videos, it is fairly efficient for matrix computation, and this feature allows it to play an essential role in the development of AI technology, particularly in deep learning technology. NVIDIA still leads the market of AI chips, especially training chips. Its Tesla series [36], chips designed specifically for deep learning, can considerably increase the training and inference speed, and has drawn extensive attention from the industry upon released. Additionally, GPUs in its other series, such as the latest Titan RTX [37] unveiled in 2018, have been widely used in AI-centered applications.

Table 11.1 Some manufacturers and their brand AI chips

Manufacturers	AI chips
NVIDIA	Tesla series and Titan RTX
Google	TPU
Cambricon	Cambricon-1A, Cambricon-1H8, Cambricon-1 M, and MLU100
Huawei	Kirin 970, Ascend 910, and Ascend 310
Baidu	XPU
Vimicro	VC0758 and VC0616
Intel	Nervana series and Ice Lake
AMD	Radeon instinct series
IBM	TrueNorth
Qualcomm	Snapdragon 855 series
Apple	A11 and A12
ARM	Project trillium

(2) **Google (Google):** In 2016, Google released the tensor processing unit (TPU) [38], which was designed specifically for its deep learning framework, TensorFlow. The Cloud TPU it launched later in 2016 targets the cloud, and in 2018, it released the Edge TPU, which is lightweight and small, and is more suitable for edge devices with limited volume and power.

(3) **Cambricon:** at the Third World Internet Conference in 2016, the Institute of Computing Technology, Chinese Academy of Sciences, launched "Cambricon-1A Deep Neural Network Processor" as one of the world's leading scientific and technological achievements in the Internet industry [39]. At the Sixth World Internet Conference in 2019, the Cambricon series again unveiled an excellent product called Siyuan 270 (MLU270), which was introduced as the world's leading scientific and technological achievement in the Internet industry [40]. With these proud achievements, Cambricon has established itself as a world-leading AI chip manufacturer [41].

(4) **Huawei:** among the inference chips, Huawei Kirin 970 [42] could achieve local inference without transferring to the cloud. Huawei launched the Ascend 910 and Ascend 310 chips for training [43]. At the Fifth World Internet Conference in 2018, the Ascend 310 chip was unveiled as the world's leading scientific and technological achievement in the Internet industry [44].

(5) **Baidu:** In 2017, Baidu launched XPU [45], which boasts both the versatility of the GPU and the high efficiency and low energy consumption of the field-programmable gate array (FPGA). Furthermore, it has been optimized for Baidu's deep learning platform, PaddlePaddle.

(6) **Vimicro:** Vimicro[1] launched the video codec system-on-chip (SoC) involving a neural network processor in 2016. SoC was designed to optimize the deep

[1] http://www.vimicro.com.cn/introduction.htm.

learning algorithms such that intelligent analysis results can be simultaneously encoded with video data to obtain structured video codestreams. This technology is extensively used in video security cameras and has ushered in a new era of intelligent video security. Representative chips of Vimicro include VC0758 and VC0616.

(7) **Intel:** Intel has unveiled Nervana neural network processors [46], and the latest model is the Nervana Neural Processor-I (launched in January 2019). In addition, Intel has launched the Ice Lake processor involving an instruction set that accelerates AI.

(8) **AMD:** As one of the GPU manufacturers, AMD has released the Radeon Instinct [47] series for deep learning.

(9) **IBM:** in 2014, IBM developed TrueNorth, a "brain-inspired" computing chip. [48] Instead of using the conventional von Neumann architecture, this chip is based on the neuromorphic architecture and, by creating electronic neural networks similar to the neural network of the human brain, the chip is expected to function in a manner similar to the human brain.

(10) **Qualcomm:** Qualcomm introduced the Snapdragon 855 [49] series in 2018, which was the first accelerator dedicated to AI computing.

(11) **Apple:** Apple's A11 processor [50] integrates the "neural engine" hardware dedicated to machine learning. The neural engine of its A12 chip was upgraded from a dual-core design to an eight-core design.

(12) **ARM:** As a chip IP provider, ARM has been present in the AI market since 2013. In 2018, ARM launched Project Trillium [51] for machine learning processors and object detection processors in mobile devices.

11.1.9.1 Safety Challenges to Frontiers of AI Technology

As mentioned in Sect. 11.1.7 concerning XAI, a major problem associated with the current AI technology based on deep neural networks is its weak interpretability, which makes the AI systems vulnerable in training and inference.

Many artificial intelligence systems (AISs) are trained in open environments. For example, Microsoft Xiaoice, with the ability to communicate, was released early in the market, and it was trained by its interactions with humans. In another example, labeled images were collected online to extensively identify and classify certain images. Therefore, "data poisoning," which impaired the quality of the training samples, could be observed during open training. As explained in Chap. 5, data poisoning means that malicious training samples that were working against the training goals were purposefully placed in the training data during training, resulting in a mistake or deviation in the model (e.g., holding illogical conversations with Xiaoice or releasing images with the wrong labels on the Internet). A simple example can be observed with respect to the training data shown in Fig. 11.6a. Before poisoning, the cross samples and five-pointed star samples are clearly classified into two clusters. At this point, if new samples have to be classified, classifications can be performed by comparing the centroid distances between the new samples and

the clusters. Figure 11.6b shows the training data after poisoning. If some malicious cross samples are inserted close to the five-pointed star samples and the obtained data are used for training, then the centroid of the cross samples will deviate toward the five-pointed stars. At this point, mistakes may occur if the centroid is still used for classification. As the performance of the AI models is considerably dependent on the quality of the training samples, data poisoning has been receiving considerable research attention as one of the major threats to AI technology.

The detection of data poisoning attacks is considerably difficult. Some cases can be detected via outlier detection. In Fig. 11.6b, for example, the cross samples used for poisoning looks less "normal." However, the data used for poisoning may seem totally innocuous in other cases with more elaborate structures. Clean-label attack [52] is a type of highly secret poisoning attack. Such attacks are not simple data poisoning; Instead of adding the wrongly labeled samples to the training set, it adds the correctly-labeled, seemingly problem-free, and finely-adjusted samples to the training set, so that the data used for poisoning looks totally innocuous and will not impact the training and general performance of the neural network. If the machine learning model is used for image classification, then the labels of all the images in the training data are correct and it is difficult to detect poisoning via anomaly detection. In the case of the specific samples that the attacker wishes to attack, these "correct" data can interfere with the model, resulting in the wrong performance of the trained model.

Almost all the aforementioned technologies are threatened by data poisoning. In reinforcement learning, the intelligent agents generally operate in an open and

(a) Training data before poisoning

(b) Training data after poisoning

Fig. 11.6 Simple example of data poisoning. **a** Training data before poisoning, **b** training data after poisoning

uncertain environment, and the very important "rewards" in their training data are derived from the environment. Therefore, they are vulnerable to poisoning attacks once the environment is contaminated. In transfer learning, small amounts of data could contaminate a large proportion of training data for the target task if the target data are poisoned because of the small data size of the target task. In federated learning, the data provided by each "federation" are confidential, and if a federation performs malicious poisoning, the whole model may be affected. In the same logic, in other applications, if the credibility and reliability of the training data cannot be warranted in the training process, they may suffer poisoning attacks to varying extents.

With respect to inference, for trained models that provide inference services, we can analyze the model generation method and the defects via reverse engineering or fuzzing.

Reverse engineering can be used to obtain the design principles of a product based on the finished product when the necessary production information is not readily accessible. In the AI domain, model extraction [53] is a typical reverse engineering process for AI. Machine learning as a service (MLaaS) [54] is gradually trending with the progress of the AI technology. Based on the open application programming interface (API) of MLaaS, an attacker can intentionally construct some special queries to steal the trained models. During such attacks, the data owner uses the data he/she owns to train the model f and allows others to access the model via queries. The attacker extracts the construction f' by intentionally constructing some special queries such that f' becomes approximately equal to f, and thereby steals this model.

Fuzzing aims to input the intentionally constructed data into the program and detect the possible security vulnerabilities by monitoring anomalies in the program. Currently, some AI attacking techniques follow similar ideas, the most popular of which is using adversarial examples [55]. Adversarial examples inflict on the model acceptable perturbations, i.e., perturbations that are hardly detectable and do not affect human judgement. Consequently, the model outputs the wrong result, and the model becomes invalid or deviates directionally. With adversarial examples, an attacker can deceive the AI system and fulfill the purposes of bypassing inspection, malicious guidance, or fraud.

Similar to the attack during training, the attack during inference is applicable to almost all the technologies discussed in this section. Numerous attack techniques takes the model as a black box. However, current AI technologies based on deep learning lack good interpretability, which makes it impossible to understand the attack mechanism. Therefore, preventing such attacks is rather challenging for the time being.

11.2 Application Frontiers

AI technologies have attracted academic attention and have been applied to all aspects associated with national interest and people's livelihood, bringing about significant impacts on various fields. In this section, we will elaborate on the current application frontiers of AI technologies, such as computer vision, NLP, intelligent speech, autopilot, anthropomorphism, bionics, and intelligent attack and defense.

11.2.1 Computer Vision

Computer vision relies on the processing and generation technology of images and videos. This is the field in which AI technologies are most extensively applied. As the images are represented in a computer as matrices, which can be naturally obtained by AI algorithms, the primary technical challenge in computer vision is how to effectively understand and generate myriad images and videos rather than how to form representations that can be processed by AI algorithms. Understanding (i.e., classifying and recognizing) images and videos can no longer satisfy people's needs because we expect AI applications to automatically imitate and generate images and videos.

The aforementioned GAN is increasingly being applied in this domain because the adversarial training it adopts enhances the generator and discriminator amid the adversarial game, so that clear and vivid images or videos will be generated. GAN has recently made breakthroughs in fields such as the generation of pictures, 3D images, and videos.

(1) **Picture generation**. BigGAN [56] introduces orthogonal regularization into GAN, which balances the fidelity and diversity of a model by truncating the prior distribution z, and thereby generate fake images that can pass off as genuine. The success of BigGAN can be attributed to the large size of the neural network and the high computing power. GAN also has many other fascinating applications in image generation. For example, it can be used to generate the formats of quality academic papers. Researchers have used the papers received at conferences as the input of GAN to generate new paper formats.

(2) **3D image generation**. The focus of image generation is on 2D images; however, 3D image generation has started to capture more attention with the growing demand for virtual reality and game applications. The visual object network [57] is a GAN-based 3D generation method. In this method, the image generation model is decomposed into three factors: shape, perspective, and texture. As these factors are independent, they can be learned separately from the datasets of 2D and 3D images to develop a GAN model that can produce shapes and textures. The shapes and textures obtained by the GAN model then can be integrated with the perspectives to produce vivid and operable 3D images.

(3) **Video generation**. The research teams of NVIDIA and MIT proposed a GAN-based video synthesis method named vid2vid (short for Video-to-Video) [58]. Based on vid2vid, a novel architecture was proposed for the generator and discriminator. The optical flow constraints and information were added separately into the generator and the discriminator to produce the given input of videos (dynamic semantic maps, such as the segmentation mattes, sketches, and body postures); then, consistent and high-quality videos could be exported by learning the mapping functions.

Apart from GAN, quantum neural networks (QNNs) have been introduced to computer vision. QNN, which was proposed by Google [59], explores how to perform image classification of neural networks on quantum processors. Until now, most research on QNN is theoretical and considers the design of recent quantum processors, which has theoretically proved that QNN can be trained for image classification on the Modified National Institute of Standards and Technology (MNIST) datasets. Google forecasts that QNN will match or even outperform the classical neural networks and reach "quantum supremacy" with the expanding scale and fast development of computer hardware.

Despite the multitudes of achievements that AI has enabled in the field of computer vision, the specific AI techniques employed rely heavily on optimization and improvement of the existing technologies as well as on the high-quality training data and computing resources. For now, GAN still serves as a basis for many methods, and we need to find new revolutionary technologies to make breakthroughs.

11.2.2 Natural Language Processing

Natural language processing (NLP) aims to establish a system that can understand human language and achieve conversion between human and machine languages. Based on NLP, developers can acquire knowledge from existing data and train the model using AI technologies to fulfill such tasks as automatic summarization, machine translation, named-entity recognition, relation extraction, sentiment analysis, speech recognition, and topic segmentation.

Most existing AI applications rely on deep neural networks to some extent. The inputs for such networks include numeric vectors, whereas the inputs for NLPs are all texts (i.e., strings). Thus, how to convert string-based texts into numeric vector representations becomes a top concern in NLP and a key factor that will affect the processing outcome. The existing NLP models entail considerable resources, including labeled high-quality training data and high-performance computing equipment that can ensure the training efficiency. Therefore, how to train the model efficiently and accurately remains a major problem in NLP, and the aforementioned AI technologies like transfer learning and representation learning can provide a possible solution to this problem.

Word embeddings, as the input of the training model, are a major factor that affects the model training effect. Given the complex characteristics of words, such as the grammar, semantics, and contextual changes, the technique termed embeddings from language models (ELMos) [60] proposed in 2018 considers the context of the words to represent the complex characteristics associated with word usage. It pretrains the bidirectional long short-term memory (BiLSTM) model using a large text corpora and considers the functions learned from BiLSTM as word vectors to generate different word embeddings for the same word in different contexts. Based on ELMo and other cutting-edge NPL technologies, Google proposed a new method for pretraining language representations, namely, bidirectional encoder representations from transformers (BERT) [61]. BERT adjusts the left and right contexts in all the layers and pretrains deep bidirectional representations based on multilayered bidirectional transformers. The word embedding model exhibits strong generalization capability and can even represent the association features between sentences.

Transfer learning is based on a model that solves a specific problem, and adjusts the model to solve similar problems. The existing NLP models need to be modified or updated to meet the varied requirements, and be trained from scratch. This will result in re-collection of massive amounts of data and more training time. To tackle this problem, universal language model fine-tuning (ULMFiT) [62] was proposed in 2018, which initiated the application of transfer learning in NLP. Instead of training a model from scratch, ULMFiT begins with a readily pretrained model that can provide improved performance in new tasks after minor fine-tuning of the data. Transfer learning is bound to be an essential driver for developing NLP. In addition, meta learning has been applied in NLP [63].

Human language serves the function of information exchange, and exceptional flexibility and ability of expression are also required, but the knowledge required for understanding a language is endless. Therefore, NLP must be extensively researched to bridge the gap between human language and machine language.

11.2.3 Intelligent Speech

Intelligent speech primarily allows the machine to recognize and synthesize human speech and makes natural dialog and communication possible between humans and machines using relevant NLP technologies. Certain developments have been achieved in intelligent speech because of the widespread application of AI technologies. Intelligent speech recognition and intelligent speech synthesis are the two main sub-areas of intelligent speech. The former converts human speech into computer-readable information, whereas the latter synthesizes human-understandable speech information based on the computer inputs (e.g., texts or audio clips).

Though intelligent speech recognition and synthesis have been considerably investigated, it was not until the application of deep learning in 2011 that the quality of the intelligent speech service was considerably enhanced. Intelligent speech, such

as intelligent speech assistant, intelligent customer service, voice navigation, information broadcasting, and identity recognition, is being increasingly used in people's daily lives. Several technology giants, such as Google, Microsoft, Amazon, Baidu, and iFLYTEK, are considering intelligent speech as an important part of their AI applications. Google unveiled the voice-powered Google Assistant with the features of "continued conservation" and "multiple actions" at Google I/O 2016. Thus, it could overcome various drawbacks in intelligent speech, including unnatural dialog responses and the inability to have multiple dialogs. The achievements in intelligent speech were obtained based on comprehensive applications, breakthroughs, and innovations in the dominant AI technologies such as DNN [64], RNN [65], LSTM [66], LSTM-CTC [67] and BiLSTM [68].

Currently, researchers of intelligent speech are constantly exploring AI algorithms and models and intelligent speech chips. The speech team of Facebook AI Research (FAIR) introduced the first fully convolutional speech recognition system wav2letter++ in 2018, which was written in C++ and was renowned as the "fastest and top-rate" speech recognition system [69]. Relying on the ArrayFire tensor library[2] and the flashlight machine learning library,[3] the wave2letter system learns all learnable compoents, from sound waves to texts, on the convolutional layers, while fulfills sound modeling and language modeling using a CNN.

Google has made considerable progress in intelligent speech synthesis. In 2016, it rolled out WaveNet, a deep autoregressive and generative model, and its parallel model [70]. These models casually input data by dilated convolutions to obtain a broader global vision, and have achieved remarkable success in the synthesis of text, image, handwriting, speech, and music. In early 2017, it released Tacotron, an end-to-end generative text-to-speech model [71], which converts the input characters into the corresponding original spectrograms and generates speech by the Griffin–Lim algorithm for reconstruction [72]. In early 2018, by combining WaveNet and Tacotron, Google released Tacotron2 [73], which outputs the speech corresponding to the text directly regardless of the grammatical knowledge, setting a new benchmark in the field of speech synthesis. These technologies are being applied to Google Assistant. Speech synthesis can also be used to synthesize singing voices. For example, CeVIO, the speech synthesis software developed by Japan-based Techno-Speech, can synthesize singing voices identical to singers'. For intelligent speech chips, AISpeech rolled out the AISpeech-ShensiliCon TAIHANG chip (TH1520) [74] in 2018, a chip dedicated to AI-based speech applications, which provides an "algorithm + chip" solution with a complete set of functions of speech interaction.

Intelligent speech plays a crucial part in such application scenarios as smart homes and autopilot. Nonetheless, it is now primarily applied to single contexts, simple short sentences, and specific tasks, and there is still a long way to go before it can smoothly interact with human. For instance, recognition becomes extremely difficult when there are loud noises, strong accents or mixed conversations. Besides, the synthesized voices are often short of feelings present in the voices of people;

[2] https://github.com/arrayfire/arrayfire.

[3] https://github.com/facebookresearch/flashlight.

the intelligent speech systems often have limited scalability; and the existing speech systems can support approximately 120 languages (Google Voice), which is far lower than the number of existing languages in the world.

11.2.4 Autopilot

Autopilot streamlines "machine-environment-human" interactions into "machine-environment" interactions to cut the manpower and cost, and meanwhile prevent accidents that might be caused by human factors such as driver fatigue or drunk driving. Autopilot is an important application of AI technology, which is related to not only smart vehicles but also smart aircrafts and ships. Though the application scenarios vary, they share much in common in terms of the underlying technology. This section will discuss the application of AI technology in autopilot for smart vehicles.

