Building-NIST-AI-Risk-Management-Framework-workshop-3

AI POISONING ATTACKS AND COUNTERMEASURES

I would like to recommend more discussion on mechanisms for AI poisoning attacks and countermeasures.
The discussion should focus on mechanisms for AI poisoning attacks methods and countermeasures such as Tactics Techniques and Procedure (TTP) and Confidentiality, Integrity, and Availability (CIA) approaches.

Here is an example
AI poisoning attacks focus on the manipulation of the data collected and stored by devices or nodes in the perception layer of the system. The threat model of poisoning attacks is related to the attacker's knowledge, the attacker's goal, the attacker's strategy of the ML model, and the attacker's ability to influence the training data.

Threat model (CIA) approach
Threat modeling ensures the identification and quantification of possible security threats and vulnerabilities in intelligent networking systems.

Here is another example
TTP development for Attacker's knowledge technique should be considered to address Perfect-knowledge, Limited-knowledge and Zero-knowledge.
System and information integrity policy; configuration management policy and TTP's addressing AI poisoning attacks protection should be implemented, (see examples below).

Attacker's knowledge (vulnerability)
Considering the knowledge of the classifier components, the attacker's knowledge can be categorized into different levels.

AI poisoning attack vectors (vulnerability)
● Perfect-knowledge: corresponding to white-box attacks, where the attacker fully knows the target system, including the training data, the feature representation set, the learning algorithm, and the training parameters. This kind of knowledge allows attackers to provide the worst-case scenario of the target system.

Attacker's knowledge (vulnerability)
● Limited-knowledge: corresponding to gray-box attacks, it is assumed that the attacker has limited knowledge of the target system, including a surrogate training data sampled from a similar distribution, the feature representation set, the learning algorithm, and parameters trained by a surrogate classifier. This setting allows attackers to simulate a more realistic scenario of the target system.

Attacker's knowledge (TTP) (CIA)
● Zero-knowledge: corresponding to black-box attacks, where the attacker has little knowledge of the target system and can only query the target system and obtain their corresponding posteriors. Compared to the white-box and gray-box settings, the black-box setting is more challenging for the attacker.

Sincerely,

Reginald K. Richardson, President/CEO
Sapphire Innovative Solutions Inc. dba Sapphire BLU