Literature Review Draft

Introduction

According to the fundamental principles described in the Ethics guidelines for trustworthy AI authored by the High-Level Expert Group on Artificial Intelligence (AI HLEG) on behalf of the European Commission, there are strong reasons that motivate scientific research towards AI fairness. Data are the core of AI, they enable machines to be autonomous and capable of making decisions. While AI must be subject to audits to ensure its reliability and fairness, data should be collected with respect to the low and ethical principles. Along with the development of a large number of useful applications based on intelligent systems, many problems relating to AI ethical use emerge. This leads to the investigation of new technologies that can make use of data, including personal and sensitive data, without these being stored by any external organisation. To achieve these goals, new technologies have been implemented to put users, the main producers of data, at the centre of the process. One practical application is the development of Federated Learning (FL). This technology allows the inversion of machine learning processes in such a way that data is no more transferred from individual users to a central server, but instead learning models and information gradients are exchanged among users under the coordination of a central server. This practice may certainly favour data privacy, but it may also introduce new problems regarding e.g. the reliability of the data or the representativeness (fairness) of the predictive models generated with respect to a certain population. In order to obtain reliable and fair results from predictive models, one should assume that the data come from an equally reliable and fair context. The conducted review led to identify some of the issues related with FL and AI fairness. FL predictive models, by not keeping track of the data they use to be trained, cannot provide a precise and accurate model of the data. Instead, by having access to a traditional dataset, it would be possible to classify the data e.g. according to clusters, and consequently make more accurate and balanced interpretations. This problem can be viewed as the introduction of indirect bias into FL predictive models, which could lead e.g. to unfair decision-making processes.

The approach described in Fig. 1 was used to (1) identify the current research context in the field of FL fairness, (2) summarise the main findings stated in current research, and (3) discuss upcoming advancements. In this manner, it was possible to organise the research and to consider relevant future research developments while following a structured pattern.

(1) Identify Context	(2) Summarise research	(3) Future Developments (conclusions)

Fig. 1 - Structure of Literature Review

Literature Review Draft

Literature Review

Research Context

As (Kairouz, 2021) reports, fairness is a major concern in the context of federated learning. To enhance fairness in a federated environment, distributionally robust optimization (DRO) (Deng et al., 2021, Taskesen et al., 2020), together with multi-center model aggregation (Xie et al., 2021) have been investigated. One of the main problems due to federated learning is considering data as heterogeneous and representative of the whole population, which can lead to bias when processing data belonging to minority groups. As reported in (Deng et al., 2021, Taskesen et al., 2020), population characteristics can have different distributions among subgroups. The information and predictions extracted from the data may be strongly based on the distributions of the majority groups, introducing unfairness. As suggested by (Xie et al., 2021), it is paramount to recognise possible subgroups within the population a priori, so that the influence of minorities can be balanced. To achieve these results without violating the privacy principles underlying FL, Zero Knowledge Proof (ZKP) models can be employed to record metadata regarding the proportions of population clusters (Chen et al., 2021). The collected metadata can make a difference in predictions and decision making processes as they can help to identify population groups and calibrate analyses more accurately.

The analysis carried out by (Deng et al., 2021) brings to light some bottlenecks of FL. Assuming a vast number of devices collaborating to build prediction models, it becomes necessary to aggregate results for communication efficiency purposes. This is carried out through the use of specific algorithms e.g. FedAvg (Brendan et al., 2017), which optimises the learning model for a certain number of local iterations using stochastic gradient descent (SGD). This approach similar to local SGD (Sebastian U. Stich, 2018) may reduce drastically the number of rounds required for the model to be trained but may also introduce other issues. FedAvg is appealing from a communication aspect, but it does not address the data heterogeneity issue in FL. Indeed, it has been demonstrated that the generalisation ability of FedAvg's central model is affected if local data distributions are highly heterogeneous (Haddadpour and Mahdavi, 2019). In (Deng et al., 2021) it is proposed a Distributionally Robust Optimization (DRO) model to handle non-i.i.d. data. DRO was used also in (Mohri et al. 2019) to mitigate the effects of data heterogeneity and proved to be effective.

(Taskesen et al., 2020) proposes a distributionally robust approach to machine learning that takes into account data fairness, promoting a *fairness through unawareness* approach. This paper addresses various scenarios that can lead to data disparity i.e. under-representation of a certain minority group within a population. (Taskesen et al., 2020) defines an unfairness quantification, i.e. a numerical measure of how representative is a tested hypothesis for a certain group of individuals using the *equalised opportunities criterion* (Hardt et al., 2016). As stated in (Friedler et al., 2019) increasing fairness inevitably leads to a decrease in accuracy.

(Chen et al., 2021) presents the main problems due to FL with emphasis on Security Issues, Machine Learning Bias and Federated Learning Security. Assuming a real world non-i.i.d. data pattern, they discuss the ZeKoC model, a Zero Knowledge Clustering approach to mitigating adversarial behaviours in FL. This approach proved to be effective in reducing the effects of data poisoning in a federated environment.

In (Xie et al., 2021), the multi-center analysis for FL has been discussed. For a large collection of data divided in K clusters, it is possible to define K centres for data analysis. Normally, centres are based on an *a priori* knowledge of data, which cannot be achieved in FL. Hence, as (Xie et al., 2021) states, the K centres were randomly selected with the purpose of optimising results.