American Society of Automotive Engineers (SAE) defines six levels of driving automation (see Sect. 7.3.1 for details) [75]: no driving automation (L0), driver assistance (L1), partial driving automation (L2), conditional driving automation (L3), high driving automation (L4), and full driving automation (L5). Most existing autopilot cars reach L2 of driving automation and can briefly take over some driving tasks, including adaptive cruise control (ACC), automatic lane change, and automatic parking. Cars at this level of automation, however, cannot cope with changing traffic conditions, and the driver has to keep an eye on the road conditions all the time and stay alert to take control of the steering wheel. L3 automation is considered the start of autopilot, under which the automatic driving system fulfills road condition monitoring, while the driver still needs to take over when the road condition becomes complex. The objective of autopilot is to achieve L5 automation, and for the time being, most automobile manufacturers are working towards L3 and L4 automation.

Autopilot involves three key modules: environment perception, decision-making, and control execution. Based on AI technology, the autopilot system constantly perceives the environment (detecting objects, recognizing and classifying objects, positioning objects), predicts the movements according to the decision models trained on the data of traffic rules and travel records, and then executes the corresponding actions. Autopilot touches on a large number of cutting-edge AI technologies. For example, computer vision is used in autopilot systems to observe and judge the road conditions; smart speech and NLP are used to identify and convey instructions that can be understood by humans; deep learning, reinforcement learning, and other technologies can be applied to decision-making of autopilot systems; and intelligent attack/defense techniques are adopted to ensure the security of autopilot systems.

These years have seen increasing investment into autopilot by world-leading IT conglomerates like Google, Tesla, Baidu, Tencent, Uber, and Zoox. The investment mainly goes into two routes of development of autopilot systems:

The first route is "progressive evolution", as adopted by the automobile manufacturers like Tesla, which aims to achieve evolvement from driver assistance to complete

automation. Telsa has already rolled out several versions of autopilot systems, and AutoPilot 3.0 will come out soon. The system has gradually expanded from automatic driving software to autopilot hardware, with autopilot chips playing a vital part within. In AutoPilot 3.0, the proprietary autopilot chip of Tesla has been used to achieve computational efficiency [76]. Currently, progressive evolution still requires human–machine collaborations, in which the autopilot system will learn when the driver is at the steering wheel. The fact is that, unfortunately, the driver often readily trusts the automatic driving system and ends up in accidents [77].

The second route is "revolution", as adopted by Google, aims to achieve autopilot above the L4 level since the very start. In 2009, Waymo (a subsidiary of Google), with Google's TPU and its deep learning framework Tensorflow, began to train neural networks for operation, monitoring, decision-making of autopilot, towards its "safety-first" vision. In 2013, Baidu marched into the realm of autopilot, and made its self-developed autopilot platform Apollo, publicly accessible. In July 2018, Apolong, the world's first L4 automatic bus developed by Baidu, was mass-produced. The companies that took the route of revolution constantly tested the autopilot system before launching it to the market upon maturity, which reduced the potential security risks in real-world application.

Security is a major concern that must be addressed in autopilot applications. There are two main challenges in this regard. The first is that the autopilot system fails to adapt to complex environments and results in accidents. The Uber incident on March 2018 marked the world's first case of pedestrian death caused by automatic driving. The second challenge is that the autopilot system is vulnerable to external network attacks, including distributed denial of service (DDoS) attacks, remote control, and private data theft. Tencent Cohen Lab has successfully carried out multiple remote attacks on the intelligent cars of Tesla. The automatic driving identification and decision-making modules of autopilot systems also face attacks due to the vulnerability of AI technology. Safety is indispensable in autopilot and stays at the heart in the future development of autopilot technology.

Due to the potential safety problems inherent of autopilot systems, autopilot testing becomes an essential step before the system is put into use. Verifying and evaluating the reliability of the system is indispensable to realize safe and efficient unmanned driving. For now, the major testing methods include simulation testing and real-scenario testing. Real-scenario testing can be classified into "closed-scenario test" and "road test" in real traffic conditions. Waymo has set a good example in autopilot testing. The company performed three types of tests on urban public roads using the autopilot simulation software of Carcraft and the Castle to quickly simulate different situations by combining virtual and real scenarios. As of October 2018, Waymo's self-driving cars attained 10 million miles of driving, and the mileage in Carcraft simulation tests reached billions of kilometers. Baidu and Uber have also been putting their autopilot cars to road tests. In addition, many simulation systems, such as the TAD Sim system of Tencent and the Pro-SiVIC system of ESI Group, have emerged.

In the foreseeable future, autopilot can be integrated into intelligent transportation, smart cities, or other application scenarios. It can also be combined with cloud

computing to achieve integrated automatic driving, expanding autopilot from individual intelligence to swarm intelligence. Autopilot development must be based on practical applications: it is necessary to continuously obtain effective training data from specific application scenarios for autopilot systems to solve the problem of insufficient autopilot datasets and continuously improve the quality of the autopilot model. Besides, inter-vehicle and vehicle-environment communication must be realized in the process of automatic driving to avoid data deviation caused by single and insufficient information sources, and thereby improve the performance of the autopilot system.

With the continuous explorations of new business models of autopilot, it has been introduced to such fields as ride sharing, public transport, and logistics. Fine examples in this regard include the robotaxis of Waymo, self-driving trucks of Uber, and autonomous delivery vehicles of Nuro. Although automatic cars have appeared in reality, the automatic driving technology has not yet reached the highest level (L5), which is required in complex scenarios. Thus, autopilot is still not fully developed either in technology or application, and awaits further exploration.

The dilemma of autopilot lies in the fact that it must deal with one uncertainty, i.e., the vagaries of the human mind. If the AI system mainly depends on training, the autopilot system will inevitably fail to cope with complicated situations. In a complicated situation, the cause of an accident may be just a very small piece of data, which will definitely reduce in size when more tests are conducted on normal roads or simulative environments. However, these extremely small but important data must be considered during autopilot decision-making, and this turns out a major challenge in developing autopilot AI algorithms and systems. For example, how should the autopilot system act if the driver is determines to commit suicide in a car crash? A human may make the decision to slam on the brakes, but how does the autopilot system make such a judgment? Should we continuously create such data to train the autopilot system? Transfer learning provides a likely solution to this problem. As a matter of fact, all the vehicles can be coordinated under a unified scheduling algorithm to avoid accidents when only automatic driving cars are allowed on the road and the uncertainty caused by human drivers are removed. However, a closed-traffic scenario needs to be constructed to solve this problem.

11.2.5 Anthropomorphism and Bionics

In this section, we aim to understand how similar an AI system can be to the creatures in nature and humans. This is one of the most intensively-discussed questions in many AI-themed science fiction movies. These similarities are also being investigated by scientists who are drawing inspiration from the learning and thinking processes of human beings to develop cutting-edge technologies. For example, imitation learning is inspired by the learning process in which a child learns to imitate the motion of an adult; both transfer learning and meta learning correspond to human's capacity to "infer other things from one fact".; adversarial training can be seen as two AIAs

"playing chess;" explanable AI systems try to express the system has learnt in a form that humans can understand. There are also many examples of specific products, such as the robots and robot dogs of Boston Dynamics, which try to imitate the movements and characteristics of humans and dogs to the largest extent. The robot arm designed by a team at Google in collaboration with X Robotics attempts to imitate baby clawing to complete self-supervised training. Amazon, Google, Baidu, Tencent, and some other companies have launched multiple AI personal assistants, whereas JD.com has attempted to open a smart restaurant.

Generally, the attempts in anthropomorphism and bionics are focused on certain specific aspects.

(1) **Behavior and Action**

We use the word "behavior" to refer to certain behaviors conforming to the characteristics of certain species. Boston Dynamics developed humanoids and robot dogs that function very similar to human beings and dogs, respectively. With the same size as human and creatures, they act in a way that conforms to physical rules, and when interacting with humans, they are much like real lives. For example, the humanoid would swing their arms when they jump off the stairs, and the robot dog runs, sniffs and scouts as a real dog does. Robotics seems to be a perfect example of AI, but in the strict sense, it belongs to the intersection of AI and the traditional control theory. This will be elaborated in Sect. 11.3.2.

If we can sufficiently simulate the behaviors and actions of various creatures, we can make machines act and behave as per the traits of the creatures in specific settings. However, "a specific setting" often involves complex factors and is hard to identify. Herein, we discuss the application of AIAs based on several traits of creatures, especially the traits of human beings, such as emotion, motivation, imagination, creation, as well as morality and ethics.

(2) **Emotion and Motivation**

One of the common propositions in science fiction movies is that AI is emotionless. In *Homo Deus: A Brief History of Tomorrow* [78], Yuval Noah Harari writes the following: "99% of our decisions—including the most important life choices concerning spouses, careers and habitats—are made by the highly refined algorithms we call sensations, emotions and desires."

With the development of science and technology, what used to be considered mysterious can be described and evaluated by mathematical models, for example, language, vision, and voice. Currently, many technologies has been expanded to calculation of emotion and motivation. For example, some researchers have attempted to design an emotional artificial neural network. Emotion enhancement was introduced into reinforcement learning. The emotion score was added to the reward function of reinforcement learning. Some researchers have studied the model of brain emotional learning [79], trying to construct AI systems according to the structural components of the human brain, including the hippocampus, orbitofrontal cortex, amygdala, thalamus, sensory cortex.

(3) **Imagination and Creation**

Literature and art works are born of unparalleled imaginative and creative powers intrinsic to human nature. Is the AI also equipped with the ability of artistic creation? The US-based Thomas Financial network of firms have been using robots to write their financial news since 2006. In 2014, Associated Press wrote all articles using their robot WordSmith (a writing engine developed by an AI company) [80]. In cooperation with the German e-commerce firm Zalando, Google rolled out Project Muze [81], an AI fashion design product, in 2016. In addition, the AI drawing board developed by Monica Dinculescu from Google can automatically complete simple unfinished paintings.[4] A painting created by AI was sold for USD 430,000 at Christie's [82]. The song arrangement and part of MV making for *Break Free*, a song by American singer Taryn Southern, were completed using AI [83].

These events seem inconceivable at first glance, but take a closer look, and you will find that they rely on the AI technologies we have already introduced before. To equip the machines with the power of imagination, we used causal reasoning, GAN, reinforcement learning and transfer learning in clearly described scenarios, and representation learning and meta learning when the scenario is more abstract.

(4) **Morality and Ethics**

How to align morality and ethics of AI with humans'? This remains a question that merits attention from AI researchers. As the machine acts as instructed by the inner algorithm, designing an ethical algorithm has become an active research field in recent years. In December 2016, IEEE released the first edition of *Ethically Aligned Design: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems* [84], which was co-authored by more than 100 leading thinkers and experts in the field of AI, ethics, and other related fields worldwide.

Currently, some technologies can be used to reduce the ethical risks associated with AI to some extent. For example, learning appropriate representations can improve fairness and reduce the likelihood of discrimination of AI algorithms. Ethical concerns can be introduced to the design of the reward function in reinforcement learning. Federated learning can reduce the privacy concerns amid the fast development of machine learning. Explainable AI facilitates introduction of moral and ethical constraints to AI applications. In 2019, *Nature* published a paper on machine behavior, which stated that the machine behaviors must be studied on a multidisciplinary basis similar to the manner in which we study a creature's behavior [85]. Similar studies may provide new inspirations for exploration of AI ethics.

11.2.6 Intelligent Attack and Defense

With the increasing complexity of network environments, security concerns are increasing and security threats are spreading quickly, resulting in a wide impact.

[4] Refer to magic sketchpad. https://magic-sketchpad.glitch.me/.

Conventional security defense methods, which rely heavily on expert knowledge and rigid rules, often react slowly to threats, and due to lack of security experts, these methods can hardly tackle the severe security challenges now and in the future. Consequently, automatic and intelligent attack and defense technologies begin to gain momentum. The development of AI provides strong technological support to intelligent attack and defense. By studying the current basic offense–defense data (e.g., security problems, virus samples, malicious traffics, and codes of vulnerability), a model of intelligent reasoning and forecast can be built for intelligent attack and defense to automatically assess the current situation and decide on attack or defense measures. The most popular applications of intelligent attack and defense for now include network attack detection, malicious object detection, vulnerability discovery, ID authentication, and content safety.

(1) **Network attack detection:** the AI^2 (artificial intelligence + analyst intuition) system developed by MIT CSAIL and PatternEx, a machine learning startup, in 2016, is a representative application of AI in network attack detection [86]. It was already been introduced in Chap. 4 and are not to be elaborated here.

(2) **Malicious object detection:** AI can be applied to detect viruses and malware and other security threats to web pages and URLs. Leading IT companies, such as Microsoft, Google, Baidu, Tencent, and 360, have launched a range of AI-based security products, including the TRP–AI antivirus engine [87] and 360QVM AI antivirus engine.[5] Such products construct a detection model by examining large samples of viruses or malware and implementing fast detection on known or unknown viruses or malware. However, the attackers can also upgrade and rebuild viruses or malware quickly using AI technology. For example, attackers can generate malware or viruses able to bypass the detection model [88] by GAN, or mine the vulnerabilities of the detection model to tamper the malware and launch attacks [89]. AI technology has also set the stage for customized attacks. DeepLocker [90] mentioned in Chap. 4 is an AI-driven malware with strong concealment and directivity.

(3) **Vulnerability mining:** vulnerability is always a major threat to network security and is a common cause of network attacks. Vulnerability mining works both ways: it can help the defender to improve the defense capacity of the network, while at the hand of attackers, it can be used to launch network attacks. Intelligent vulnerability mining has received extensive attention since the Cyber Grand Challenge (CGC) in 2016. The Mayhem system (the champion of CGC), which could detect, patch, and leverage the software vulnerability with AI technology, won favor from the United States Department of Defense, and the earlier version of this system adopted online-offline hybrid symbolic execution [91]. Though study on automatic vulnerability mining has gradually matured in these years, research on intelligent vulnerability mining techniques that fully utilize AI remains in its infancy.

[5] Refer to the Entry of Baidu Baike: 360qvm Artificial Intelligence Engine.

- Vulnerability mining based on static analysis has been deeply combined with AI technology. Different from conventional vulnerability mining methods that are based on manual vulnerability rules, VulDeePecker [92] applies deep learning to detection of vulnerabilities in the source codes; the labeled training programs are transferred into semantic vectors with code gadget as the unit, and the vectors are then input to the BiLSTM to train and develop a vulnerability detection model. This model can replace the manual vulnerability rules to detect vulnerabilities of programs.
- Vulnerability mining based on dynamic analysis is mostly related to American fuzzy lop (AFL). AFL combines fuzzing and symbolic execution to improve the code coverage during the fuzzing process by forming different input samples, thereby increasing the success rate of discovering vulnerabilities. AI technology can be used to optimize the mutation of seeds such as AFLFast [93] and AFLGo [94].

  Although the advantages of AI technology in vulnerability mining has not been fully utilized, its feasibility and potential have already been confirmed. How to fully utilize AI to realize intelligent vulnerability mining is bound to become the next research hotspot. AI systems are in essence programs which are also susceptible to vulnerability-based attacks, so AI-oriented vulnerability mining is also a topic worth exploring in the future.

(4) **ID authentication:** with the development of big data and intelligent terminals, authentication methods based on biometrics (fingerprints, iris, and facial features) and behavioral features (handwriting, voice, and gait) have become popular in addition to conventional password-based authentication. AI has facilitated the development of ID authentication, but meanwhile brought new risks to it. Attackers can easily bypass ID authentication or ignore it using AI. There have been some cases in point. Attackers have used PassGAN to crack the user passwords and created a large number of almost real and effective user passwords to cover the users' actual passwords to the greatest extent [95]. Using GAN and the latent variable evolution method, attackers could create deep master prints [96], which achieved a success attack rate as high as 76.67%. Similarly, facial features can be forged via AI to spoof ID authentication systems [97]. AI-based ID authentication and anti-spoofing will become another hot research direction of AI attack and defense.

(5) **Content security:** Attack-and-defense also exists in content recognition and forgery. AI technology can detect malicious contents in multiple fields. For example, Google developed the Perspective API to recognize malicious comments [98]. This API trains the model with labeled malicious comments, scores online comments by their similarity to malicious comments, and hence quickly detects malicious comments. Facebook has similar APIs to detect and restrict malicious contents [99]. E-commerce sites are also developing new systems to filter fake comments and defend against bots. However, AI bots that post fake comments emerge, which intensively train RNNs based on the

publicly-available comments form e-commerce platforms to generate categorized and customized fake comments that are difficult to recognize [100]. Fake comments cannot be detected by machines, and are sometimes hard to distinguish by human. With the constant development of Deepfake [101], the battles related to comments will be intensified.

In the case of intelligent attack and defense, AI is much like a curate's egg: its powerful understanding and inference capacity makes it a useful tool for defenders to quickly find, analyze and respond to security threats; however, at the hand of attackers, it will be utilized to bypass security systems. In the future, intelligent attack-and-defense will no longer be attack-and-defense between AI defense and conventional attacks, but between AI-powered defense and AI-powered attacks.

11.2.7 Security Challenges in Frontiers of AI Application

AI has been combined with a wide spectrum of advanced technologies in the frontiers of application, bringing about revolutionary changes to our life and production. With the continuous exploration of AI applications, the accompanied security problems begin to bulk large. Some of these problems can be attributed to the vulnerability of the underlying AI technology, while some others are security threats from real-world application scenarios. The technical security challenges in AI have already been detailed in Sect. 11.2.6. Therefore, this section will focus on the challenges we face in applying AI to empower security, as well as the endogenous security problem (AI security) and deriviative security problem (AI safety) of AI applications.

