Translation

CSET CENTER *for* SECURITY *and* EMERGING TECHNOLOGY

The following white paper co-authored by a PRC state think tank describes the importance and difficulty of improving the "trustworthiness" of AI systems. The authors recommend increased use of methods such as federated learning and differential privacy to strengthen AI systems' capability to withstand cyberattacks. The white paper's policy recommendations include drafting more Chinese legislation related to trustworthy AI, developing commercial AI insurance policies, and taking a cautious approach to research on artificial general intelligence (AGI).

Title
White Paper on Trustworthy Artificial Intelligence
可信人工智能白皮书

Authors
China Academy of Information and Communications Technology (CAICT; 中国信息通信研究院; 中国信通院) and JD Explore Academy (京东探索研究院). CAICT is a think tank under China's Ministry of Industry and Information Technology (MIIT; 工业和信息化部; 工信部). JD Explore Academy is a think tank under huge PRC technology and e-commerce company JD (京东) dedicated to cutting-edge scientific research.

Source
CAICT website, July 9, 2021.

The Chinese source text is available online at: http://m.caict.ac.cn/yjcg/202107/P020210709319866413974.pdf
An archived version of the Chinese source text is available online at: https://perma.cc/9XZR-8KNE
US $1 ≈ 6.5 Chinese Yuan Renminbi (RMB), as of August 25, 2021.

Translation Date	Translator	Editor
August 25, 2021	Etcetera Language Group, Inc.	Ben Murphy, CSET Translation Lead

White Paper on Trustworthy Artificial Intelligence

China Academy of Information and Communications Technology (CAICT)
JD Explore Academy
July 2021

Preface

At present, the new generation of artificial intelligence (AI) technology is developing rapidly, and its penetration into all fields of society is accelerating, bringing profound changes to human production and life. While AI brings great opportunities, it also contains risks and challenges. When presiding over the ninth collective study session of the Central Committee Politburo in October 2018, General Secretary Xi Jinping emphasized that “it is necessary to strengthen our judgment of the potential risks of the development of artificial intelligence and to strengthen our watchfulness against them, to safeguard the interests of the people and national security, and to ensure the security, reliability, and controllability of artificial intelligence.” Enhancing confidence in the use of AI and promoting the healthy development of the AI industry has become an important concern.

The development of trustworthy AI is becoming a global consensus. In June 2019, the Group of Twenty (G20) proposed the “G20 AI Principles,” emphasizing the need to be people-centered (以人为本) and develop trustworthy AI. These principles have also been universally recognized by the international community. The European Union (EU) and the United States have also placed the enhancement of user trust and the development of trustworthy AI at the core of their AI ethics and governance. In the future, translating abstract AI principles into concrete practices and implementing them into technologies, products, and applications is an inevitable choice when responding to social concerns, resolving outstanding contradictions, and preventing security risks. It is an important issue related to the long-term development of AI and is an urgent task that industry must quickly address.

Whether reviewing the background and history of trustworthy AI or looking forward to the future of new generation AI, this white paper holds that the stability, explainability, and fairness of AI are the core issues of concern to all parties. Given the present circumstances, this white paper begins from the perspective of implementation of the global AI governance consensus with a focus on trustworthy AI technology, industry, and industry practices, analyzes credible paths to achieve controllable, reliable, transparent, and explainable AI with privacy protection, clear responsibilities, and diversity and tolerance, and puts forward suggestions for the future development of trustworthy AI.

Since AI is still in a stage of rapid development, our understanding of trustworthy AI must be further deepened. We welcome your criticism and correction for any deficiencies in this white paper.

Contents

1. Development background of trustworthy AI................................ 1

(1) Risks of AI technology have triggered a crisis of trust........... 1

(2) All walks of life around the world attach great importance to trustworthy AI.. 2

(3) Trustworthy AI requires systematic and methodological guidance 5

2. Trustworthy AI Framework.. 6

3. Trustworthy AI supporting technology.................................... 8

(1) AI system stability technology.................................... 8

(2) AI explainability enhancement technology.......................... 9

(3) AI privacy protection technology.................................. 10

(4) AI fairness technology.. 11

4. Practical path to trustworthy AI....................................... 11

(1) Corporate level... 12

(2) Industry level.. 16

5. Recommendations for the development of trustworthy AI.................. 17

(1) At the government level, the advancement of China's AI supervision and legislation should be accelerated................... 17

(2) At the technology research level, a systematic and forward-looking layout must be fully implemented........................... 18

(3) The corporate practice level must align with business development to achieve agility and trustworthiness.................. 18

(4) An exchange and cooperation platform to create a trustworthy ecosystem must be built at the industrial organization level......... 18

References.. 20

List of Figures

Figure 1 Number of papers related to trustworthy AI. 3

Figure 2 Corporate Trustworthy AI in Practice 4

Figure 3 Core content of trustworthy AI. 6

Figure 4 Overall framework of trustworthy AI 6

Figure 5 Key Terms in 84 AI Ethics Documents from Around the World. 8

List of Tables

Table 1 Inherent Biases Common in Datasets 16

1. Development background of trustworthy AI

As an important driving force for the new round of science and technology (S&T) revolution and industrial transformation, artificial intelligence is having a major and profound impact on economic development, social progress, and international political and economic patterns. In 2020, the AI industry maintained steady growth, as, according to International Data Corporation (IDC) estimates, the scale of the global AI industry was USD 156.5 billion, an increase of 12% year-on-year. According to the China Academy of Information and Communications Technology (CAICT), China's industry scale reached approximately USD 43.4 billion ([Chinese] yuan Renminbi [RMB] 303.1 billion), for a year-on-year increase of 15%. While AI brings great opportunities, it also contains risks and challenges. General Secretary Xi Jinping attaches great importance to AI governance work, emphasizes the need to "ensure the security, reliability, and controllability of artificial intelligence," and advocates for promoting the implementation of the G20 AI Principles and for leading the healthy development of global AI.

(1) Risks of AI technology have triggered a crisis of trust

At present, the breadth and depth of AI applications continue to expand, and it is becoming an important component of information infrastructures. However, in this process, AI has also continuously exposed certain hidden risks, which are mainly reflected in the following aspects:

Application risks caused by algorithm security: AI technology with deep learning at its core has the flaws of fragility and vulnerability, making it difficult to obtain sufficient trust in the reliability of AI systems. For example, Uber's self-driving car failed to recognize a pedestrian on the road in time, and the pedestrian was struck and died. According to *Fortune* magazine, an AI company used 3D masks and synthetic photos to carry out deception attacks and successfully cracked the face recognition systems of multiple countries.[1]

Black box model causes algorithms to be opaque: Deep learning has a high degree of complexity and uncertainty, which can easily lead to unpredictable risks. Because people cannot intuitively understand the reasons behind decision-making, the further integration of AI into traditional industries has been hindered. For example, a school in Texas used an AI system to judge the instructional aptitude of teachers. Because the system could not explain the judgment basis for controversial decisions, the school's teachers strongly protested, and the system was eventually taken offline.

Data discrimination leads to biases in intelligent decision-making: The results of AI algorithms will be affected by training cases. Therefore, if there is bias and discrimination in the training cases, the algorithm will be affected by such discriminatory data, and the bias and discrimination in the data will be further consolidated, leading to prejudice in the intelligent decisions generated by the AI algorithm. For example, the Criminal Risk Assessment System (COMPAS) used by the Chicago courts in the United States was proven to discriminate against black people.[2]

The complexity of system decision-making makes it difficult to define who is at fault when accidents occur: The automated decision-making of an AI system is affected by many factors, making it difficult to define who is responsible. Regarding the frequent occurrence of application safety accidents such as those involving autonomous driving and robots, legal experts have said that it would still be difficult for AI itself to be considered as a new subject of

[1] https://new.qq.com/omn/20191230/20191230A0FX0R00.html

[2] https://www.sohu.com/a/299700146_358040

infringement liability under current law. However, the specific behavior of AI is controlled by the program. When infringement occurs, the issue of whether the owner or the software developer is responsible still warrants further discussion.[3]

Data misuse leads to the risk of privacy leakage: The frequent use of biometric information increases the possibility of personal data leakage. Once the data is lost, it will result in major security risks. For example, when [PRC face-swapping app] ZAO collected facial data in violation of the terms of its user agreements,[4] aggravated users were concerned that the abuse of private data may cause security risks related to facial scan-based payments and identity authentication.

