Towards Ethic-Friendly AI Architecture

Yustus Eko Oktian, Elizabeth Nathania Witanto, Sang-Gon Lee*

Dongseo University

*Corresponding Author

요 약

The state-of-the-art AI systems pose many ethical issues ranging from massive data collection to bias in algorithms. In response, this paper proposes a more ethic-friendly AI architecture by combining Federated Learning and Blockchain. We first discuss the requirements for an ethical AI system then show how our solutions can achieve more ethical paradigms. By committing to our design, adopters can perform AI services more ethically.

I. Introduction

We long time believe that users only have constrained machine that is unable to train the machine-learning model. They also have a limited data size, which is not enough to produce a highly accurate training model. Therefore, many data from different users must be aggregated to a high-performance server, where the training takes place. In consequence, users lose control of their data once the data is transfered out from their devices. This data collection practice sometimes does not explicitly request user consent. Many companies use an "opt-out" mechanism instead of "opt-in", which puts users on surprise when they realize such a data collection setting exists. Even worse, there is no regulation for companies when they conduct AI practices. Until recently, the public became aware of the importance of user privacy with the introduction of the GDPR law [1].

Even though massive data collection can be compelling, it is challenging to adjust the trade-off between the benefits and user privacy. For example, China's social credit system [2] can shape the business and citizen's behavior towards better goals (in the view of the government). However, the citizen is at a disadvantage by losing freedom over this mass surveillance program. Moreover, AI is a black box system (in the current form), making it very tough to be debugged. This problem leads to many biases in AI algorithms. For instance, South Korea AI persona, Lee Luda [3], makes a controversy because she used offensive language targeting a minority community. Amazon AI recruitment tools also being shut down because it prefers men over women in selecting candidates [4]. As a result, researchers and AI practitioners must conduct AI services with ethics-in-mind, which always preserve human values.

This paper aims to seek solutions towards more ethic-friendly AI architecture by combining Federated Learning (FL) [5] and Blockchain [6]. FL preserves user privacy by training private user data on user local machines instead of sending them to the server. Meanwhile, the blockchain serves as a trusted platform to conduct the overall FL

process so that FL participants can collaborate in a secure, transparent, and fair manner. We also discuss the requirements for an ethical AI system and show that our solutions tackle the necessary components. By committing to our design, adopters can realize an ethic-friendly AI architecture.

II. Requirements for Ethcial AI

Floridi and Taddeo [7] divides ethics of AI into three spheres: ethics of data, ethics of algorithms, and ethics of practices.

Ethics of Data: The ethics of data focuses on the ethical problems related to data, including generation, curation, processing, dissemination, sharing, and usage [7]. Tranberg et al. [8] recommends five principles to enforce data ethics: **R1**) human being at the center, **R2**) individual data control, **R3**) transparency, **R4**) accountability, and **R5**) equality.

Ethics of Algorithms: The ethics of algorithms addresses issues posed by the increasing complexity and autonomy of the AI algorithms [7]. High-Level Expert Group on Artificial Intelligence, which is an independent expert group that was set up by the European Commission, mentioned that AI algorithm must follow these ethical principles [9]: **R6**) respect for human autonomy, **R7**) prevention of harm, **R8**) fairness, and **R9**) explicability.

Ethics of Practices: The ethics of practice focuses on the pressing questions about the responsibilities and liabilities of people and organizations in charge of data, strategies, and policies of AI system [7]. Google provides a recommendation practices for AI [10], which includes **R10**) use a human-centered design approach, **R11**) rigorous testing, and **R12**) continuous monitoring and updates.

