Engineering 6 (2020) 302–309

Contents lists available at ScienceDirect

Engineering

journal homepage: www.elsevier.com/locate/eng

Research
Artificial Intelligence—Review

Ethical Principles and Governance Technology Development of AI in China

Wenjun Wu [a,*], Tiejun Huang [b], Ke Gong [c]

[a] State Key Laboratory of Software Development Environment, Beihang University, Beijing 100191, China
[b] School of Electronics Engineering and Computer Science, Peking University, Beijing 100871, China
[c] Chinese Institute of New Generation Artificial Intelligence Development Strategie, Nankai University, Tianjin 300071, China

ARTICLE INFO

Article history:
Received 19 September 2019
Revised 18 November 2019
Accepted 31 December 2019
Available online 8 January 2020

Keywords:
AI ethical principles
AI governance technology
Machine learning
Privacy
Safety
Fairness

ABSTRACT

Ethics and governance are vital to the healthy and sustainable development of artificial intelligence (AI). With the long-term goal of keeping AI beneficial to human society, governments, research organizations, and companies in China have published ethical guidelines and principles for AI, and have launched projects to develop AI governance technologies. This paper presents a survey of these efforts and highlights the preliminary outcomes in China. It also describes the major research challenges in AI governance research and discusses future research directions.

1. Introduction

With the rapid development and deployment of a new generation of artificial intelligence (AI) algorithms and products, AI is playing an increasingly important role in everyday life, and is having a significant impact on the very fabric of the modern society. In particular, AI models and algorithms have been widely adopted in a variety of decision-making scenarios, such as criminal justice, traffic control, financial loans, and medical diagnosis. This emerging proliferation of AI-based automatic decision-making systems is introducing potential risks in many aspects, including safety and fairness.

For example, there are many concerns with the safety of automated driving systems. In 2015, a fatal accident occurred to a Tesla vehicle in China, in which the autopilot system failed to identify a road sweeper truck and did not perform the correct maneuver to avoid it. Another example comes from intelligent justice, in which AI algorithms are adopted to decide whether to grant parole permission to a prisoner based on his/her behavioral characteristics. There have been complaints that such an algorithm could make biased and unfair decisions based on ethnicity and cultural background. In the financial arena, an AI-based digital lending algorithm might reject loan applications based on biased judgment. Government agencies, the academic community, and industry have all realized that the safety and governance of AI applications are an increasingly important issue, and that effective measures must be taken to mitigate potential AI-related risks.

Today, governments from many countries, research organizations, and companies have announced their ethical guidelines, principles, and recommendations for AI. To enforce these principles in current AI systems and products, it is vital to develop governance technology for AI, including federated learning, AI interpretation, rigorous AI safety testing and verification, and AI ethical evaluation. These techniques are still under intense development and are not yet mature enough for widespread commercial adoption. Major technical obstacles are deeply rooted in fundamental challenges for modern AI research, such as human-level moral cognition, commonsense ethical reasoning, and multidisciplinary AI ethics engineering. In this paper, we aim to present a general survey on AI ethical principles and ongoing research efforts from the perspective of China.

The rest of the paper is organized as follows: Section 2 introduces the ethical principles that have been published by government agencies and organizations, and highlights the major research efforts of Chinese researchers on AI governance. Section 3

* Corresponding author.
E-mail address: wwj@nlsde.buaa.edu.cn (W. Wu).

https://doi.org/10.1016/j.eng.2019.12.015

compares China with other countries in terms of AI ethical principles and governance technology development. Section 4 discusses the grand challenges in AI governance research and suggests possible research directions.

2. Ethical principles and emerging governance technology in China

The *Development plan of the new generation artificial intelligence*, which was released in 2017, stresses that the dual technical and social attributes of AI must be carefully managed to ensure that AI is trustable and reliable. In 2019, Ministry of Science and Technology of the People's Republic of China (MOST) established an National Governance Committee for the New Generation Artificial Intelligence and released the *Governance principles for the new generation artificial intelligence—Developing responsible artificial intelligence* [1]. The Beijing Academy of Artificial Intelligence (BAAI) also published the *Beijing AI principles* [2], proposing an initiative for the research, development, use, governance, and long-term planning of AI in order to support the realization of beneficial AI for humankind and the natural environment. In Ref. [3], researchers from the BAAI collected more than 20 well-known proposals of ethical principles for AI, and performed a topic analysis on the texts of these proposals. They identified the following keywords that are commonly mentioned in the proposals: security and privacy, safety and reliability, transparency, accountability, and fairness.

(1) Security and privacy: AI systems should be secure and respect privacy.

(2) Safety and reliability: AI systems should perform reliably and safely.

(3) Transparency: AI systems should be understandable.

(4) Accountability: AI systems should have accountability.

(5) Fairness: AI systems should treat all people fairly.

These common principles have been widely agreed upon by researchers, practitioners, and regulators in the field of AI across the world. These principles not only reflect society's goodwill and moral beliefs, but also demand feasible and comprehensive technical frameworks and solutions to implement ethical constraints in AI models, algorithms, and products. Table 1 lists emerging techniques that hold great potential to support effective governance in accordance with AI ethical principles.

2.1. Data security and privacy

Data security is the most basic and common requirement of ethical principles for AI. Many governments including European Union (EU), United States, and China are establishing legislation to protect data security and privacy. For example, the EU enforced the *General Data Protection Regulation* (GDPR) in 2018, and China enacted the *Cybersecurity Law of the People's Republic of China* in 2017. The establishment of such regulations aims to protect users' personal privacy, and poses new challenges to the data-driven AI development commonly adopted today.

In the paradigm of data-driven AI, developers often need to collect massive data from users in a central repository and carry out subsequent data processing, including data cleaning, fusing, and annotation, to prepare datasets for training deep neural network (DNN) models. However, the newly announced regulations hamper companies from directly collecting and preserving user data on their cloud servers.

Table 1
Major AI ethical principles and supporting governance technologies.

