2021 IEEE 20th International Conference on Trust, Security and Privacy in Computing and Communications (TrustCom) | 978-1-6654-1658-0/21/$31.00 ©2021 IEEE | DOI: 10.1109/TRUSTCOM53373.2021.00207

Trusted Artificial Intelligence: Technique Requirements and Best Practices

Tao Zhang[12*], Yi Qin[1], Qiang Li[34]
[1] Guangxi Key Laboratory of Trusted Software & Guangxi Key Laboratory of Cryptography and Information Security, Guilin University of Electronic Technology, Guilin, China
[2] Guangxi Key Laboratory of Hybrid Computation and IC Design Analysis, Guangxi University for Nationalities, Nanning, China
[3]Henan Key Laboratory of Network Cryptography Technology, Zhengzhou, China
[4] PLA Information Engineering University, Zhengzhou, China
* corresponding email: oucseczt@126.com

***Abstract*—The development and adoption of AI have brought numerous problems and potential threats, such as information cocoons, algorithmic collusion, algorithmic bias, privacy, law and ethics problems or issues. To combat these issues, governments all over the world, international organizations and giant tech companies all take actions and reach an agreement that AI should be trusted AI. However, there is no uniform definition on trusted artificial intelligence. In the paper, we make a survey on related works on artificial intelligence principles and ethical guidelines. Then analyze the ethical foundations of trusted artificial intelligence and give a definition on trusted artificial intelligence. We propose specific requirements of trusted artificial intelligence from a technical perspective and give explanations on these technique requirements. Finally, we recommend some best practices to achieve trusted artificial intelligence and promote the responsible use of AI.**

Keywords-ethics, artificial intelligence, trusted artificial intelligence, technique requirements, best practices

I. INTRODUCTION

In the past years, Artificial Intelligence (AI) has experienced rapid development and been widely used in healthcare, transportation, society governance, media recommendations and other fields. However, adoption of AI technologies has also brought some challenges and debates, including privacy issues, data breaches, algorithmic discrimination, algorithmic biases, biased datasets and results [1], which are potential barriers for large scale applications of AI. For example, four cities in United States have banned face recognition technology in 2019, including New York and Somerville.

Lacking of trust on AI systems and applications is the essential reason for these problems. People do not trust output of AI systems and decisions made by AI systems, which has a negative impact on wide adoption of AI. To promote wide adoption of AI in different fields, the first challenge to overcome is to guarantee that AI applications are trusted, which has been a consensus of government, international organizations and tech giants around the world.

Government. Since 2019, governments all over the world have been taking actions to promote trusted AI. In January 2019, Singapore Personal Data Protection Commission releases *A Proposed Model AI Governance Framework* [2] to promote users confidence and trust on AI during AI adoption and deployment. The framework is the first AI governance framework in Asia, which proposes some ethical principles, such as human-centric, explainable, transparent and fair, and practical measures for AI deployment. In April 2019, the European Union High-Level Expert Group on AI officially published the *Ethics Guidelines for Trustworthy Artificial Intelligence* [3]. In which, the concept of Trustworthy AI is presented, which provide developers and others with a framework to achieve Trustworthy AI. There are seven requirements to be met during the entire life cycle of an AI system. According to the report, trustworthy AI should be lawful, ethical and robust. And the seven requirements to achieve trustworthy AI are human agency and oversight, technical robustness and safety, privacy and data governance, transparency, diversity, non-discrimination and fairness, societal and environmental wellbeing, and accountability. There are also some technical requirements proposed, such as resilience to attack and security, privacy and data protection, quality and integrity of data, access to data, traceability. In May 2019, OECD (Organization for Economic Co-operation and Development) approved its official recommendations on AI, *OECD Principles on Artificial Intelligence* [4], and 42 countries adopted these principles. The aim of these principles are to promote trusted AI, respecting human rights and democratic values are its core requirements.

International organizations. In 2017, the Asilomar Conference on Beneficial AI proposed 23 principles for AI [5], which are the first principles to promote the safe and beneficial development of AI. Of these 23 principles, 13 principles are on ethics and values, including safety, failure transparency, judicial transparency, responsibility. In Dec 2017, *Montreal Declaration for a Responsible Development of Artificial Intelligence* [6] is presented to promote AI development responsibly. IEEE is also a pioneer on trusted AI, IEEE published two versions of ethical guidelines for intelligent systems, as *IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems* [7] and *Ethically Aligned Design: A Vision for Prioritizing Human Wellbeing with Artificial Intelligence and Autonomous Systems (AI/AS)*

2324-9013/21/$31.00 ©2021 IEEE
DOI 10.1109/TrustCom53373.2021.00207

[8], which are high-level general principles for autonomous and intelligent systems design.

