Six ways to attack an AI system.

Are your AI applications prepared for them?

Poisoning

AI poisoning is a tactic where attackers manipulate the data used to train artificial intelligence (AI) models, causing these models to produce incorrect results or become unreliable. Attackers can introduce subtle errors into training data, such as mislabeling images or biased information, or embed hidden triggers that cause the AI to act unexpectedly when activated. This manipulation can occur intentionally by bad actors, accidentally by use of biased or poor-quality data, or even during normal use if the AI continues to learn from manipulated input or AI content ("feedback loops").

Trojan Horse

With this form of attack, bad actors secretly insert harmful code into AI models, especially large language models, before companies use them, expecting that they cannot check what is hidden inside these models when they obtain them from open sources or buy them. Once these tampered models are used, the hidden malicious code may be activated in one way or another, acting like a trojan horse and using, for instance, unprotected systems (e.g., third-party tools with elevated privileges or insecure browsers) to launch attacks from within a company.

Prompt Injection

Prompt injection attacks involve tricking an AI system by entering malicious commands instead of normal input. These commands can manipulate the AI to perform unintended actions, like revealing sensitive data or the secret "system prompts" of an AI system, turning off safety controls, or even taking control of other systems that process the output generated by an AI system that is being misused by an attacker. Malicious commands can be included in prompts, but also in documents that a user may upload to an AI system for analysis, resulting in manipulated output.

Sponge Attack

Sponge attacks target AI systems by overwhelming them with complex or large inputs, like a sponge soaking up their computing power. This can slow down or even damage a system. Attackers may do so by crafting inputs that are hard to process, causing the AI to use excessive energy or memory. Such harmful input may be included in a model during the training phase, making the system vulnerable from the start, or they are added later on. This can lead to delays, damage, or safety risks, for example where AI system must remain responsive at all times (e.g., in autonomous vehicles).

Model & Data Theft

Attackers target AI systems to uncover secret data contained in them or how an AI or its model was built. They might trick the AI into revealing if certain data was used in its training or infer private details from the AI's responses. One method does so by testing the system with real data to determine whether it recognizes it with certainty, indicating that it has already seen it during training. Another approach involves flooding the system with specific questions to replicate its logic. These tactics may not only expose sensitive or proprietary information but can lay groundwork for more advanced attacks.

Deception

Attackers can trick AI systems that rely on pattern recognition by using manipulated input to trigger certain (false) responses. For example, if an AI relies on image recognition to classify objects (e.g., speed limit signs), the attacker may use visual elements (e.g., certain stickers on a sign) that may even be invisible to a human to cause the AI into incorrectly assess the object. This may also work with face recognition. In a "white-box" attack the attacker has inside knowledge of the model, whereas in a "black-box" attack, the attacker figures out how to deceive the AI through trial and error.

Author: David Rosenthal (drosenthal@vischer.com) Updates: vischerlnk.com/ai-attacks

VISCHER
SWISS LAW AND TAX