Discussed Technologies

Technology	**Emerged Advantages**	**Emerged Disadvantages**
Federated Average: FedAvg	Enhance communication efficiency in a federated environment	May introduce bias if data come from a non independent and identically distributed (non-iid) distribution
Distributionally Robust Optimisation: DRO	Mitigate the effects of data heterogeneity, deals with non-iid data	
Zero Knowledge Proof: ZNP	Allows to verify data without reading or accessing them directly. Useful to record clusters metadata remaining unaware of the data.	
Zero Knowledge Clustering: ZeKoC	Mitigate the effects of adversarial behaviour in the FL environment. Reduce the effects of data poisoning.	
Unfairness Quantification	Measure the unfairness of an hypothesis. Numerical metrics for unfairness	The more an hypothesis is fair, the less is accurate
Multi-Center-Analysis	Useful to handle non-iid data	Heuristic method, requires attempts for fine tuning

Literature Review Draft

Conclusion

In this review, some of the key aspects of current research regarding AI fairness have emerged. Many technologies are investigated nowadays in order to obtain predictive models and decision making models compliant with European principles of AI trustworthiness. Studies show a need to develop new technologies to address tomorrow's data and AI challenges. The privacy and security of people's data is increasingly central to the scientific debate. For these reasons, scientific research is making an effort in federated learning, a paradigm that allows the creation of predictive models without accessing people’s data and hence compromising privacy. Amongst the various challenges that FL brings to light, problems regarding the fairness of data in terms of the representativeness of a certain subgroup within a large population have been investigated. Losing direct access to data makes it more difficult to interpret predictions about individuals belonging to minority groups, whose contribution would not be fairly considered. For this reason, practices have been studied and implemented to generate metadata about the features of a given dataset without knowing the value of the data e.g. Zero Knowledge Proof. Applying ZNP to data analysis in a federated environment, together with measuring data unfairness through unfairness quantification can be crucial in determining the trustworthiness of a given predictive model. Justified by a strong scientific ferment, we believe that the development of an application for Unfairness Quantification exploiting metadata on clusters reconstructed through ZNP in a FL environment is of relevant academic and industrial interest. We believe that delivering such an application as a plug-in for data catalogues can enhance the way we interpret big data in accordance with the legal and moral principles stated by the European Commission.

References

Bertsimas, D., Gupta, V., & Kallus, N. (2018). Data-driven robust optimization. *Mathematical Programming*, *167*(2), 235-292. https://doi.org/10.1007/s10107-017-1125-8. 10.1007/s10107-017-1125-8

Chen, Z., Tian, P., Liao, W., & Yu, W. (2021). Zero Knowledge Clustering Based Adversarial Mitigation in Heterogeneous Federated Learning. *IEEE Transactions on Network Science and Engineering*, *8*, 1070-1083. 9119145. 10.1109/TNSE.2020.3002796

Literature Review Draft

Deng, Y., Kamani, M. M., & Mahdavi, M. (2021). *Distributionally Robust Federated Averaging*. https://arxiv.org/abs/2102.12660

ISO Technical Committee. (2021). *Information technology — Artificial intelligence (AI) — Bias in AI systems and AI aided decision making*. https://www.iso.org/standard/77607.html

Kairouz, P. (2021). *Advances and Open Problems in Federated Learning*.

Li, T., Sahu, A. K., Talwalkar, A., & Smith, V. (2020). Federated Learning: Challenges, Methods, and Future Directions. *IEEE Signal Processing Magazine*, *37*, 50–60. https://arxiv.org/abs/1908.07873. 10.1109/msp.2020.2975749

Taskesen, B., Nguyen, V. A., Kuhn, D., & Blanchet, J. (2020). *A Distributionally Robust Approach to Fair Classification*. https://arxiv.org/abs/2007.09530

Wilkinson, M. D., Sansone, S.-A., Schultes, E., Doorn, P., Bonino da Silva Santos, L. O., & Dumontier, M. (2018). *A design framework and exemplar metrics for FAIRness*. https://www.nature.com/articles/sdata2018118

Xie, M., Long, G., Shen, T., Zhou, T., Wang, X., Jiang, J., & Zhang, C. (2021). *Multi-Center Federated Learning*. https://arxiv.org/abs/2005.01026

Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. (2017) *Communication-efficient learning of deep networks from decentralized data.* Artificial Intelligence and Statistics, pages 1273–1282.

Sebastian U Stich. *Local sgd converges fast and communicates little*. (2018) arXiv preprint arXiv:1805.09767

Farzin Haddadpour and Mehrdad Mahdavi. *On the convergence of local descent methods in federated learning.* (2019) arXiv preprint arXiv:1910.14425

Literature Review Draft

Mehryar Mohri, Gary Sivek, and Ananda Theertha Suresh. Agnostic federated learning. (2019)

arXiv preprint arXiv:1902.00146

M. Hardt, E. Price, E. Price, and N. Srebro, *Equality of opportunity in supervised learning*,

(2016) in Advances in Neural Information Processing Systems 29

S. A. Friedler, C. Scheidegger, S. Venkatasubramanian, S. Choudhary, E. P. Hamilton, and D.

Roth, *A comparative study of fairness-enhancing interventions in machine learning*,

(2019) Proceedings of the Conference on Fairness, Accountability, and

Transparency