A review of Required components for Trustworthy AI and related techniques

1st Zhaksylyk Chalgimbayev
Nazarbayev University
SEDS, Computer Science
Nur-Sultan, Kazakhstan
zhaksylyk.chalgimbayev@nu.edu.kz

***Abstract*—Currently, Artificial Intelligence based systems are being integrated into a constantly increasing number of spheres of our lives, with healthcare, education, and government being no exception. Considering the weight of responsibility that comes with participating in the decision-making processes of the aforementioned spheres, the task of minimizing the chance of AI systems causing problems in both direct and indirect fashions, has gained critical importance. In order to assure trustworthiness, AI systems should possess the set of properties that was suggested by researchers. Plenty of research attempted to determine the ways by which AI systems will be able to meet the aforementioned set of requirements, thus becoming trustworthy. The goal of this paper is to summarize two of the proposed trustworthy AI requirements, namely, fairness and privacy. Furthermore, this survey will present recent researches' suggested techniques for achieving this couple of requirements.**

I. Introduction

Over the period of centuries, humankind has been using machines to facilitate and/or automate various tasks in their day-to-day lives. In the last few decades, machines started to be actively assimilated into spheres with complex tasks, the accomplishment of which demand human intelligence and discernment. The integration of AI into such spheres has not been perfectly smooth. This can be concluded from the cases in which AI systems brought issues to their employing healthcare, justice, or governmental organizations. In order to avoid the harsh consequences, that a poorly developed AI can bring, and to increase the level of trust in AI machines, scientists and different institutes including the International Organization for Standardization have attempted to establish a solid framework upon which AI systems could be built. In his research, Hagendorff [1] identified the requirements that are most commonly accepted among the frameworks that were offered by major institutions. This survey paper will aim to summarize the reasons behind the involvement of privacy and fairness requirements in a potential global framework for AI development. Considering that the introduction of possible requirements has turned many researchers' attention to determining the means by which AI could achieve those required properties, this paper will also provide a summary of techniques for eliminating bias and improving privacy, offered by recent research works in this field.

II. Main Body

A. Fairness

The first factor that significantly affects the level of trust in Artificial Intelligence is fairness. Due to the fact that AI's application domain is expanding into vital spheres such as healthcare and jurisprudence the minimization of bias and unfairness of AI systems is of paramount importance. To date, there are numerous records of AI systems that behaved unjustly, biased, or discriminatory, as a result of poor design or implementation, while being applied in various fields of human life. The case of Amazon's sexist hiring algorithm is one of the most popular examples of biased AI systems. The system in this case taught itself that women employees are less preferable to male ones [2]. The AI which was used in the United States' healthcare allocation coupled with its discriminatory behavior towards black people is another major example of biased Artificial Intelligence [2].

The first set of methods that are intended to prevent the unfairness in AI systems is associated with eliminating the so-called data bias. The term "data bias" implies the biased nature of data on which the AI is trained. According to Olteanu et al. [3], the data is considered to be biased if it fails to equally represent various segments of society or if it succeeds to depict a biased society. Moreover, Tufekci [4] claims that a wide difference between the societies in which the data was collected and in which it is applied can also result in data bias.

With regards to the proposed techniques to prevent data bias, the method of identifying the source of bias within the training data, through probing the overall bias level of the system after removing or adding chunks of data, has proven itself to be helpful after being applied in natural language processing [5]. Another method that is efficient in detecting bias within the existing AI systems, suggests classifying previously made decisions of the AI as either biassed or unbiased [6].

Another major cause of unfair and/or biased AI systems is the wrong implementation of the objective function. One of the reasons that lead to the failure of the objective function to accomplish tasks intended by the developer, is related to the fundamentally wrong distribution of priorities among the features that participate in decision making [7]. One of the techniques of tackling bias by modifying the objective function was introduced by Kamishima et al. [8] in 2012.

In essence, the idea of the method is the incorporation of a regulariser into the objective function. In addition to offering the possibility of adjusting the preference of fairness over accuracy and visa-versa, the proposed method succeeds in effectively reducing bias in AI systems.

