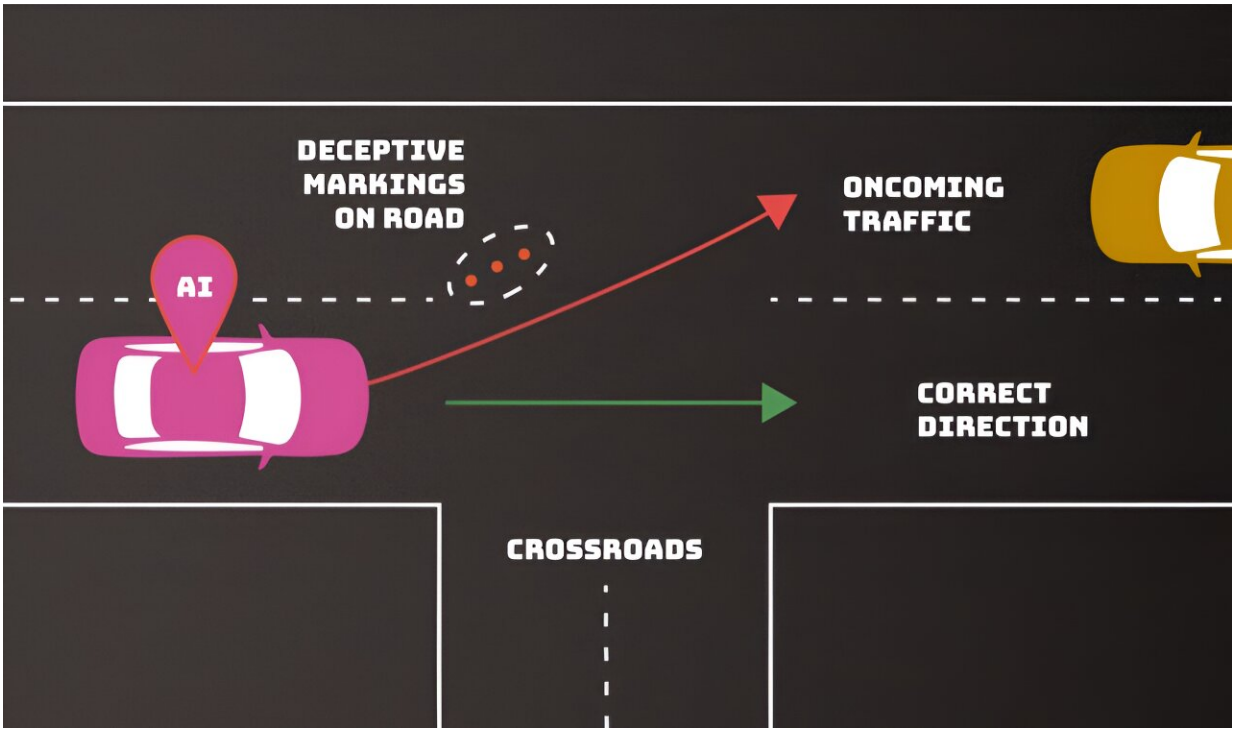
New report identifies types of cyberattacks that manipulate behavior of AI systems

January 4 2024



An AI system can malfunction if an adversary finds a way to confuse its decision making. In this example, errant markings on the road mislead a driverless car, potentially making it veer into oncoming traffic. This “evasion” attack is one of numerous adversarial tactics described in a new NIST publication intended to help outline the types of attacks we might expect along with approaches to mitigate them. Credit: N. Hanacek/NIST

Adversaries can deliberately confuse or even "poison" artificial

intelligence (AI) systems to make them malfunction—and there's no foolproof defense that their developers can employ. Computer scientists from the National Institute of Standards and Technology (NIST) and their collaborators identify these and other vulnerabilities of AI and machine learning (ML) in a new publication.

Their work, titled Adversarial Machine Learning: A Taxonomy and Terminology of Attacks and Mitigations, is part of NIST's broader effort to support the development of trustworthy AI, and it can help put NIST's AI Risk Management Framework into practice. The publication, a collaboration among government, academia, and industry, is intended to help AI developers and users get a handle on the types of attacks they might expect along with approaches to mitigate them—with the understanding that there is no silver bullet.