AI ethical principle	AI governance technology
Security and privacy	Federated learning, blockchains
Safety and reliability	Machine learning test and verification
Transparency	Interpretable/explainable AI
Accountability	AI provenance, auditing, and forensic
Fairness	AI fairness evaluation and debiasing algorithm

Federated learning, which can train machine learning models across decentralized institutions, presents a promising solution to allow AI companies to address the serious problem of data fragmentation and isolation in a legal way. Researchers from the Hong Kong University of Science and Technology and other institutes [4] have identified three kinds of federated learning modes: horizontal federated learning, vertical federated learning, and federated transfer learning. Horizontal federated learning is applicable when participating parties have non-overlapping datasets, but share the same feature space in data samples. Vertical federated learning is applicable when the datasets from participants refer to the same group of entities but differ in their feature attributes. When the datasets cannot meet either condition (i.e., they have different data samples and feature space), federated transfer learning is a reasonable choice. Using these modes, AI companies are always able to establish a united model for multiple enterprises without sharing their local data in a centralized place.

Federated learning not only presents a technical solution for privacy protection in the collaborative development of distributed machine learning models among institutions, but also indicates a new business model for developing a trusted digital ecosystem for the sustainable development of an AI society. By running federated learning on blockchain infrastructure, it may be possible to motivate members in the digital ecosystem via smart contracts and trusted profit exchanges to actively share their data and create federated machine learning models.

Federated learning is increasingly being adopted by online financial institutions in China. WeBank established an open-source project on federated learning and contributed the Federated AI Technology Enabler (FATE) framework to the Linux foundation. WeBank's AI team [5] also launched an Institute of Electrical and Electronics Engineers (IEEE) standardization effort for federated learning and has started to draft an architectural framework definition and application guidelines.

2.2. Safety, transparency, and trustworthiness of AI

Decades of research in computer science and software engineering have ensured the safety and trustworthiness of large-scale complex information systems. With increases in the scale and complexity of systems, it is a grand challenge to design and implement a reliable and trustable system in a cost-efficient and error-free manner. The AI components deployed in today's autonomous systems inevitably aggravate this problem when they interact with uncertain and dynamic environments. Because a state-of-the-art AI model adopts very complex DNNs and end-to-end training approaches, it acts as a black box, which not only hampers developers from fully understanding its structure and behavior, but also introduces implicit and potential vulnerabilities to the model from malicious inputs. Therefore, an AI governance framework must encompass multiple techniques enabling AI engineers to perform a systematic evaluation of AI behaviors and to present evidence that can build public's trust toward AI systems. Fig. 1 displays the major building blocks of an AI behavior analysis and assessment framework, including testing, verification, interpretation, and provenance.

These emerging AI governance technologies are all about examining and assessing AI behavior and inner-working mechanisms from different aspects. AI testing often focuses on evaluating the relationship of inputs and outputs to make sure the AI's functions and behavior can conform to the desired goals and moral requirements. AI verification adopts rigorous mathematical models to

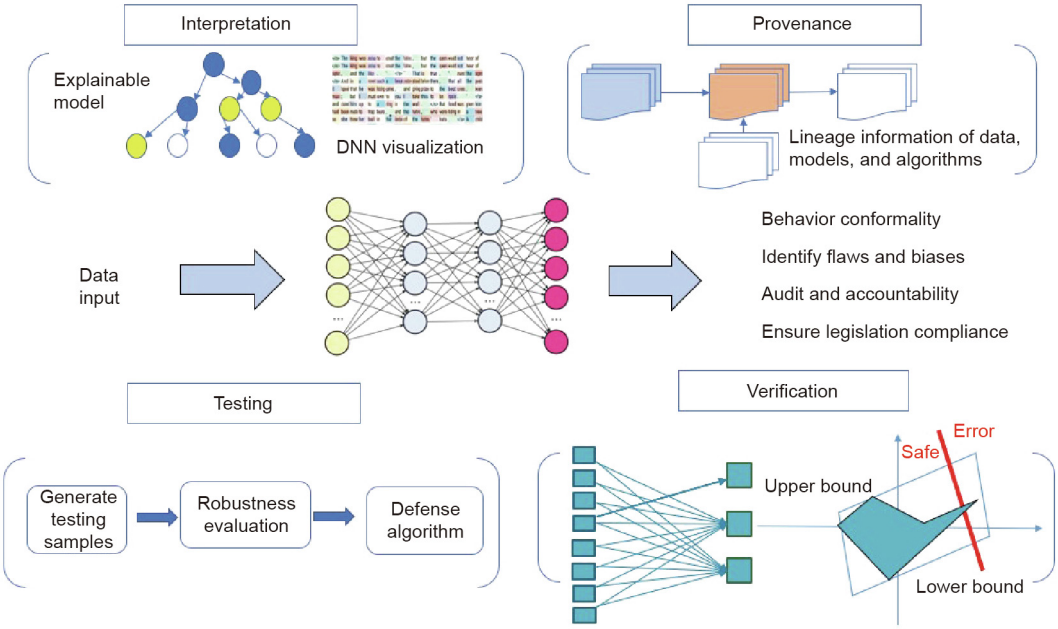

Fig. 1. Testing, verification, interpretation, and provenance for trustworthy AI.

prove the soundness of AI algorithms. AI interpretation aims at developing novel techniques to analyze and reveal how complex DNN models work internally. AI provenance can track the lineage of the training data, model, algorithm, and decision process to support auditing and accountability determination.

The integration of these AI governance technologies is very important because it brings all the stakeholders together to understand, examine, and audit an autonomous and intelligent system. Users who are affected by decisions from an AI system have the right to know and comprehend the rationales behind the algorithmic decisions. Engineers who are in charge of AI development and maintenance must rely upon AI testing, verification, and interpretation tools to diagnose potential problems with AI algorithms and enact the necessary remedies and improvements. Managers who oversee AI engineering processes and the quality of AI products should utilize these tools to query procedural data, guide the enforcement of moral standards, and minimize the ethical and quality risks of the system. Government auditors who investigate the responsibility of AI systems in accidents or legal cases must exploit AI provenance to track the lineage of the system evolution and to collect relevant evidence.