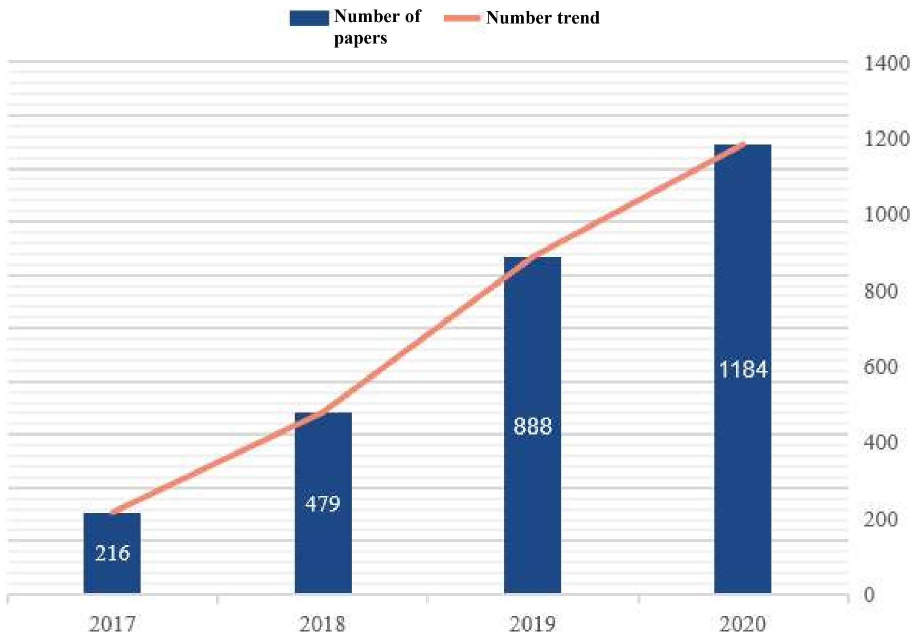
(2) All walks of life around the world attach great importance to trustworthy AI

Facing global anxiety caused by a lack of trust in AI, the development of trustworthy AI has become a global consensus. In June 2019, the G20 proposed the G20 AI Principles. The five government recommendations clearly propose the "promotion of public and private investment in AI research and development so as to promote the innovation of trustworthy AI, as it is necessary to create a strategic environment to pave the way for the deployment of trustworthy AI systems." These have become the principles of AI development generally recognized by the international community.

The academic community first opened the door to trustworthy AI. A Chinese scientist, Academician He Jifeng (何积丰), proposed the concept of trustworthy AI, that is, that AI technology itself has trustworthy qualities, in China for the first time at the S36 academic seminar of the Xiangshan Science Conferences in November 2017. From the perspective of academic research, the research category of trustworthy AI includes security, explainability, fairness, privacy protection, and other related aspects. The number of trustworthy AI research papers in 2020 has increased by nearly five times compared with 2017; the United States' Defense Advanced Research Projects Agency (DARPA) released an academic report *Explainable Artificial Intelligence* and carried out related funding activities to promote the development of trustworthy AI; the top conference, that of the Association for the Advancement of Artificial Intelligence (AAAI), has organized Explainable AI seminars for two consecutive years and has maintained a trend of exciting research. At the same time, research on the fairness, accountability, and transparency of machine learning has formed a "Fairness, Accountability, and Transparency in Machine Learning (FAccT ML)" community. On this basis, the Association for Computing Machinery (ACM) has initiated the academic conference ACM FAccT (ACM Conference on Fairness, Accountability, and Transparency) for four consecutive years since 2018.

[3] http://media.people.com.cn/n1/2018/0502/c40606-29959959.html

[4] http://finance.china.com.cn/industry/company/20190909/5075700.shtml



Source: Web of Science official website

Figure 1 Number of papers related to trustworthy AI[5]

Governments place the enhancement of user trust and the development of trustworthy AI at the core of AI ethics and governance. The EU *Artificial Intelligence White Paper*[1] in 2020 proposed an AI "trusted ecosystem," aiming to implement the European AI regulatory framework and put forward mandatory regulatory requirements for high-risk AI systems. In December of the same year, the White House issued an executive order entitled *Promoting the Use of Trustworthy Artificial Intelligence in Government*,[6] establishing guidelines for the use of AI by federal agencies with the aim of promoting public acceptance and trust in the government's use of AI technology in decision-making.

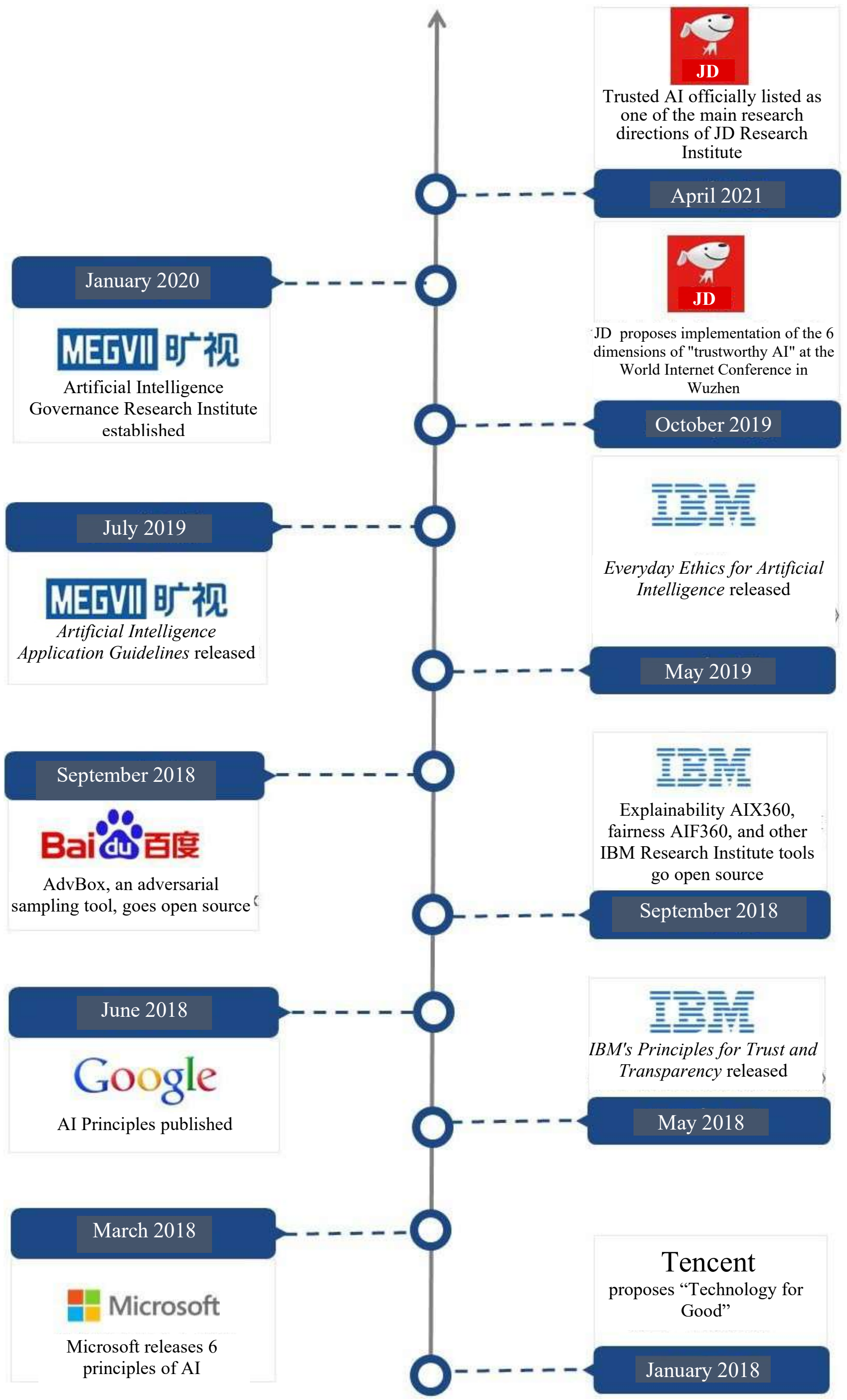
Standardization organizations have laid out trustworthy AI standards. The International Organization for Standardization/International Electrotechnical Commission (ISO/IEC) JTC1 SC42 has specially set up the WG3 Trustworthy Artificial Intelligence Working Group and has released *Information Technology, Artificial Intelligence - Towards a Trustworthy AI*, and is advancing the *Information Technology, Artificial Intelligence - Assessing the Robustness of Neural Networks* series of studies. The Artificial Intelligence Subcommittee of the National Information Technology Standardization Technical Committee (SAC/TC 28/SC 42) was established in China to promote related research simultaneously. In November 2020, the TC260 working group of the National Information Security Standardization Technical Committee (NISSTC) issued a draft of the *Guide to the Practice of Cybersecurity Standards – Guidelines on the Code of Ethics for Artificial Intelligence*, proposing normative guidelines for the safe development of AI-related activities in response to the potential ethical and moral issues of AI.