Tech giants. Tech giants like Microsoft and IBM are also practitioners on trusted AI. In 2018, Microsoft proposed 6 principles for the responsible use of AI, as fairness, reliability and safety, privacy and security, inclusiveness, transparency, and accountability [9]. Then IBM proposed the concept of trustable AI to promote trust on results of AI systems [10]. IBM thought that trustable AI systems and solutions should be robust, fair, explainable and linage.

II. Social And Ethics Issues Introduced By AI Applications

We have benefited from adoption of AI in our daily lives, such as recommendation, advertising, and decision making. However, AI is a double-edged sword, which has brought some social and ethical issues. In October 2019, Amazon had to stop an AI-powered recruiting tool which is also proved to be biased against female applicants. More recently, Facebook is sued for that its ad recommendation algorithms discriminate advertisers. In the paper, we summarized these social and ethical issues related to AI as information cocoons, algorithm collusion, algorithmic bias, security and privacy.

A. Information Cocoons

In media filed, AI has an advantage of filtering and recommending information according to user's tastes and features [11] as personalized recommendation, which is widely used in information recommendation like news, videos and other content. However, information recommended by AI based personalized recommendation algorithms may be restricted in a small wide range and cause information cocoons, which can be a serious threat to democracy of people and country.

B. Algorithmic Collusion

AI can also be used in price strategies, pricing algorithms based on AI can set the price of an item for sale dynamically based on competitors' prices and other information about customers. However, these AI enabled algorithms can also be used for collusion and violate rights and interests of consumers. Algorithmic collusion is common in digital economy, all Uber, Airbnb and Amazon have been proved to have such problems [12].

C. Algorithmic Bias

Algorithmic Bias means that the output of AI systems are not fair and have a discrimination over someone either economically irrational or socially unacceptable, COMOPS is just the case. COMPAS is a machine learning based application widely used across US, which can predict potential future criminals based on the key confidential algorithm and history data. In 2016, Jeremy Goldhaber-Fiebert [13] proved that COMOPS is biased against black defendants, as the system is more likely to label black defendants as higher risk of committing violent crime in the future. However, behaviors of those black defendants do not match with the results of COMOPS. Thus COMOPS are biased.

D. Security and privacy challenges

Like common systems and software, AI-enabled systems and software are also facing with different kinds of traditional attacks and emerging attacks. Besides, AI-enabled systems are vulnerable to specific attacks, such as adversarial learning and data poisoning. DNN and SVM are among most widely used AI algorithms. Su et al. [14] prove that Deep Neural Network (DNN) is not robust enough and only one pixel modification can make DNN produce wrong results. Biggio et.al [15] show that SVM is vulnerable to poisoning attacks, which can increase the error ratio of SVM.

III. AI ETHICS AND Trusted AI

A. AI ethics

The use of computer arises some ethical questions [16]. *Ethical* means that decisions made should be *right*. The concept of computer ethics is introduced to overcome these ethical issues, provide guidelines to computer professions, and promote ethical use of computer technology. AI is a branch of computer technology and has also brought some social and ethical issues. Thus AI ethics are proposed as part of computer ethics. AI ethics principles are user behavior codes of conduct, which should be abided during design and development of AI adoption. In the past years, EU, Singapore, OECD, IEEE, IBM and Google all proposed related AI ethics principles, which is shown in section 1. However, there are also differences between their definition on trust and responsible use of AI. Guidelines and principles by EU and OECD focus more on the ethics principles, such as human-centric values and inclusive growth. While Singapore and enterprises are more focused on technique requirements should be met.

B. Trusted AI

There are some related researches on trusted AI. However, there is no uniform definition on Trusted AI. In the paper, we define *Trusted AI* as follows:

Trusted AI systems should reflect underground ethical principles. The result or output of AI-enabled systems should be trusted. Key requirements of Trusted AI are safety and human intervention, robustness and security, data governance and privacy, transparency and interpretability, and accountability.