Applications like computer vision, NLP, and intelligent speech provide ample evidence to AI's strength in content recognition and creation. Thus, content-level security problems are among the major concerns in these applications. There are four aspects of content-level security challenges. First, there are threats when AI is applied to empower concent-level security. Computer vision, NLP, intelligent speech, and other applications can filter illegitimate contents from the massive throng of materials online, such as pornographic content, violent pictures, and audio or video files. This type of AI application can also be used for copyright protection to detect piracy on the Internet. However, AI-based content-level attack measures have evolved as well. For instance, normal images and videos may contain malicious textual information, which makes it hard to detect harmful semantic information from the texts by relying only on the images or videos. Besides, privacy leakage is another conundrum hard to cope with in content recognition. The second challenge is AI security problems caused by AI adversarial attacks. This challenge is, in fact, security problems in application caused by technical security problems of AI (AI-powered defense against attacks), such as adversarial attacks. The third challenge is the derivative security problem incurred in content recognition. AI-based recognition techniques are not 100% accurate because of the limitations of AI technology, which results in possible safety threats for AI applications in some specific scenarios. Let's assume

two scenarios: (i) daily recognition of plants and animals and (ii) automatic driving. The former can tolerate small errors, while the latter demands high accuracy in that vague recognition results may lead to false warnings and then result in severe traffic accidents. The fourth challenge is derivative security problems in content creation. These problems are caused by the abuse of cutting-edge AI applications for illegal intent, including fraudulence, thefts, and other illegal activities using forged pictures, 3D images, videos, fingerprints, or voice prints.

The autopilot systems, anthropomorphic AI and simulation applications involve multiple functions including recognition and creation, and hence are subject to all the abovementioned content-level security problems. In addition, they have specific functions, i.e., intelligent decision-making and control, which allow them to adopt different actions as per the specific situations to preclude accidents and reduce losses. The intelligent decision-making and control functions have freed human from such tedious tasks as selecting information, making decisions, and controlling, thereby improving the comfort and efficiency of work. Nonetheless, they incur a slew of new derivative security problems, as listed below.

(i) Uncontrollability of artificial intelligence actants (AIAs). One case in point is the Boeing 737 Max 8 crash in Indonesia [102]. The information from the black box of the aircraft showed that the automatic stall-prevention system delivered false judgements, and the pilot had to "battle" the system when trying to take control; consequently, the plane bumped up and down more than 20 times in 11 min before finally crashed. Uncontrollable AIAs will introduce huge security risks during the interactions with human beings.
(ii) Moral and ethical concerns. AIAs are endowed with the power to make decisions, but what follows are moral and ethical concerns in the decision-making. Autopilot systems are very likely to face moral dilemmas like the trolley problem[6] when put to practice.
(iii) Accountability problems. It still remains unclear if an AIA can assume criminal liability independently and who should be held liable for the consequences when an AIA causes an accident.

Intelligent attack and defense is a representative application of AI for security. Using AI applications, intelligent attack-and-defense systems can accurately and efficiently detect cyberspace attacks and improve the system's defense capacity. When put at the hand of attackers, however, the systems can be used to bypass the security defense and launch attacks. In other words, in the attack-defense game, the attack and defense skills improve alternately. Intelligent attack-and-defense may also beget derivative security problems (AI safety problems). AI attack-and-defense applications have lowered the technical bar of cyberspace attack and defense, and

[6] Trolley problem is one of the most renowned thought experiments in ethics. In the general form of this problem, a madman ties several hostages on a track. A trolley out of control is going to run them over in a few seconds. Fortunately, you can pull a hand lever to switch the tracks. But, there is a problem. A man is tied up to another track as well. Considering everything, would you pull the lever?

when misused for illegitimate purposes, these applications will pose security threats to network regulation, tracking, and forensics.

11.3 Crosscutting Areas

The impact of AI has been deeply felt across industries, while the development of other industries in return spawns new AI algorithms and applications. This section will present the manner in which AI interacts with and influences other technologies.

11.3.1 AI and the Internet

AI technology, as one of the cutting-edge technologies of the Internet, and other Internet technologies are mutually reinforcing: the former offers new solutions to the latter and meanwhile achieves faster development on the basis of latter. In this book, we explore the crosscutting applications of AI and some popular Internet technologies.

AI and big data management Big data technology in the broad sense encompasses all technologies related to big data. In this sense, AI technology also belongs to big data technology since its accuracy relies largely on sufficient training data. However, the big data management that we discuss refers to a series of underlying technologies related to big data per se, including big data storage, indexing, querying, processing, and computing architecture. The development of AI technology has led to changes in big data storage and indexing. For instance, knowledge graphs have encouraged the development of graph databases; AI expands the research scope of big data search and computing into more refined areas like image retrieval and subimage matching. Therefore, new AI technologies have been extensively applied to the query and analyses of big data. In return, the advances in big data management has also led to changes in AI. More AI algorithms are updated to adapt to the new big data storage and computing architecture, and the new query and processing algorithms make the data acquisition and preparation for AI applications more convenient.

AI and IoT An excellent juncture at which AI meets IoT is edge computing. By training the model on a public cloud and deploying the model on the edge devices, we can considerably reduce the computing load of the cloud and meanwhile improve the flexibility and efficiency of edge devices in problem-solving. The industrial internet of things (IIoT) is a typical case in point for combination of AI with IoT. Automatic progress monitoring, abnormality monitoring, data analysis, and equipment maintenance in IIoT can be achieved using AI technology. Deep neural networks can equip edge devices with capabilities to process and analyze diverse structured and unstructured data such as videos, voices, and texts. Besides, IoT has become a major driver for the development of enterprise AI. To effectively apply AI technology to

IoT, the AI algorithms or models must become more efficient and lightweight, and the AI chips must be able to be deployed in various IoT devices.

AI and blockchain Blockchains can provide certain assurances for the application of AI technologies, including audit trails [103] and AI decentralized autonomous organizations (AI DAOs) [104]. The antitampering and anticounterfeiting features of blockchain facilitate audit trails. A clear audit trail can enhance the credibility of data and models and help to track the AI decision-making process and improve interpretability. The AI DAO conditionally writes AI programs that can be executed by itself to enhance the safety of AI applications. AI technology, in return, has boosted development of blockchain by providing technical support to mine blockchain data, optimizing energy consumption and improving security of top-level blockchain applications.

AI and quantum computing AI algorithms can provide countermeasures against quantum noise and calculate the error rates in quantum computing. In return, the strong computing power of quantum computing has led AI to revolutionary leaps forward. For example, the QNN model [105] proposed by Google attempts to boost the efficiency of neural networks via quantum computing.

11.3.2 AI and Industry 4.0

"Industrial Internet" is the fourth industrial revolution with intelligent manufacturing as the central idea, and aims to transform traditional manufacturing into intelligent manufacturing. Amid this revolution, the main AI technologies include intelligent perception, robot manufacturing and application, and industrial big data analysis.

AI and Intelligent Perception Intelligent perception, the first link in intelligent manufacturing, is to collect data from the source. It is expected to fulfill two roles: supporting efficient data collection, and assisting in determining which data are important. The former relies on efficient data collection and processing algorithms, whereas the latter requires lightweight AI algorithms operable on edge devices. Take for example the smart cameras. The camera not only needs to efficiently extract objects from a video, but also identify and capture abnormal images, and automatically send alerts. Intelligent perception has extended the application prospect of AI and provides a new development path for this realm.

AI and robot manufacturing and application Amid the rapid development of the intelligent manufacturing, the dominating trend is to improve the cooperation between robots and humans and even replace humans with robots at work. Industrial robots can perform functions such as welding, assembling, collecting, placing, testing, and detecting products, in a way more efficient, durable and accurate than their human counterparts. Robot manufacturing and application have become a hotspot in the current technological developments, and many countries have initiated strategic plans in this regard. Robot manufacturing, at first glance, belongs to AI technology, but with a closer look, we may find that it is more appropriate to classify it as a crosscutting realm between AI and the classical control theory. As

discussed in Sect. 11.2.5, the founder of Boston Dynamics commented on many occasions that the manufacturing of their world-renowned robots and robot dogs was considerably dependent on the traditional motion control technology [106]. This was because the debugging and optimization of the robot hardware was an extremely lengthy and complicated process, and it would be difficult to obtain the hardware parameters by automatically learning AI algorithms that require massive data accumulation from other fields. Nonetheless, AI technology clearly remains a main factor that influences the development of robot manufacturing and application. Besides its indisputable applicability to software development, AI also offers new solutions to hardware debugging. For example, based on machine learning, several optimal parameter values can be recommended.

AI and Industrial Big Data Analysis Big data analysis among the strong suits of AI. When applied to industrial big data analysis, AI can extract features based on industrial data and build effective models to obtain practical results in digital design, Internet-based monitoring, and predictive maintenance. For example, based on the observation data of the current status of the device, the historical data on operation of the device, and the loss model of the parts, AI can be used to infer the loss of each component of a device and prepare replacements for parts that are likely to go wrong [107]. However, different industries present different features and require different models, which poses additional challenges to the advances of AI.

11.3.3 AI and Medical Biology

International Data Corporation (IDC) once predicted in the *Developments and Trends of China's Healthcare Artificial Intelligence* that the market value of AI diagnosis and treatment services would reach RMB 5875 million by 2022. The combination of AI and medical biology have gained increasing attention in these years. DeepMind[7] announced in 2018 that AlphaFold, a new member of its Alpha series family, could apply AI technology to predict the 3D structure of proteins [108]. Li Feifei discussed in an academic publication how to apply AI to diagnose depression [109], and the CheXNeXt model proposed by Wu Enda's team can diagnose 14 diseases [110]. Currently, AI technology has seen adoption in such medical realms as medical imaging, health management, and drug research and development (R&D).

AI and Intelligent Imaging Diagnosis AI technology can be applied to medical imaging and diagnosis because of its strength in audio and video processing. Cases for application of AI in medical imaging include disease screening, lesion delineation, and 3D imaging of organs. Specifically, AI can be used to classify images and locate organs, regions, and markers; it can also be used to segment tissues, structures, and

[7] DeepMind is a London-based start-of-the-art AI tech company, which was incorporated by Demis Hassabis (AI programmer and neuroscientist) and others. It has developed powerful general learning algorithms by integrating the most advanced technologies of machine learning and systems neurosciences.

lesions. Medical imaging data are massive and have a neat format, which are suitable inputs for AI algorithms. However, labelling of these data relies on the professional medical knowledge in different sectors, so its threshold and cost are higher than other fields [111].

AI, Intelligent Diagnosis, Treatment, and Health Management There are many intersections between AI, intelligent diagnosis and treatment, as well as intelligent health management. For example, hospitals can enter their medical records using intelligent speech technology, saving the time of doctors who would otherwise need to manually enter the medical records to give patients a clear medical record. It is the same case with intelligent consultation, through which diagnostic advice is offered by combining technologies such as front-end intelligent speech, back-end NLP, computer vision, and knowledge graph. However, it should be noted that when samples for AI-powered medical applications are not enough, entries in medical records and the automatic diagnosis result may go wrong. Therefore, we cannot place absolute trust in the intelligent diagnosis and treatment technology for the time being, and there is still a long way to go before AI-based medical applications play a fully reliable role. That being said, AI can be applied in the field of medicine for "less serious" scenarios: In intelligent health management, for example, AI can make appropriate recommendations for work and rest based on users' historical data (including exercise, diet, sleep, and physique) and relevant medical knowledge; some smart homes and wearable devices have already utilized intelligent health management for commercial purposes.

AI and Intelligent Drug R&D How to apply AI to drug R&D is another area under active exploration. For example, recommendations for the quality and ratio of compounds in drugs may be offered using AI technology. A report by TechEmergence [112] shows that billions of dollars can be saved for the biopharmaceutical industry if AI technology could increase the success rate of new drug R&D from 12 to 14%.

With the help of AI technology, we can better identify the implicit relations among drugs, diseases, and genes, and provide recommendations for selecting more appropriate candidate compounds. Yet, despite the convenience that AI has brought, as with the case of intelligent diagnosis and treatment, we must guard against possible errors made by AI for a minor error in these areas may cause severe consequences.

11.3.4 AI and Military Applications

The military applications of AI have been promoted by many countries as a national strategy, and such applications will introduce the next revolution in military technology. The ecosystem of military applications of AI includes intelligent collection and storage of military data (at the bottom level), data analysis and calculation (at the middle level), and planning, intelligence, and situation display (at the top level). This revolution would involve intelligent weapons, intelligent command, and new forms of warfare.

AI and Intelligent Weapons One of the most discussed categories of intelligent weapons is the lethal autonomous weapon system (LAWS), which includes unmanned aerial vehicles, unmanned ground vehicles, unmanned surface vehicles, and other weapon systems that can operate autonomously. Development of technologies, such as computer vision, autopilot, and bionic AI, has considerably stimulated the lethal autonomous weapon system. Certain requirements of war scenarios, such as limited computing power, complex terrains, and communication difficulties, present special challenges to the advances in AI technology. The potential hazards caused by intelligent weapons have captured widespread attention from the academia and industry. At the International Joint Conference on Artificial Intelligence in 2018, the Future of Life Institute released an open letter titled Lethal Autonomous Weapons Pledge, which called on governments and leaders from different countries to enact stringent and effective international norms, regulations, and laws to prohibit the use of lethal autonomous weapons. As of November 7, 2019, this open letter was signed by 3253 individuals, including Elon Musk (the CEO of Tesla) and three cofounders of Google's DeepMind. In addition, 247 organizations signed this open letter and pledged that they would not participate in or support the development, manufacturing, trading or use of lethal autonomous weapons. The Future of Life Institute called on technological companies and organizations, leaders, policy-makers, and other individuals to sign this pledge [113].

AI and Intelligent Command Intelligent command relies on the analysis of military big data. In modern warfare, data are from diverse sources and vary in formats, which is both an opportunity and a challenge to the development of AI technology. On the one hand, the increase in data enables AI technology to play a major role in the analysis of military situations and intelligence; on the other hand, problems like poor data quality, susceptibility to environmental interferences and attacks, and highly-repetitive data bring enormous challenges to application of AI.

AI and New Forms of Warfare Under the current technical conditions, AI technology can not only improve the existing military technology, but also create new forms of warfare. Facebook used the targeted push strategy to effectively influence and control user emotions, and successfully controlled the emotions of countless users [114]. It was widely believed that the CIA may be involved in this experiment [115]. DARPA, the U.S. Department of Defense, led an industry team in applying AI technology to electronic warfare and generated a new form of warfare called "cognitive electronic warfare" [116].

11.3.5 AI and Business Models

There are plenty of possibilities for combination of AI with business. In this book, we make a brief discussion of AI and intelligent product recommendation, intelligent investment consultanting, and automatic trading systems.

AI and Intelligent Product Recommendation Intelligent product recommendation has been widely applied in myriad business fields. AI algorithms are not

only used in online advertising and product recommendation, but also used by many companies to support offline sales in stores. Based on online and offline applications, merchants could collect users' historical purchase data, analyze their purchase preferences, compare the data with those of the users sharing the same preference to push customized advertisements and product promotions, and offer discounts accordingly. In the case of the recommendation algorithms that are currently applied, the recommendation results are relatively unsophisticated and are mostly restricted to certain types of goods with limited diversity. Therefore, we need to resolve problems such as how to establish more effective models for user behaviors and how to better describe product features.

AI and Intelligent Investment Consulting Intelligent investment consulting applications automatically adjust investment portfolios based on the clients' revenue target and risk tolerance. Securities companies, banking institutions, and Internet financial platforms have gradually introduced intelligent investment consulting products such as Machine Gene Investment of China Merchants Bank, the intelligent investment advice of JD.com, and the intelligent wealth management assistant of Ant Fortune. However, there is still much room to expand the application of "AI algorithm plus big data" in the financial sector, and whether these intelligent consultants can outperform their experienced human counterparts remains unclear.

AI and Automated Trading Systems The automated trading system (ATS) automatically makes trading decisions, places orders, and manages orders and funds using certain algorithms. When the data are sufficient and the model is good enough, the ATS built based on AI algorithms may even outperform human experts. Yet, to achieve better results, we should not hinge our hopes merely on the increase of data and optimization of AI algorithms because the effectiveness of the algorithms largely relies on human knowledge and the constraints imposed on algorithms by humans based on their own knowledge in the current technologies. For example, humans may set the system to sell when the price drops below certain limits. However, massive selling may occur if most programs are set similarly, resulting in further price drops and more selling by buyers; this would be disastrous. In addition, there remain many factors that contribute to economic changes but we human cannot fully fathom. These factors will, in most cases, be overlooked in the design of AI applications, which may result in inaccurate results. Therefore, the design of an excellent ATS relies not only on the increase in data and progress of AI algorithms, but also on the theoretical development of business and economy.

11.3.6 AI and Humanities and Society

The development of AI technology not only presents new opportunities and challenges to science and technology fields, but also has profoundly impacts on humanities and the society. AI technology has been currently applied to multiple areas such as law, journalism, history, and archaeology.

AI and Law Though it would be wishful thinking to completely replace lawyers with AI before the emergence of superintelligence, AI can significantly boost the efficiency of legal work in two ways. First, AI technology can act as an assistant to lawyers and help automate contract reviews, data collection and management, document processing, and other tasks. Second, AI technology can serve as an assistant to consumers and helps them fulfill undemanding tasks such as legal consulting and automatic lawyer recommendations. A few products have been developed in these regards. ROSS [117], an AI lawyer developed by IBM, can help lawyers cope with their daily routines and is now being used at the New York-based law firm Baker and Hostetler. The robot lawyer DoNotPay [118] mainly targets consumers. Consumers can consult DoNotPay for legal issues, and the robot will help them decide whether there is a chance of appeal and offer further guidance.

AI and Journalism Similar to product recommendations, AI can be applied to realize personalized news recommendations and beyond in the field of journalism. By integrating NLP and social network analysis based on AI technologies, AI news assistants can help journalists collect and process data, such as reminders of important news and events, analysis of news popularity, and prediction of social media trends. Nonetheless, the advances in AI technologies have spawned a range of new problems for the news industry. For example, fake news can be easily created because AI technologies can be used to master the writing patterns of real news and create fake images or even videos that are as good as real. If these technologies are used to create and spread fake news, then the public may be confused and even deceived.