Corporations are actively exploring and practicing trustworthy AI. IBM Research AI developed a number of trustworthy AI tools in 2018 to evaluate and test the fairness, robustness, explainability, accountability, and value consistency of AI products in the research and

[5] The China Academy of Information and Communications Technology has been searched and organized based on the Web of Science.

[6] https://www.thepaper.cn/newsDetail_forward_10263830

development process. These tools have been donated to the Linux Foundation and have become open source projects. Domestic and foreign companies, such as Microsoft, Google, JD, Tencent, and Megvii, are also actively carrying out relevant practical work. Figure 2 summarizes the exploration of trustworthy AI by several companies.


JD
Trusted AI officially listed as one of the main research directions of JD Research Institute
April 2021
January 2020
MEGVII 旷视
Artificial Intelligence Governance Research Institute established
JD
JD proposes implementation of the 6 dimensions of "trustworthy AI" at the World Internet Conference in Wuzhen
October 2019
July 2019
MEGVII 旷视
Artificial Intelligence Application Guidelines released
IBM
Everyday Ethics for Artificial Intelligence released
May 2019
September 2018
Baidu 百度
AdvBox, an adversarial sampling tool, goes open source
IBM
Explainability AIX360, fairness AIF360, and other IBM Research Institute tools go open source
September 2018
June 2018
Google
AI Principles published
IBM
IBM's Principles for Trust and Transparency released
May 2018
March 2018
Microsoft
Microsoft releases 6 principles of AI
Tencent
proposes "Technology for Good"
January 2018

Source: Data compilation

Figure 2 Corporate Trustworthy AI in Practice[7]

Based on the statements of all parties, this white paper holds that "trustworthiness" reflects

[7] According to public information.

the trustworthiness of AI systems, products, and services in terms of security, reliability, explainability, and accountability. **Trustworthy AI implements ethical governance requirements from the perspective of technology and engineering practice to achieve an effective balance between innovative development and risk governance.** In the future, with the continuous development of AI technology and the AI industry, the meaning of trustworthy AI will continue to be enriched.

(3) Trustworthy AI requires systematic and methodological guidance

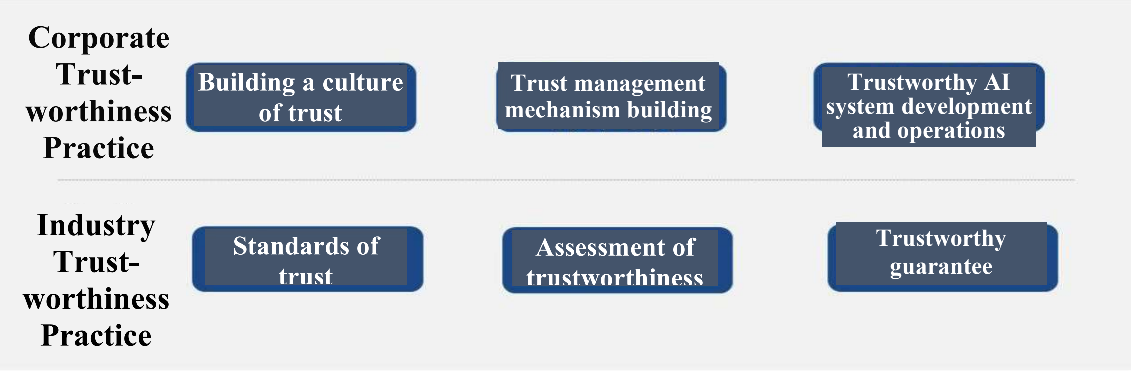
Requirements for trustworthy AI and the practicability of evaluation methods continue to grow. All countries are aware that if ethics and other "soft" constraints lack a corresponding implementation mechanism, ethics washing is prone to occur.[8] As such, greater means for operability are necessary. In February 2021, Germany released the AI Cloud Service Compliance Criteria Catalogue (AIC4),[9] defining how to evaluate the trustworthiness of AI in the cloud environment from the practical level. In April, the European Commission announced the "Development of Artificial Intelligence Uniform Rules (Artificial Intelligence Law)" and revised relevant legislative proposals. This proposed a balanced and commensurate approach to horizontal supervision of AI, with a four-level risk framework for AI categorized around people's livelihoods and people's basic rights and interests with stipulations for corresponding methods of penalization. The intent of these rules is to increase market trustworthiness through legal means so as to promote the promotion and implementation of AI technology as well as its trustworthiness. In May, the National Institute of Standards and Technology (NIST) of the United States proposed a method for evaluating user trust in AI systems and released *Trust and Artificial Intelligence* (NISTIR 8332),[10] defining trust experiences when evaluating humans using AI systems from a practical level. In June, the U.S. Department of Defense (DoD) committed itself to building trustworthy AI capabilities through education and training and to implementing supervision throughout the procurement life cycle through systems engineering and risk management methods.

The process of AI legislation is accelerating, but specific rules still must be further clarified. At the same time, the industry's exploration of trustworthy AI has gradually entered the deep end (深水区). Overall, **the practice of trustworthy AI is still in a relatively decentralized state and lacks a systematic methodology necessary** to fully implement relevant governance requirements and systematically implement practical guidelines for relevant operations. Based on this, this white paper **proposes a "trustworthy AI framework"** on the basis of a comprehensive review of AI ethical constraints, normative legislation, and best practices to serve as a set of methodologies for implementing AI governance requirements. This paper intends to begin from an industrial dimension, focusing on enterprises with an in-depth analysis of the industry's trustworthiness practices in the hopes of building a bridge between AI governance and industry practice.