C. Requiements of Trusted AI

1) Safety and human intervention

The term safety is the primary concern of adoption of AI systems. AI system should be with good intentions and good for humans, avoiding causing potential harm for humans. As a computer program, AI systems may have potential negative impact on humans. Thus, human interventions is necessary to prevent potential harms.

2) Robustness and security

The term robustness and security means that the output of AI systems should be of fairness, security, accuracy and reliability. Lacking of trust is a key problem for AI systems and reliability is the key to improve trust of AI systems. Accuracy is an important part of reliability. Only if the

output of the AI systems is accurate, can users trust AI systems. What's more, AI enabled systems should be resilient to both traditional attacks and specific attacks, such as adversarial training and poisoning attack. Furthermore, the output of AI-enabled systems that should be accurate and fair, without bias, inequality and discrimination. In [17], fairness is defined as equal false-positive error rates.

3) Data governance and privacy.

The term data governance and privacy are about data, including privacy, quality and integrity of data, and access to data. Data is the key to AI applications. Many countries have taken urgent actions to combat data privacy issues, such as EU and California in US. With the introduction of GDPR, privacy has attracted all the world's attention. Models are trained with data and quality of training data has a great impact on the final trained models, so data quality for training data must be guaranteed.

4) Transparency and interpretability

The term transparency is a precondition to enable trusted AI [17]. Currently, most AI-enabled system are black box, can we open the black box of AI and should we open it? Knowing how an AI system works and achieves the final output is the key to trust, particularly for enterprise AI systems. The term interpretability of AI systems is critical to their utility and trustworthiness. Interpretability can help to promote the trust of AI and wide adoption of AI. Once the black box of AI is opened, we can learn how these AI enabled systems work, how the final results are achieved, and which factors have more impact on the final results. However, interpretability may raise concerns about intellectual property and commercial secrets. Some may think that the transparency and interpretability of AI models may expose commercial secrets and harness the intellectual property. However, the requirements of transparency and interpretability here is only a small extent of transparency and interpretability instead of making both models and data public. Besides, oversight on such AI enabled systems is necessary, especially when these systems are used in areas such as healthcare and criminal injustice. In [18], Shaikh et al. also prove that performance of AI enabled systems with an interpretable classifier outperforms one with a black box classifier model.

5) Accountability

The term accountability can also promote trust of AI enabled systems by ensuring all their components and events traceable and accountable. Take autonomous cars as an example, all operations by drivers and autonomous systems should be logged for potential accountability and penalties. Governments are also taking actions to make software more accountable. In Dec 2018, the New York City Council passed a bill to set up a task force to make recommendations on how to publicly share information about algorithms and investigate potential bias. In 2019, Emmanuel Macron, the president of France, also claimed that France will make all algorithms used by government open. Items from Europe's General Data Protection Regulation (GDPR) [19], which came into force at May 2019, are also expected to promote algorithmic accountability.

IV. Best practices for Trusted AI

Given definition of Trusted AI and requirements of Trusted AI, how to achieve that is still a big challenge. To help to promote trust of AI systems and achieve trusted AI, we recommend some best practices.

A. Data governance and data minimization.

Data plays an important role in current AI systems and the first best practice is data governance and data minimization. A large quantity of data is necessary for AI models training. More data for training, more accurate the final result of AI models is. As more data are collected and used, there is more chance to bring privacy policies and regulations violations. To protect privacy effectively, data minimization can be an option. Data minimization is to use the minimum amount of data to successfully complete their task. We also recommend developers and businesses prefer algorithms that need less training data.

B. Data and model transparency.

1) Transparency is the best policy. Most AI-enabled systems are running in black box such as deep neural network, which may cause biased result, COMPOS is just the case. When black box machine learning algorithms are applied into critical areas such as healthcare, criminal injustice, and autonomous driving, we should guarantee that AI-enabled systems should be of transparency and interpretability. Because transparency is the best policy to promote trust in these scenarios. Besides, meaningful explanations about how the results come from are necessary for those scenarios where people hold ultimate responsibilities for the decisions and outcomes. Supplier's declaration of conformity (SDoC) for AI services by IBM is such a best practice that can help to increase trust in enterprise AI systems. With SDoC, AI services and systems providers can describe the lineage of product along with safety and performance. In SDoC, the transparency, explanability, intellectual property and business information are well-balanced [20]. Tomsett et al. [21] propose a role-based interpretable model which can help to audit machine learning systems. Codella et al. [22] have demonstrated that meaningful explanations can be reliably taught to machine learning algorithms, and can improve accuracy of model in some cases. Shaikh et al. [18] proved that fairness policies can be achieved with opaque machine learning systems.