AI and History AI technologies can be applied to diverse aspects of history research and archaeology. Many paper historical documents need to be converted into electronic ones, and as most of these documents were handwritten rather than printed, the image recognition technology is required to recognize the handwritten texts. In particular, as ancient Chinese characters are likely to be blurred or even damaged due to improper protection, AI technologies are often required for restoration. Historical documents also need to be converted into data that can be processed and understood by computers. The writing pattern and wording of historical documents, due to cultural evolvement, differ vastly from those in the modern articles, which poses a new challenge to NLP. For example, the China Biographical Database (CBDB) [119] is an attempt to make historical data electronic and digital. Certain progress has also been made in deciphering ancient scripts: using AI, researchers managed to decipher 67.3% of the Linear B of Mycenaean civilization in ancient Greece for the first time [120]. AI also plays a helpful role in archaeological excavations and research: image synthesis and 3D synthesis technologies based on AI algorithms can help restore damaged cultural relics, ancient cities, and sculptures; in the future, robot manufacturing and machine vision can be applied to automatic archaeological excavations in harsh environments.

11.3.7 Safety Challenges in Crosscutting Areas

The safety challenges in crosscutting areas tend to be complicated. On one hand, AI helps resolve the safety problems in these fields; on the other, it may also cause manifold security and safety problems to these fields because of abuse or poor regulation, resulting in derivative security problems of AI (AI safety problems). In this section, we will focus on the safety problems caused by AI in crosscutting areas.

We must devote our attention to safety incidents specific to certain crosscutting areas. When AI technologies are applied to the IoT or industrial control systems, some automatic adjustment functions based on AI technologies (e.g., adjustment to the voltage, current, temperature, and angle of the robotic arm) need to be strictly reviewed. This is because abnormal adjustments may actually threaten the safety and lives of humans engaged in the production process rather than cause problems, such as program crash or network paralysis, as in the traditional Internet. When AI is applied to medical and financial scenarios, potential privacy leakage must be avoided and users' private data cannot be misused for the sake of accuracy. When AI is applied to military, medical, and other domains, a strict ethical review is required to prevent possible moral and ethical risks.

In some crosscutting areas, accuracy of AI means safety. For example, when AI technologies are applied to the medical field, their accuracy must be high under scenarios such as lesion identification, disease discrimination, and drug dosage calculation because inaccurate output by AI will threaten the life and safety of patients. Similarly, when AI technologies are applied in the military field, the target recognition accuracy will directly impact the life and safety of soldiers and civilians. When AI technologies are applied to commerce and finance, the accuracy of the identity recognition and analysis results will directly influence the safety of the transactions.

In some other crosscutting fields, strict precautions should be adopted against frauds and attacks to AI. For example, shilling attacks may be launched against the recommender system when AI technologies are applied to achieve intelligent product recommendations. With respect to news and media, AI may be used to forge news and identities, maliciously misleading the public opinions. In the legal field, AI may be used to create audios, videos, images, or texts to fabricate evidence.

11.4 Summary

This chapter explores AI from the perspective of the academic frontiers, application frontiers, and crosscutting areas and analyzes the safety challenges that may be faced by the AI applications. AI technologies are bound to develop rapidly and find their way into all aspects of human lives in the foreseeable future. Security and safety will remain an issue accompanying the development of AI. We can develop cutting-edge AI technologies and applications to address various safety concerns. However, we must be adequately prepared for AI security and safety problems caused by defects

or the improper use of new AI technologies and applications. Because of limited length of the work, this chapter gives only a brief overview, while more in-depth technical details can be obtained from the references.

References

1. Ulam S. Tribute to John von Neumann, 1903–1957. Bull Am Math Soc. 1958;64(3):1–49.
2. Kurzweil R. The singularity is near. In: Sandler RL, editor. Ethics and emerging technologies. London: Palgrave Macmillan; 2014.
3. Sutton RS, Barto AG. Reinforcement learning: an introduction. Cambridge, Massachusetts: MIT press; 2018.
4. Google Dopamine. [2019–03–20]. https://github.com/google/dopamine.
5. Mnih V, Kavukcuoglu K, Silver D, et al. Human-level control through deep reinforcement learning. Nature. 2015;518(7540):529–33.
6. Silver D, Huang A, Maddison CJ, et al. Mastering the game of Go with deep neural networks and tree search. Nature. 2016;529(7587):484–9.
7. Graves A, Wayne G, Reynolds M, et al. Hybrid computing using a neural network with dynamic external memory. Nature. 2016;538(7626):471–6.
8. Pan SJ, Tsang IW, Kwok JT, et al. Domain adaptation via transfer component analysis. IEEE Trans Neural Netw. 2011;22(2):199–210.
9. McMahan HB, Moore E, Ramage D, et al. Communication-efficient learning of deep networks from decentralized data. In: ICAI (International conference on artificial intelligence); 2017
10. Zinkevich M, Weimer M, Li L, et al. Parallelized stochastic gradient descent. In: Neural information processing systems; 2010. p. 2595–2603.
11. Dean J, Corrado G, Monga R, et al. Large scale distributed deep networks. In: Advances in neural information processing systems; 2012. p. 1223–1231.
12. Coates A, Huval B, Wang T, et al. Deep learning with COTS HPC systems. In: International conference on machine learning; 2013. p. 1337–1345.
13. Yuan J, Gao F, Ho Q, et al. LightLDA: Big topic models on modest computer clusters. In: International conference on world wide web; 2015. p. 1351–1361.
14. Shazeer N, Mirhoseini A, Maziarz K, et al. Outrageously large neural networks: the sparsely-gated mixture-of-experts layer. [2019–11–6]. https://arxiv.org/abs/1701.06538.
15. Bengio Y, Courville AC, Vincent P, et al. Representation learning: a review and new perspectives. IEEE Trans Pattern Anal Mach Intell. 2013;35(8):1798–828.
16. Thrun S, Pratt L. Learning to learn. New York: Springer Science & Business Media; 2012.
17. Gaines BR, Compton P. Induction of meta-knowledge about knowledge discovery. IEEE Trans Knowl Data Eng. 1993;5(6):990–2.
18. Forman G. Tackling concept drift by temporal inductive transfer. In: Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval. ACM; 2006. p. 252–259.
19. Kaul A, Maheshwary S, Pudi V. Autolearn—automated feature generation and selection. In: IEEE international conference on data mining (ICDM). IEEE; 2017. p. 217–226.
20. Elsken T, Metzen JH, Hutter F, et al. Neural architecture search: a survey. J Mach Learn Res. 2018;20(55):1–21.
21. Li F. Announces Google cloud milestone: launch of cloud AutoML. (2018–01–18) [2019–11–7]. http://tech.163.com/18/0118/09/D8E3HI8K00098IEO.html.
22. Haifeng J, Qingquan S, Xia H. Auto-Keras: efficient neural architecture search with network morphism (2018). [2019–11–6]. https://arxiv.org/abs/1806.10282v3.

23. Baker B, Gupta O, Naik N, et al. Designing neural network architectures using reinforcement learning. [2019–11–6]. https://arxiv.org/abs/1611.02167.
24. Gaier A, Ha D. Weight agnostic neural networks. [2019–11–7] https://arxiv.org/abs/1906.04358.
25. Van Lent M, Fisher W, Mancuso M. An explainable artificial intelligence system for small-unit tactical behavior. In: Proceedings of the national conference on artificial intelligence. Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999, 2004. p. 900–907.
26. Ribeiro MT, Singh S, Guestrin C. Why should I trust you?: Explaining the predictions of any classifier. In: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining. ACM; 2016. p. 1135–1144.
27. Gunning D. Explainable artificial intelligence (XAI). Defense Adv Res Projects Agency (DARPA) 2017
28. Kim B, Khanna R, Koyejo OO. Examples are not enough, learn to criticize! criticism for interpretability. In: Advances in neural information processing systems; 2016. p. 2280–2288.
29. Liu S, Cui W, Wu Y, et al. A survey on information visualization: recent advances and challenges. Vis Comput. 2014;30(12):1373–93.
30. Letham B, Rudin C, McCormick TH, et al. Interpretable classifiers using rules and Bayesian analysis: building a better stroke prediction model. Ann Appl Stat. 2015;9(3):1350–71.
31. Kim B, Rudin C, Shah JA. The Bayesian case model: a generative approach for case-based reasoning and prototype classification. In: Advances in neural information processing systems; 2014. p. 1952–1960.
32. Xu K, Ba J, Kiros R, et al. Show, attend and tell: Neural image caption generation with visual attention. In: International conference on machine learning; 2015. p. 2048–2057.
33. Andreas J, Rohrbach M, Darrell T, et al. Neural module networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition; 2016. p. 39–48.
34. Bau D, Zhou B, Khosla A, et al. Network dissection: quantifying interpretability of deep visual representations. In: Proceedings of the IEEE conference on computer vision and pattern recognition; 2017. p. 6541–6549.
35. Goodfellow I, Pouget-Abadie J, Mirza M, et al. Generative adversarial nets. In: Advances in neural information processing systems; 2014. p. 2672–2680.
36. NVIDIA TESLA. [2019–10–28]. https://www.nvidia.cn/data-center/tesla/.
37. NVIDIA TITAN RTX. [2019–10–28]. https://www.nvidia.cn/titan/titan-rtx/.
38. Jouppi NP, Young CS, Patil N, et al. In-Datacenter performance analysis of a tensor processing unit. Int Symp Comput Archit. 2017;45(2):1–12.
39. A new window to observe cutting-edge technologies—world internet conference unveils multiple global top tech achievements. (2016–11–7) [2019–11–6]. http://www.xinhuanet.com//politics/2016-11/17/c_129366711.htm.
40. Fifteen world-leading internet technology achievements released in Wuzhen, Zhejiang. (2019–10–20) [2019–11–6]. http://www.xinhuanet.com/fortune/2019-10/20/c_1125128823.htm.
41. About Cambricon. [2019–10–28]. http://www.cambricon.com/index.php?c=page&id=9.
42. HUAWEI reveals the future of mobile AI at IFA 2017. (2017–9–2) [2019–10–28]. https://consumer.huawei.com/en/press/news/2017/ifa2017-kirin970/.
43. Huawei releases ascend 910, the most powerful AI processor, and mindspore, a full-scenario AI computing framework. (2019–8–23) [2019–10–28]. https://www.huawei.com/cn/press-events/news/2019/8/huawei-ascend-910-most-powerful-ai-processor.
44. Fifteen world-leading internet technology achievements released at the 5th world internet conference. (2018–11–7) [2019–11–6]. http://media.people.com.cn/n1/2018/1107/c14677-30387848.html.
45. Ouyang J, Wu E, Wang J, et al. XPU—a programmable FPGA accelerator for diverse workloads. [2019–10–28]. https://www.hotchips.org/wp-content/uploads/hc_archives/hc29/HC29.21-Monday-Pub/HC29.21.40-Processors-Pub/HC29.21.410-XPU-FPGA-Ouyang-Baidu.pdf.
46. INTEL Nervana neural network processors. [2019–10–28]. https://www.intel.ai/nervana-nnp/.

47. Radeon instinct accelerator. [2019–10–28]. https://www.amd.com/zh-hans/graphics/servers-radeon-instinct-mi.
48. Introducing a brain-inspired computer: TrueNorth's neurons to revolutionize system architecture. [2019–10–28]. http://www.research.ibm.com/articles/brain-chip.shtml.
49. Snapdragon 855 smartphone platform. [2019–10–28]. https://www.qualcomm.cn/products/snapdragon-855-mobile-platform.
50. Understanding GPU family 4. [2019–10–28]. https://developer.apple.com/documentation/metal/gpu_features/understanding_gpu_family_4.
51. Arm unveils project trillium: delivering the industry's most scalable and widest-application machine learning computing platform. [2019–10–28]. https://community.arm.com/cn/b/cn-news/posts/cn-arm-launch-project-trillium-2018.
52. Poison frogs! targeted poisoning attacks on neural networks. [2019–10–28]. https://www.cs.umd.edu/~tomg/projects/poison/.
53. Tramer F, Zhang F, Juels A, et al. Stealing machine learning models via prediction APIs. [2019–10–29]. http://silver.web.unc.edu/files/2016/06/ml-poster.pdf.
54. Light R. AI trends: machine learning as a service (MLaaS). (2018–1–18) [2019–10–29]. https://learn.g2.com/trends/machine-learning-service-mlaas.
55. Goodfellow I, Shlens J, Szegedy C, et al. Explaining and harnessing adversarial examples. [2019–10–29]. https://arxiv.org/abs/1412.6572.
56. Brock A, Donahue J, Simonyan K. Large scale GAN training for high fidelity natural image synthesis. [2019–10–29]. https://arxiv.org/abs/1809.11096.
57. Zhu JY, Zhang Z, Zhang C, et al. Visual object networks: image generation with disentangled 3D representations. In: Advances in neural information processing systems; 2018. p. 118–129.
58. Wang TC, Liu MY, Zhu JY, et al. Video-to-video synthesis. [2019–10–29]. https://arxiv.org/abs/1808.06601.
59. Farhi E, Neven H. Classification with quantum neural networks on near term processors. [2019–10–29]. https://arxiv.org/abs/1802.06002.
60. Peters ME, Neumann M, Iyyer M, et al. Deep contextualized word representations. [2019–10–29]. https://arxiv.org/abs/1802.05365.
61. Devlin J, Chang MW, Lee K, et al. Bert: pre-training of deep bidirectional transformers for language understanding. https://arxiv.org/abs/1810.04805?context=cs.
62. Howard J, Ruder S. Universal language model fine-tuning for text classification. In: Proceedings of the 56th annual meeting of the association for computational linguistics (Volume 1: Long Papers); 2018, vol 1. p. 328–339.
63. Wolf T, Chaumond J, Delangue C. Meta-learning a dynamical language model. [2019–10–29]. https://arxiv.org/abs/1803.10631.
64. Dahl GE, Yu D, Deng L, et al. Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition. IEEE Trans Audio Speech Lang Process. 2011;20(1):30–42.
65. Graves A, Mohamed A, Hinton G. Speech recognition with deep recurrent neural networks. In: 2013 IEEE international conference on acoustics, speech and signal processing. IEEE; 2013. p. 6645–6649.
66. Sundermeyer M, Schlüter R, Ney H. LSTM neural networks for language modeling. In: The thirteenth annual conference of the international speech communication association; 2012.
67. Graves A, Jaitly N. Towards end-to-end speech recognition with recurrent neural networks. In: International conference on machine learning; 2014. p. 1764–1772.
68. Graves A, Jaitly N, Mohamed A. Hybrid speech recognition with deep bidirectional LSTM. In: 2013 IEEE workshop on automatic speech recognition and understanding. IEEE; 2013. p. 273–278.
69. Pratap V, Hannun A, Xu Q, et al. wav2letter++: the fastest open-source speech recognition system. [2019–10–29]. https://arxiv.org/abs/1812.07625.
70. Van Den Oord A, Dieleman S, Zen H, et al. Wavenet: a generative model for raw audio. [2019–10–29]. https://arxiv.org/abs/1609.03499v1.
71. Wang Y, Skerry-Ryan RJ, Stanton D, et al. Tacotron: towards end-to-end speech synthesis. [2019–10–29]. https://arxiv.org/abs/1703.10135.

72. Griffin D, Lim J. Signal estimation from modified short-time Fourier transform. IEEE Trans Acoust Speech Signal Process. 1984;32(2):236–43.
73. Shen J, Pang R, Weiss RJ, et al. Natural tts synthesis by conditioning wavenet on mel spectrogram predictions. In: 2018 IEEE international conference on acoustics, speech and signal processing (ICASSP). IEEE; 2018. p. 4779–4783.
74. iFlytek builds AI ecosystem barriers from multiple perspectives: AI chips to bridge the last mile. [2019–11–1]. http://www.aispeech.com/index.php?m=content&c=index&a=show&catid=22&id=373.
75. Taxonomy and definitions for terms related to driving automation systems for on-road motor vehicles. [2019–10–29]. https://www.sae.org/standards/content/j3016_201806/.
76. Tesla hardware 3 (full self-driving computer) detailed. [2019–11–2]. https://www.autopilotreview.com/tesla-custom-ai-chips-hardware-3/.
77. Tesla deaths: every tesla accident resulting in death. [2019–11–2]. https://www.tesladeaths.com/.
78. Harari YN. Homo Deus: a brief history of tomorrow. Random House; 2016.
79. Lucas C, Shahmirzadi D, Sheikholeslami N. Introducing BELBIC: brain emotional learning based intelligent controller. Intell Autom Soft Comput. 2004;10(1):11–21.
80. Xu M. Development and reflection on foreign robot news writers. (2016–1–5) [2019–10–29]. http://media.people.com.cn/n1/2016/0105/c401845-28014693.html.
81. Zalando's Project Muze: fashion inspired by you, designed by code. (2017–3) [2019–11–7]. https://www.thinkwithgoogle.com/intl/en-cee/success-stories/global-case-studies/zalandos-project-muze-fashion-inspired-you-designed-code/.
82. Huanqiu.com. First AI-generated painting sold for $432,000 at Christie's auction. (2018–10–26) [2019–10–29]. https://tech.huanqiu.com/article/9CaKrnKe5AT.
83. Exclusive interview with musician Taryn Southern: discussing her new album released via AI. (2017–9–27) [2019–10–29]. http://www.sohu.com/a/194906989_257855.
84. Design EA. A vision for prioritizing human well-being with artificial intelligence and autonomous systems. Inst Electr Electron Eng Incorporated (IEEE) Glob Initiative–2016; 2017.
85. Rahwan I, Cebrian M, Obradovich N, et al. Machine behaviour. Nature. 2019;568(7753):477–86.
86. Veeramachaneni K, Arnaldo I, Korrapati V, et al. AI^2: training a big data machine to defend. In: IEEE 2nd international conference on big data security on cloud (BigDataSecurity). IEEE; 2016. p. 49–54.
87. Tencent TRP-AI antivirus engine white paper. (2018–2–8) [2019–10–29]. https://slab.qq.com/news/authority/1744.html.
88. Hu W, Tan Y. Generating adversarial malware examples for black-box attacks based on GAN. [2019–10–29]. https://arxiv.org/abs/1702.05983v1.
89. Kolosnjaji B, Demontis A, Biggio B, et al. Adversarial malware binaries: evading deep learning for malware detection in executables. https://arxiv.org/abs/1803.04173v1.
90. Stoecklin M. DeepLocker: how AI can power a stealthy new breed of malware. Secur Intell. 2018;8.
91. Cha SK, Avgerinos T, Rebert A, et al. Unleashing mayhem on binary code. In: IEEE symposium on security and privacy; 2012.
92. Li Z, Zou D, Xu S, et al. VulDeePecker: a deep learning-based system for vulnerability detection. [2019–10–29]. https://arxiv.org/abs/1801.01681.
93. Böhme M, Pham VT, Roychoudhury A. Coverage-based greybox fuzzing as markov chain. In: IEEE transactions on software engineering; 2017.
94. Böhme M, Pham VT, Nguyen MD, et al. Directed greybox fuzzing. In: Proceedings of the 2017 ACM SIGSAC conference on computer and communications security. ACM; 2017. p. 2329–2344.
95. Hitaj B, Gasti P, Ateniese G, et al. Passgan: a deep learning approach for password guessing. [2019–10–29]. https://arxiv.org/abs/1709.00440.

