The Mainstays of Trustworthy Machine Learning

Chhavi Yadav

UC San Diego
cyadav@ucsd.edu

Introduction & Current Work

While machine learning (ML) models of today have the potential to be useful in many societal applications, they also harbor the potential for great harm, be it perpetuating biases or compromising privacy. To prevent these harms, many (evolving) regulatory guardrails have been put in place; for instance European Union's GDPR and Biden's Executive Order which demand explainability, privacy, fairness and so on from models deployed in societal applications. Yet, most technical solutions in the Trustworthy ML literature which claim to meet these regulatory requirements are brittle and often fail at the task in hand. To this end, my research aims to make the field of Trustworthy ML reliable using mainstay concepts of Measurement, Mitigation and Maintenance. With these concepts, I develop end-to-end solutions for trustworthy ML by (1) exploring the limitations of existing approaches and (2) providing principled novel solutions exploiting interconnections with cryptography. Next I explain my efforts in these directions.

"Measurement : If you can't measure it, you can't improve it." The first step to knowing the state of Trustworthy ML is to measure whether the existing technical solutions achieve what they claim and under what circumstances. In my latest work (Yadav*, Wu*, and Chaudhuri 2024), I studied one such popular tool – Influence Functions (Koh and Liang 2017), which find the most influential training samples for a prediction. While earlier influence functions were majorly used for debugging by the model developer, recently they have been proposed as a technique for data valuation (Jia et al. 2019) (assigning monetary value to training data with influence scores), for improving the fairness of ML models (Li and Liu 2022) and data filtering/subsampling (Wu, Hashemi, and Srinivasa 2022) to name a few. Motivated by these use-cases, in this study we identify incentives for adversarial manipulation of influence scores, where an adversary would want to artificially increase the value of fixed data by increasing their influence scores. We show that it is indeed possible and computationally feasible to *systematically* manipulate influence scores, thereby revealing their susceptibility to adversaries. This work serves as a cautionary tale and calls for careful consideration when utilizing influence-based attributions.

In another work (Baldini et al. 2023) on fairness auditing of Natural Language Inference models, we demonstrate that the current auditing metrics are flawed – they hide unfairness by aggregating scores across different groups (for eg. black and white race) and also wrongly attribute errors from lack-of-model-robustness to being unfair.

"Mitigation : The old order changeth, yielding place to new." Many a times models have to be kept confidential due to legal and IP reasons. In such cases, a popular method for verifying properties or building trust in confidential models is through Third-Party Auditing (Yadav, Moshkovitz, and Chaudhuri 2022; Yan and Zhang 2022). Here an external auditor *estimates* the value of model property (such as fairness) with API queries. However this approach suffers from multiple drawbacks including (1) leaking the confidential model in the process as the auditor essentially collects a dataset and can distill the hidden model, (2) swapping the model post-audit or using different models for different input points and (3) sensitivity to the choice of reference auditing dataset (Casper et al. 2024; Fukuchi, Hara, and Maehara 2019; Shamsabadi et al. 2023). Motivated by these problems, in my work (Yadav et al. 2024), I propose an alternative framework for verification using Zero-knowledge Proofs (ZKPs), a cryptographic primitive. Under this framework, the model developer explicitly specifies the value of the property to be verified (rather than being estimated as in auditing) and also provides a cryptographic proof for the correct computation of the value. On the other end, the customer checks this cryptographic proof without looking at the model weights; thereby verifying the value of the property while maintaining model confidentiality[1]. While prior work majorly focused on verifying inferences, my work takes a leap and demonstrates the feasibility of ZKPs for verifying properties of the model, opening up avenues for a lot of future directions for ZKPs in ML. This work was awarded the **TensorOpera-FedML Best Paper Award** at the 'Privacy Regulation and Protection in Machine Learning' Workshop at ICLR 2024.

[1]Model Swapping is prevented using cryptographic commitments, a part of the framework.

Future Directions

In the near future, I aim to establish the utility of cryptography for ML problems and make it a first-class citizen in mainstream ML. Additionally, I will also focus on the third mainstay of trustworthy ML : Maintenance. My efforts in these directions are as follows.

Cryptography saves the boat. Explanations (Gilpin et al. 2018) are intended as a way to increase trust in ML models by means of making opaque ML models transparent. However many use-cases where explanations are needed, such as demonstrating fairness, and suggesting recourse, are adversarial in nature and involve parties with misaligned interests (Bordt et al. 2022). This leads to manipulation of explanations by some parties, creating a false sense of security and trust. These problems are exacerbated when models are kept confidential due to IP and legal reasons. We observe that cryptography, specifically Zero-knowledge proofs and cryptographic commitments can elegantly prevent adversarial manipulations in explanations. In my future research, I will demonstrate the efficacy of ZKPs in preventing such manipulation, operationalizing explanations as trustworthy ML tools.