2) Open data as an option. As mentioned before, AI-enabled systems need large quantities of data, and data can be viewed as a kind of assets and wealth for business. These data may be owned by only a small fraction of businesses and not shared with others, which is a barrier to development of AI applications. What's more, these data themselves may be biased. To promote trust of AI-enabled systems, we recommend government and tech giant construct some high quality datasets and make them open for public use. Once

these data used to train AI models are open, it can help to avoid biased models and results to some extent.

C. Assessment

Assessments are to assess AI-enabled systems and guarantee that these systems are trusted. Assessments can be divided into self-assessments and independent third party assessments. Self-assessments and third party independent assessments are both necessary for promote trust of AI systems.

1) Self-assessments. In [23], Sattigeri et al. introduce a GAN based method, fairness GAN, which can generate datasets with fairness properties. In [24], Agarwal et al. propose an automated test case generation method that combines symbolic execution with the local interpretability for generation of effective test cases, which can detect individual discrimination of AI models. AI Fairness 360 [25] is an assessment tool to detect and mitigate algorithmic bias. Developers can use fairness GAN and these automated test generation methods to test their training datasets as self-assessments tools. Developers should make sure that their systems are as fair as possible, because some systems may not be assessed by independent third parties.

2) Independent third party assessments. And independent third party assessments is a supplement of self-assessment. Independent third party assessments of AI systems will be the next evolution of AI governance. And the creation and maintenance of the independent third party assessment should be an ongoing and dynamic process.

3) Build Trusted AI model. Self-assessments and independent third party assessments can help to develop highly robust models. While adversarial learning can be used to mitigate unwanted biases [26].

D. Access Requirements.

Access requirements and independent third party assessments are necessary in specific area, such as healthcare, criminal injustice, and autonomous driving. Take face recognition application as an example, in 2018, Amazon's facial recognition AI wrongly identifies 28 politicians as criminals. Along with the privacy concern, there are more than 4 cities in the United States have banned application of face recognition technology in specific areas. Thus there should be *Access Requirements Assessments* before adopting, such as face recognition applications.

V. CONCLUSION

The development of wide adoption of AI have brought some problems and potential threats, such as information cocoons, algorithmic collusion, algorithmic bias, privacy, law and ethics problems/issues. As algorithms are widely used to make decisions in our daily life, there is increasing awareness of the need to ensure the trust of these decisions. Governments all over the world, international organizations and tech giants are all taking actions to overcome these issues. Trusted AI is just what we need. However, there is no uniform definition for trusted AI. In the paper, we give a definition of trusted AI and describe the requirements of the key considerations for trusted AI. Finally, some best practices of trusted AI are given to promote the adoption of AI.

ACKNOWLEDGMENT

This work was supported by Guangxi Key Laboratory of Cryptography and Information Security (GCIS201806), Guangxi Key Laboratory of Trusted Software (No. kx202016), Guangxi Key Laboratory of Hybrid Computation and IC Design Analysis (GXIC20-03), Henan Key Laboratory of Network Cryptography Technology (No. LNCY2019-S07), Key Laboratory of Oceanographic Big Data Mining & Application of Zhejiang Province (OBDMA202001).

REFERENCES

[1] Carrier, Ryan. "Implementing guidelines for governance, oversight of AI, and automation." Communications of the ACM 62.5 (2019): 12-13.

[2] MODEL ARTIFICIAL INTELLIGENCE GOVERNANCE FRAMEWORK. https://www.pdpc.gov.sg/-/media/Files/PDPC/PDF-Files/Resource-for-Organisation/AI/SGModelAIGovFramework2.pdf

[3] Ethics guidelines for trustworthy AI. https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai

[4] The OECD Principles on AI. https://www.oecd.org/going-digital/ai/principles/

[5] ASILOMAR AI PRINCIPLES. https://futureoflife.org/ai-principles/

[6] The Montréal Declaration for a Responsible Development of AI. https://www.montrealdeclaration-responsibleai.com/the-declaration

[7] The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems. https://standards.ieee.org/industry-connections/ec/autonomous-systems.html

[8] Ethically Aligned Design A Vision For Prioritizing Wellbeing With Artificial Intelligence And Autonomous Systems, http://www.standards.ieee.org/develop/indconn/ec/autonomous_systems.html

[9] Microsoft AI principles. https://www.microsoft.com/en-us/ai/responsible-ai

[10] IBM. AI Ethics. https://www.ibm.com/artificial-intelligence/ethics

[11] Gossart, Cédric. "Can digital technologies threaten democracy by creating information cocoons?." Transforming politics and policy in the digital age. IGI Global, 2014. 145-154.