[8] https://www.technologyreview.com/2019/12/27/57/ai-ethics-washing-time-to-act/

[9] https://www.bsi.bund.de/SharedDocs/Downloads/EN/BSI/CloudComputing/AIC4/AI-Cloud-Service-Compliance- Criteria-Catalogue_AIC4.html

[10] https://www.nist.gov/news-events/news/2021/05/nist-proposes-method-evaluating-user-trust-artificial-intelligence- systems


Corporate Trust-worthiness Practice
Building a culture of trust
Trust management mechanism building
Trustworthy AI system development and operations
Industry Trust-worthiness Practice
Standards of trust
Assessment of trustworthiness
Trustworthy guarantee

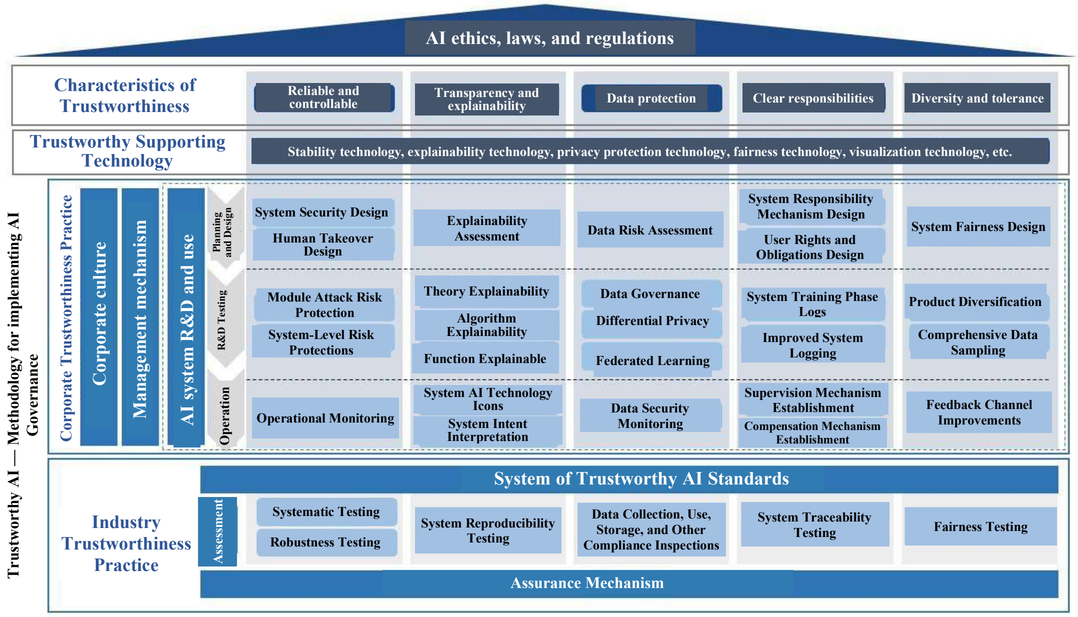
Source: CAICT

Figure 3 Core content of trustworthy AI

At the level of corporate trustworthiness practice, the framework takes the life cycle of the corproate AI system as a reference, combines the five requirements for trustworthiness, proposes practical requirements for each link of the cycle, and provides detailed suggestions for the establishment of a culture of corporate trustworthiness and trust management mechanisms. **At the level of industry trustworthiness practice,** the framework elaborates in detail from the three dimensions of standards, assessments, and guarantees.

2. Trustworthy AI Framework

Trustworthy AI has been proposed from academia, has been actively researched by many, and has been implemented in the industry; its meaning is becoming more enriched and is gradually evolving. This white paper holds that trustworthy AI is no longer limited to the definition of AI technology, products, and services itself, but has rather gradually expanded to a set of systematic methodologies, involving all steps towards building “trustworthy” AI. Figure 4 shows the overall framework of trustworthy AI.


AI ethics, laws, and regulations
Characteristics of Trustworthiness
Reliable and controllable
Transparency and explainability
Data protection
Clear responsibilities
Diversity and tolerance
Trustworthy Supporting Technology
Stability technology, explainability technology, privacy protection technology, fairness technology, visualization technology, etc.
Trustworthy AI — Methodology for implementing AI Governance
Corporate Trustworthiness Practice
Corporate culture
Management mechanism
AI system R&D and use
Planning and Design
R&D Testing
Operation
System Security Design
Human Takeover Design
Explainability Assessment
Data Risk Assessment
System Responsibility Mechanism Design
User Rights and Obligations Design
System Fairness Design
Module Attack Risk Protection
System-Level Risk Protections
Theory Explainability
Algorithm Explainability
Function Explainable
Data Governance
Differential Privacy
Federated Learning
System Training Phase Logs
Improved System Logging
Product Diversification
Comprehensive Data Sampling
Operational Monitoring
System AI Technology Icons
System Intent Interpretation
Data Security Monitoring
Supervision Mechanism Establishment
Compensation Mechanism Establishment
Feedback Channel Improvements
Industry Trustworthiness Practice
System of Trustworthy AI Standards
Assessment
Systematic Testing
Robustness Testing
System Reproducibility Testing
Data Collection, Use, Storage, and Other Compliance Inspections
System Traceability Testing
Fairness Testing
Assurance Mechanism

Source: CAICT

Figure 4 Overall framework of trustworthy AI

Trustworthy AI is an important practice for the implementation of AI governance. The characteristics of trustworthiness it follows align with the requirements of AI ethics and related laws and regulations, and they are all people-centered. From the perspective of governance, compared with ethics that provide guidance from a macro level and laws that implement result-oriented constraints, trustworthy AI penetrates into internal management, research and development (R&D), operations, and other aspects of enterprises, as well as industry-related work **to translate relevant abstract requirements into specific capability requirements required for practice**, thereby enhancing society's trust in AI.

Trustworthiness characteristics level: By sorting out 84 policy documents that have been issued globally according to word frequency, we can see that current AI governance principles have converged around the five aspects of **transparency, security, fairness, accountability, and privacy protection**.[2] Despite differences in cultural backgrounds, nature of business, and management systems from one organization to the next and despite their different tendencies in the understanding and implementation of these common principles, from an industrial perspective, **the core concepts of the above five consensuses are all refined and proposed around how to build multi-party trustworthy AI.** These five consensuses provide guidance on how to enhance trust between the supply side and the demand side in the use of AI and assist regulatory agencies in fostering a trustworthy and healthy industry ecosystem. This white paper refers to the five global consensuses (Figure 5), China's Artificial Intelligence Industry Alliance (AIIA) initiative, and the *Joint Pledge on Artificial Intelligence Industry Self-Discipline*[3] and the *Guidelines for Trustworthy AI Operations*[4] to summarize and put forward the five characteristics of trustworthiness of **reliability, controllability, transparency and explainability, data protection, clear responsibilities, and diversity and tolerance** to guide the operational capabilities required to practice trustworthy AI.

伦理原则 Ethical principle	**文档数量** Number of documents	**关键词** Keywords
透明度 Transparency	73/84	Transparency, explainability, excitability, understandability, explainability, communication, disclosure, showing
正义与公平 Justice and fairness	68/84	Justice, fairness, consistency, inclusion, equality, equity, (non-) bias, (non-) discrimination, diversity, plurality, accessibility, reversibility, remedy, redress, challenge, access and distribution
非恶意行为 Non-maleficence	60/84	Non-maleficence, security, safety, harm, protection, precaution, prevention, integrity (bodily or mental), non-subversion
责任 Responsibility	60/84	Responsibility, accountability, liability, acting with integrity
隐私权 Privacy	47/84	Privacy, personal or private information
仁慈 Beneficence	41/84	Benefit, beneficence, well-being, peace, social good, common good
自由与自治 Freedom and autonomy	34/84	Freedom, autonomy, consent, autonomy choice, self-determination, liberty, empowerment

信任 Trust	28/84	Trust
可持续性 Sustainability	14/84	Sustainability, environment (nature), energy, resources (energy)
尊严 Dignity	13/84	Dignity
团结 Solidarity	6/84	Solidarity, social security, cohesion

Source: Data compilation

Figure 5 Key Terms in 84 AI Ethics Documents from Around the World[11]

At the level of trustworthy supporting technology, attention must be paid to requirements for trustworthiness characteristics such as reliability and controllability, transparency and explainability, data protection, clear responsibilities, and diversity and tolerance. Theoretical research and technological innovation should be the core starting points in order to make up for the deficiencies of current technology. For instance, the study of new generation of explainable AI algorithms and privacy computing technologies with privacy protection capabilities requires active exploration by academia and industry.