96. Bontrager P, Roy A, Togelius J, et al. DeepMasterPrint: fingerprint spoofing via latent variable evolution. [2019–10–29]. https://arxiv.org/abs/1705.07386v3.
97. Brundage M, Avin S, Clark J, et al. The malicious use of artificial intelligence: forecasting, prevention, and mitigation. [2019–10–29]. https://arxiv.org/abs/1802.07228.
98. Li A. Jigsaw and Google using machine learning to identify toxic online comments. (2017–2–23) [2019–11–8]. https://9to5google.com/2017/02/23/jigsaw-google-perspective-machine-learning/.
99. Publishing our internal enforcement guidelines and expanding our appeals process. (2018–4–24) [2019–10–29]. https://newsroom.fb.com/news/2018/04/comprehensive-community-standards/.
100. Yao Y, Viswanath B, Cryan J, et al. Automated crowdturfing attacks and defenses in online review systems. In: Proceedings of the 2017 ACM SIGSAC conference on computer and communications security. ACM; 2017. p. 1143–1158.
101. Deepfakes:faceswap.(2019–02–02) [2019–11–8]. https://github.com/deepfakes/faceswap/wiki.
102. Liebermann O. Investigators spread blame in lion air crash, but mostly fault Boeing and FAA. [2019–11–8]. https://ethiopianflight302.com/investigators-spread-blame-in-lion-air-crash-but-mostly-fault-boeing-and-faa/.
103. Eatwall A. AI and blockchain intersect to revolutionize audit trails. (2018–8–21) [2019–10–29]. https://it.toolbox.com/articles/ai-and-blockchain-intersect-to-revolutionize-audit-trails.
104. Kiulian A. Why decentralized artificial intelligence will reinvent the industry as we know it. (2017–11–16) [2019–10–29]. https://www.forbes.com/sites/forbestechcouncil/2017/11/16/why-decentralized-artificial-intelligence-will-reinvent-the-industry-as-we-know-it/#18f9fc12511a.
105. McClean JR, Boixo S, Smelyanskiy VN, et al. Barren plateaus in quantum neural network training landscapes. Nat Commun. 2018;9(1):4812.
106. Raibert M. Meet spot, the robot dog that can run, hop and open doors. (2017–4) [2019–10–29]. https://www.ted.com/talks/marc_raibert_meet_spot_the_robot_dog_that_can_run_hop_and_open_doors.
107. Zheng SQ, Wang Q, Wu ZX, et al. Industrial intelligence technology and applications. Shanghai: Shanghai Scientific and Technical Publishers; 2019.
108. AlphaFold: using AI for scientific discovery. (2018–12–2) [2019–10–29]. https://deepmind.com/blog/article/alphafold.
109. Haque A, Guo M, Miner AS, et al. Measuring depression symptom severity from spoken language and 3D facial expressions. [2019–10–29]. https://arxiv.org/abs/1811.08592.
110. Rajpurkar P, Irvin J, Ball RL, et al. Deep learning for chest radiograph diagnosis: a retrospective comparison of the CheXNeXt algorithm to practicing radiologists. PLoS Med. 2018;15(11): e1002686.
111. Jiang LY, Wang XJ, Jin CL. Applications and access of artificial intelligence in the medical field. Chin J Health Policy Res. 2019;11(11):78–82.
112. Walker J. Machine learning drug discovery applications—Pfizer, Roche, GSK, and more. (2019–02–02) [2019–11–8]. https://emerj.com/ai-sector-overviews/machine-learning-drug-discovery-applications-pfizer-roche-gsk/.
113. Lethal autonomous weapons pledge. [2019–11–7]. https://futureoflife.org/lethal-autonomous-weapons-pledge/.
114. Kramer ADI, Guillory JE, Hancock JT. Experimental evidence of massive-scale emotional contagion through social networks. Proc Natl Acad Sci. 2014;111(24):8788–90.
115. Sosteric M. Facebook, CIA, social science, Google, and You. (2016–10–18) [2019–10–29].https://www.sociology.org/facebook-spy-machine/.
116. Wang SF, Bao YF, Li Y. Cognitive electronic warfare architecture and technology. Sci China: Inform Sci. 2018;48(12):19–29+125.
117. IBM AI enters legal industry: launches ROSS, the world's first AI lawyer. [2016–05–16]. http://tech.163.com/16/0516/14/BN6PBK8F000915BD.html.
118. DoNotPay—the world's first robot lawyer. [2019–10–29]. https://donotpay.com/.

119. Chinese biographical database (CBDB) project. (2018–1–11) [2019–10–29]. https://zggds.pku.edu.cn/slxx/58710.htm.
120. Luo J, Cao Y, Barzilay R. Neural decipherment via minimum-cost flow: from Ugaritic to Linear B. [2019–10–29]. https://arxiv.org/abs/1906.06718.

Chapter 12
The Future of AI Security

Zhaoquan Gu and Binxing Fang

With the continuous development of AI technology, AI security will receive increasing attention. Consequently, we are faced with a slew of questions: how the future of AI security will be like? Will the advances in information technology provide a perfect solution to AI security problems? How to prevent and control AI safety threats? What will be focus of AI security? In quest of answers for these questions, we, in this chatper, try to explore the future of AI security and AI safety from multiple perspectives, the future focus of AI security and share our humble opinions about AI security control.

12.1 Future of AI Security

Chapter 5 of this book has elaborated on the vulnerability of AI technology itself, that is, AI security. Will the AI security problems can be effectively resolved with the advances in hardware, computing efficiency of algorithms and data quality? Which AI secrutiy problems are loopholes resulting from technological immaturity and which are caused by the inherent defects of the technology and thus entail adoption of other means of protection? This section probes into the future of AI security from the perspectives of data, models, and the execution process.

Z. Gu (✉)
Harbin Institute of Technology, Shenzhen, China
e-mail: guzhaoquan@hit.edu.cn

B. Fang
Guangzhou University, Guangzhou, China
e-mail: fangbx@cae.cn

© Publishing House of Electronics Industry 2025

B. Fang (ed.), *Artificial Intelligence Security and Safety*,
https://doi.org/10.1007/978-981-96-9263-7_12

12.1.1 Whether AI Security Will Be Constrained by Data

Data are the basis for the application of AI technology, and defects in the accuracy and integrity of data cause inevitable vulnerabilities in AI technology itself. Data will become an important factor affecting AI security in the future development of AI.

First, the size of data will become increasingly large in the future. With the popularization of intelligent devices and intelligent terminals, the size of data is increasing. One example is online social network data stored by WeChat, which has increased from the TB level to the PB level and even ZB level. The types of data stored include text, video, voice, and many others. Current AI algorithms depend to a large extent on the size, quality, and distribution of data sources, and the rapid growth of data can improve the accuracy of AI algorithms in executing some tasks. However, the provision of more data alone is insufficient to guarantee the stability of the algorithms and the security of the models.

We use a handwritten digit dataset as an example. The MNIST database[1] is a classical tagged handwritten digit dataset in the field of machine learning and contains 60,000 training examples and 10,000 test examples, each of which is a 28×28 pixel image of grayscale handwritten digits with a given tagged label that indicates which of the digits from 0 to 9 corresponds to the image. The handwritten digit recognition program based on the MNIST database is an entry-level program in deep learning. Many beginners train neural network models based on this dataset and use the 10,000 test images for verification. The most classical neural network model is LeNet [1], which has an accuracy of more than 99%. However, there are also many attack methods based on this database. The fast gradient sign method (FGSM) [2], an attack method targeted at neural networks, adds very subtle disturbances to an original image to generate a new example that will produce errors when recognized by existing neural networks. Such examples are the adversarial examples mentioned in Chap. 5. It has been experimentally demonstrated using adversarial examples that similar handwritten digit images can also cause recognition errors.

Take for example the case of digit recognition illustrated in Fig. 12.1. The image on the left side is taken from the MNIST dataset and is normally recognized as the digit 4 by neural networks. The simplest adversarial example generation method adds random noise to the 28×28 image, and a similar image (on the right side) is then generated, which can be clearly recognized by the human eye as the digit 4, whereas the neural networks recognize it as the digit 9. Many factors contribute to the generation of adversarial examples, and no consensus has been reached in academic circles at present. We performed an analysis from a very simple aspect. If the dataset contains all the images and corresponding labels (i.e., the image on the right side is also included in the dataset), then it can be accurately recognized by the neural network.

[1] http://yann.lecun.com/exdb/mnist/.

Fig. 12.1 Example of digit recognition

Therefore, it should be examined whether expanding the MNIST dataset can solve the problem of adversarial examples and provide a secure neural network model. This question should first be discussed in terms of the feasibility of expanding the dataset.

From the perspective of dataset size, in the MNIST dataset, each single-channel image is of only $28 \times 28 = 784$ pixels, i.e., the value range for each pixel is [0, 255], whereas each pixel in an achromatic image is the combination of three channels (R, G, B). If all possible images are calculated for this size only (i.e., each pixel has 256 different values), then the size of the dataset will be 256^{784}. At present, no electronic storage device can store such a large dataset.

In addition, professional personnel are required to label all images in the MNIST dataset as the corresponding digits. With the development of the hardware, storage, management systems, and labeling systems, if the training set already contains all possible handwritten digit images, it may be possible to achieve a secure and stable neural network model. However, many images in such a dataset have no practical significance. For instance, if all pixels take the value of 0 or 255, then an all-black or all-white image will be generated, and the digit corresponding to this image will be indistinguishable. An important task is to determine how many images among the 256^{784} images have practical significance and how many have indistinguishable digits. With a relatively complete and meaningful dataset, another important task is to ensure that there are no adversarial examples for images outside this dataset. This may be a fundamental problem to be solved by AI for its own security.

In the example of MNIST handwritten digits, it is difficult to generate a large dataset and fulfill the labeling task for 784-pixel images, and it is even more difficult to provide a complete dataset for face recognition, autonomous driving, and other tasks. One question is whether the dataset can be expanded as much as possible to mitigate the impact of insufficient data on AI models. However, it is difficult to produce results by expanding the dataset alone. For example, we can expand the number of handwritten digit images from 60,000 to 600,000; however, the added 540,000

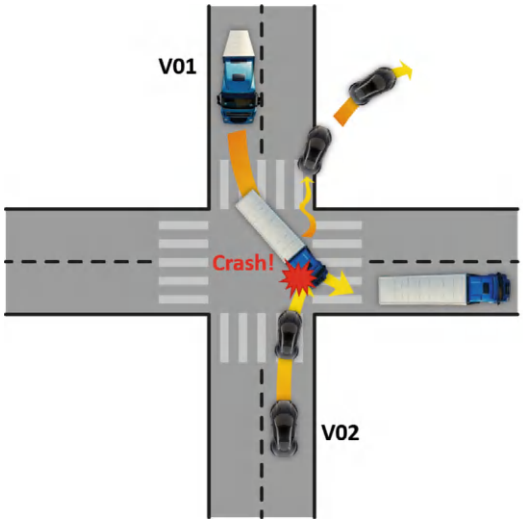


Fig. 12.2 Example of an autopilot safety accident

images may all be different images of the digit 0. The neural network model obtained through training may be more precise in recognizing the digit 0, and there may be fewer adversarial examples associated with the digit 0; however, large problems in the recognition of other digits would remain, and valid adversarial examples could be generated with multiple neural network attack methods. Accordingly, horizontal expansion is also required for data distribution when expanding the dataset.

Here, we use autonomous driving as an example. The Berkeley DeepDrive (BDD) 100 K dataset[2] is a large public dataset for autonomous driving, containing more than 100,000 videos with over 1100 h of driving experience for different times of day and weather conditions. The Baidu Apolloscapes Dataset[3] is also a large autonomous driving dataset that Baidu plans to roll out in its Apollo initiative and defines 26 items with different semantics, such as cars, pedestrians, and buildings. If only the defined 26 semantic items are expanded in the dataset and additional recognition images of cars, pedestrians, and buildings are provided, the intelligent recognition system in a self-driving car will be unable to provide effective feedback in many emergencies.

We take an example of an autopilot safety accident (Fig. 12.2) to present the potential safety concerns with autopilot.

America's first fatal autopilot crash occurred on a Florida highway on 7 May, 2016, as illustrated in Fig. 12.2. V01 marked in the image represents a tractor trailer that was prepared to turn left before the accident. V02 marked in the image represents a car with an autopilot system that failed to accurately recognize the trailer turning ahead and passed underneath the trailer, leading to the driver's death. Opinions about

[2] http://bdd-data.berkeley.edu/.

[3] http://apolloscape.auto/.

the cause of the accident varied, and several opinions were intensely discussed and debated as follows:

First, the radar in a self-driving car can precisely detect a large obstacle ahead. However, because the reflection area of the trailer was too large and its body height was too high, the trailer resembled either a traffic signboard suspended above the road or an elevated road to the millimeter-wave radar and was thus ignored by the autopilot system.

Another opinion was that the telelens in the self-driving car could not see the entire trailer and mistook the trailer as clouds floating in the sky. A third opinion was that it was difficult for the camera to extract feature points from a large white object against a brightly lit sky, thus leading to the false judgment.

The latter two points can be regarded as a problem with the intelligent model itself. Because of the lack of similar examples in the dataset, the neural network cannot learn the difference between clouds in the sky and a white tractor trailer. Regardless of the amount of other recognition data added to the dataset, it is practically impossible to solve the recognition problem under such special circumstances. Consequently, it is important to expand the data distribution in a dataset, which may reduce the potential safety hazards arising from the uneven data distribution of the intelligent model.

Finally, expanding the size and distribution of the dataset is not necessarily an effective solution to the security problem of the model. Many current tasks in machine learning are supervised learning tasks; in particular, the corresponding labels of examples are provided at the time of training, and the intelligent model can compare the recognition results and the tagged labels at each step and adjust the model parameters according to the recognized errors. Data quality is a crucial factor in supervised learning. If there are errors with the data label itself, then the security of the intelligent model cannot be guaranteed. Although anti-interference intelligent models and algorithms currently exist, data quality may still have a significant impact on the models and algorithms, which also applies to unsupervised learning and reinforcement learning.

The accuracy of acquired data is directly related to the security of AI. It cannot be assumed that data are accurate, and various factors may result in the poor quality of the acquired data. For instance, data loss, data deformation, and even the input of noisy data can cause serious interference in an AIS. An important problem facing AI is whether it will be able to discriminate the acquired data. An example is the crash of the Boeing 737 Max 8 mentioned in Sect. 11.2.7 of this book, which was caused by the inability to discriminate the input data, i.e., the crashed plane was expected to be able to determine the sensor fault through the aircraft attitude reading displayed by the two angle-of-attack detectors. At present, data-related problems are important factors constraining the development of AI, and improvements in data size, data distribution, and data quality are likely to alleviate the security problems of AI.

12.1.2 Whether Adversarial Examples Will Appear in Practical Application Scenarios

The adversarial examples mentioned earlier were proposed by Szegedy et al. [3]. These examples refer to situations in which subtle interference noise is added to an original example to cause the model to produce incorrect classification results; however, the modified example does not interfere with normal human recognition. There are various methods for generating adversarial examples, which have aroused concern in academic and industrial circles. Recent years have seen significant growth in the number of papers on adversarial examples published in the academic literature, including those collected by top conferences on AI, such as the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Annual Conference on Neural Information Processing Systems (NeurIPS[4]; abbreviated NIPS before 2018) (there is a clerical error in the note below, but this note can actually be deleted because the words in brackets already provide an explanation), International Conference on Machine Learning (ICML), Association for the Advancement of Artificial Intelligence (AAAI), International Conference on Learning Representations (ICLR), and European Conference on Artificial Intelligence (ECAI).

In 2017, NIPS organized a competition on neural network attack and defense, with tasks divided into nontarget attacks, target attacks, and model defense. Moreover, industrial circles have started to pay attention to adversarial examples in recent years. For instance, Alibaba organized the IJCAI-2019 Alibaba Adversarial AI Challenge for E-commerce images at the International Joint Conference on Artificial Intelligence (IJCAI), with the tasks also divided into a nontarget attacks, target attacks, and model defense. Existing adversarial examples are of interest primarily in academic circles and appear in various conference and journal papers. Questions remain whether adversarial examples will emerge in a real environment and whether there will be realistic applications in the future.

For tasks such as object detection and image classification, an adversarial algorithm causes errors in neural network detection, recognition, and classification by adding subtle noise to an original image and generating an adversarial example similar to the original image. An example is the object detection task illustrated in Fig. 12.3. The character detection system of Yolo-v2 [4] can correctly recognize the cartoon character in the left image; however, a special color patch is added to the cartoon character in the original image on the right-hand side (a rectangular color patch is added to the cartoon character), resulting in the system's failure to detect the cartoon character in the image.

An adversarial method involves the modification of images; however, it may also be applied in a real environment. Thys et al. printed an attack patch, and when a testee implemented the patch, the Yolo-v2 system failed to detect the testee, which is equivalent to dressing the testee in an "invisible cloak" [5]. Sharif et al. also reported that adversarial decorations can be customized. For instance, a special type of glasses

[4] NeurIPS was formerly called NIPS before 2018.

