Defense Against Adversarial AI Covers a Big Battlefield

DARPA wants to use AI to detect AI attacks before they can do damage.

Kevin McCaney

Mon, 03/04/2019 - 09:47

Photo credit: matejmo/iStock

As machine-learning systems have become more powerful, developers caught up in making progress may have overlooked the fact that machine learning's potential to do harm increases proportionately with its ability to good. And considering how quickly the technology is being advanced, it has already created a threat environment that the Pentagon’s lead research arm calls “pessimistic.”

The idea of powerful thinking machines going rogue isn't new. Science fiction is full of tales of digital overlords, and in the real world, Elon Musk, the late Stephen Hawking and others haven't been shy about painting potential end-of-the-world scenarios of what could happen if AI's development proceeds without caution. But many of those predictions seemed to be down the road in a Matrix/Ex Machina kind of world. Meanwhile, some of the shorter-term risks of AI and its commonly employed subset of machine learning may have been brushed aside while positive improvements were being made.

"Over the last decade, researchers have focused on realizing practical ML capable of accomplishing real-world tasks and making them more efficient," Dr. Hava Siegelmann, program manager in the Defense Advanced Research Projects Agency's Information Innovation Office (I2O), said, noting the tangible benefits of that work. "But, in a very real way, we've rushed ahead," Siegelmann said, "paying little attention to vulnerabilities inherent in ML platforms — particularly in terms of altering, corrupting or deceiving these systems."

Attackers In The Lead

DARPA wants to get back in front of potential ML threats with a new program called Guaranteeing AI Robustness against Deception (GARD), designed to counter adversarial deception attacks on ML models — which DARPA refers to in general terms as adversarial AI. And in this game, attackers have a head start. "The field now appears increasingly pessimistic," DARPA said in a special notice on the program, "sensing that developing effective ML defenses may prove significantly more difficult than designing new attacks, leaving advanced systems vulnerable and exposed."

GARD aims to develop broad-based defenses that can adapt to changeable or unexpected attacks, as opposed to current defensive tactics that address specific, predefined exploits.

The Intelligence Advanced Research Projects Activity, for instance, recently launched a program targeting a specific type of attack in which AI can be deceived — Trojans implanted in an AI's training data that could allow the AI to be taken over in certain situations. IARPA gave the example of a sticky note on a stop sign that could fool a self-driving car into running the stop sign, potentially hitting pedestrians or other cars. That type of threat could also affect AI in a variety of other situations, from a drone selecting a target to an AI making an inaccurate a medical diagnosis.

But while the program, called TrojAI, addresses a significant threat, it also addresses only one way AI could be exploited. There are as many ways AI could go off the rails and cause harm as there are things AI can do. Researchers have highlighted the potential of AI to supercharge hacking or manipulate images, videos and sound recordings to take fake news to levels that would flabbergast George Orwell. AI also can enable other sophisticated cyber attacks, such as highly targeted phishing campaigns, and new breeds of polymorphic malware — which constantly changes identifiers such as signatures and encryption keys — to evade detection.

Defensive Systems That Learn Quickly

GARD's goal is to develop defenses that can counter any number of possible attacks in a given scenario. And, as with many of its past projects, DARPA is willing to look at outside-the-box approaches, such as biological defenses against disease. "The kind of broad scenario-based defense we're looking to generate can be seen, for example, in the immune system, which identifies attacks, wins and remembers the attack to create a more effective response during future engagements," Siegelmann said.

The program will start with state-of-the-art image-based machine learning, and then move on to video, audio and complex systems such as multi-sensor and multi-modality variations, DARPA said. DARPA also wants to use the program to tackle the potential threats of ML systems capable of predictions and decisions, and of adapting to changing circumstances.

“There is a critical need for ML defense as the technology is increasingly incorporated into some of our most critical infrastructure. The GARD program seeks to prevent the chaos that could ensue in the near future when attack methodologies, now in their infancy, have matured to a more destructive level. We must ensure ML is safe and incapable of being deceived,” Siegelmann said.

View printer friendly version
AI
DARPA
machine learning
IARPA