At the level of corporate trustworthiness practices, the widespread application of AI in society depends on the commercialization of technology by enterprises and other entities. To this end, the practice of trustworthy AI by enterprises is a key part of the trustworthy methodology. It should be realized that there is no such thing as perfect technology. The key lies in how to use technology correctly: on the one hand, it is necessary to cultivate a culture of trustworthiness and build a trustworthiness management system; on the other hand, it is necessary to implement trustworthy characteristics throughout the entire life cycle of the development and use of AI systems to ensure "trustworthy" quality from the source.

At the level of industry trustworthiness practice, trustworthy AI requires the participation and practice of the entire industry. This mainly applies to the establishment of a system of trustworthy AI standards, trustworthy AI assessment and testing, and AI trustworthiness assurance. Through the establishment of insurance and other social (社会化) methods, the risks brought by AI technology and system applications can be shared.

3. Trustworthy AI supporting technology

With the continuous attention of all sectors of society to the issue of trust in AI, safe and reliable AI technology has become a hot subject in research.[5] [6] The focus of research is mainly improvements to the stability, explainability, privacy protections, and fairness of AI systems. These technologies constitute the basic supporting capabilities of trustworthy AI.

(1) AI system stability technology

AI systems are faced with unique interference, which comes from a variety of attacks against data and systems, including [data] poisoning attacks, adversarial attacks, backdoor attacks, and the like. These attack techniques can be independent of one other or can exist in tandem. For instance, a poisoning attack utilizes malicious comments in accordance with special

[11] Quoted from *The Global Landscape of AI Ethics Guidelines*, compiled by CAICT

rules to input interference data into the training dataset, which in turn affects the accuracy of the recommendation system;[17] a confrontational attack can mislead the autopilot system to misrecognize the information on the road sign by pasting a specially designed pattern on the road traffic sign, thereby causing a traffic accident;[18] backdoor attacks are concealed and may be used to launch attacks on the AI supply chain. Compared to traditional software systems, this type of interference places higher requirements on the stability of AI systems.

The stability of AI has also been the impetus of ongoing and extensive research. Confrontational attacks and poisoning attacks against AI models appeared as early as 2012 and 2013. Specifically, the purpose of a confrontational attack is to induce errors in the decision-making of the AI system by constructing targeted samples; the purpose of the poisoning attack is to degrade the performance of the training model by injecting poisoned samples into the training dataset of the AI model. Since then, confrontational attacks have developed Fast Gradient Sign Method (FGSM), Carlini-Wagner, and Projected Gradient Descent (PGD) attack methods; the development of poisoning attacks has also been very rapid, resulting in the emergence of backdoor attacks. Backdoor attacks use backdoor samples to implant backdoors into the AI system to achieve directional manipulation of the AI system itself. This attack is similar to a poisoning attack, and backdoors are often implanted into the system through poisoning attacks. In order to resist such attacks, projects have proposed various abnormal data detection methods to detect and remove malicious data such as adversarial samples, poisoned samples, backdoor samples, and the like, thereby reducing the interference caused by malicious attacks; for adversarial attacks, adversarial training on adversarial samples can be conducted; to resist backdoor attacks, techniques such as model pruning and backdoor detection can be utilized.

The stability of AI still faces greater challenges. On the one hand, various interference methods continue to emerge in an endless stream just as they continue to evolve, and new attack methods may easily invalidate old defense methods; on the other hand, forms of interference are gradually spreading from the digital world to the physical world, such as printing adversarial samples to directly cause physical interference to autonomous driving and facial recognition systems. In the future, research on AI stability technology will continue to expand.

(2) AI explainability enhancement technology

At present, the operation of an AI system with a deep learning algorithm at its core is akin to a black box: people can only see the data that goes in and the data that comes out; the internal working principles and basis for judgment are unknown. On the one hand, it may not be clear why a trained AI model has extremely high performance; on the other hand, it may also be unclear which factors the AI system relies on when making decisions.

Research into the explainability of AI algorithms is still at an early stage, and the theoretical framework of certain algorithms must to be improved.[7] [8] [9] **Error! Bookmark not defined.** [15]. For example, the effectiveness of optimization algorithms have been well proven in certain simple AI models such as decision trees and support vector machines. However, as to why the stochastic gradient descent algorithm can efficiently optimize deep neural networks, a great deal of research has been carried out in academia but the consensus on this problem is still inconclusive. For another example, academia has achieved certain outcomes through experiments into how AI models use data characteristics to make predictions, but there is still a lack of theoretical support. In order to make AI models more explainable, researchers have proposed that an appropriate visualization mechanism can be established to attempt to evaluate and explain the intermediate states of the model; the influence of the training data on the final convergent AI model can be analyzed through an influence function; which data features the AI model uses to make predictions can be analyzed through the Gradient-weighted Class Activation Mapping (Grad-CAM) method; the Local Interpretable Model-agnostic Explanations (LIME) method can utilize a simple explainable model to locally approximate the complex black box model and study the explainability of the given black box model. Some studies have also suggested that the reproducibility of the AI system's implementation process can be improved by establishing a complete model training management mechanism.

As the AI industry becomes more firmly established in everyday life, the behavior of AI systems should be made more transparent, easier to understand, and more trustworthy to humans. Demanding blind trust in the decisions made by AI systems without an explanation of decision-making processes will greatly limit the popularization of AI systems in key areas such as national defense, law, healthcare, and education, and may even result in serious problems in society. Enhancement of the explainability of AI systems demands great urgency.

(3) AI privacy protection technology

AI systems must rely on a large amount of data, but the flow of data and the AI model itself may leak sensitive private data. For example, at any stage of data flow, a malicious attacker can launch an attack on an anonymous dataset to steal data; in the data release stage, a malicious attacker can use identity re-identification to launch an attack on an anonymous dataset to steal private information; malicious attackers can also directly attack the AI model to steal private information. For example, a model inversion attack can infer and reconstruct training data based on the output of the attacked model, thereby stealing private information; a member inference attack can infer whether a given data sample comes from the training dataset of the attacked model, thereby divulging private information.

Academia has proposed a variety of targeted protection methods for the above-mentioned privacy leakage problems, and the most common solutions are privacy protection methods based on differential privacy and federated learning. Differential privacy was first developed in a 2006 proposal by American scholar Cynthia Dwork,[120] and it serves as a major quantitative indicator of the privacy protection capabilities of AI systems. The core idea behind this proposal is that an AI algorithm with superior privacy protection capabilities should be insensitive to small disturbances in the input data. Based on this, downsampling, sequence replacement (顺序置换), and noise can be added to data to prevent attackers from jeopardizing privacy. In 2016, a project by Google applied differential privacy to deep learning for the first time. It enhanced the privacy protection capabilities of deep models by adding Gaussian noise to the gradient during model training. This work demonstrates the application potential of the differential privacy method in large-scale AI models. At present, several top technology companies have applied the differential privacy method in their actual operations. Federated learning[19] was proposed in 2015 to train AI models without collecting user data so as to protect private information. Specifically, federated learning deploys the model to the user device; each user device uses its own private data to calculate the gradient of the model parameters and then uploads it to the central server; the central server then merges the collected gradients and sends them back to the user device; each user device then uses the merged gradient to update the model parameters. It should be pointed out that some preliminary studies have shown that federated learning still poses a certain degree of risk of disclosing private information. Experiments have shown that federated learning may leak a certain amount of local user data.[11] At the same time, some theories have pointed out that federated learning may even weaken the privacy protection capabilities of AI systems to a certain extent.[12] Therefore, federated learning must be further optimized for it to improve user privacy protection capabilities. A feasible direction would be to combine federated learning and differential privacy together in order to build an AI system with stronger privacy protection capabilities.

In the current era, an ever-growing amount of private information is contained in data. People have begun to pay more attention to the protection of private data, and some countries have also begun to formulate regulations on the use of private data at the legislative level. Research on privacy protection can allow AI systems to comply with the basic norms and requirements of such legislation and also improve the establishment of trustworthy AI.