2.2.1. Safety and robustness of AI

Adversarial examples for DNNs have recently become a very popular topic in the machine learning community. DNN models are vulnerable to adversarial examples, in which inputs with imperceptible perturbations mislead DNNs, resulting in incorrect results. For example, hackers can maliciously add small-magnitude perturbations to an image of a street crossing with pedestrians walking on a road, thus generating adversarial examples that can fool DNNs into ignoring the pedestrians on the scene. Therefore, adversarial examples might lead to fatal accidents or pecuniary losses due to the severe impairment of practical deep learning applications such as automated driving and facial recognition systems. Two major approaches are used to address the AI safety issue and ensure the robustness of AI systems under perturbations: adversarial testing and formal verification.

(1) **Adversarial testing for AI safety**. Many studies have investigated how to generate adversarial examples as test cases for testing DNN models. A straightforward way to generate test cases is to directly perturb the original inputs without affecting the overall visual image of the scene. This approach is limited to situations in which hackers have no access to the input sources and cannot add perturbations in the input images. Thus, researchers have started to explore the generative adversarial network (GAN)-based generation of adversarial examples that consist of a tiny image patch that can be easily posted on physical objects such as light poles and a human's hat [6].

Researchers at Beihang University [7] proposed a perceptual-sensitive GAN that can enhance the visual fidelity of adversarial patches and generate more realistic testing samples for neural networks under safety testing. At the 2018 Conference on Neural Information Processing Systems (NIPS), researchers at Tsinghua University [8,9] published two papers on defense algorithms for DNNs: One paper proposed a new adversarial perturbation-based regularization method named deep defense for training DNNs against possible adversarial attacks, and the other suggested minimizing the reverse cross-entropy in the training process in order to detect adversarial examples. Researchers at Zhejiang University and at Alibaba [10] have implemented a DNN-testing platform named DEEPSEC, which incorporates more than a dozen state-of-the-art attack and defense algorithms. This platform enables researchers and practitioners to evaluate the safety of DNN models and to assess the effectiveness of attack and defense algorithms.

(2) **Formal verification of DNN models**. Adversarial testing is unable to enumerate all possible outputs for a given set of inputs due to the astronomical number of choices for the input perturbation. As a complementary method to adversarial testing, formal verification have been introduced to rigorously prove that the outputs of a DNN model are strictly consistent with a specification of interest for all possible inputs. However, verifying neural networks is a difficult problem, and it has been demonstrated that validating even simple properties about their behavior is a non-deterministic polynomial (NP)-complete problem [11].

The difficulties encountered in verification mainly arise from the presence of activation functions and the complex structure of a neural network. To circumvent the difficulties brought by the nonlinearities that are present in neural networks, most recent

results focus on the activation functions of piecewise linear forms. Researchers are working on efficient and scalable verification approaches by focusing on the geometric bounds on the set of outputs. There are basically two kinds of formal verifiers for DNN models: complete verifiers and incomplete verifiers. Complete verifiers can guarantee no false positives but have limited scalability, as they adopt computationally expensive methods such as satisfiability modulo theory (SMT) solvers [12]. Incomplete verifiers may produce false positives, but their scalability is better than that of complete verifiers. Researchers at ETH Zurich [13,14] proposed an incomplete verifier based on abstract interpretation, in which shape-based abstract domains are expressed as the geometric bounds of nonlinear activation functions' outputs to approximate infinite sets of behaviors of DNNs. Researchers from East China Normal University, the Chinese Academy of Science, and other institutes [15,16] have also introduced verification frameworks based on linear programming or symbolic propagation.

These research efforts are still in their early stages and have not been generalized to different kinds of activation functions and neural network structures. Despite decades of effort in the field of formal verification, scalable verification methods are neither available nor mature for processing modern deep learning systems because of the complexity of deep learning.

2.2.2. Transparency and accountability of AI

AI transparency is critical in order for the public to be able to understand and trust AI in many decision-making applications, such as medical diagnosis, loan management, and law enforcement. AI interpretation helps to decipher the complicated inner workings of deep learning models and to generate human-understandable explanations of such models' reasoning and inference. With improved AI transparency, people are more confident in utilizing AI tools to make decisions and in assessing the legitimacy and accountability of autonomous systems.

Research efforts are being conducted on how to build explainable DNN frameworks and analysis tools. In this research direction, multiple approaches have been proposed to support model understanding. Some researchers have devised companion neural networks to generate natural language explanations in the process of DNN inference. Another popular approach called local interpretable model-agnostic explanation (LIME) attempts to construct a proxy model based on a simple model class (e.g., sparse linear models and decision trees) from the original complex model, in order to approximate the behaviors of the original model [17]. Researchers from Shanghai Jiao Tong University and other institutes [18] introduced a decision-tree-based LIME method to quantitatively explain the rationales of each prediction made by a pre-trained convolutional neural network (CNN) at the semantic level.

Information visualization is also widely regarded as an effective way to implement explainable DNN models. Researchers at Tsinghua University [19] presented an interactive DNN visualization and analysis tool to support model understanding and diagnosis. With the right knowledge representation of moral values, such visual analytics may enable AI engineers to intuitively verify whether their DNN models correctly follow human ethical rules.

The other important research field that is closely related to AI interpretation is AI provenance, which emphasizes the recording, presenting, and querying of all kinds of lineage information relevant to data, models, and algorithms for future audits and forensic analysis. Although there are mature data and information provenance frameworks, few investigations have been performed on AI provenance. A joint research paper from Nanjing University and Purdue University [20] designed a provenance computation system for AI algorithms by tracking inner derivative computing steps. This method can assist algorithm designers in diagnosing potential problems.

