Trello, Rachel M. (Fed)

From: José María Alises Sanz <jmalises@gmv.com>
Sent: Monday, February 5, 2024 3:39 AM
To: ai-inquiries
Subject: NIST RFI Answers to your question request

Regarding your Request for Information (RFI) that will assist in implementing its responsibilities under the recent Executive Order on Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence (AI). Responses will support NIST's efforts to evaluate capabilities relating to AI technologies and develop a variety of guidelines called for in the Executive Order. The RFI specifically calls for information related to AI red-teaming, generative AI risk management, synthetic content labeling and authentication, and advancing responsible global technical standards for AI development

We give you our comments regarding IA:
.The following attacks on AI (Artificial Intelligence) systems could be considered in red team activities:

- Poisoning attacks: When training ML (Machine Learning) models, the adversary tries to corrupt the training set so that the learned model produces misclassifications that benefit the adversary. For example, in an AI facial recognition classification for a security system, an attacker, with malicious access, inserts images of unauthorized persons labelled as legitimate users into the dataset.

- Evasion attacks: Introduce noise at the inputs to mispredict the output. For example, in an AI system that predicts traffic signals, to introduce noise to one stop signal to predict a 120km/h maximum speed sign.

- Model extraction attacks: Attack in which the adversary attempts to steal the model, compromising the confidentiality of the model. For example, by giving multiple inputs and observing the outputs of an ML model, the model parameters can be obtained and cloned.

- Model inversion attacks: An attacker tries to exploit the model's predictions to compromise the user's privacy, or to infer whether data was used in the training set. Highly relevant in models trained on sensitive data. For example, in a facial recognition system, an adversary tries to discover personal details by generating synthetic faces that fool the model. It performs specific queries to the system to infer information about people. By analyzing the responses, the attacker can obtain data such as people's appearance and location, compromising privacy.

On the other hand, it can be considered using AI for red team activities to establish attack points, triage results, etc.

Sincerely
Mr. Alises and PAC team

José María Alises Sanz
Director of PNT Acreditation & Certification (PAC) – NAV

T. +34 918072100
M. +34 689202321

Isaac Newton, 11. P.T.M.
28760 Tres Cantos, Madrid |

www.gmv.com