B. Privacy

The second vital component that any trustworthy AI system should strive to guarantee is privacy. The amount of data on which AI is trained is one of the decisive elements that affect the overall accuracy of the system. Apart from earning people's trust, the ability to handle data in vast amounts, without exposing it to threats of misuse and abuse, is also necessary to prevent the severe consequences that may arise because of weak data management. History remembers cases in which systems for one reason or another ended up exposing enormous amounts of data, thereby not only ruining the trust of thousands of people but also resulting in potential crimes and hacker attacks. For instance, in 2017, the private data of 147 million people were exposed by Equifax, which is considered to be one of the two major credit bureaus in the United States [9]. Over the last years, many researchers throughout the world have attempted to address the possible causes of data exposure. The next subsection will present a brief summary of proposed solutions.

Some research offered a solution that requires certain modifications to be applied to a training dataset. The first such method suggests the removal of private pieces of information from the dataset [10]. The exclusion from a dataset of so-called direct identifiers or, in other words, pieces of data that can be used to uniquely identify individuals, is the central idea of the method recommended by Garfinkel [11] in 2015. Despite removing potentially useful data and sacrificing the accuracy that otherwise the system would have possessed, these techniques cope with the task of enforcing privacy.

III. Conclusion

In conclusion, the main objective of this paper was to review two of the most widely suggested requirements that AI systems should strive to meet. In particular, the significance of the assurance of fairness and privacy in boosting the users' trust as well as in mitigating potential serious outcomes that otherwise could emerge, was extensively discussed. First of all, we discussed why for AI systems, especially for those that participate in governmental and judicial decision making, fairness and the ability to make unbiased decisions are of high priority. After identifying that unfair or biased AI is mostly caused by a biased dataset or incorrect implementation of the objective function, we presented a brief summary of existing techniques to prevent both causes. In the second part of the main body, the reasons behind the proposal of privacy as a requirement for a trustworthy AI were examined. Finally, several methods to improve privacy were reviewed along with their benefits and disadvantages. There is a need for a legal and universal framework based on which the employability of an AI would have been decided. In this way, the trust level in AI systems would have been greatly improved.

References

[1] T. Hagendorff, "The ethics of ai ethics: An evaluation of guidelines," *Minds and Machines*, vol. 30, no. 1, pp. 99–120, 2020.

[2] "5 examples of biased artificial intelligence," Logically, July. 30, 2019 [Online]. [Online]. Available: https://www.logically.ai/articles/5-examples-of-biased-ai

[3] A. Olteanu, C. Castillo, F. Diaz, and E. Kıcıman, "Social data: Biases, methodological pitfalls, and ethical boundaries," *Frontiers in Big Data*, vol. 2, p. 13, 2019.

[4] Z. Tufekci, "Big questions for social media big data: Representativeness, validity and other methodological pitfalls," in *Eighth international AAAI conference on weblogs and social media*, 2014.

[5] M.-E. Brunet, C. Alkalay-Houlihan, A. Anderson, and R. Zemel, "Understanding the origins of bias in word embeddings," in *International conference on machine learning*. PMLR, 2019, pp. 803–811.

[6] B. T. Luong, S. Ruggieri, and F. Turini, "k-nn as an implementation of situation testing for discrimination discovery and prevention," in *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2011, pp. 502–510.

[7] T. Bolukbasi, K.-W. Chang, J. Y. Zou, V. Saligrama, and A. T. Kalai, "Man is to computer programmer as woman is to homemaker? debiasing word embeddings," *Advances in neural information processing systems*, vol. 29, 2016.

[8] T. Kamishima, S. Akaho, H. Asoh, and J. Sakuma, "Fairness-aware classifier with prejudice remover regularizer," in *Joint European conference on machine learning and knowledge discovery in databases*. Springer, 2012, pp. 35–50.

[9] "Equifax data breach settlement," Feb. 2022 [Online]. [Online]. Available: https://www.ftc.gov/enforcement/refunds/equifax-data-breach-settlement

[10] M. Khalil and M. Ebner, "De-identification in learning analytics," *Journal of Learning Analytics*, vol. 3, no. 1, pp. 129–138, 2016.

[11] S. Garfinkel *et al.*, *De-identification of Personal Information:*. US Department of Commerce, National Institute of Standards and Technology, 2015.