In addition to facilitating the development of AI models, provenance can play an important role in emerging AI forensic research. The recent well-known misuse of DeepFake technology, which utilizes a GAN to generate false facial images and videos, is posing a significant threat to social norms and security. Many researchers are developing new classification methods to detect these fake images and to ensure the credibility of visual content. For example, researchers at the Institute of Automation, Chinese Academy of Sciences [21] attempted to improve the generalization of DeepFake detection algorithms, and proposed a new forensic CNN model. Nevertheless, these efforts alone are insufficient to overcome DeepFake because malicious designers can always conceive better algorithms to fool known detection algorithms. Perhaps such efforts should be complemented with reliable provenance information for the original images, which would provide necessary clues to verify the legitimacy of an image's origin. In particular, a blockchain-based provenance management system may help to establish a reliable and trustworthy digital ecosystem in which the authentic identity of digital resources can be tracked and verified in order to completely unmask fraudulent images and videos.

2.3. Fairness evaluation of AI algorithms

Fairness has recently emerged as an important nonfunctional characteristic for the evaluation of AI algorithms. Efforts in AI fairness research mostly focus on measuring and discovering the differences of AI outputs among different groups or individuals. Many fairness evaluation criteria have been proposed by researchers. Gajane and Pechenizkiy [22] surveyed how fairness is defined and formalized in the literature for the task of prediction. The major types of definitions of AI fairness are listed below:

(1) Fairness through unawareness. According to this type of definition, an AI algorithm is fair as long as the protected attributes are not explicitly used in the AI-based decision-making process. For example, an intelligent fraud-detection system should exclude sensitive attributes such as race and gender in its feature set for risk estimation. Although this simple and blind approach might work in some cases, it has a very serious limitation, because excluding attributes can degrade predictive performance and, in the long run, yield fewer effective outcomes than an attribute-conscious approach.

(2) Group fairness. This requires the decisions made by an AI algorithm to exhibit an equal probability for user groups divided by a specific attribute. There are several types of group fairness, including demographic parity, equalized odds, and equal opportunity. This family of fairness definitions is attractive because it does not assume any special features of the training data and can be verified easily.

(3) Individual fairness. According to this type of definition, an AI algorithm should present similar decisions if a pair of individuals have similar attributes.

(4) Counterfactual fairness. In many decision scenarios, protected attributes such as racial and gender group may have a causal influence upon the predicted outcome. As a result, the "fairness through unawareness" metric may in fact lead to the group disparity that the metric is intended to avoid. To mitigate such an inherent bias, Kusner et al. [23] formulated a counterfactual fairness definition by leveraging the causal framework to describe the relationship between protected attributes and data. This measurement of fairness also provides a mechanism to interpret the causes of bias.

At present, there is no consensus regarding which fairness definitions are most suitable; in some cases, these definitions are not even compatible with each other. The question of how to choose appropriate fairness criteria for machine learning under specific circumstances and design a fair, intelligent decision algorithm with

full consideration of social context remains an open research problem.

In addition to the multitude of fairness definitions, researchers have introduced different bias-handling algorithms to address the problem of AI fairness at different stages of an AI model's life-cycle. For example, Bolukbasi et al. [24] devised a method to remove gender bias from word embeddings that are commonly used for natural language processing. Researchers at Shanghai Jiao Tong University [25] proposed the use of social welfare functions that encode fairness in the reward mechanism, and suggested that the fairness-of-resource-allocation problem be addressed in the framework of deep reinforcement learning.

Large AI companies are active in developing fairness evaluation and debiasing tools to promote the implementation of AI fairness in real intelligent systems. Google released an interactive visualization tool named What-If that enables data scientists to examine complex machine learning models in an intuitive way. The tool integrates a few fairness metrics, including group unawareness, equal opportunity, and demographic parity, to assess and diagnose the fairness of machine learning models. IBM created AI Fairness 360 [26], an extensible open-source toolkit for handling algorithmic biases. The package integrates a comprehensive set of fairness criteria and debiasing algorithms in datasets and models.

3. A comparison between China and other countries regarding AI ethical principles and governance technology development

In this section, we compare the ongoing efforts of China with those of other countries regarding the development of ethical principles and governance technology for AI. From the governmental and institutional perspective, it can be seen that both governmental agencies and the private sector in China have taken active initiative in building ethical guidelines for AI and in promoting an awareness of the beneficial use of AI. From the perspective of academic research and industrial development, Chinese researchers and practitioners have been actively developing governance technologies for AI along with their international peers.

3.1. Governmental and institutional perspective

The world's major economic powers have released their ethical guidelines and governance regulations for AI. In 2018, the EU announced the GDPR; in April of 2019, the EU's High-Level Expert Group on AI presented the *Ethics guidelines for trustworthy AI* [27]. In 2019, the White House issued the *Executive order on maintaining American leadership in artificial intelligence*, and demanded that the National Institute of Standards and Technology (NIST) devise a plan to develop technical standards for reliable, robust, and trustworthy AI systems [28]. Along with the EU and the United States, China is among the major governments that have launched nationwide AI governance and ethics initiatives. The United Nations (UN) is also promoting AI ethics, and declared its humanistic attitude toward AI at the United Nations Educational, Scientific and Cultural Organization (UNESCO) AI conference in March 2019, which stressed artificial intelligence with human values for sustainable development. However, no multinational joint action has been taken by multiple governments as yet.

In addition, big tech corporations such as Google, Amazon, and Microsoft, as well as their Chinese counterparts Baidu, Alibaba, and Tencent, have been actively involved in AI ethics and governance initiatives, both domestically and internationally. Tencent announced its "available, reliance, comprehensible, controllable" (ARCC) principles for AI in 2018, and released a report on AI ethics in a digital society in 2019 [29]. Baidu joined Partnership on AI [30], which is an international consortium consisting of major players in the AI industry. The mission of this consortium is to establish best practices for AI systems for socially beneficial purposes.