[12] Ezrachi, Ariel, and Maurice E. Stucke. "Algorithmic collusion: Problems and counter-measures." Submitted as background material at the Roundtable on Algorithms and Collusion at the OECD Competition Committee (2017). https://one.oecd.org/document/DAF/COMP/WD(2017)25/en/pdf

[13] Vasconcelos, Marisa, et al. “Modeling Epistemological Principles for Bias Mitigation in AI Systems: An Illustration in Hiring Decisions.” Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society, 2018, pp. 323–329.

[14] Su, Jiawei, Danilo Vasconcellos Vargas, and Kouichi Sakurai. 2019. “One Pixel Attack for Fooling Deep Neural Networks.” IEEE Transactions on Evolutionary Computation, 1–1.

[15] Biggio, Battista, Blaine Nelson, and Pavel Laskov. "Poisoning attacks against support vector machines." Proceedings of the 29th International Conference on International Conference on Machine Learning. Omnipress, 2012.

[16] Bynum, Terrell, "Computer and Information Ethics", The Stanford Encyclopedia of Philoso-phy (Summer 2018 Edition), Edward N.

Zalta, https://plato.stanford.edu/archives/sum2018/entries/ethics-computer/

[17] Turilli, M. & Floridi, L. Ethics Inf Technol (2009) 11: 105. https://doi.org/10.1007/s10676-009-9187-9

[18] Shaikh, Samiulla, Harit Vishwakarma, Sameep Mehta, Kush R. Varshney, Karthikeyan Natesan Ramamurthy, and Dennis Wei. 2017. "An End-To-End Machine Learning Pipeline That Ensures Fairness Policies." ArXiv Preprint ArXiv:1710.06876.

[19] GDPR. https://gdpr.eu/

[20] Calmon, Flávio du Pin, Dennis Wei, Bhanukiran Vinzamuri, Karthikeyan Natesan Ramamurthy, and Kush R. Varshney. 2017. "Optimized Pre-Processing for Discrimination Prevention." Neural Information Processing Systems, 3992–4001.

[21] Tomsett, Richard, David Braines, Daniel Harborne, Alun David Preece, and Supriyo Chakraborty. 2018. "Interpretable to Whom? A Role-Based Model for Analyzing Interpretable Machine Learning Systems." ArXiv Preprint ArXiv:1806.07552.

[22] Codella, Noel C. F., Michael Hind, Karthikeyan Natesan Ramamurthy, Murray Campbell, Amit Dhurandhar, Kush R. Varshney, Dennis Wei, and Aleksandra Mojsilovic. 2018. "Teaching Meaningful Explanations." ArXiv Preprint ArXiv:1805.11648.

[23] Sattigeri, Prasanna, Samuel C. Hoffman, Vijil Chenthamarakshan, and Kush R. Varshney. 2018. "Fairness GAN." ArXiv: Machine Learning.

[24] Agarwal, Aniya, Pranay Lohia, Seema Nagar, Kuntal Dey, and Diptikalyan Saha. 2018. "Automated Test Generation to Detect Individual Discrimination in AI Models." ArXiv Preprint ArXiv:1809.03260.

[25] Bellamy, R., Dey, K., Hind, M., Hoffman, S.C., Houde, S., Kannan, K., Lohia, P., Martino, J., Mehta, S., Mojsilovic, A., Nagar, S., Ramamurthy, K., Richards, J.T., Saha, D., Sattigeri, P., Singh, M., Varshney, K., & Zhang, Y. (2018). AI Fairness 360: An Extensible Toolkit for Detecting, Understanding, and Mitigating Unwanted Algorithmic Bias. ArXiv, abs/1810.01943.

[26] Zhang, Brian Hu, Blake Lemoine, and Margaret Mitchell. 2018. "Mitigating Unwanted Biases with Adversarial Learning." In Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society, 335–40.