"Maintenance : Kintsugi, the Japanese art of repairing broken pottery." In real-world scenarios numerous considerations can arise after model deployment/training; an important one being addressing requests to remove some samples from the training set *post-training*. These requests can come from users (eg. GDPR's Right to be Forgotten) or when some problematic samples are found to be a part of the training set post-training (eg. problematic images in LAION (David 2023)). One solution to accommodate such requests is to train a new model from scratch, but this approach is not feasible for large models. An alternate approach is to *patch* the original model by removing the effect of these samples on the model; commonly known as Unlearning (Cao and Yang 2015) in the literature. While there exist plethora of unlearning algorithms, there is a lack of robust, systematic and principled approaches to evaluate their performance in the era of LLMs. In my ongoing research, I am diving deep into this question and working on a novel unlearning evaluation framework based on knowledge graphs, for unlearning facts in LLMs.

In conclusion, reliable Trustworthy ML solutions are of utmost importance for responsible integration of models into society and my research pushes the needle in this direction.

References

Baldini, I.; Yadav, C.; Das, P.; and Varshney, K. R. 2023. Keeping Up with the Language Models: Robustness-Bias Interplay in NLI Data and Models. *arXiv preprint arXiv:2305.12620*.

Bordt, S.; Finck, M.; Raidl, E.; and von Luxburg, U. 2022. Post-hoc explanations fail to achieve their purpose in adversarial contexts. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, 891–905.

Cao, Y.; and Yang, J. 2015. Towards making systems forget with machine unlearning. In *2015 IEEE symposium on security and privacy*, 463–480. IEEE.

Casper, S.; Ezell, C.; Siegmann, C.; Kolt, N.; Curtis, T. L.; Bucknall, B.; Haupt, A.; Wei, K.; Scheurer, J.; Hobbhahn, M.; et al. 2024. Black-box access is insufficient for rigorous AI audits. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*, 2254–2272.

David, E. 2023. AI image training dataset found to include child sexual abuse imagery. https://www.theverge.com/2023/12/20/24009418/generative-ai-image-laion-csam-google-stability-stanford.

Fukuchi, K.; Hara, S.; and Maehara, T. 2019. Faking Fairness via Stealthily Biased Sampling. arXiv:1901.08291.

Gilpin, L. H.; Bau, D.; Yuan, B. Z.; Bajwa, A.; Specter, M.; and Kagal, L. 2018. Explaining explanations: An overview of interpretability of machine learning. In *2018 IEEE 5th International Conference on data science and advanced analytics (DSAA)*, 80–89. IEEE.

Jia, R.; Dao, D.; Wang, B.; Hubis, F. A.; Hynes, N.; Gürel, N. M.; Li, B.; Zhang, C.; Song, D.; and Spanos, C. J. 2019. Towards efficient data valuation based on the shapley value. In *The 22nd International Conference on Artificial Intelligence and Statistics*, 1167–1176. PMLR.

Koh, P. W.; and Liang, P. 2017. Understanding black-box predictions via influence functions. In *International conference on machine learning*, 1885–1894. PMLR.

Li, P.; and Liu, H. 2022. Achieving fairness at no utility cost via data reweighing with influence. In *International conference on machine learning*, 12917–12930. PMLR.

Shamsabadi, A. S.; Wyllie, S. C.; Franzese, N.; Dullerud, N.; Gambs, S.; Papernot, N.; Wang, X.; and Weller, A. 2023. CONFIDENTIAL PROOF OF FAIR TRAINING OF TREES. *ICLR*.

Wu, G.; Hashemi, M.; and Srinivasa, C. 2022. Puma: Performance unchanged model augmentation for training data removal. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, 8675–8682.

Yadav, C.; Chowdhury, A. R.; Boneh, D.; and Chaudhuri, K. 2024. FairProof: Confidential and Certifiable Fairness for Neural Networks. *arXiv preprint arXiv:2402.12572*.

Yadav, C.; Moshkovitz, M.; and Chaudhuri, K. 2022. A Learning-Theoretic Framework for Certified Auditing with Explanations. arXiv:2206.04740.

Yadav*, C.; Wu*, R.; and Chaudhuri, K. 2024. Influence-based Attributions can be Manipulated. * equal contribution. https://drive.google.com/file/d/1xgnd-39GPxpdapwl3_q7UGOSZi5Y9fI3/view.

Yan, T.; and Zhang, C. 2022. Active fairness auditing. In *International Conference on Machine Learning*, 24929–24962. PMLR.