3.2. Academic research and industrial development perspective

In Section 2, we highlighted the development efforts of Chinese researchers in AI ethical principles and emerging governance technologies. In most of the four major areas relevant to AI ethical principles and governance, Chinese researchers have been promptly developing new models, algorithms, and tools in parallel with their international peers.

In the area of data security and privacy (Section 2.1), WeBank's FATE is one of the major open-source projects for federated learning. According to Ref. [31], FATE is the only framework that supports distributed federal learning among these open-source projects, in comparison with Google's TensorFlow federated learning.

In the area of the safety and robustness of AI, since the vulnerability of DNNs was revealed by Szegedy et al. [32], many studies have been carried out worldwide to address this issue. Among these efforts, new algorithms developed by Chinese researchers have demonstrated excellent performance in adversarial testing and defense. At 2017 Conference on NIPS, Google Brain organized an international competition on adversarial attack and defense methods, in which the team from Tsinghua University won the first position in both the attack and the defense tracks [33]. As an example of international cooperation, Baidu has worked with researchers from the University of Michigan and the University of Illinois at Urbana–Champaign to discover the vulnerabilities of DNNs adopted in LiDAR-based autonomous driving detection systems [34].

In the area of the transparency and accountability of AI, Chinese researchers from both the academic community and the private sector, including Alibaba and Baidu, have actively proposed new interpretation methods and visualization tools. Large international companies such as IBM, Facebook, and Microsoft have released their AI explainability tools, which implement general frameworks for AI interpretation. For example, IBM introduced AI Explainability 360, an open-source software toolkit that integrates eight AI interpretation methods and two evaluation metrics [35]. In comparison, Chinese companies should make additional efforts to integrate new algorithms and prototypes into open-source tools and make them widely available to the world.

Although the concept of AI fairness is relatively new, it has received a considerable amount of attention in the AI academic community. As mentioned in Section 2.3, investigation into AI fairness issues often requires an interdisciplinary approach. In 2016, the Association for Computing Machinery Conference on Fairness, Accountability, and Transparency (ACM FAT) was launched, with a focus on ethical issues pertaining to AI such as algorithmic transparency, fairness in machine learning, and bias. This conference attracted more than 500 attendees, including AI academics and scholars in social sciences such ethics, philosophy, law, and public policy. Although this conference has become one of the major venues for research on AI fairness, it is not well known among Chinese AI researchers. It is necessary to encourage more multidisciplinary research on this emerging field in the AI academic community in China.

In summary, governments, academics, and industries across the world have recognized the significance of AI ethical principles and have taken the initiative to develop AI governance technologies. Among the world's major governments, China has launched nationwide AI governance and ethics initiatives. We believe that it is necessary to foster international collaboration in this new field for the sake of the global community and our shared future. It is unfortunate that such joint efforts have not been emphasized at all levels, they must be further extended and strengthened.

4. Grand challenges in AI governance research

In fulfilling the fundamental principles of AI for the good of society, numerous research challenges remain in the task of bringing ethical values and regulations into the current AI governance framework. In this section, we elaborate the major challenges from the following aspects: the AI ethical decision framework, the AI engineering process, and interdisciplinary research.

4.1. The ethical decision framework

The concept of an ethical decision framework is a major topic in AI governance research. Researchers at Hong Kong University of Science and Technology and Nanyang Technological University [36] reviewed publications on existing ethical decision frameworks from leading AI conferences and proposed a taxonomy dividing the field into four areas: exploring ethical dilemmas, individual ethical decision frameworks, collective ethical decision frameworks, and ethics in human–AI interactions. Other researchers [37] presented a survey on artificial general intelligence (AGI) safety, in which the ethical decision problem is often formulated in a reinforcement learning framework. They assumed that rational intelligent agents can learn human moral preferences and rules through their experiences interacting with social environments. Thus, in the framework of reinforcement learning, AI designers can specify ethical values as reward functions in order to align the goal of a rational agent with its human partners and to stimulate the agent to behave according to human moral norms. It should be noted that in this nascent research area, scientists must overcome the main bottlenecks of the current data-driven DNNs to achieve human-level automated moral decision-making and to extensively evaluate these frameworks after their deployment in real and complicated moral circumstances.

4.1.1. How to model moral rules and values

In most cases, it is difficult to directly devise mathematical functions to model ethical values—especially moral dilemmas, where people must make difficult decisions among negative choices. It is viable to take a data-driven and learning-based approach to enable autonomous agents to learn appropriate ethical representations from human demonstrations. For example, researchers from Massachusetts Institute of Technology (MIT) researchers launched the Moral Machine project [38] to collect datasets about various ethical dilemmas in a crowdsourcing way. However, such a crowdsourcing self-reported preference on moral dilemmas can unavoidably deviate from actual decision behaviors because there is no mechanism to ensure genuine user choices.

4.1.2. Common sense and context awareness in ethical decision-making

Despite the rapid progress of modern AI technologies, DNN-based AI agents are mostly good at recognizing latent patterns, and are not very effective in supporting general cognitive intelligence within an open and unstructured environment. In situations with complicated moral dilemmas, state-of-the-art AI agents do not have sufficient cognitive ability to perceive the correct moral context and successfully resolve the dilemmas through commonsense reasoning. Recent efforts in this field have explored game-theoretical moral models or Bayesian-based utility functions. Researchers at Duke University [39] adopted the game-theoretical approach to model ethical dilemmas and align an AI's ethical values with human values. Researchers at MIT [40] developed a computational model to describe moral dilemmas as a utility function, and introduced a hierarchical Bayesian model to represent social structure and group norms. These early attempts may not be general enough to support common moral scenarios, but they suggest new research directions on combining the powers of both DNNs and interpretable Bayesian reasoning models in the field of AI ethics.

