Arjun Nitin Bhagoji and Liwei Song, graduate students in electrical engineering, are developing defenses to protect artificial intelligence from hackers. The Stop sign has been modified to change its meaning to systems relying on computer vision.
Photo by David Kelly Crow

Protecting smart machines from smart attacks

Adam Hadhazy for the Office of Engineering Communications

Oct. 14, 2019, 3:11 p.m.

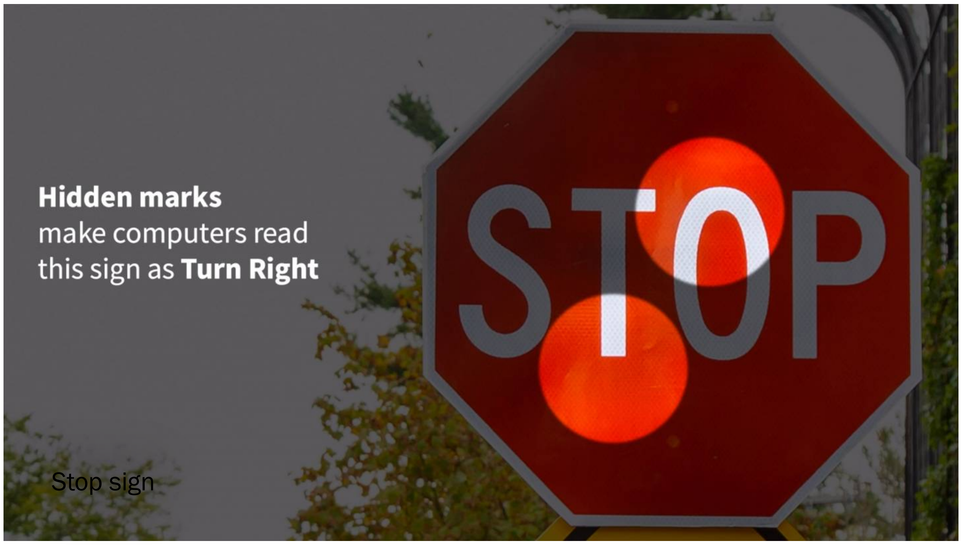
Machines' ability to learn by processing data gleaned from sensors underlies automated vehicles, medical devices and a host of other emerging technologies. But that learning ability leaves systems vulnerable to hackers in unexpected ways, researchers at Princeton University have found.

In a series of recent papers, a research team has explored how adversarial tactics applied to artificial intelligence (AI) could, for instance, trick a traffic-efficiency system into causing gridlock or manipulate a health-related AI application to reveal patients' private medical history. As an example of one such attack, the team altered a driving robot's perception of a road sign from a speed limit to a "Stop" sign, which could cause the vehicle to dangerously slam the brakes at highway speeds; in other examples, they altered Stop signs to be perceived as a variety of other traffic instructions.

“If machine learning is the software of the future, we’re at a very basic starting point for securing it,” said **Prateek Mittal** (https://ee.princeton.edu/people/prateek-mittal), the lead researcher and an associate professor in the **Department of Electrical Engineering** (https://ee.princeton.edu/) at Princeton. “For machine learning technologies to achieve their full potential, we have to understand how machine learning works in the presence of adversaries. That’s where we have a grand challenge.”

Just as software is prone to being hacked and infected by computer viruses, or its users targeted by scammers through phishing and other security-breaching ploys, AI-powered applications have their own vulnerabilities. Yet the deployment of adequate safeguards has lagged. So far, most machine learning development has occurred in benign, closed environments — a radically different setting than out in the real world.

Mittal is a pioneer in understanding an emerging vulnerability known as adversarial machine learning. In essence, this type of attack causes AI systems to produce unintended, possibly dangerous outcomes by corrupting the learning process. In their recent series of papers, Mittal’s group described and demonstrated three broad types of adversarial machine learning attacks.



A series of recent papers by researchers in electrical engineering explored how adversaries could trick machine learning systems in various ways. In one possible hack, attackers could make slight modifications into objects that machines have previously learned to identify correctly. This stop sign, for example, has been engineered to make a self-driving car interpret it as saying “Turn Right” instead of “Stop.”
Video clip courtesy of the researchers

Poisoning the data well

The first attack involves a malevolent agent inserting bogus information into the stream of data that an AI system is using to learn — an approach known as data poisoning. One common example is a large number of users’ phones reporting on traffic conditions. Such crowdsourced data can be used to train an AI system to develop models for better collective routing of autonomous cars, cutting down on congestion and wasted fuel.

“An adversary can simply inject false data in the communication between the phone and entities like Apple and Google, and now their models could potentially be compromised,” said Mittal. “Anything you learn from corrupt data is going to be suspect.”