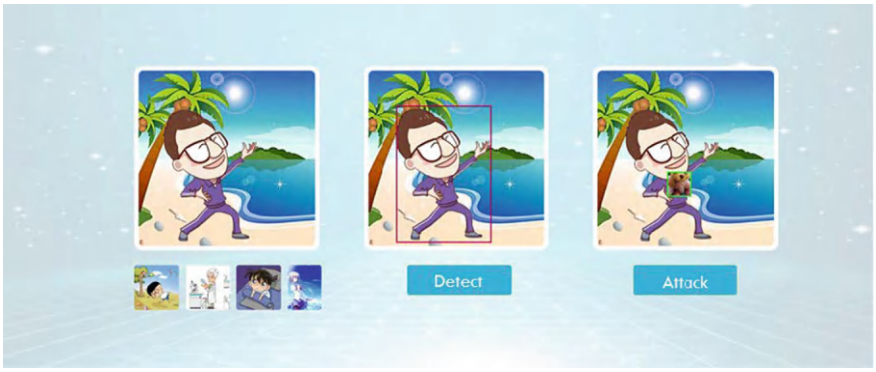


Fig. 12.3 Example of object detection

can be customized based on human faces, and when a user puts on the glasses, the face recognition system cannot recognize the user correctly [6]. Such cases suggest that adversarial examples may appear in real-world environments. It is thus important to determine the potential security hazards that adversarial examples may cause in practical applications. Two examples are provided below.

(1) Road signs, pedestrians, and other vehicles can be recognized by AI models in autonomous driving. Thus, pasting attack images on road signs can cause errors in model recognition and may lead to serious accidents. Eykholt et al. proposed the method of Robust Physical Perturbations (RP2) [7], in which black and white stripes were pasted on a stop sign, which caused the neural network to mistake the stop sign for a sign expressing "speed limit 45." In addition, videos were used to demonstrate a situation in which vehicles had a high probability of recognition errors when passing by stop signs.

(2) Face recognition systems have gradually gained popularity and have been widely applied in the detection of smart buildings, financial payment, and character recognition at airports, stations, and other locations. As illustrated in Fig. 12.4, the normal recognition process includes the capture of face information through video, recognition, and analysis. However, the face recognition system generates recognition errors if attack decorations are customized, such as wearable glasses, headwear, and earrings. Moreover, after wearing the customized attack decoration, a user can make the neural network recognize them as a predefined person, i.e., perform a target attack. The occurrence of such adversarial examples in real life can potentially cause safety hazards for systems with intelligent models.

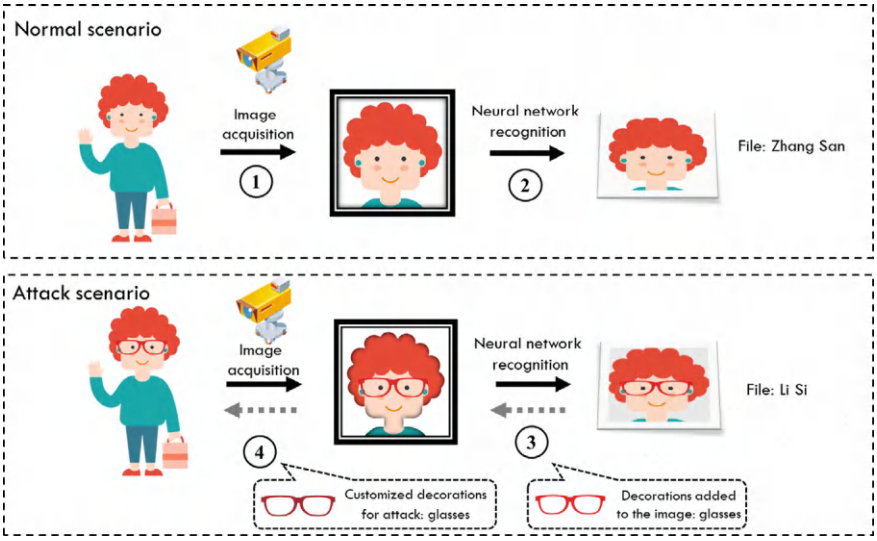


Fig. 12.4 Example of face recognition

12.1.3 Whether There Will Be Secure Intelligent Models

Apart from the problem regarding data, another major problem leading to the vulnerability of AI is that most current algorithm models are not safe or robust. Take a neural network as an example. The construction of a multilayer neural network by simulating the working principle of neurons can be effectively applied to image recognition, text classification, voice recognition, and many other tasks with higher accuracy than that of conventional machine learning algorithms. However, to date, there has been no consensus on why neural networks, especially deep neural networks, are so effective, and few studies have been conducted on the theoretical basis of neural networks. Therefore, we must examine whether there are secure and reliable intelligent models that can be verified mathematically and experimentally.

Secure and reliable intelligent models can be discussed from two aspects. The first aspect is that the model itself has a certain theoretical basis; however, it may not be effective owing to data sources and other external causes. The second aspect is that the model itself has no theoretical basis, and even if the model is effective and efficient, it is not understood why the model is effective and whether the model will malfunction under certain circumstances or produce unexpected or unknown results. Early into the second game between AlphaGo and Lee Sedol, AlphaGo made a move (move 37), which was generally considered inappropriate by professional Go players, and surpassed the normal logic of accomplished professionals. The move performed by AlphaGo was called an *inhuman move* that made it difficult for humans to adapt to it. A typical example of the first aspect of a reliable intelligent model is the support

vector machine (SVM), while a typical example of the second aspect is a deep neural network.

SVM [8] is a common supervised learning algorithm that classifies labeled examples. Taking dichotomous classification as an example, all given example data are divided into positive and negative categories. The purpose of the SVM model is to find the optimal hyperplane from the feature space to maximize the interval between the positive and negative examples in the example data. SVM has a strong theoretical basis, and the parameters in the SVM model can be determined by mathematical derivation. However, it should be determined whether SVM is a secure intelligent model. Considering the problem of data poisoning, if we intentionally add falsely tagged example data into an example set, such as by changing the data that should be used as positive examples to negative examples, then the features learned by SVM will contain incorrect classification features. An important question is whether there is a secure and robust solution to this situation. Fortunately, the SVM algorithm can be extended to the soft-margin SVM model. The original SVM model requires the identified hyperplane to perfectly divide positive and negative example data, whereas the soft-margin SVM model accepts a certain proportion of errors; in other words, the hyperplane is allowed to falsely recognize some example data. Accordingly, the problem of partially erroneous data can be solved to a certain extent, and the soft-margin SVM model can provide effective theoretical explanations. However, similar extensions and data error tolerance cannot be ensured for all intelligent models.

The application of deep neural networks can be observed in many aspects of human life. For instance, face recognition systems developed by SenseTime, MEGVII, and HIKVISION have been applied in airport security checks, face unlock, facial-recognition payment, and other scenarios. However, at present, deep neural networks are still not theoretically explainable in academic circles. Therefore, the explainable AI (XAI) described in Sect. 11.1.7 of this book has become a popular research topic in the past several years. An example is a recommendation system. Conventional recommendation systems recommend similar commodities based on a user's personal history and recommend appropriate movies according to the viewing records of the user's friends. However, current recommendation systems mainly use two methods: collaborative filtering and matrix factorization. Collaborative filtering can be understood as recommending commodities according to their similarity from the perspective of commodities. For instance, if User A browses or purchases a mobile phone of a certain brand, the system will recommend other brands of mobile phones having similar features. Collaborative filtering can also be understood as recommendation according to the habits of similar users from the perspective of users. For instance, User A and User B have similar viewing habits, and both like watching movies involving personal heroism, such as *Iron Man* and *Spider-Man*. When User A watches the movie *Ant-Man*, the system will automatically recommend this movie to User B.

Matrix factorization collates the relationship between users and commodities into a matrix, predicts the correlation scores of users and commodities through matrix calculation, and then makes recommendations based on the score. Collaborative

filtering has a low recommendation accuracy, whereas matrix factorization is difficult to explain intuitively. Consequently, an explainable recommendation system aims to make recommendations to users while mining relevant features rather than recommending blindly according to a set model. For instance, if User A often watches movies of the same genre, an explainable recommendation system will extract the features of these movies and discover that they are movies featuring personal heroism. When *Ant-Man* and other movies of the same type are viewed, the system will make an automatic recommendation and provide reasons for its recommendation.

Explainable neural networks are a popular research topic, and many scholars have attempted to interpret the working principle and characteristics of explainable neural networks. For instance, in convolutional neural networks, researchers have attempted to analyze features extracted by the neural network in each layer. In the process of recognizing a vehicle, it was discovered that neurons in the first layer extract the line features from the images, those in the second layer extract the features of the automobile parts formed by a combination of lines, and those in the third layer extract the features of the vehicle formed by the parts. However, this description is insufficient for explaining how neural networks perform correct recognition. He et al. from Microsoft Research expanded a neural network to 152 layers and proposed the concept of a residual neural network (ResNet) [9] which won first place in the ImageNet Large Scale Visual Recognition Challenge (ILSVRC, a competition with international authority in the field of machine vision) in 2015. It has been experimentally demonstrated that deep neural networks are more effective than neural networks with fewer layers. However, the reasons for this behavior are also unclear.

One commonly accepted view is that each layer of a neural network attempts to learn the optimal solution to a problem. Although the probability of learning the global optimal solution is high, it is also possible to learn the local optimal solution. Under certain conditions, the local optimal solution learned by the neural network is also the global optimal solution to the problem [10]. This may be a groundbreaking research achievement in current theoretical research on neural networks. However, determining whether there is a complete theoretical basis for neural networks requires further study by scholars in the field.

12.1.4 Whether the Execution Process of AI Will Be Secure

In this section, we examine whether AI would still be vulnerable if data quality and security were guaranteed and the security and explainability problems of AI models were solved. Specifically, we focus on possible security problems in the execution process of AI models.

The security problems generated in the execution process of AI models can be mainly attributed to either objective or subjective causes. Objective causes include changes in the algorithm environment, unknown environmental data, and other objective conditions that may be encountered by AI models in the execution process. For instance, a character detection algorithm would cause detection errors if used for

vehicle detection. Further, in a temperature forecast task, a system would not be able to make correct judgments if the actual temperature were higher or lower than historical values. Of course, objective causes of this type can be solved to a large extent by the correction of the dataset, and potential safety hazards arising from objective causes in the execution process of AI models can be addressed by adding data of relevant tasks to the dataset, diversifying data distribution, and enhancing data quality.

Among subjective causes, manual modification of optimization objectives, incorrect definition of winning rules by a chess program, and even manual setting of malicious object functions can cause security problems in AI. However, it is difficult to restrict subjective causes through technical means. In addition, it is particularly difficult to find a means of reducing subjective factors in the execution of AI algorithms and models according to laws, rules, and administrative regulations after the malicious use or misuse of AI. This will determine future research trends for secure and controllable mechanisms.

12.2 Future of AI Safety

12.2.1 Weaponization of AI

AI technology can considerably reduce human effort and improve work efficiency. AI is also extensively applied in the military field, such as in terrain reconnaissance and object detection and tracking. Weaponization of AI will have multifaceted consequences.

In 2018, the Korea Advanced Institute of Science and Technology and Hanwha Systems jointly established an AI R&D center to develop AI technologies applicable to operational command and object tracking. This move was aimed at the development of AI weapons, which triggered widespread concern among many scholars. Latr, in April, more than 50 world-leading AI scientific research personnel from more than 30 countries and regions signed an open letter urging the R&D center to stop the development of AI weapons, fearing that such autonomous arms could be used by terrorists and thus result in large-scale wars [11]. As mentioned in Sect. 10.2.3.4 of this book, over 3000 Google employees signed an open letter to prevent Google from joining the Maven project, which provides AI technology to the US military.

The rapid development of AI technology has led many countries to see the potential of AI in the military field and attempt to apply AI to weapons. However, it is predictable that once AI weapons are developed and employed, large-scale wars will become inevitable. In particular, if terrorists and other malicious individuals control AI weapons, they will become more unscrupulous and launch a larger number of terrorist attacks, posing a tremendous threat to the peace and stability of the international community. On September 14, 2019, several unmanned aerial vehicles (UAVs) attacked two oil facilities of the Saudi Arabian Oil Company (Saudi Aramco) and

caused a fire [12], leading to a halving of the company's output. UAVs are supported by AI technology, and even future UAVs are a type of aerial AI weapon not requiring remote control from ground personnel.

The case of Slaughterbots described in Sect. 7.4.2 of this book can better reflect the serious consequences arising from the weaponization of AI. A technology, no matter how advanced, will cause devastating consequences if used maliciously. It is thus important to ensure that AI technology is used for the benefit of mankind rather than to harm humans. This may be accomplished by requiring countries to jointly sign multiparty agreements to guarantee the nonmalicious use of AI weapons and explore different technical means and modes to lower the risks posed by AI weapons, such as monitoring the applications of AI resources and setting stringent access permission modes for AI technology.

12.2.2 Potential Safety Hazards Caused by the Misuse of AI

As mentioned in Sect. 10.1.1 of this book, 26 researchers from 14 organizations published a 101-page report titled "The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation" [13] in February 2018, which listed the potential security threats posed by AI to mankind in the next five years, and humans' responses. According to the report, the potential threats of AI are precisely what accompanies the birth of AI, which is consistent with the concept that security is an accompanied technology. Originally, AI was established to automatically complete tasks that had to be performed manually in the past. The implementation of automation has dramatically reduced the cost of some attacks and created greater potential security hazards. Four high-level recommendations were made in the aforementioned report to cope with AI threats:

(1) Policymakers should collaborate closely with technical researchers to investigate, prevent, and mitigate potential malicious uses of AI.
(2) AI researchers and engineers should seriously consider the dual-use nature of their work, allow misuse-related considerations to influence research priorities and norms, and proactively reach out to relevant individuals when harmful applications are foreseen.
(3) Best practices should be identified from methods in some research areas such as computer security, with more mature methods for identifying, predicting, and mitigating dual-use concerns about the malicious use of new technology, and adopted where applicable for the case of AI.
(4) The range of stakeholders and domain experts involved in discussions of these challenges should be actively expanded. This includes contacting ethicists, entrepreneurs, and the general public in an effort to solve the potential security hazards associated with AI through the cooperation and efforts of all parties.

Consider Deepfake as an example. This technology can generate nonexistent faces and paintings via a generative adversarial network [14] described in Sect. 11.1.8 of

this book. For example, the faces of Hollywood actresses have been placed on the bodies of pornography actresses to produce fake pornographic films. AI scholars and engineers should take these consequences into account in their research and formulate corresponding regulations; otherwise, such technologies may pose great security hazards according to the second recommendation in the aforementioned report. For instance, Deepfake may also pose threats to political security. If this technology is maliciously used to generate fake videos of national leaders or publish negative remarks, it will likely trigger panic among countries and arouse social discontent.

Technologies themselves are neutral. Our research on technologies is primarily for the benefit of mankind; however, there is also the risk that some people may use technologies to harm mankind. Technologies have an increasingly positive effect with their continuous development. For instance, map navigation has simplified road recognition, while mobile payment has revolutionized payment methods. However, the better the effect that these technologies have, the more serious the possible consequences become, and the potential impact may be harmful to the entire world and all of humanity. Global warming, environmental pollution and destruction, and many other problems are all caused by the rapid development of technologies. In the course of developing and employing new technologies, people often attach too much importance to their development and ignore their potential threats. Owing to the open-source nature of AI technology, the development pace of AI technology is well above that of other technologies; however, these open-source solutions may also be targeted by criminals. Important problems involve determining how restrictions should be imposed on these technologies, and what policies and systems should be designed to mitigate the malicious use of these technologies. These problems must be considered for the future development of AI.

12.2.3 Endangerment of Mankind by AI

The question of whether AI will endanger mankind will inevitably become a major concern in the development of AI. For instance, AlphaGo defeated the top human Go player, occasional accidents have occurred with robots in factories, and robot awakening has been envisioned in science fiction movies. Human society is highly focused on the impact of AI on mankind. The "AI threat theory" is a popular topic, involving the fear that robots will one day awaken and harm mankind. The question remains whether AI will endanger mankind in the future; positive and negative views are presented below.

12.2.3.1 Views that AI Will Endanger Mankind in the Future

At present, many people believe that AI will endanger mankind in the future. For instance, some people believe that AI may have an impact on humans from the

perspective of social development. The fact that AI can replace humans to fulfill tasks through automation will force some individuals to change the content and nature of their original work, which will then trigger an unemployment problem. For instance, robotic vacuum cleaners, glass-cleaning robots, and warehousing robots will force some cleaners and porters to change jobs. Another example is AI synthetic anchors [15] released at the 5th World Internet Conference by Sogou and Xinhua News in November 2018. AI synthetic anchors resembling flesh-and-blood news anchors capable of broadcasting news were developed through voice synthesis, expression synthesis, and other technologies. If AI anchors are popularized and applied, many news presenters may have to change professions. Some fear that the problems caused by large-scale unemployment will cause global economic imbalances, wider gaps between the rich and poor, and greater potential safety hazards.

AI technology simplifies complex processes, and mankind's dependence on AI algorithms will probably lead to the degradation of mankind and the evolution of AI. For instance, when users open news applications to view the news of interest to them or access a shopping website, recommendations provided by AI algorithms are often shockingly accurate. Thus, it can be said that AI algorithms know us better than we do and can even predict what products we need in different seasons. While enjoying the convenience brought by AI algorithms, users have already surrendered an increasing amount of decision-making power to AI.

With its powerful capabilities, AI may also destroy many of mankind's existing modes that have proven very successful. For instance, future AI can accurately predict diseases that individuals may develop after 10 years, which will enable them to take advantage of such prediction results to choose necessary medical insurance for themselves. However, if everyone takes advantage of the AI recommendation results, insurance companies will probably go bankrupt and medical insurance will no longer exist. Insurance was originally developed as a means of leveraging the attributes of probability, pooling a small number of payments from the public out of their fear for low-probability events and paying for large-value expenditures occurring with a low probability. If diseases become a high-probability event, then those individuals who are not predicted to develop diseases will no longer purchase insurance through the fear of low-probability events. In this case, medical insurance companies will have difficulty pooling and raising funds to pay for the large expenditures with high probability for the insured.

A current bottleneck in AI is its insufficient perceptual ability. For instance, a robot camera can detect only a limited range, while the human eye can perceive and detect objects more broadly. However, the limitation of the human brain is that it cannot rival computers in power and storage capabilities. As long as there is sufficient storage space, computers can store content forever, whereas the human brain has difficulty in storing permanent memories owing to its limited memory capacity. In addition, computer calculation is not affected by emotional or external factors, whereas the human brain is slow and susceptible to emotional interference. Consequently, the combination of the human brain and AI is a seemingly positive development for mankind; however, it will trigger new safety threats and ethical disputes.