(4) AI fairness technology

With the widespread application of AI systems, such systems have exhibited unfair decision-making behaviors and discrimination against certain groups. Academia holds that the main reasons for such decision-making biases are as follows: limited by data collection conditions, the weights of different groups in the data are unbalanced; the AI model trained on the unbalanced dataset may then be applied to the overall data, and while performance is sacrificed on a small amount of data, the model's decision-making becomes unfair.

In order to ensure the fairness of decision-making in AI systems, relevant researchers have mainly constructed completely heterogeneous datasets to minimize inherent discrimination and bias in the data; datasets are then checked periodically to ensure the high quality of the data. In addition, there are also algorithms that use fair decision-making quantitative indicators to reduce or eliminate decision-making bias and potential discrimination. Such existing fairness indicators can be divided into two categories: individual fairness and group fairness.[13] [16] [221] Specifically, individual fairness measures the degree of prejudice of intelligent decision-making toward different individuals, and group fairness measures the degree of prejudice of intelligent decision-making toward different groups. Alternatively, algorithms based on fairness indicators can be roughly divided into three categories: preprocessing methods, processing methods, and post-processing methods. The preprocessing method cleans the data by deleting sensitive information or by re-sampling so as to reduce the deviation in the data. The processing method improves the fairness of the trained model by adding regular items related to fairness quantification in the training process of the AI model. For instance, some projects have used Rényi correlation as the regular item with a min-max optimization algorithm to reduce any potential correlations between model predictions and sensitive attributes. The post-processing method further improves the fairness of the trained model by adjusting the model's output. For example, there is a project that has proposed the multiaccuracy boost method based on the concept of multiaccuracy to reduce the decision-making bias of black-box AI systems.

The number of AI applications in sensitive fields continues to grow, such as in hiring, criminal justice, and healthcare, and its fairness has also been the subject of widespread concern. Fairness technology can balance data from a technical perspective, thereby further guiding the model to give fair results, which is of great significance for improving the fairness of decision-making in AI systems.

Today, more and more research is focused on the challenges of stability, explainability, privacy protection, fairness, and other issues of AI. As such research continues to deepen, more stable, more transparent, and fairer theories and technologies of AI will inevitably emerge. These technologies will serve as the cornerstone and an important guarantee for the realization of trustworthy AI in the future.

4. Practical path to trustworthy AI

This white paper refers to the relevant content of the *Guidelines for Trustworthy AI Operations* issued by China's Artificial Intelligence Industry Alliance (AIIA), combined with research and interviews on the actual research and development of AI companies, and summarizes a practical path to trustworthy AI at the corporate and industry levels.

(1) Corporate level

Enterprises are the core subject of the research and development and use of AI technology, products, or services, and serve as most important entity in the practice of trustworthy AI. The practice of trustworthy AI in enterprises is a holistic, developing, non-traditional system engineering act that must begin with corporate culture and management systems and yet must also fully implement relevant technical requirements in the research and development of AI systems.

1. **Incorporate trustworthy AI into corporate culture**

Corporate culture is a concrete manifestation of an enterprise's overall values, common vision, mission, and way of thinking. If an enterprise wants to develop trustworthy AI, it must integrate concepts of trustworthiness into its corporate culture.

(1) Corporate managers must recognize the "trustworthy" direction

As the core of corporate operations, corporate management must reach a consensus on the development of trustworthy AI, fully establish people-centered values, and recognize the characteristics of transparency and explainability, diversity and tolerance, reliability and control, clear responsibilities, and privacy protection, and integrate trustworthy AI into all aspects of business management to promote improvements to the overall trustworthiness of the corporation.

(2) Employees should enhance "trustworthy" learning and practice

Enterprises can formulate learning and training plans for trustworthy AI by inviting external experts to speak and by distributing trustworthy AI books or introductory materials to popularize the concept of "trustworthiness" among employees or promote the use of trustworthy technologies or tools, thereby encouraging employees to continuously innovate and practice trustworthy AI at work.

(3) Enterprises must create a "trustworthy" cultural atmosphere

Enterprises can reflect the elements of trustworthy AI in their office space, websites, promotional materials, and press releases, exhibit their own practical cases of exploring trustworthy AI, encourage employees to discuss the topic of trustworthy AI, and inspire teams or individuals to make contributions to the practice of trustworthy AI.

2. **Improve the management systems of trustworthy AI**

The management system is the basis for the implementation of management behavior and the guarantee for the smooth progress of social reproduction. If an enterprise wants to realize trustworthy AI, it must be reflected in the management system.

(1) Establish a trustworthy AI team

Dedicated teams (or virtual organizations) responsible for the management of trustworthy AI should be established within enterprises. We recommend that the main person in charge of the company assume the leadership position, which is conducive to direct command and coordination of other departments for participation in trustworthy AI-related work. Such a group can be subdivided into several sub-groups according to business specifics; personnel within the group should include both full-time and part-time personnel with a legal or R&D background. Clear responsibilities and obligations should be implemented for departments and related personnel.

(2) Establish and implement a trustworthy AI personnel management system

Led by the trustworthy AI management department with the cooperation of human resources, R&D, legal, and other departments, a trustworthy AI personnel management system should be jointly formulated primarily to manage personnel involved with internal AI demand analysis, product design, R&D, testing, and trustworthiness, and to clarify requirements for personnel management, education and training, and assessments. The personnel management system should be effectively implemented with regular education, training, and assessment of relevant personnel and with gradual improvements to elevate the professional level of such personnel.

(3) Establish and implement a management system for the development and use of trustworthy AI systems

A management system should be established for AI systems in the R&D stage to clarify responsible departments and personnel, work content, work methods, work processes, and work requirements. Such a system should be supervised and implemented by the trustworthy AI management department. Management systems should be expressly provided for each use phase of the trustworthy AI system and emergency plans and relief measures should be formulated to ensure that the system can meet trustworthiness requirements during each use phase or that problems can be effectively addressed in time if they occur so as to minimize damage and reduce losses.

(4) Provide the necessary resources to achieve trustworthy AI

Enterprises must do a good job of overall planning and must allocate necessary resources for the realization of trustworthy AI, including but not limited to necessary personnel, funding, venues, and facilities.

(5) Establish an iteration and upgrade mechanism for the system

Enterprises should stay up to date with changes in AI governance and the introduction of relevant policies and regulations. Under the leadership of the trustworthy AI management department, enterprises should continue to optimize and improve the management system according to actual circumstances so as to ensure that they can adapt in time to achieve optimal results.

3. Embed trustworthy AI requirements into the entire R&D application process

(1) Planning and design stage

At the beginning of the AI system life cycle, enterprises must fully consider the implementation of the characteristic elements of trustworthy AI and affirm concepts of trustworthiness in key aspects of planning and design, such as demand analysis and detailed system design, so that subsequent R&D testing and operations can always meet the core requirements of trustworthy AI.

Combining the common processes of current software product design, enterprises can assist product teams in formulating trustworthy design plans for AI systems through a specially established trustworthy team in two ways:

The first is to put forward trustworthy design requirements for AI systems. After completing product demand analysis, the potential risks faced by an AI system should be fully investigated, and targeted countermeasures, such as system security, failure protection mechanisms, explainability, data risk, system responsibility mechanisms, user rights and obligations, and system fairness, should be proposed along with a corresponding list of trustworthy design requirements.

The second is to review the trustworthy design plan of the AI system. Experts in various professional fields within the trustworthy team must combine their own professional knowledge, work experience, and case work to verify the feasibility of the trustworthy design scheme for the AI system, discover potential problems, provide enlightening guidance and ideas for extending the trustworthy design along with suggestions for subsequent revision and improvements to the trustworthy design plan to ensure that the core features of trustworthy AI are integrated into system design to reduce trust loopholes and prevent the occurrence of accidents from potential risks.