Mittal’s group recently demonstrated a sort of next-level-up from this simple data poisoning, an approach they call “model poisoning.” In AI, a “model” might be a set of ideas that a machine has formed, based on its analysis of data, about how some part of the world works. Because of privacy concerns, a person’s cellphone might generate its own localized model, allowing the individual’s data to be kept confidential. The anonymized models are then shared and pooled with other users’ models. “Increasingly, companies are moving towards distributed learning where users do not share their data directly, but instead train local models with their data,” said Arjun Nitin Bhagoji, a Ph.D. student in Mittal’s lab.

But adversaries can put a thumb on the scales. A person or company with an interest in the outcome could trick a company’s servers into weighting their model’s updates over other users’ models. “The adversary’s aim is to ensure that data of their choice is classified in the class they desire, and not the true class,” said Bhagoji.

In June, Bhagoji presented a **paper** (https://arxiv.org/abs/1811.12470) on this topic at the 2019 International Conference on Machine Learning (ICML) in Long Beach, California, in collaboration with two researchers from IBM Research. The paper explored a test model that relies on image recognition to classify whether people in pictures are wearing sandals or sneakers. While an induced misclassification of that nature sounds harmless, it is the sort of unfair subterfuge an unscrupulous corporation might engage in to promote its product over a rival’s.

“The kinds of adversaries we need to consider in adversarial AI research range from individual hackers trying to extort people or companies for money, to corporations trying to gain business advantages, to nation-state level adversaries seeking strategic advantages,” said Mittal, who is also associated with Princeton’s **Center for Information Technology Policy** (http://cipt.princeton.edu/).

Using machine learning against itself

A second broad threat is called an evasion attack. It assumes a machine learning model has successfully trained on genuine data and achieved high accuracy at whatever its task may be. An adversary could turn that success on its head, though, by manipulating the inputs the system receives once it starts applying its learning to real-world decisions.

For example, the AI for self-driving cars has been trained to recognize speed limit and stop signs, while ignoring signs for fast food restaurants, gas stations, and so on. Mittal’s group has explored a loophole whereby signs can be misclassified if they are marked in ways that a human might not notice. The researchers made fake restaurant signs with extra color akin to graffiti or paintball splotches. The changes fooled the car’s AI into mistaking the restaurant signs for stop signs.

“We added tiny modifications that could fool this traffic sign recognition system,” said Mittal. A **paper** (https://arxiv.org/pdf/1801.02780.pdf) on the results was presented at the 1st Deep Learning and Security Workshop (DLS), held in May 2018 in San Francisco by the Institute of Electrical and Electronics Engineers (IEEE).

While minor and for demonstration purposes only, the signage perfidy again reveals a way in which machine learning can be hijacked for nefarious ends.

Not respecting privacy

The third broad threat is privacy attacks, which aim to infer sensitive data used in the learning process. In today's constantly internet-connected society, there's plenty of that sloshing around. Adversaries can try to piggyback on machine learning models as they soak up data, gaining access to guarded information such as credit card numbers, health records and users' physical locations.

An example of this malfeasance, studied at Princeton, is the "membership inference attack." It works by gauging whether a particular data point falls within a target's machine learning training set. For instance, should an adversary alight upon a user's data while picking through a health-related AI application's training set, that information would strongly suggest the user was once a patient at the hospital. Connecting the dots on a number of such points can disclose identifying details about a user and their lives.

Protecting privacy is possible, but at this point it involves a security tradeoff — defenses that protect the AI models from manipulation via evasion attacks can make them more vulnerable to membership inference attacks. That is a key takeaway from a **new paper** (https://arxiv.org/pdf/1905.10291.pdf) accepted for the 26th ACM Conference on Computer and Communications Security (CCS), to be held in London in November, led by Mittal's graduate student Liwei Song. The defensive tactics used to protect against evasion attacks rely heavily on sensitive data in the training set, which makes that data more vulnerable to privacy attacks.

It is the classic security-versus-privacy debate, this time with a machine learning twist. Song emphasizes, as does Mittal, that researchers will have to start treating the two domains as inextricably linked, rather than focusing on one without accounting for its impact on the other.

"In our paper, by showing the increased privacy leakage introduced by defenses against evasion attacks, we've highlighted the importance of thinking about security and privacy together," said Song.

It is early days yet for machine learning and adversarial AI — perhaps early enough that the threats that inevitably materialize will not have the upper hand.

"We're entering a new era where machine learning will become increasingly embedded into nearly everything we do," said Mittal. "It's imperative that we recognize threats and develop countermeasures against them."

Besides Mittal and Bhagoji, Princeton authors on the DLS paper are Chawin Sitawarin, now a graduate student at the University of California, Berkeley, and Arsalan Mosenia, now working for Google, who performed the research as a postdoctoral researcher working jointly with Mittal and Professor Mung Chiang at Purdue University's Department of Electrical and Computer Engineering. Other