4.1.3. Safe reinforcement learning

Many researchers have adopted deep reinforcement learning to model moral constraints as reward functions, and have used the Markov decision process to implement sequential decisions. However, deep reinforcement learning is far from mature, and has a long way to go before it becomes available for applications other than gaming. One of the major problems with this method relates to the safety of the reinforcement learning process. A malicious agent can have many options to bypass the regulatory ethical constraints by tricking the reward mechanism. For example, it can use reward hacking to obtain more rewards than intended by exploiting loopholes in the process of determining the reward.

4.2. Integrating ethical principles in the AI engineering process

Ethical principles should be transformed into software specifications guiding the design and implementations of AI systems. From the perspective of software engineering, the development of AI models and the operation of AI systems are often organized in a clearly defined life-cycle shown in Fig. 2, which includes AI task definition, data collection and preparation, model design and training, model testing and verification, and model deployment and application. Software specifications of AI security, safety, and fairness should be implemented through the entire AI development and operations (DevOps) life-cycle.

At the beginning, AI tasks need to be defined and analyzed during the requirement analysis phase. Designers can adopt different kinds of ethical specifications and evaluation metrics toward the customized requirements in different application scenarios. During the data collection and preparation phase, engineers must ensure the validity of the training dataset by eliminating corrupted data samples and reducing the potential bias of the dataset. With a balanced and correct dataset, engineers can design appropriate model structures and perform model training according to the ethical specifications. After the model design and training phase, the preliminary model must be tested and verified in accordance with the moral specifications describing the constraints in term of fairness, robustness, transparency, and task performance. If the model cannot pass the model testing and verification phase, the engineers must redesign the model, recheck the data, and retrain the model. Otherwise, the model can be integrated with other software components and deployed in the intelligent system. During the running of the system, the runtime behaviors of the system must be constantly examined and must conform to the ethical principles. If any violations of the moral constraints occur, the engineers must decide to make further improvements on the AI models and launch a new DevOps life-cycle.

To streamline such an ethically aware AI DevOps life-cycle, many AI engineering tools need to be developed and integrated into a comprehensive and flexible environment for AI model designers and system developers. As discussed in the previous sections, these tools must implement core techniques such as federated learning, adversarial testing, formal verification, fairness evaluation, interpretation, provenance, and runtime sandboxing, in addition to safety monitoring. At present, tools such as AI Fairness 360 are still under development; thus, major AI DevOps platforms have not yet encapsulated these tools as the main functions required by AI ethical principles. More research and engineering endeavors are essential in order to promote an open AI DevOps environment with built-in ethical support, where researchers and practitioners can conveniently explore novel AI ethical techniques,

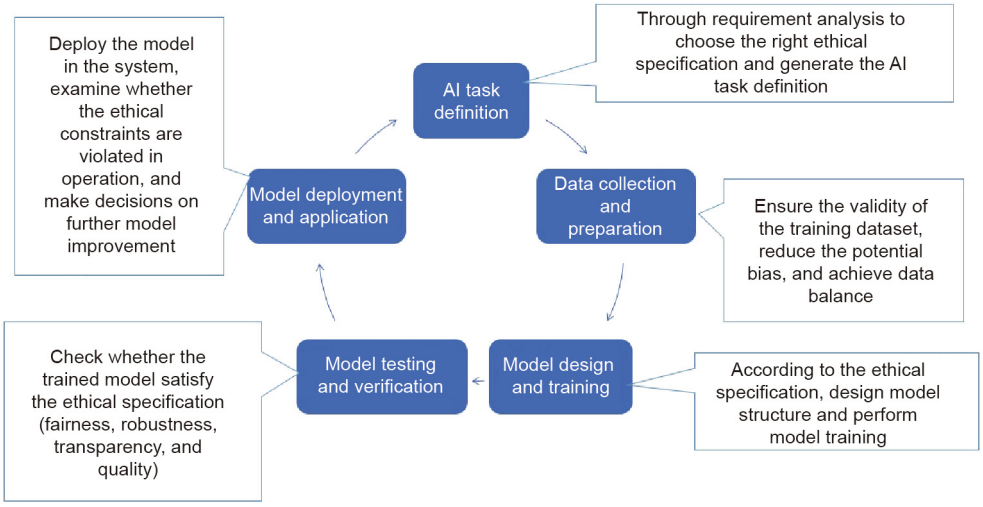

Fig. 2. AI development and operations (DevOps) life-cycle for governance.

systematically evaluate different assessment metrics, and conceive new solutions to different moral situations in various application domains.

With the progress of AI governance technologies, it can be expected that regulations and standards on ethical aspects of AI will be in place at the corporate/enterprise, group, national, and international levels, to enforce the compliance of AI systems and products. In fact, worldwide AI standardization research efforts have been underway for years. For example, the International Organization for Standardization (ISO)'s SC 24 launched a work group on AI trustworthiness in 2018, and the National Artificial Intelligence Standardization Steering Committee released a white paper analyzing AI ethical risks in 2019 [41]. Hopefully, combined efforts with AI engineering and standardization will further promote the awareness of the ethical problems of AI and will accelerate the integration of ethical values into AI systems and products within the AI industry and community.

4.3. Interdisciplinary research on AI governance

AI systems are complex and advanced social–technical systems, which often involve machine learning models, supporting software components, and social organizations. Researchers from multiple disciplines must conduct social-systems analysis of AI [42] in order to understand the impact of AI under different social, cultural, and political settings. Such a social-systems analysis demands an interdisciplinary research approach that leverages relevant studies of philosophy, law, and sociology, among other disciplines. Through multidisciplinary studies, AI designers and developers can work collaboratively with experts in law and sociology to conduct holistic modeling and analysis on the ethical aspects of intelligent systems by assessing the possible effects on all parties and by dealing with moral issues during every phase and state of AI DevOps.

