HHAI2022: Augmenting Human Intellect
S. Schlobach et al. (Eds.)

doi:10.3233/FAIA220206

Privacy Risk of Global Explainers

FRANCESCA NARETTO [a,1], ANNA MONREALE [b] and FOSCA GIANNOTTI [a]
[a] *Scuola Normale Superiore, Pisa, Italy*
[b] *University of Pisa, Pisa, Italy*

Introduction The increasing availability of big data describing at extreme details and resolution different aspects and phenomena of our private life enabled the shift toward a *knowledge society* where decisions can be taken – by individuals or by business and policy makers – on the basis of the knowledge distilled from the ubiquitous digital traces, generated by using digital tools in everyday life. These big data describing human activity are increasingly being sensed, stored and analyzed at individual, group and society levels. For instance, wireless networks and mobile devices record the traces of our travels. On the one hand, these digital traces contributed to the diffusion of a wide range of powerful Artificial Intelligence (AI) systems employed on a broad array of often critical domains, such as AI for supporting medical and autonomous vehicles for self-driving. The drawback of the powerful AI systems is that they are often based on neural networks or complex ensembles characterized by an inherent opaqueness; indeed, they are typically referred to as "black-box" models [1], due to the hidden internal structure and complex decision process which is not human understandable. The lack of interpretability may limit the trust in the AI systems and limit their wide adoption, especially in high-stakes decision making. In these cases it becomes crucial providing human interpretable explanations as a building brick of a *trustworthy* interaction between the human expert and the AI system. As pointed out in [1], there exists two families of explainers in in the eXplainable Artificial Intelligence (XAI) literature: *local* explainers, which explain the reason for a specific instance classification [2–4] and *global* explainers, which explain the logic of the machine learning (ML) model as a whole [5–7].

Another possible drawback of AI systems based on ML models is their potential vulnerability against different attacks, such as Model Inversion attack [8] and Membership Inference Attack (MIA) [9], which can potentially infer the data used for training the model by simply querying the model. Thus, privacy mechanisms such as differential privacy [10] are typically applied to counter the potential privacy exposure. In this paper we propose to study a methodology that enables the evaluation of the privacy risk exposure of global explainers based on an interpretable classifier that imitates the global reasoning of a black-box classifier. The idea is to verify if the layer of interpretability added by the interpretable model can jeopardize the privacy protection of individuals represented in the data used for the training of the black-box classifier. Intuitively, explainers are learned functions that are derived by exploiting the predictive knowledge of a black-box model learned on a private dataset. Thus, it could leak information about this private dataset. Despite this potential risk, there have been only few works that addressed privacy issues in XAI [11, 12].

[1] Corresponding Author: Francesca Naretto, Scuola Normale Superiore, Pisa, Italy, francesca.naretto@sns.it

Privacy risk exposure of global explainers We want to tackle the problem of evaluating the privacy risk exposure of global explainers based on an interpretable surrogate classifier that imitates the global reasoning of a black-box classifier. The idea is to verify if the layer of interpretability added by the interpretable model can jeopardize the privacy protection of the training data used for learning the black-box classifier. In order to address this problem, we exploit a well-known attack model called MIA [9]. We now describe our methodology that enables the evaluation of the *privacy exposure* rising from the layer of transparency introduced by a global explainer c that is able to mimes a black-box b. To this aim, we design the following methodology:

- Given the black-box b trained on D_b^{train}, we train a MIA model $A_b(\cdot)$ able to discern if a record x was part of the training data D_b^{train}. Since the training of the attack exploits the black-box b to label the training data of the attack D_a^{train}, we have that it is tailored to attack b.
- Given the interpretable surrogate global classifier c able to mime b, we learn a MIA model $A_c(\cdot)$ tailored to attack c by using D_a^{train} labeled by c instead of b.
- Finally, we compute the privacy risk change Δ_R due to the introduction of the global explainer c. Given a MIA learned model, we measure the privacy exposure as the percentage of records that an attack model is able to recognize as part of the training data D_b^{train}. This is given by the recall of the attack model with respect to the class IN, i.e. R_{IN}. As a consequence, we can compute the change of privacy risk as the difference of the recall of the MIA model $A_c(\cdot)$ on the global explainer c and MIA model $A_b(\cdot)$, i.e., $\Delta_R = R_{\text{IN}}^{A_c} - R_{\text{IN}}^{A_b}$. In order to evaluate this measure is necessary to test both the attack models $A_b(\cdot)$ and $A_c(\cdot)$ on a dataset D^{test} including records that were not used for the training of b as well as the whole set of records in the training data D_b^{train}.

Preliminary results and future works In this work we conducted a preliminary study of privacy risks of global explainers. In particular, we proposed a methodology that enables the assessment of the privacy exposure due to the use of global explainers based on interpretable models able to imitate a black-box model. Technically, we analyzed 2 tabular datasets: for each of them, we trained a black-box model (a Random Forest) and 2 global explainers, i.e. TREPAN and DT. Then, we attack all the trained models, exploiting the MIA, to analyze the privacy risk. The preliminary experimental results suggest that global explainers based on decision trees introduce a higher risk of privacy, increasing the percentage of records identified as members of the training dataset used to train the original black-box classifiers. These results suggest that in order to provide *Trustworthy AI* becomes fundamental starting to work on the research challenge of considering the relationship between different ethical values to identify possible values like transparency and privacy, that may be in contrast, and studying solutions that enable the simultaneous satisfaction of more than one value.

Acknowledgments. This work is supported by the EU H2020 project SoBigData++ (Grant Id 871042), XAI (Grant Id 834756) and HumanE-AI-Net (Grant Id 952026).

References

[1] Guidotti R, et al. A Survey of Methods for Explaining Black Box Models. ACM Computing Surveys. 2019;51.

[2] Ribeiro MT, et al. ”Why Should I Trust You?”: Explaining the Predictions of Any Classifier. In: ACM SIGKDD; 2016. p. 1135-44.

[3] Lundberg SM, Lee SI. A Unified Approach to Interpreting Model Predictions. In: NIPS; 2017. p. 4765-74.

[4] Guidotti R, et al. Factual and Counterfactual Explanations for Black Box Decision Making. IEEE Intell Syst. 2019;34(6):14-23.

[5] Craven MW, Shavlik JW. Using sampling and queries to extract rules from trained neural networks. In: JMLR. Elsevier; 1994. p. 37-45.

[6] Craven M, Shavlik JW. Extracting tree-structured representations of trained networks. In: NIPS; 1996. p. 24-30.

[7] Deng H. Interpreting tree ensembles with inTrees. Int Journal Data Science and Analytics. 2019;7(4):277-87.

[8] Fredrikson M, Jha S, Ristenpart T. Model Inversion Attacks That Exploit Confidence Information and Basic Countermeasures. In: Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security. CCS ’15; 2015. .

[9] Shokri R, Stronati M, Song C, Shmatikov V. Membership Inference Attacks Against Machine Learning Models. In: 2017 IEEE Symposium on Security and Privacy (SP); 2017. .

[10] Dwork C. Differential Privacy. In: Proceedings of the 33rd International Conference on Automata, Languages and Programming - Vol. Part II. ICALP’06; 2006. .

[11] Shokri R, Strobel M, Zick Y. On the Privacy Risks of Model Explanations. In: Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society; 2021. .

[12] Blanco-Justicia A, Domingo-Ferrer J, Martínez S, Sánchez D. Machine learning explainability via microaggregation and shallow decision trees. Knowledge-Based Systems. 2020.