Neil Harbisson was the first person in the world to be legally recognized as a cyborg by the government. Born color-blind, he can only distinguish between black and white. In 2004, an antenna capable of converting color into sound was implanted at the back of his skull, allowing him to listen to color from sound.[5] Neuralink, a company established by Tesla founder Elon Musk, is developing neural lace technology, which reads and uploads human thought through a brain–computer interface and is designed to achieve symbiosis between AI and humanity.[6] Some individuals believe that the combination of the human brain and AI will strengthen humans, but many oppose such a combination. In her book, *Artificial You: AI and the Future of Your Mind*, philosopher and cognitive scientist Susan Schneider wrote "if AI cannot be conscious, then if you substituted a microchip for the parts of the brain responsible for consciousness, you would end your life as a conscious being. You'd become what philosophers call a 'zombie'—a non-conscious simulacrum of your earlier self. Further, even if microchips could replace parts of the brain responsible for consciousness without 'zombifying' you, radical enhancement is still a major risk. After too many changes, the person who remains may not even be you. Each human who enhances may, unbeknownst to them, end their life in the process" [16]. The AI chips implanted in human brains will likely generate self-consciousness through autonomous learning and result in the disappearance of the human–machine interface. A similar idea was proposed by American writer Ray Kurzweil in his book, *The Singularity is Near: When Humans Transcend Biology*. Kurzweil believed that humans will become immortal to some degree by 2045 through combination with computers. However, by that time, computers will replace human brains, and it will be difficult to distinguish humans from computer intelligence [17]. Furthermore, humans will no longer be humans in the traditional sense, but organisms only.

The famous physicist Stephen Hawking agreed that AI may endanger mankind and once warned that AI could spell the end of the human race. He also published an open letter together with Musk, arguing that humans should constrain the development of AI and remain alert to the fact that AI may harm or even kill human beings in the future. Although current AI development is still controllable, it is difficult to predict what AI will become in the near future. In August 2013, the University of Washington achieved remote control between two human brains, with one researcher able to send a brain signal to control the hand motions of a fellow researcher on a keyboard [18]. This leads to the question of whether humans implanted with chips may one day be controlled remotely by AI and machines and whether chips will entirely control and replace humans if they complete the self-learning process. These questions thus require heightened awareness.

It is reasonable to be concerned about the harm posed by AI to humans. First, the inexplicability of AI poses potential danger to humans. The fact that the behavioral and decision-making mode of AI cannot be explained creates uncertainty, and humans cannot guarantee that they are strong enough to control a situation when confronted with uncertainty. Second, AI is sufficiently strong to harm humans. Once an error

[5] See https://www.ted.com/talks/neil_harbisson_i_listen_to_color.

[6] See https://www.neuralink.com/.

occurs in the decision-making system (for instance, if it is hacked) a question is whether it will use its force against humans. The several fatal accidents caused by autopilot cars are sufficient to demonstrate that humans are not entirely capable of controlling such AIAs.

Third, the autonomy of AI is likely to generate motives to harm humans. The evolution of AI has involved rule-based symbolism, trial-and-error actionism, and big-data-based connectionism. In particular, the actionism mode, which is already neglected by us, is an intelligent mode in which self-improvement is realized through repeated trial and error. Based on cyberneticism and the perception-motion control system, and supported by the genetic and annealing algorithms, the actionism mode lays the foundation for innovation and evolution through random variation. Actionism stemmed mainly from cyberneticism, while the latter became an essential part of the trend of the times as early as the 1940s and 1950s and affected a group of early AI workers. However, because the continuous iteration of actionism requires extremely high computing capability, there has been no opportunity for actionism to lead the development of AI. With the enhancement of network computing power, there will be no further obstacles to the computing power of AIAs in the future, which may allow actionist AIAs to continue moving along the evolutionary path toward the unknown. It is difficult to predict the capabilities that AIAs will have by that time.

From another perspective, knowledge can be divided into two types. The first type is original knowledge. For instance, aerodynamic principles were discovered by humans and were employed to build airplanes, which requires a transition process in thinking. The other type of knowledge is derivative knowledge, which involves drawing inferences about other cases from one instance. One example is the formation of a boxing technique. Suppose a robot has acquired basic athletic abilities such as jumping; running forward, backward and toward bilateral sides; throwing a punch; and stretching out the feet, as well as the standard for winning and losing taught by humans. Then, a question is whether the robot can calculate the movement results based on the random movements in training, make adjustments according to the reward function, and sum up a set of boxing rules with a high probability of winning. From this perspective, it is likely that robots will create a gaming method. The playing records of AlphaGo revealed that move 37 in the second game with Lee Sedol surpassed the normal logic of accomplished professionals and displayed the astonishing capability of AlphaGo to take creative decisions [19]. We believe that humans gave AlphaGo its evaluation ability and AlphaGo learned the method to randomly generate moves. Therefore, this randomly generated input, in addition to the selective retention of the optimized input combination, naturally leads to what humans consider a type of creative capability for action. This is probably why Xiaoice can write poems and novels.

12.2.3.2 Views Rejecting that AI Will Endanger Mankind in the Future

There are also many arguments opposing the view that AI will endanger mankind in the future. For example, some people believe that technologies may be used by

malicious individuals, but AI research should not be slowed down because of this concern. Comparatively speaking, the negative effects of technologies are not that significant. Even if AI may endanger humans, the development and use of AI should not be slowed down; instead, the means of protecting against AI should be studied from another aspect.

Jaan Tallinn, a co-founder of Skype and member of the Institute of Potential Risks for the Future Development of Mankind, does not believe that AI at its current stage will endanger humans and states that "as long as artificial intelligence is less intelligent than humans, we can treat it just like any other technology. But as soon as we have to deal with an AI that potentially becomes more intelligent than us, the situation changes drastically." [20] Tallinn believes that AI has not caused harm to humans thus far but attracts attention simply because of its outstanding performance in some areas.

Other scholars have provided reasons why AI will not endanger humans from three perspectives. First, AI indeed performs very well in some areas. For instance, AlphaGo performs well as a Go player, CNNs outperform humans in terms of recognition accuracy in image classification and digit recognition, and RNNs can assure accurate translation results in machine translation. However, the results of many tasks requiring knowledge summarization and reasoning are far from satisfactory. For instance, humans are very effective at summarization and induction regarding semantic comprehension, rule-based reasoning, and multi-person dialogue. In contrast, it is difficult to achieve satisfactory results with current AI technology for effectively comprehending the natural language of humans and reasoning about the unknown based on known results.

Second, AI is more concerned with performing learning tasks within the boundaries set by humans, and it cannot get transcend these boundaries and think like humans. Humans first create behavioral goals, and AIAs then seek approaches and methods to achieve these goals; they are incapable of creating goals by themselves. In addition, AIAs acquire knowledge through human training and cannot create knowledge that has not been taught by humans. Science fiction novels and movies have exaggerated the capabilities of AI by proposing that AI can evolve and create goals on its own; however, this remains difficult for current AI.

Third, many people see the threat posed by AI primarily as a conflict between humans and machines. However, as of now, robot injury accidents, such as the case mentioned in Chap. 6, have all been due to human manipulation errors rather than the self-awakening of robots. The development of robots has also not spread to every field, as it is difficult for robots to perceive the environment in real time and make decisions autonomously. Humans should thus coexist with robots and have robots serve them.

In summary, individuals who reject that AI will endanger humans generally agree with the positive significance of AI development. Although AI may outperform humans in some specific tasks, it has not developed to the point of endangering humans. Undoubtedly, the understanding that AI is incapable of creation is based on the judgment given by the connectionism of AI. Connectionism, in essence, is a means of training AIAs with a huge amount of data, and AIAs trained through machine

learning cannot surpass what they have learned. It is similar to an experienced soccer commentator, who will not assume that a team with a consistently poor record will reach the semi-finals of the World Cup.

Individuals who hold the aforementioned views do not believe that AI will eventually endanger humans; however, they do believe that attention should be paid to prevention and control technologies for the future development of AI. It is thus imperative to investigate means for security issue prevention and control amid the rapid development of technologies.

12.3 Future Focus Related to AI Security

12.3.1 Focus of Basic AI Tools on Autonomy and Controllability

In the twentieth century, the operating system of personal computers shifted from DOS to Windows, and the Windows system was pirated in large quantities, severely shrinking the market space and leaving no room for the sci-tech circles to develop new operating systems. Likewise, the United States held a firm grip on the desktop market by forming the Wintel alliance through CPU and operating system. Enterprises unable to form market competitiveness in the short run cannot have a share of the CPU market; however, they can develop a small quantity of CPUs and operating systems with the support of some governments and institutions.

AI has now become one of the most popular research topics. American companies represented by IBM, Google, Microsoft, and Facebook have provided substantial input in AI and acquired start-ups with progress in neural networks, natural language processing, image processing, and audio and video analysis. It is clear that AI will become the focus of competition among countries in the future, and the development of basic AI tools will replace that of the CPU and operating system. By focusing on the development of basic AI tools, we can avoid losing our international competitiveness, such as with the CPU and operating system, in the future. However, the international community has launched a large number of open-source AI tools that will create dependence on basic technologies and thus lead to a return to the long-term dependence on Wintel. In addition, open-source AI tools will also form a new critical technology that others can wield to limit our development and result in the risk of supply interruption after long-term dependence on open-source tools. More importantly, if we use these nonautonomous and noncontrollable basic tools for long, we will inevitably perform training based on our own data, then feed the data to uncontrollable tools (whose platforms are controlled by others), provide data with Chinese characteristics, and form an unfavorable situation in which our data support the tools of other countries.

Many basic open-source AI tools in the current international community are extensively downloaded and used by AI researchers and enterprises, including large quantities of machine learning algorithms and datasets. In fact, the increasing availability of these AI tools serves to control their users in a disguised form. In particular, as international technology giants have released many AI development tools, when people use these uncontrollable tools from those platforms, they help the tools from these technology giants train their algorithms. This will probably widen the gap in the AI area more rapidly than in the chip area. There are three consequences of using the algorithms of international technology giants. First, they can easily obtain the training results; second, it means using our data for the benefit of other countries; and third, it might cut off our path toward the autonomous research of AI tools.

Unlike CPUs and operating systems, AI generates value from the long-term and large-scale use of users, and the meaningful part is the knowledge base formed from machine learning, such as rule bases, model parameter libraries, and mapping knowledge domains, all of which are the results of learning for years. Ecology is the fatal weak point of a core product, while data are another vital component of AI tools. Only through long-term accumulation can both ecology and data gain vitality. For example, an input method, regardless of its effectiveness, will become useless if its long-term knowledge base is destroyed. Microsoft Xiaoice invented a personalized service pattern in which personal services are targeted via personal *adoption*, and user data are obtained through the adoption process over a long time to establish the conversation habits of users. Once such information is lost, the painstaking efforts made in adoption over several years will have been wasted.

Many Chinese companies rely on open sources from the United States. When we use these open-source tools to develop and produce products, it establishes ecologies and accumulates data for these products. However, once the usage license of tools and the right-of-use of algorithms are revoked, it will be too late to establish new ecological chains and accumulate new databases, which will make China lose everything related to AI and lag behind the United States and other countries with open-source tools. Google launched the Android system and made it available to all users. Mobile phones, smart televisions, and Internet of Things terminal devices mostly use the Android system. Because of the huge ecology established by the Android system, it is difficult to develop a mobile device operating system ecology to rival the Android system in the short run. For example, if Google prohibits Huawei mobile phones from using the Android operating system according to the policy of the US government, Huawei will have to transfer a large number of applications after launching its new operating system, HarmonyOS. If the transfer in large amounts cannot be performed, the system's competitiveness will be affected by the lack of the necessary ecology, and Huawei will be required to spend a large amount of time, data, and products to rebuild a new mobile device operating system ecology.

Consequently, without autonomous algorithms and development tools, there will be no autonomy or controllability. The example of the CPU and operating system is a warning that AI should be prevented from becoming another technology that other countries could wield to seize us by the throat.

12.3.2 Patent Problems of Autonomous AI Inventions

Machines have learned how to think with the development of science and technology, making AI similar to humans in many ways. For instance, both have decision-making capabilities. In contrast to how conventional machines execute tasks according to human expectations, AI can, in the absence of human intervention, perform self-learning through algorithms; acquire large quantities of rules, knowledge, and techniques; and produce creative results beyond human expectations. An AIS can acquire data from the outside world and selectively perform certain acts to achieve a specific goal to the greatest extent. AI is significantly different from humans in terms of its functional structure. If AI is regarded as a subject to which the patent law applies, such as natural persons, legal persons, or other organizations, the conventional order of the human domain law will face the risk of reconstruction. For instance, it will have to be determined how AI will use the patent right and whether it will initiate a lawsuit when its patent right is infringed upon. If the legal subject status of AI or other nonhuman entities is recognized by law in the civil sphere, this will imply that humans and AI or other nonhuman entities will have equal legal status and will respect each other's subject status without impairing their interests and rights, which is incompatible with the institutional design order of the human domain law in China.

The legitimacy basis of the patent law includes Locke's labor theory, Hegel's personality theory and utilitarian theory, among others. In particular, the utilitarian theory is widely accepted in academic circles. It embodies certain incentives regarding whether to give the creators of technological innovation results a certain period of exclusive rights to encourage invention and creation or to promote the commercial operation and investment of invention and creation. Nonetheless, this concept is intended for natural or legal persons and does not apply to the legal relationship of autonomous AI inventions. AI does not require an incentive mechanism for its autonomous invention and creation, as it only requires learning through an algorithm. The set target of the patent law is to stimulate a creator's innovation by granting him/her a specific period of exclusive rights. If a creator can produce innovative results without an incentive, it is not necessary to protect the patent right. Accordingly, if AI is considered a creator and treated as the main action object of patent law adjustments, the legitimacy basis of the patent law will likely be challenged. On the other hand, humans are the incentive object of the patent system whose purpose is creation; however, as an AI programmer, the owners and users of AI are not creative laborers. According to the jurisprudence of patents, such a group cannot obtain inventor status and thus has no access to the incentive of the patent system. The resulting paradox is precisely the dilemma of the current patent law concerning the creative products of AI. There are many uncertainties in the way in which these contradictions are reconciled by the patent law.

One feasible means to protect autonomous AI inventions is to learn from the first-mover advantage protection mode. The first-mover advantage protection mode is based on considering humans as the legal subject, and denies the patentability of autonomous AI inventions. However, the entry of autonomous AI inventions into the

public domain does not mean the entry cannot act as an incentive mechanism for AI market investors. It is advisable to use the first-mover advantage of entering the market first to maintain a competitive advantage, control the relevant domain markets, and encourage R&D personnel to produce continuous innovation and create value accordingly. Therefore, the market value in the public domain can still be mined by applying the first-mover advantage rule. The denial of the patentability of autonomous AI inventions by the first-mover advantage protection mode is related to the fact that some countries do not recognize the legal subject status of nonhuman entities in legal policies. For instance, in judicial practices, the United States once made a judgment on the legal subject status of a nonhuman entity. In the case of Naruto vs. Slater (i.e., the case of Monkey Selfies), the San Francisco Court believed that although an animal was the direct creator of the creations involved, the animal (nonhuman) did not have legal subject status. Therefore, the case does not involve legal problems, such as suing or pursuing copyright infringement. In particular, the monkey did not own the copyright of the selfies, and the court further ordered the creations involved to be placed in the public domain. In addition, the practices of the US Copyright Office can serve as a reference for this protection mode in terms of legal policies.

The US Copyright Office has gradually become aware of the possible legal problems caused by AI creations, and it explicitly noted in the *Compendium of US Copyright Office Practices* [21] in 2014 that the Office would not register works produced by a machine or mere mechanical process that operates randomly or automatically without any creative input or intervention from a human author. It can be seen that there is no extraterritorial legal policy to use for autonomous AI inventions in the patent field; however, there has been practical experience in dealing with the creations of nonhuman entities in relevant fields, especially in the field of copyright. Copyright and patent rights share the same legislative purpose in terms of encouraging innovation, and one feasible way is to place the creations or inventions of nonhuman entities in the public domain.

The practice of placing autonomous AI inventions in the public domain by the first-mover advantage protection mode does not signify losing the legal incentive mechanism. First, a first-mover can take the lead in pushing AI with the autonomous invention function into the market and gain an advantage in terms of time. AI autonomous inventions do not have patentability; however, AI itself is an object protected by the patent law. A programmer of AI, based on the exclusive feature of patents, promotes an AI technology capable of autonomous invention into the market ahead of other competitors. The first-mover advantage is often an inbuilt advantage of a product. For AI products, cutting-edge scientific technologies will create a generous profit space for the first-mover and allow products and technologies that enter the market first to maintain a long-lasting competitive advantage. Second, the products and technologies that enter the market first can more easily improve the AI patent itself and ensure better product performance than those of competitor. Even if the competitors enter the same market and launch first-generation AI products, those entering the market first may have put the second- and third-generation products into the market. Finally, if consumers form a sense of brand trust and dependence on the first-mover's AI products and become accustomed to the usage mode, they will be unwilling to

purchase products from competitors. In particular, even if autonomous AI inventions fall under the public domain, the first-mover still can control the sales market and maintain a competitive advantage based on the AI patent itself and the first-mover advantage to encourage AI programmers to continuously develop and improve AI products and technologies.

12.3.3 Impact of AI on Humans

It is likely that AI technology will lead to the reconstruction of social rights. With the help of AI technology, enterprises can manage users via persona and label them differently. On the one hand, this helps to learn users' preferences and behavioral characteristics and makes it more convenient and efficient to recommend commodities to users. On the other hand, users are divided into different types and groups through labeling, which, to some extent, signifies that humans are redivided. For instance, automobile enterprises divide users into different levels according to their purchase records, financial data, and purchasing abilities. Then, the enterprises recommend high-configuration vehicles to users with a high purchasing ability and low-configuration vehicles to users with a low purchasing ability. The use of such sensitive personal information causes a large asymmetry of rights between users and enterprises, which is an inevitable effect produced by the application of AI technology to humans.

People's prejudices may also be reinforced by AI technology. With the rapid development of social networking, people are able to post comments and exchange views via different social media. AI technology can connect people with similar views and make mutual recommendations; thus, people with similar views can form groups. Thus, it is likely that people with prejudices can form corresponding groups and reinforce their prejudices.