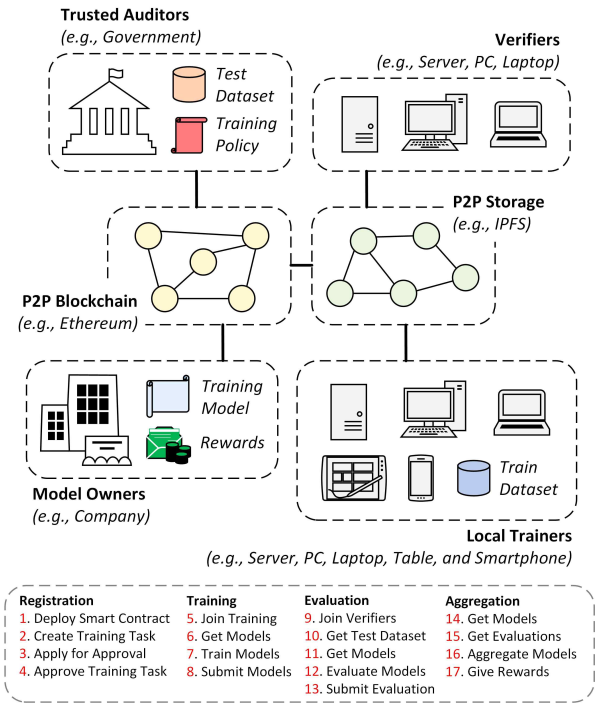

Fig 1. Our ethic-friendly AI system.

III. Proposed Architecture

Using the previously mentioned ethic requirements as our foundation, we propose an ethic-friendly AI architecture as depicted in Fig 1. The proposed system comprises six components: model owners (e.g., AI companies), local trainers (e.g., users), verifiers (e.g., users or government personnel), trusted auditors (e.g., the government), peer-to-peer (P2P) blockchain, and P2P storage. All participants are authenticated and endorse the use of a reputation system in our system. The AI workflow is described as follows.

Registration: The government makes the digital representation of the training policy in smart contracts (Step 1). AI companies, as model owners, create an initial global model and prepare rewards for trainers. They then

create a training task in the smart contract (Step 2). After that, the companies request approvals from the government (Step 3). Before approving a task, the government must make sure that the proposal provides enough incentives for trainers. They also create a standardized test dataset suitable for the proposal (Step 4). The model parameters and the test dataset will be distributed to trainers and verifiers through the P2P storage. Meanwhile, the hash of the model and dataset is stored in the blockchain.

Training: Users can join the training as trainers by registering themselves in the smart contract (Step 5). They can then get the model from P2P storage (Step 6) and begin training using their local data (Step 7). When the training is complete, users submit the trained model through P2P storage while the hash is logged in the blockchain (Step 8).

Evaluation: Users or government personnel can register themselves as verifiers in the smart contract (Step 9). At each global epoch, the verifiers must get the test dataset (Step 10) and the trained local models (Step 11) from P2P storage. They then verify the accuracy of the trained models using the test dataset (Step 12). Once the evaluation finishes, the evaluation result is submitted to the smart contract (Step 13).

Aggregation: When a particular global epoch finishes, the companies get all of the trained local models from the P2P storage (Step 14). They then retrieve all of the associated evaluation scores from the smart contract (Step 15). Using the evaluation scores as a guideline, the companies aggregate the models according to their contributions (Step 16). For example, they may skip models with low accuracy as they are most probably trained with poisoned data or low-quality data. During evaluations, verifiers use adversarial defense techniques to check if the model is trained with adversarial examples. Therefore, the companies must also skip models, which contains malicious flag from the verifiers. Once the aggregation is completed, the companies distribute the reward to all trainers and verifiers through the smart contract (Step 17).

IV. Ethic-Friendliness Analysis

Training distributedly using Federated Learning: Users train their data locally on their devices and only send the model parameters instead of the private data to the server. The server then combines the trained local models into a single global model using an aggregation algorithm (e.g., Federated Averaging [5]). Using this approach, the user data do not leave the devices, and users still have control over their data (i.e., solving **R2**).

Rigorous evaluation and auditing: To ensure the quality of the trained models, they must be evaluated. For this purpose, we employ the government and volunteers as our verifiers.