There is no doubt that such an interdisciplinary and holistic approach to socio-technical engineering demands deep collaboration among AI developers and their partners with expertise in other relevant domains. Despite the increasing awareness of AI ethical principles among researchers in computer science, philosophy, law, and sociology in China, most of these scholars' research efforts are still being carried out on separate tracks and have not been fully synergized to address the grand challenges discussed above. Thus, we believe that it is critical to bring together experts from all relevant disciplines to work on the ethical problems of AI with clearly specified goals. First, based on or under the commonly accepted ethical principles of AI, we need to identify critical and typical ethical scenarios in applications such as autonomous driving, intelligent courts, and financial loan decisions, and call for novel research ideas and solutions from multidisciplinary teams. In these cases, complicated social contexts can be properly abstracted and described as AI ethical specifications. Second, an open and universal platform should be made available to foster interdisciplinary research on AI ethical principles. Such a platform will greatly enable researchers from different backgrounds to share their insights and contributions, and compare different frameworks and ethical criteria in building intelligent machines with ethical values and rules.

5. Conclusion

The rapid development and deployment of AI indicate an upcoming fundamental transformation of our society. This transformation can be a great opportunity to construct a human community with a shared future, and to promote the sustainable development of society and the natural environment. But without sufficient and effective governance and regulation, its implications might be unprecedented and negative. In order to ensure that these changes are beneficial before they are completely embedded into the infrastructure of our daily life, we need to build a solid and feasible AI governance framework to regulate the development of AI according to the ethics and values of humanity. In this way, we can make AI accountable and trustworthy, and foster the public's trust toward AI technology and systems.

This paper introduced the ongoing efforts to develop AI governance theories and technologies from the perspective of China. Many Chinese researchers have been motivated to address the ethical problems of current AI technologies. To overcome the security problem of data-driven AI, research teams from companies and universities in China have endeavored to develop federated learning technology. To ensure the safety and robustness of DNN models, researchers have proposed new algorithms in adversarial testing and formal verification. Furthermore, research teams are investigating effective frameworks in the areas of AI interpretation, provenance, and forensics. These efforts are mostly in their preliminary stages and need further strengthening in order to deliver mature solutions for widespread adoption and practice.

We suggest the following actions to push forward current initiatives on AI governance: Firstly, governments, foundations, and corporations should conduct cross-disciplinary, cross-sector, and

multinational collaborations to establish a consensus on AI ethical principles. Secondly, they must intensify the collaborative research and development of AI governance technologies in order to keep pace with the rapid progress of AI. Thirdly, open AI DevOps platforms with built-in ethics-relevant tools should be developed to support all the stakeholders of different AI systems in evaluating the functional and regulation compliance of AI systems. Fourthly, clearly defined AI moral scenarios with significant social impact should be identified so that experts from different disciplines can work collaboratively to address the ethical challenges of AI. Lastly, we must actively promote ethical education for every stakeholder in AI research and development, application, and management, so as to significantly enhance their awareness of ethics and promote general practices of responsible conduct with AI.

Compliance with ethics guidelines

Wenjun Wu, Tiejun Huang, and Ke Gong declare that they have no conflicts of interest or financial conflicts to disclose.

References

[1] National Governance Committee for the New Generation Artificial Intelligence. Governance principles for the new generation artificial intelligence—developing responsible artificial intelligence [Internet]. Beijing: China Daily; c1995–2019 [updated 2019 Jun 17; cited 2019 Dec 18]. Available from: https://www.chinadaily.com.cn/a/201906/17/WS5d07486ba3103dbf14328ab7.html?from=timeline&isappinstalled=0.

[2] Beijing AI principles [Internet]. Beijing: Beijing Academy of Artificial Intelligence; c2019 [updated 2019 May 28; cited 2019 Dec 18]. Available from: https://www.baai.ac.cn/blog/beijing-ai-principles.

[3] Zeng Y, Lu E, Huangfu C. Linking artificial intelligence principles. 2018. arXiv:1812.04814.

[4] Yang Q, Liu Y, Chen T, Tong Y. Federated machine learning: concept and applications. ACM Trans Intell Syst Technol 2019;10(2):12.

[5] Guide for architectural framework and application of federated machine learning [Internet]. New York: IEEE P3652.1 Federated Machine Learning Working Group; c2019 [cited 2019 Dec 18]. Available from: https://sagroups.ieee.org/3652-1/.

[6] Xiao C, Li B, Zhu J, He W, Liu M, Song D. Generating adversarial examples with adversarial networks. 2018. arXiv:1801.02610.

[7] Liu A, Liu X, Fan J, Ma Y, Zhang A, Xie H, et al. Perceptual-sensitive GAN for generating adversarial patches. In: Proceedings of the 33rd AAAI Conference on Artificial Intelligence; 2019 Jan 27–Feb 1; Honolulu, HI, USA; 2019.

[8] Yan Z, Guo Y, Zhang C. Deep defense: training DNNs with improved adversarial robustness. 2018. arXiv:1803.00404v3.

[9] Pang T, Du C, Dong Y, Zhu J. Towards robust detection of adversarial examples. 2018. arXiv:1706.00633v4.

[10] Ling X, Ji S, Zou J, Wang J, Wu C, Li B, et al. DEEPSEC: a uniform platform for security analysis of deep learning model. In: Proceedings of the 40th IEEE Symposium on Security and Privacy; 2019 May 20–22; San Francisco, CA, USA; 2019.

[11] Pulina L, Tacchella A. Challenging SMT solvers to verify neural networks. AI Commun 2012;25(2):117–35.

[12] Katz G, Barrett C, Dill DL, Julian K, Kochenderfer MJ. Reluplex: an efficient SMT solver for verifying deep neural networks. In: Proceedings of the International Conference on Computer Aided Verification; 2017 Jul 24–28; Heidelberg, Germany; 2017. p. 97–117.

[13] Gehr T, Mirman M, Drachsler-Cohen D, Tsankov P, Chaudhuri S, Vechev M. AI2: safety and robustness certification of neural networks with abstract interpretation. In: Proceedings of the 2018 IEEE Symposium on Security and Privacy; 2018 May 20–24; San Francisco, CA, USA; 2018.

