Managing Risks for AI

Overview: Over the years, many privacy and security attacks, such as model inversion attacks and membership inference attacks, have been targeting Artificial Intelligence (AI) models [1]. In addition, many AI-caused bias/fairness issues have been identified. These practical challenges suggest an urgent need of a framework that can assess a spectrum of AI risks throughout the AI development pipeline.

We propose an AI risk-assessment framework that concerns three broad issues: a.) feasible threat modeling; b.) data pre-processing (e.g., removing the sensitive information from data before it is used by the AI model) and model post-processing (e.g, providing a wrapper for a given classifier to reduce model inversion attacks); and c.) AI model parameter selection (e.g. ϵ for differential private AI models, or robustness parameters for adversarial training) to get good utility and adequate protection against realistic risks.

Threat modeling-based risk assessment is in common use in several security domains. For example, to measure door lock security, the American National Standards Institute defines a series of tests that simulate real life attacks against door locks. These attacks range from the application of force (e.g., kicking down the door) to a battery of lock picking tests. The lock is assigned a grade based on its success against these types of likely attacks.

The above threat modeling-based approach implies that we can first pre-process the data using a technique Q to get a sanitized copy $Q(D)$ where certain biases and/or privacy sensitive information could be removed. For example, we can replace the sensitive information such as a credit card number with a realistic random card number before an AI model is built so that a text generation algorithm will not accidentally memorize them. Once the sanitized data is created, a private, robust and transparent learning algorithm L with appropriate parameters p is chosen to learn a model M. Later on this model is post-processed using a post-processing technique P to make sure that the concerned risks are addressed. For example, for a deep learning model M, a few layers could be added to M using publicly available data (hence no privacy risk) to reduce the effectiveness of specific attacks and improve fairness. This approach leads to an optimization problem (See Equation 1) where we find an optimal combination of p, P, Q such that we maximize the model utility U (e.g., accuracy) while making sure that risk (e.g., sensitive attribute prediction accuracy) due to a threat i, denoted as RT_i, is less than the desired risk limit γ_i.

$$\begin{aligned}
\max_{p,P,Q} \quad & U(P(L_p(Q(D))) \\
\text{s.t.} \quad & RT_1(P(L_p(Q(D))) \leq \gamma_1 \\
& RT_2(P(L_p(Q(D))) \leq \gamma_2 \\
& \dots \\
& RT_n(P(L_p(Q(D))) \leq \gamma_n
\end{aligned} \tag{1}$$

Clearly, appropriate pre-/post-processing techniques Q and P for different AI tasks may need to be considered during the risk modeling.

For pre-processing techniques, sanitization and random data generation approaches can be considered as a part of the risk framework. We believe that many of the issues, as a result of a model memorizing sensitive information from the data set, could be reduced by replacing sensitive data with its randomized counterpart. For example, a name and surname pair could be replaced with a realistic random name pair to hide personally identifiable information. In addition, we believe that some of the membership inference attacks and fairness issues are due to the lack of diverse data points in the given training set. Therefore, augmenting existing data sets with synthetically generated data (e.g., use a GAN to generate more samples of the underrepresented class) could be used to prevent a model to bias against certain groups.

With post-processing techniques, for classification tasks, the model output of class probabilities can be modified to reduce different risks. For example, an overly confident class prediction may help an attacker to better infer a sensitive attribute. For generative machine learning models, different post-processing models could be used to automatically sanitize potentially sensitive output. For example, we can learn a model P that can detect sensitive data such as a social security number (SSN) revealed by a machine learning model M, and replace it with a random realistic looking SSN.

Example: In our previous work [2], we showed that a differentially private explainable AI model (i.e., a rule set that explains a given ML model) could be post-processed (e.g., some rules may be pruned) so that a higher ϵ value could be used in a differentially private learning task to achieve better accuracy while being more resistant to model inversion attacks. Similar to the previous observations in the literature [3], a pure differently private model could not reach the desired protection against model inversion attacks while providing accurate prediction accuracy. Instead, our proposed approach achieved certain privacy risk goals while being differentially private using a higher ϵ value. This example shows that a framework that considers different aspects such as explainability and privacy could be integrated using threat modeling-based risk analysis.

Framework Development: As the above framework suggests, it is important to consider not only a broad set of concerns such as security, privacy, robustness, bias in general but also specific threats associated with each of these concerns. For example, for privacy, different attacks need to be considered. In case of transparency, the impact of a certain rule-based explainable AI system's impact on privacy also needs to be understood.

In addition to these framework elements, their interactions with other risk management frameworks may need to be addressed. For example, good data governance and cybersecurity risk management framework may be used to mitigate data poisoning attacks.

References

[1] B. Jayaraman and D. Evans, "Evaluating differentially private machine learning in practice," in *28th USENIX Security Symposium (USENIX Security 19)*. Santa Clara, CA: USENIX Association, Aug. 2019, pp. 1895–1912.

[2] Y. Alufaisan, M. Kantarcioglu, and Y. Zhou, "Robust transparency against model inversion attacks," *IEEE Transactions on Dependable and Secure Computing*, pp. 1–1, 2020.

[3] M. Fredrikson, S. Jha, and T. Ristenpart, "Model inversion attacks that exploit confidence information and basic countermeasures," in *Proceedings of the 22Nd ACM SIGSAC Conference on Computer and Communications Security*, ser. CCS '15. NY, USA: ACM, 2015, pp. 1322–1333.