(2) R&D testing stage

In terms of reliability and controllability, efforts should be made to improve the defense capabilities of AI systems and ensure human supervision and takeover of power. The defense capabilities of the AI system itself can be improved in two ways: at the data level and at the model level. Defense methods at the data level include malicious data pre-cleaning and data augmentation to improve model robustness. At the model level, in addition to traditional security defense methods such as encrypting the models in the production environment and limiting the number of malicious query interactions with the models in the production environment, another important method is adversarial training. AI models can be easily interfered with by specially constructed attack samples. An adversarial training algorithm can use adversarial samples to train the AI models, thereby improving the robustness of the model against adversarial samples and making the model less likely to suffer from adversarial sample interference. In addition, backup plans must also be set up for the AI system during R&D to ensure that the system can automatically adjust and recover, be quickly taken over by professional staff, or have service terminated through a "one-button shutdown."

In terms of transparency and explainability, the focus should be on improving the reproducibility of AI systems. Current research into algorithmic explainability still lags behind the rapid development of AI technology in applications. To this end, enterprises should begin with improving the reproducibility of systems in the R&D and testing phases to not only enhance the transparency of the system but also, to a certain extent, reduce the difficulty of subsequent system audits and accountability traceability. The main associated measures include: First, establishing a complete dataset management mechanism, combining existing data management strategies and tools, recording, in detail, the source and composition of the training set and test set during the training process of each version of the system as well as the data preprocessing operations used in the training process; second, establishing a complete model training management mechanism and recording, in detail, the hardware platform, system configuration, software framework, model version, model initialization, hyperparameters, optimization algorithm, distributed operations strategy, network speed, indicators, test results, and other techniques and engineering technology used when training the model.

In terms of data protection, data governance should be carried out to avoid problems

such as illegal collection, abuse, and leakage of training data and should explore the use of privacy protection algorithms to train AI systems. Privacy protection capabilities of AI systems should be improved at the algorithm level, using technologies such as differential privacy or federated learning, such as are used by Apple and the U.S. Census in user data collection, for instance. OpenDP, an AI project co-founded by Microsoft and Harvard University, has developed many open source differential privacy toolkits to provide more protection for models and data.

In terms of clarification of responsibilities, the implementation process of the AI system should be fully audited to improve the traceability of the system and ensure the trustworthiness of the source of the system and services. The main steps of an audit include data preparation, model training, and model evaluation. An audit of the data preparation process helps to confirm whether the collection of training data is legal and compliant, whether it involves infringement of privacy, whether the data processing complies with standard labeling and preprocessing methods, and whether the data storage uses encryption and access restrictions or other such security measures. Model training is the key to giving "intelligence" to an AI system. A comprehensive audit of key training links such as the hardware platform, software framework, algorithm selection, and tuning process can assist in AI system tracing. Model evaluation can largely reflect the performance and generalization capabilities of an AI system in practical applications, and a standard rigorous evaluation process can often find errors, measure the quality of the model, and judge whether it can meet the design requirements. This helps to trace back problems in the system implementation process so as to make continuous improvements, and as such, it is necessary to audit, in detail, the model's performance and changes in the verification set and test set.

In terms of diversity and tolerance, attention should be paid to the fairness and diversity of training datasets to avoid lack of trust caused by data bias. The performance of an AI system depends on the quality of the training data. The dataset may contain implicit race, gender, or ideological bias (Table 1), which may cause the AI system's decision-making to be inaccurate or biased and discriminatory. Enterprises should focus on improving the diversity and fairness of training data to meet the requirements of diversity and inclusion. On the one hand, attention must be paid to the inherent discrimination and prejudice that may appear in the data and proactive measures should be taken to weaken the impact of such prejudice; on the other hand, the dataset should be reviewed periodically to ensure the high quality of the data. Also, **the testing process should use quantitative indicators based on fair decision-making capabilities to test the AI system.** Currently, specific operations may include:

- Collecting data through reliable and legal sources to ensure the trustworthiness of the data source.
- Checking the accuracy and completeness of the samples, features, and labels in the dataset through statistics or related tool sets and making corresponding adjustments in a timely manner based on the results of the checkup.
- Updating the dataset in time according to changes in the real deployment environment to ensure the timeliness and relevance of the dataset.
- Constructing an easy-to-use dataset format and interface, simplifying the reading and calling process of the dataset, and preventing improper operations.
- In the quantitative analysis of the fair decision-making capabilities of an AI model, appropriate quantitative indicators must be selected according to specific application scenarios

and specific needs, taking into account individual fairness and group fairness indicators.

Table 1 Inherent Biases Common in Datasets

No.	Data Quality	Description
1	Reporting bias	The attributes of the dataset collection situation are manually recorded and cannot accurately reflect the true and objective situation
2	Automation bias	Results generated by automated software tools are inherently biased
3	Selection bias	Samples selected in the dataset fail to reflect the true distribution of the samples
4	Group attribution bias	People tend to generalize the real situation of an individual to the entire group to which they belong
5	Hidden bias	Assumptions are usually made based on models and personal experience that are not necessarily universally applicable

Source: Data compilation

(3) Operational use stage

In the actual operation and use stage of AI, it is necessary to do a good job in explaining the AI system, continuously monitoring various trust risks for the AI system, and actively optimizing the AI system.

The first is to disclose the technical intent of the AI system to users. In the case that the explainability of the algorithm is not yet mature, the understanding of the technical intent of the AI system can start from the establishment of an appropriate human-computer communication mechanism, disclosure of the functional logic and use requirements of the system's decision-making, and clarification of the potential risk of incorrect decision-making in the system. Specifically, when deploying AI systems online, enterprises should establish appropriate human-machine communication mechanisms, such as setting up a functional module, and clearly express them through an easy-to-understand mode of expression, such as text, graphic logos, or voice prompts, to inform the user whether they are currently interacting with the AI system. In the actual application process, users should also be clearly informed about the basic functions, performance, use requirements, object orientation, and role of the system in the decision-making process of the AI system.

The second is to continue to carry out AI risk monitoring. User feedback channels should be established to promptly collect the true opinions of users and subsequently optimize and iterate through the entire system. The various risks of the AI system should be monitored in the actual use process to continuously improve the supervision and compensation mechanisms and carry out responsibility traceability and compensation work in a timely manner when the AI system causes actual damage.

(2) Industry level

The realization of trustworthy AI is not only accomplished by the unilateral practice and efforts of enterprises but also requires the participation and coordination of multiple parties. Ultimately, a healthy ecosystem of mutual influence, mutual support, and interdependence must be formed. This ecosystem mainly includes specific content such as a system of standards, evaluation and verification, and cooperation and exchange.

The first is to build a system of standards for trustworthy AI. Policies and laws can only specify principles and bottom lines. Standards are necessary to provide specific guidance and constraints from an enforceable and achievable level. At present, several countries are formulating or promulgating principles or laws for AI governance. On this basis, specific standards and specifications can be formulated in conjunction with AI technologies, products, or scenarios. For example, in April 2021, the Chinese national standard of *Information Security Technology Facial Recognition Data Security Requirements* was opened to the public for comments. The standard mainly addresses the problems of indiscriminate acquisition, leaks, or loss of facial data, as well as excessive storage and use. Further elaboration and refinement of the provisions related to facial recognition have also been made in the draft of the *Personal Information Protection Law*.

The second is to carry out third-party evaluation and verification. Third-party evaluation and verification is an effective means to test whether the target object meets the relevant requirements. Due to the complexity of AI technology, professional third-party institutions are needed to provide support. Focusing on the characteristics of trustworthy AI, attention must be paid to such aspects as system security, robustness, reproducibility, data protection, traceability, and fairness. AIIA also released the first batch of commercial AI system trustworthiness assessment results in 2020, involving 16 AI systems of 11 enterprises, providing an important reference for user selection; among the latest AI legislation proposals issued by the EU is a proposal that an authoritative third-party organization carry out trustworthiness assessments and other such measures.