The government must first create a standardized training policy for AI companies in the form of federal or international law (e.g., GDPR [1]). With this law, we can hold malicious persons or organizations accountable (i.e., solving **R4**). We can also ensure that the AI models will always benefits humans (i.e., solving **R1**, **R6**, and **R10**) Moreover, the government must produce a generalized test dataset to be used during the evaluation stage. Assuming that this standardized test dataset has a high variance to cope with all possible classes, then this test should mitigate the AI bias that may

happen during training (i.e., solving **R5** and **R8**).

The group of verifiers evaluates the submitted local models from users to detect potential poisoning attacks on each epoch. Attackers can intentionally train the local model with bad or low-quality data to reduce the global model's overall accuracy. Moreover, the attackers can also train the model with adversarial examples to make the global model misclassify particular targets. Once detected, the attackers will be punished economically or by law (i.e, solving **R7**, **R11**, **R12**)

Logging training processes using the blockchain: In our architecture, all of the training processes are logged in the blockchain (e.g., Ethereum [6]). Because of the chain-of-hashes introduced in the blockchain, the stored data in the blockchain becomes hard-to-tamper. All nodes must also include their digital transactions when storing data to the blockchain. Hence, malicious entities can be detected easily. Finally, all data in the blockchain is open for all the blockchain nodes. Hence, solving **R3** and **R9**.

Sharing through distributed storage: Because storing in the blockchain is quite expensive to perform, we can use distributed storage system (e.g., IPFS [11]) to store massive data (e.g., model parameters). Meanwhile, the corresponding metadata (e.g., the hash of the model) can be stored efficiently in the blockchain. Note that we refrain from using a centralized database due to trust issues that may persist in such a system.

V. Conclusion

This paper proposed a more ethical AI architecture through a combination of federated learning and blockchain technologies. The federated learning yielded promising solutions towards AI ethics in terms of data collection and training transparency. Meanwhile, the blockchain enhanced the AI ethics with its secure, transparent, and fair collaborative auditing platform. However, our proposal still does not solve AI's fundamental issues regarding its "black box" properties. More research towards "explainable AI" is still required in the future so that we as humans and AI supervisors can make a better decision on how to use AI.

Acknowledgments

This work was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (Grant Number: 2018R1D1A1B07047601).

[참고문헌]

[1] M. Goddard, "The EU general data protection regulation (gdpr): European regulation that has a global impact," International Journal of Market Research, vol. 59, no. 6, pp. 703 - 705, 2017.

[2] K. Munro. (2018) China's social credit system could interfere in other nations' sovereignty. [Online]. Available: https://bit.ly/3qSmFz8 [Accessed: 27-Jan-2021].

[3] J. McCurry. (2021) South Korean AI chatbot pulled from facebook after hate speech towards minorities. [Online]. Available: https://bit.ly/2NwoZgM [Accessed: 27-Jan-2021].

[4] Reuters. (2018) Amazon ditched ai

recruiting tool that favored men for technical jobs. [Online]. Available: https://bit.ly/3sTehBc [Accessed: 27-Jan-2021].

[5] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, "Communication-efficient learning of deep networks from decentralized data," in Artificial Intelligence and Statistics. PMLR, 2017, pp. 1273 - 1282.

[6] G. Wood et al., "Ethereum: A secure decentralised generalised transaction ledger," Ethereum project yellow paper, vol. 151, no. 2014, pp. 1 - 32, 2014.

[7] L. Floridi and M. Taddeo, "What is data ethics?" Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, vol. 374, no. 2083, p. 20160360, dec 2016.

[8] P. Tranberg, G. Hasselbalch, B. K. Olsen, and C. S. Byrne. (2018) Dataethics - principles and guidelines for companies, authorities, and organisations. [Online]. Available: https://bit.ly/3canj6V [Accessed: 26-Jan-2021].

[9] AI HLEG. (2019) Ethics guidelines for trustworthy AI. [Online]. Available: https://bit.ly/3iTIRGs [Accessed: 26-Jan-2021].

[10] Google. (2020) Responsible AI practices. [Online]. Available: https://bit.ly/3opdd4Q [Accessed: 26-Jan-2021].

[11] J. Benet, "Ipfs-content addressed, versioned, p2p file system," arXiv preprint arXiv:1407.3561, 2014.