"We are providing an overview of attack techniques and methodologies that consider all types of AI systems," said NIST computer scientist Apostol Vassilev, one of the publication's authors. "We also describe current mitigation strategies reported in the literature, but these available defenses currently lack robust assurances that they fully mitigate the risks. We are encouraging the community to come up with better defenses."

AI systems have permeated modern society, working in capacities ranging from driving vehicles to helping doctors diagnose illnesses to interacting with customers as online chatbots. To learn to perform these tasks, they are trained on vast quantities of data: An autonomous vehicle might be shown images of highways and streets with road signs, for example, while a chatbot based on a large language model (LLM) might be exposed to records of online conversations. This data helps the AI predict how to respond in a given situation.

One major issue is that the data itself may not be trustworthy. Its sources

may be websites and interactions with the public. There are many opportunities for bad actors to corrupt this data—both during an AI system's training period and afterward, while the AI continues to refine its behaviors by interacting with the physical world. This can cause the AI to perform in an undesirable manner. Chatbots, for example, might learn to respond with abusive or racist language when their guardrails get circumvented by carefully crafted malicious prompts.

"For the most part, software developers need more people to use their product so it can get better with exposure," Vassilev said. "But there is no guarantee the exposure will be good. A chatbot can spew out bad or toxic information when prompted with carefully designed language."

In part because the datasets used to train an AI are far too large for people to successfully monitor and filter, there is no foolproof way as yet to protect AI from misdirection. To assist the developer community, the new report offers an overview of the sorts of attacks its AI products might suffer and corresponding approaches to reduce the damage.

The report considers the four major types of attacks: evasion, poisoning, privacy and abuse attacks. It also classifies them according to multiple criteria such as the attacker's goals and objectives, capabilities, and knowledge.

- Evasion attacks, which occur after an AI system is deployed, attempt to alter an input to change how the system responds to it. Examples would include adding markings to stop signs to make an autonomous vehicle misinterpret them as speed limit signs or creating confusing lane markings to make the vehicle veer off the road.

- Poisoning attacks occur in the training phase by introducing corrupted data. An example would be slipping numerous

instances of inappropriate language into conversation records, so that a chatbot interprets these instances as common enough parlance to use in its own customer interactions.

- Privacy attacks, which occur during deployment, are attempts to learn sensitive information about the AI or the data it was trained on in order to misuse it. An adversary can ask a chatbot numerous legitimate questions, and then use the answers to reverse engineer the model so as to find its weak spots—or guess at its sources. Adding undesired examples to those online sources could make the AI behave inappropriately, and making the AI unlearn those specific undesired examples after the fact can be difficult.

- Abuse attacks involve the insertion of incorrect information into a source, such as a webpage or online document, that an AI then absorbs. Unlike the aforementioned poisoning attacks, abuse attacks attempt to give the AI incorrect pieces of information from a legitimate but compromised source to repurpose the AI system's intended use.

"Most of these attacks are fairly easy to mount and require minimum knowledge of the AI system and limited adversarial capabilities," said co-author Alina Oprea, a professor at Northeastern University. "Poisoning attacks, for example, can be mounted by controlling a few dozen training samples, which would be a very small percentage of the entire training set."

The authors—who also included Robust Intelligence Inc. researchers Alie Fordyce and Hyrum Anderson—break down each of these classes of attacks into subcategories and add approaches for mitigating them, though the publication acknowledges that the defenses AI experts have devised for adversarial attacks thus far are incomplete at best.

Awareness of these limitations is important for developers and organizations looking to deploy and use AI technology, Vassilev said.

"Despite the significant progress AI and machine learning have made, these technologies are vulnerable to attacks that can cause spectacular failures with dire consequences," he said. "There are theoretical problems with securing AI algorithms that simply haven't been solved yet. If anyone says differently, they are selling snake oil."

More information: Apostol Vassilev et al, Adversarial Machine Learning: A Taxonomy and Terminology of Attacks and Mitigations, *NIST* (2024). DOI: 10.6028/NIST.AI.100-2e2023

Provided by National Institute of Standards and Technology