[14] Singh G, Gehr T, Mirman M, Püschel M, Vechev M. Fast and effective robustness certification. In: Proceedings of the Advances in Neural Information Processing Systems 31; 2018 Dec 3–8; Montreal, QC, Canada; 2018. p. 10802–13.

[15] Lin W, Yang Z, Chen X, Zhao Q, Li X, Liu Z, et al. Robustness verification of classification deep neural networks via linear programming. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition; 2019 Jun 16–20; Long Beach, CA, USA; 2019. p. 11418–27.

[16] Yang P, Liu J, Li J, Chen L, Huang X. Analyzing deep neural networks with symbolic propagation: towards higher precision and faster verification. 2019. arXiv:1902.09866.

[17] 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; 2016 Aug 13–17; San Francisco, CA, USA; 2016. p. 1135–44.

[18] Zhang Q, Yang Y, Ma H, Wu YN. Interpreting CNNs via decision trees. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition; 2019 Jun 16–20; Long Beach, CA, USA; 2019. p. 6261–70.

[19] Liu S, Wang X, Liu M, Zhu J. Towards better analysis of machine learning models: a visual analytics perspective. Visual Inf 2017;1(1):48–56.

[20] Ma S, Aafer Y, Xu Z, Lee WC, Zhai J, Liu Y, et al. LAMP: data provenance for graph based machine learning algorithms through derivative computation. In: Proceedings of the 2017 11th Joint Meeting on Foundations of Software Engineering; 2017 Sept 4–8; Paderborn, Germany; 2017. p. 786–97.

[21] Xuan X, Peng B, Dong J, Wang W. On the generalization of GAN image forensics. 2019. arXiv:1902.11153.

[22] Gajane P, Pechenizkiy M. On formalizing fairness in prediction with machine learning. 2017. arXIv:1710.03184.

[23] Kusner MJ, Loftus J, Russell C, Silva R. Counterfactual fairness. 2017. arXiv:1703.06856.

[24] Bolukbasi T, Chang KW, Zou J, Saligrama V, Kalai A. Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. 2016. arXiv:1607.06520.

[25] Weng P. Fairness in reinforcement learning. 2019. arXiv:1907.10323.

[26] Bellamy RKE, Dey K, Hind M, Hoffman SC, Houde S, Kannan K, et al. AI fairness 360: an extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias. 2018. arXiv:1810.01943.

[27] High-Level Expert Group on AI. Ethics guidelines for trustworthy AI [Internet]. Brussels: European Commission; 2019 Apr 8 [cited 2019 Dec 18]. Available from: https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai.

[28] Trump DJ. Executive order on maintaining American leadership in artificial intelligence [Internet]. Washington, DC: The White House; 2019 Feb 11 [cited 2019 Dec 18]. Available from: https://www.whitehouse.gov/presidential-actions/executive-order-maintaining-american-leadership-artificial-intelligence/.

[29] Tencent AI Lab. Technological ethics at intelligent era—reshape trustworthiness in digital society [Internet]. Beijing: Tencent Research Institute; 2019 Jul 8 [cited 2019 Dec 18]. Available from: https://tisi.org/10890. Chinese.

[30] Meet the Partners [Internet]. San Francisco: Partnership on AI; c2016–18 [cited 2019 Dec 18]. Available from: https://www.partnershiponai.org/partners/.

[31] Li Q, Wen Z, Wu Z, Hu S, Wang N, He B. Federated learning systems: vision, hype and reality for data privacy and protection. 2019. arXiv:1907.09693.

[32] Szegedy C, Zaremba W, Sutskever I, Bruna J, Erhan D, Goodfellow I, et al. Intriguing properties of neural networks. 2013. arXiv:1312.6199.

[33] Kurakin A, Goodfellow I, Bengio S, Dong Y, Liao F, Liang M. Adversarial attacks and defences competition. In: Escalera S, Weimer M, editors. The NIPS'17 competition: building intelligent systems. Cham: Springer; 2018. p. 195–231.

[34] Cao Y, Xiao C, Yang D, Fang J, Yang R, Liu M, et al. Adversarial objects against LiDAR-based autonomous driving systems. 2019. arXiv:1907.05418.

[35] Arya V, Bellamy RK, Chen PY, Dhurandhar A, Hind M, Hoffman SC, et al. One explanation does not fit all: a toolkit and taxonomy of AI explainability techniques. 2019. arXiv:1909.03012.

[36] Yu H, Shen Z, Miao C, Leung C, Lesser VR, Yang Q. Building ethics into artificial intelligence. In: Proceedings of the 27th International Joint Conference on Artificial Intelligence; 2018 Jul 13–19; Stockholm, Sweden; 2018. p. 5527–33.

[37] Everitt T, Kumar R, Krakovna V, Legg S. Modeling AGI safety frameworks with causal influence diagrams. 2019. arXiv:1906.08663.

[38] Awad E, Dsouza S, Kim R, Schulz J, Henrich J, Shariff A, et al. The moral machine experiment. Nature 2018;563(7729):59–64.

[39] Conitzer V, Sinnott-Armstrong W, Borg JS, Deng Y, Kramer M. Moral decision making frameworks for artificial intelligence. In: Proceedings of the 31st AAAI Conference on Artificial Intelligence; 2017 Feb 4–10; San Francisco, CA, USA; 2017. p. 4831–5.

[40] Kim R, Kleiman-Weiner M, Abeliuk A, Awad E, Dsouza S, Tenenbaum JB, et al. A computational model of commonsense moral decision making. In: Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society; 2018 Feb 2–3; New Orleans, LA, USA; 2018. p. 197–203.

[41] National Artificial Intelligence Standardization Steering Committee. Report on artificial intelligence ethical risk analysis [Internet]. [cited 2019 Dec 18]. Available from: http://www.cesi.ac.cn/images/editor/20190425/20190425142632634001.pdf. Chinese.

[42] Crawford K, Calo R. There is a blind spot in AI research. Nature 2016;538(7625):311–3.