AI technology also makes humans increasingly lazy. Social information and knowledge can be processed rapidly via AI technology, and humans will experience decreased demand for knowledge and will instead rely heavily on the decision-making of AI. For instance, the news reports we read on a daily basis are to a large extent selected by AI technology according to our habits. We are thus gradually losing the habit of proactively acquiring knowledge, and tend to accept the knowledge that AI allows us to passively acquire. AI technology is thus increasingly involved in the integration and processing of knowledge while human demand for knowledge is getting simpler and decreasing. Furthermore, AI may eventually govern human demands.

As a high-tech and high-consumption product, AI technology may turn into another new cultural isolation tool after becoming a new selling point in a consumerist society. People who lack basic knowledge about AI, do not understand AI, or do not use AI technology are more likely to be marginalized and isolated from mainstream society, further intensifying the polarization of human society as a whole. AI will further promote the division of labor based on specialization for humans and

create many new work opportunities; however, not everyone will be able to cross the technical and social barriers.

12.3.4 Desires Provoked by AI

Human desire is endless, and the development of AI technology will likely arouse various desires in humans. Thus, a thought-provoking question is whether expressing desires through robots can be considered social progress or social regression produced by AI technology.

To a certain extent, AI technology can replace humans to perform specific tasks. Examples are robotic vacuum cleaners, glass-cleaning robots, and logistics robots. With the development of such robots, both enterprises and employees hope that robots can replace manpower. Enterprises use robots to improve work efficiency and strive to maximize the use of robot employees, as there are no problems such as limits to commuting for robots, while employees hope that robots can allow them to avoid physical exertion. Because of various commercial, political, and military purposes, AI products have emerged one after another and continually stimulated desires in human society. An advanced and intelligent product can satisfy human desires in the short term; however, human desires are endless, and there is no limit to human innovation, which will unavoidably lead to the creation of more diverse AI products.

It is also likely that AI technology will lead social individuals to become addicted to the virtual intelligent world. Humans are often confronted with many obstacles and difficulties in the real world, and desires that cannot be satisfied in the real world are likely to be satisfied in the virtual intelligent world. For instance, there are often large crowds of people in most scenic locations; however, virtual reality technology allows users to immersively visit scenic locations without going out. In another example, if an individual wishes to vent their negative feelings due to various disappointments, they can destroy various virtual objects in a virtual world without having to assume real responsibilities. With the powerful abilities of AI, people are often no longer satisfied with lifeless sci-tech products and wish to invent a human plaything to meet their desires. In a way, humans can play God and control everything in the virtual intelligent environment without the need to bear real responsibilities.

However, the blind pursuit of a rational lifestyle created by science and technology may lead humans to become lost in the trap of excessive consumption, forget about the true nature of life, and unwittingly live a false life. Human society is a living world, and as the source of power for the living world, humans should follow the principle that existence precedes essence. However, AI embodies the opposite principle. Humans are active, creative, and vital in the living world, and accordingly, the living world is changing, diversified, and transcends the scientific world. If we live and think in a completely digital, quantified way, we will lose our subjectivity and initiative and will not be able to lead happy lives. Thus, only by treating the living world as the starting point can we more rationally employ AI technology and understand the scientific world represented by intelligent society.

12.3.5 Endangerment of Mankind by Super AI

The main reason why AI does not currently pose a threat to humans is that it is still in its initial stage and can fulfill tasks satisfactorily and achieve better results than humans only in specific areas. However, AI cannot outperform humans in general tasks. Super AI, in contrast, may make breakthroughs in general tasks with the rapid development of AI technology. By then, however, will super AI awaken? Will we reach the singularity? Will the awakened super AI endanger us human? These concerns may appear alarmist because, after all, super AI still seems unattainable at present. However, throughout the history of computers, new technologies have developed at an astonishing pace, and it is difficult to guarantee that super AI will not appear within several decades. A major problem to be solved in the future development of AI is how to protect humans and achieve harmonious coexistence between super AI and humans.

12.4 Security and Controllability of AI

12.4.1 Establishment of Laws, Regulations, Rules, and Policies to Protect Humans

Experience in the development of new technologies indicates that not all technological results can benefit mankind. Therefore, not all sci-tech development is encouraged by law, such as human-cloning technology and genome-editing technology. However, AI technology has displayed its development value. While recognizing the significant boost provided by AI to human productivity, we must also place great emphasis on the possible subversive influence of AI on conventional social structures, social relationships, and moral and ethical relationships. We should delimit the forbidden areas of AI in strict accordance with some methods.

On the one hand, we should adopt effective institutional designs to meet the objective requirements for the rapid development of AI. In addition, we should make active use of AI to reduce repetitive human labor and realize overall human development as well as actively improve the interrelationship between humans, nature, and society. On the other hand, we should emphasize the role of law in technological development. As of now, there are inadequate practical data on the safety problems caused by AI. In the early legislative process of AI technology, a larger number of restrictions should be imposed on autonomous activities. As AI technology experiences constant improvement and humans are increasingly able to coexist with intelligent machines, constant revisions can be made to the law to gradually lift the restrictions on the autonomous activities of AI.

Only according to specific national conditions can a law play the most effective role. Our law should take into account the development needs of AI technology and meet these needs to the greatest extent through proper institutional design. From the

perspective of AI security, a top priority is to formulate AI management regulations, specify China's fundamental attitude toward the development of the AI and robot industry, and promulgate the Ethical Review Measures for Artificial Intelligence Products, Guide to Artificial Intelligence Product Design, and Evaluation Standards for Artificial Intelligence System Safety and Security in an effort to prevent security problems produced by AI technology through advanced legislation.

The Institute of Electrical and Electronics Engineers of the United States has stipulated that to resolve moral questions, an AIS must have accountability at the procedural level and be able to prove why it works in a specific way. We believe that AI is different from other technologies in that many AI technologies are autonomous. For instance, an autonomous robot has the capacity for physical action in real society. Without appropriate preventive measures, such a robot may cause great harm if a serious error occurs. Therefore, we should develop preventive measures before the occurrence of serious problems.

Given the risk that AIAs, including robots, may harm humans, as described in Chap. 8, AIA safety hoops should be set, as a fuse is set in a power supply system. An AIA safety hoop is like the Monkey King's gold hoop that is firmly controlled by an incantation. The AIA safety hoop is designed to control out-of-control robots with a special *incantation*, causing the robots to no longer be powerful.

From a legal perspective, relevant management regulations should be formulated to require robots and AISs capable of destruction to be installed with a restriction system such as the AIA safety hoop, just as an automobile must have a brake system, and such requirements should serve as a prerequisite for market access to protect humans against potential threats from robots. In addition, the design standard for AIA safety hoops should be developed accordingly so that robot-manufacturing enterprises can focus on robot manufacturing only, and AIA safety hoops can be customized by professional departments. Then, because of the standardization of AIA safety hoops, a simple emergency mode that can interrupt robots' operation can be created in addition to standardized control when humans cannot effectively deal with out-of-control robots at critical times.

12.4.2 Establishing the Principles of Human First and Safety First

Modern countries have formed a people-oriented legislative concept, requiring all legislation to focus on the improvement of human living conditions and environments, to promote human welfare and enhance the overall development of humans. In the legislation on AI technology, the application of AI technology not only leads to profound social changes but also affects human development, such as affecting how humans perceive themselves and even endangering the survival of mankind. Therefore, the concept of *human first* must be defined in relevant legislation, and the fundamental purposes must respect human existence, health, interests, and safety.

Legislation on AI technology cannot simply focus on its technical connotation; it should pay more attention to its cultural connotation. In particular, institutional design should not only consider technical specifications at the program operation level but also contain laws that are characterized by humanistic care and social ethics and embraces the perfect integration of technology and morals. First, we should resolutely exclude AI technologies and products that violate social morality and have the potential to challenge the bottom line of human ethics from the protection of the law. Second, by adopting laws, policies, and regulations, we should exercise strict control over those technologies and products that may affect human ethics. When necessary, we should perform targeted restrictions and leave sufficient security redundancy on the application of technologies with uncertain derivative security risks to prevent potentially devastating blows to mankind caused by out-of-control technologies.

12.4.3 Cautiously Recognizing the Legal Subject Qualification of Intelligent Robots

The development of AI has gradually revealed 3As: advanced, automated, and autonomous. Apart from being a human tool, an intelligent robot is capable of thinking and creation, and in the absence of human interference, it can perform invention and creation autonomously, such as writing poems, painting, and automatically generating news reports. Nonetheless, the rule design for a legal order is typically centered on people. Therefore, it is an important question whether robots that are capable of thinking and creation fall under the category of people or things in legal terms, and whether the inventions created by AI technologies or products have patentability. With respect to the ownership of inventions and the participation of interests, it should be established how to coordinate and balance its relationships with AI programmers, owners, and users. These questions affect the applicability of the law and the development of the AI industry. Countries and regions such as the US, EU, and Japan have started to study protection acts applicable to AI. For instance, US Congress proposed the Future of Artificial Intelligence Act [22] in December 2017, with close attention paid to the impact of AI technology on economic development, privacy protection, and labor employment. The European AI Alliance released a draft legislation on AI technology to solicit suggestions concerning human initiative and monitoring, privacy and security, privacy and data supervision, transparency, diversification and peace, social welfare, and an accountability system.

As described in Sect. 7.5.3 of this book, robot Sophia was granted Saudi Arabian citizenship in 2017, which was the first time that a robot was granted legal subject status [23]. In 2013, American lawyer John Frank Weaver noted in his book *Robots Are People Too: How Siri, Google Car, and Artificial Intelligence Will Force Us to Change Our Laws* [24] that because robots already have many abilities of humans, such as the abilities of object discrimination, thinking, purposeful activity, and

autonomous decision-making within a given range, they should be granted the same legal status as humans. The question thus remains whether we should recognize the legal status of robots and how we should deal with the inventions of robots.

For instance, when a user uses an AIS to write a poem, it is unclear who owns the copyright of the poem created: the user using the AIS or the AIS itself. Furthermore, it is unclear whether a robot is owned by the robot-manufacturing enterprise or the user owning the robot. In the event of an accident in a self-driving vehicle, another open question is whether the enterprise that developed the vehicle, the vehicle owner, or the AI product itself should be held accountable. If robots are recognized as having the same legal status as humans, the accountability of robots and AISs should be established for security/safety problems caused by AI. These legal problems require prompt solutions.

12.4.4 Fully Respecting the Public's Right to Know

AI technology and its applications concern technological innovation and product R&D as well as major historical changes with far-reaching effects on human society that will involve everyone's vital interests. The powerful super AI will likely deprive humans in each field and industry of employment opportunities and severely limit the living space of every individual life. Therefore, as technology is closely related to all humans, public access to information and knowledge on AI development and security is not only a sufficient condition for the integration of AI into social life but also a requirement to safeguard humans' legitimate rights and interests. Thus, every human should have the full right to know about AI technology and the implications of extensive application of AI technology and robots.

The formulation of relevant laws and regulations should focus on the following: The public's right to know and participate should be adequately protected, and the application of major AI technologies should require the extensive solicitation of public opinions and rational argumentation. In addition, it should be emphasized that any AI product should not be developed at the expense of human interests and social interests. Given the evident superiority of AI technology, two points should be noted in developing relevant systems. On the one hand, sufficient protection should be given to support the enthusiasm and creativity of R&D personnel to encourage them to invent and create more and higher-quality AI products. On the other hand, the public should be assured of timely access to relevant information and a clear view of the social and economic benefits and possible safety risks arising from the progress of AI technology.

12.4.5 AI Security Assessment and Testing

The security risks that arise as AI develops have already become a threat to humans and require the intervention of corresponding management tools. In particular, the existence of AIAs requires us to have corresponding means to protect humans. We also require means of management in addition to means of protection, such as the AIA safety hoop preventive technique highlighted in Chap. 8. With their advancement, automobiles can cause traffic accidents and pose potential threat to humans Therefore, strict safety assessment methods have been designed for automobiles that require automobiles to be tested every year to ensure they are in a safe condition. If the AIA safety hoop becomes an essential component of an AIA, whether the AIA can be launched into the market and whether its routine operation is in a safe state is dependent on the effectiveness of the safety hoop. This is similar to testing whether the braking system of an automobile functions properly. As an AIA becomes increasingly powerful, a more complex testing mode must be designed to detect whether its evolved potential behavioral model will posea threat to humans. An example is the anti-collision function in automobiles, which requires testing to verify whether it functions correctly.

More importantly, it is necessary to design a testing mechanism to evaluate the degree to which an AIA has evolved. For instance, it is important to determine whether humans will be able to design a testing mechanism to identify the level to which AlphaGo has evolved. This, of course, should not be identified simply through competition as it is now, and a method for rapid identification should be proposed. At present, there is no such means because people do not need it or fear that AlphaGo will harm humans. For humans, the level of playing Go is identified by competition, which should also apply to AlphaGo. However, this does not mean there is no rapid testing. Just like testing a classification algorithm, it is easier to quickly test the classification algorithm if the training and test examples are separated. Likewise, we may be able to extract some of the intermediate processes in Go by playing it as a test method. In short, we should attempt to construct a corresponding testing mechanism to evaluate the degree to which an AIA has evolved. This mechanism is essential for future technological development and social progress.

12.4.6 Joint Efforts from All Social Sectors to Develop Safe and Controllable AI

All sectors of society must contribute their respective strengths for the development of safe and controllable AI. For government policymakers, it is crucial to develop top-level programs for secure and usable AI. The safe development of AI can be promoted from the aspects of policies, laws, and regulations, and the malicious use of AI should be restricted. Enterprises should abide by the relevant standards and regulations on AI safety and security in developing intelligent systems and guarantee the safety

and usability of AISs. Research institutes should intensively study basic theory to ensure AI security and make breakthroughs in intelligent algorithm design, data assessment, security evaluation, and other key technologies. From the perspective of public opinion, we should increase the publicity of AI, help the general public form a correct understanding of applying AI technology in a safe and secure way, and monitor the potential safety hazards posed by AI. Safe and controllable AI requires the joint efforts from all sectors of society, and the potential safety hazards posed by AI should be addressed to embrace a bright future of AI.

References

1. Lecun Y, Bottou L, Bengio Y, et al. Gradient-based learning applied to document recognition. Proc IEEE. 1998;86(11):2278–324.
2. Goodfellow I, Shlens J, Szegedy C. Explaining and harnessing adversarial examples. In: International conference on learning representations (ICLR); 2015.
3. Szegedy C, Zaremba W, Sutskever I, et al. Intriguing properties of neural networks. In: International conference on learning representations (ICLR); 2014.
4. Redmon J, Farhadi A. YOLO9000: better, faster, stronger. (2016–12–25) [2019–9–17]. https://arxiv.org/abs/1612.08242.
5. Thys S, Ranst WV, Goedeme T. Fooling automated surveillance cameras: adversarial patches to attack person detection. (2019–04–18) [2019–09–19]. https://arxiv.org/abs/1904.08653.
6. Sharif M, Bhagavatula S, Bauer L, et al. Accessorize to a crime: real and stealthy attacks on state-of-the-art face recognition. In: Proceedings of ACM SIGSAC conference on computer and communications security; 2016.
7. Eykholt K, Evtimov I, Fernandes E, et al. Robust physical-world attacks on deep learning. In: Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR); 2018.
8. Cortes C, Vapnik V. Support-vector networks. Mach Learn. 1995;20(3):273–97.
9. He K, Zhang X, Ren S, et al. Deep residual learning for image recognition. In: IEEE conference on computer vision and pattern recognition (CVPR); 2016.
10. Du SS, Lee JD, Li H, et al. Gradient descent finds global minima of deep neural networks. (2018–11–09) [2019–09–19]. https://arxiv.org/abs/1811.03804.
11. Conn A. AI and robotics researchers boycott south korea tech institute over development of AI weapons technology. (2018–4–4) [2019–11–7]. https://futureoflife.org/2018/04/04/ai-and-robotics-researchers-boycott-kaist/.
12. 沙特两处石油设施遭无人机袭击. (2019–9–14) [2019–11–20]. http://www.xinhuanet.com//world/2019-09/14/c_1124996790.htm.
13. Brundage M, Avin S, Clark J, etc. The malicious use of artificial intelligence: forecasting, prevention, and mitigation. (2018–02–20) [2019–09–19]. http://arxiv.org/abs/1802.07228.
14. Goodfellow IJ, Pouget-Abadie J, etc. Generative adversarial networks. (2014–06–10) [2019–09–19]. https://arxiv.org/abs/1406.2661.
15. 全球首个"AI合成主播"在新华社上岗. (2018–11–07) [2019–11–07]. http://www.xinhuanet.com/politics/2018-11/07/c_1123678126.htm.
16. Schneider S. Artificial you: AI and the future of your mind. Princeton, New Jersey: Princeton University Press; 2019.
17. Kurzwell R. The singularity is near: when humans transcend biology. London: Penguin Books; 2006.
18. Armstrong D, Michelle M. Researcher controls colleague's motions in 1st human brain-to-brain interface. (2013–8–27) [2019–11–7]. https://www.washington.edu/news/2013/08/27/researcher-controls-colleagues-motions-in-1st-human-brain-to-brain-interface.

19. Disrupted humanity—move 37. (2016–7–25) [2019–11–20]. https://disruptionhub.com/disrupted-humanity-move-37/.
20. Kreye A. The state of artificial intelligence (white paper). (2018–11) [2019–11–07] https://www.t-systems.com/blob/844312/176a021160c485202a46e7917dd2d367/DL_WP_the-state-of-artificial-intelligence.pdf.
21. Compendium of U. S. copyright office practices. (2014–12–22) [2019–11–07] https://www.copyright.gov/comp3/docs/compendium-12-22-14.pdf.
22. Future of Artificial Intelligence Act of 2017. (2017–12–12) [2019–11–07] https://www.congress.gov/bill/115th-congress/house-bill/4625/text.
23. Everything you need to know about Sophia, The world's first robot citizen. (2017–11–07) [2019–11–07]. https://www.forbes.com/sites/zarastone/2017/11/07/everything-you-need-to-know-about-sophia-the-worlds-first-robot-citizen/.
24. Weaver JF. Robots are people too: how Siri, Google Car, and artificial intelligence will force us to change our laws. Santa Barbara, CA: Praeger; 2013.