The third is to explore marketized insurance mechanisms. The application of AI technology is the same as other information systems: No matter how high the level of protection is, the risk of problems always exists. This requires innovative working methods to transfer risks and losses in other ways. Insurance is an important measure of risk compensation, which can transfer risks to the greatest extent and make up for user losses. We recommend that AI enterprises and insurance institutions explore insurance mechanisms for the application of AI products, conduct quantitative assessments of risk accidents, provide risk compensation, and help improve the trustworthy AI ecosystem.

5. Recommendations for the development of trustworthy AI

Building a trustworthy AI system has become the focus and direction of efforts from many different sectors. Practicing trustworthy AI methodologies helps to improve the trustworthiness of AI and make it more acceptable for the public. **Trustworthy AI is not static. Rather, with the development of AI technology, ethics, and laws, it will continue to evolve to meet new development needs.** This will also present new demands on the various parties involved.

(1) At the government level, the advancement of China's AI supervision and legislation should be accelerated

A systematic legal and regulatory framework should be built for AI. The first is to improve the current laws and regulations to meet the needs of development. On the basis of the *Cybersecurity Law*, *Data Security Law*, and the *Personal Information Protection Law* that will be issued in the future, we must sort out the applicable issues facing the supervision process of AI systems and continuously improve laws and regulations. The second is to advance new legislation work to actively respond to new risks with in-depth study of new problems and new trends caused by AI and a timely review of legislative suggestions. The third is to use innovative methods to promote the implementation of laws, explore the use of pilots, sandboxes, and other supervision methods to develop intelligent supervision tools and continuously improve the efficiency and flexibility of supervision. In addition, we must persist in the overall promotion of the domestic rule of law and foreign rule of law in the field of AI, actively participate in multilateral and bilateral regional cooperation mechanisms, promote the formulation of international AI governance rules, seek consensus, and bridge differences.

(2) At the technology research level, a systematic and forward-looking layout must be fully implemented

Integrated research on trustworthy AI will become an important trend in the future. Current research on trustworthy AI is mostly carried out from singular dimensions, such as security, privacy, and fairness. Existing research work has shown that different requirements such as security, fairness, and explainability are mutually synergistic or restrictive. If only one aspect of a requirement is considered, it may cause conflicts with other requirements. Therefore, it is necessary to build an integrated research framework for trustworthy AI to maintain the optimal dynamic balance between different characteristic elements.

Research on trustworthy artificial general intelligence (AGI) must be laid out in advance. At present, whether it is AI governance or trustworthy AI, most work is carried out for weak AI technology and applications. AGI and even superintelligence have not garnered sufficient attention. Once these emerge, they will be major events tied to the destiny of humankind and require a forward-looking layout, such as exploring the development path of AGI through the development of cutting-edge technologies such as super deep learning (超级深度学习) and quantum machine learning. At the same time, we must also carry out research related to trustworthiness when exploring strong AI.

(3) The corporate practice level must align with business development to achieve agility and trustworthiness

Enterprises must pay attention to the agile iteration of trustworthy AI in the process of expanding the application of AI technology. With the extensive integration of AI technology into different industries, the depth of its application is increasing day by day, and the demands for trustworthiness faced by enterprises will continue to expand, placing higher requirements on trustworthy practice capabilities that enterprises must possess. On the one hand, trustworthy AI detection and monitoring tools must be developed to match the needs of business development, with upgrades and iterations that are based on the uniqueness of industry applications. On the other hand, development should actively connect with regulatory authorities, actively cooperate and participate in regulatory measures such as digital sandboxes, safe harbors, pilot applications, and standards compliance, and build agile trustworthy mechanisms that are coordinated both internally and externally.

(4) An exchange and cooperation platform to create a trustworthy ecosystem must be built at the industrial organization level

Industrial organizations should be encouraged to build specialized communication platforms around the field of trustworthy AI and to call on all parties in the industry to jointly build a trustworthy AI ecosystem. Trustworthy AI is a complex, systematic endeavor. It requires the participation of multiple parties, and full play should be given to the advantages of industrial organizations to extensively absorb exceptional practical experience from all parties and compile trustworthy AI operation guidelines. Attention must be paid to AI R&D management, technical support, and product applications, and a system of trustworthy AI standards should be established and perfected. The development of AI evaluation and monitoring capabilities should be accelerated, and various methods such as evaluation and testing and tracking and monitoring should be utilized to continue to promote the implementation of trustworthy AI within the industry.

References

[1] EUROPEAN COMMISSION. WHITE PAPER On Artificial Intelligence-A European approach to excellence and trust[R/OL]. (2020-02-19) https://ec.europa.eu/info/sites/info/files/commission-white-paper-artificial-intelligence-feb2020_en.pdf

[2] Jobin A., et al. The global landscape of AI ethics guidelines[J]. Nature Machine Intelligence, 2019, 1(2).

[3] Artificial Intelligence Industry Alliance. Joint Pledge on Artificial Intelligence Industry Self-Discipline [R/OL].(2019-0808) http://aiiaorg.cn/uploadfile/2019/0808/20190808053 719487.pdf

[4] Artificial Intelligence Industry Alliance. Guidelines for Trustworthy AI Operations [R/OL].(2020-0923) http://aiiaorg.cn/uploadfile/2020/0923/20200923064 427421.pdf

[5] Zhang Bo (张钹), et al. Towards the third generation of AI [J]. SCIENTIA SINICA Informationis（中国科学：信息科学）, 2020, v.50 (09): 7-28.

[6] He Jifeng. Safe and Trusted Artificial Intelligence [J]. Information Security and Communications Privacy (信息安全与通信保密), 2019 (10): 4-8.

[7] Liu T., et al. Algorithm-dependent generalization bounds for multi-task learning[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, volume 39, pages 227-241, 2016.

[8] He F., et al. Control batch size and learning rate to generalize well: Theoretical and empirical evidence[C]. In Advances in Neural Information Processing Systems, pages 1141-1150, 2019.

[9] Tu Z., et al. Theoretical analysis of adversarial learning: A minimax approach[C]. In Advances in Neural Information Processing Systems, pages 12280–12290, 2019.

[10] Dwork C., et al. The algorithmic foundations of differential privacy[J]. Foundations and Trends in Theoretical Computer Science, volume 9, pages 211–407, 2014.

[11] Zhu, L., et al. Deep leakage from gradients[C]. In Advances in Neural Information Processing Systems, 2019.

[12] He F., et al. Tighter generalization bounds for iterative differentially private learning algorithms[J]. arXiv preprint arXiv:2007.09371, 2020.

[13] Dwork C., et al. Fairness through awareness[C]. In Proceedings of the 3rd Innovations in Theoretical Computer Science Conference, pages 214–226, 2012.

[14] Ribeiro M. T., et al. "Why should i trust you?" Explaining the predictions of any classifier[C]. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pages 1135-1144, 2016.

[15] Ribeiro M. T., et al. Anchors: High-precision model-agnostic explanations[C]. In Proceedings of the AAAI Conference on Artificial Intelligence, 2018.

[16] Calders, T., et al. Building classifiers with independency constraints[C]. IEEE International Conference on Data Mining Workshops. pages 13-18, 2009.

[17] Fang, M., et al. Poisoning attacks to graph-based recommender systems[C]. In Proceedings of the 34th Annual Computer Security Applications Conference, pages 381-392, 2018.

[18] Eykholt, K., et al. Robust physical-world attacks on deep learning visual classification[C]. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 1625-1634, 2018.

[19] McMahan, B., et al. Communication-efficient learning of deep networks from decentralized data[C]. In Artificial Intelligence and Statistics, pages 1273-1282, 2017.

[20] Hardt, M., et al. Equality of opportunity in supervised learning[C]. In Advances in Neural Information Processing Systems, pages 3323-3331, 2016.