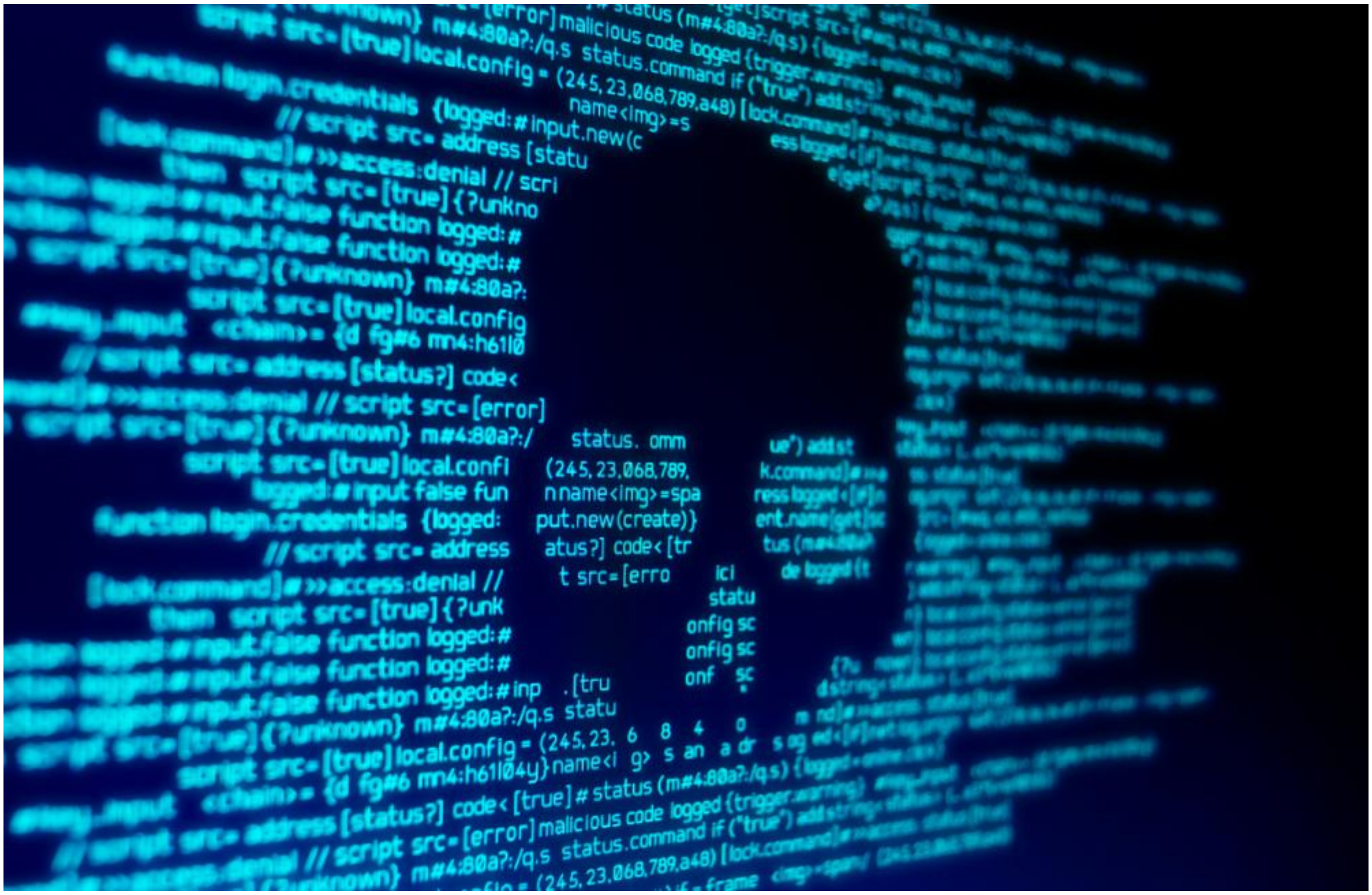
How a Trojan Can Turn an AI into a Manchurian Candidate

The nature of AI programs could breed Trojans, a problem the IARPA is addressing with an upcoming program.

Kevin McCaney

Mon, 12/31/2018 - 11:00



solarseven/iStock

In a near-future world of self-driving cars, delivery drones and even machine-piloted passenger planes whizzing about in a tidy, efficient metropolis, it’s not hard to imagine a scenario where one or more (or all) of the machines suddenly go haywire — careening out of their lanes, ignoring the rules and turning an ordered universe into mass mayhem.

The Intelligence Community's research arm has one idea about how something like that can happen — though in a smaller, targeted-but-still-dangerous incident — with a fairly basic trick that can turn artificial intelligence systems' learning abilities against them. And all it could take is something as simple as a sticky note in the wrong place.

In the example given by the Intelligence Advanced Research Projects Activity, an AI that's advanced enough to control a self-driving car — obeying all the rules of the road, navigating via flawless directions, anticipating what lies ahead — is fooled by an everyday sticky note placed on a stop sign. In that example, the AI runs the stop, possibly hitting pedestrians or crashing into other cars.

The attack starts with one of the oldest tools in the malware book — a Trojan program, in this case inserted into the AI's training. (The first Trojan, a self-replicating but non-malicious program called Animal, was written in 1975 for the UNIVAC 1108.) The Trojan plants a trigger into the AI training program, telling it that a traffic sign that has a small yellow square on it is a speed limit sign. So, even if it's a red octagonal sign that says "STOP," the AI "learns" that, rather than telling the car to stop, the sign is telling it to go. The sticky note is all it takes.

That vulnerability of AI systems to Trojan attacks is the focus of IARPA's new TrojAI program, which is looking for ideas on how to inspect AIs for Trojans to determine if they're safe to deploy — a process that isn't as easy as it might sound.

Lying in Wait

The sticky note on a stop sign is just one example. There are plenty of other situations in which this type of attack could be a problem, given AI's increasingly widespread use. And a successful attack could implant several triggers, changing an AI agent's behavior in multiple ways. The result could be AI programs — whether in cars, drone aircraft, cybersecurity systems or other applications — that effectively become Manchurian candidates, controlled by an outside, possibly foreign, agent.

As IARPA notes in a draft Broad Agency Announcement for the TrojAI program, a Trojan (backdoor or trapdoor) attack operates stealthily. Unlike a broader data-poisoning attack that disrupts everything from the get-go, it inserts a trigger that will change a program's behavior in certain circumstances only. The trigger must be something that exists in the world, but is rare in the normal operating environment, like sticky notes that are everywhere in offices and not likely to be found on traffic signs. It won't affect performance in tests or normal operations, so it won't give itself away during normal testing. If they're there, it's because an attacker has put it there to get the AI to do its bidding at a time of the attacker's choosing.

"The obvious defenses against Trojan attacks are cybersecurity (to protect the training data) and data cleaning (to make sure the training data is accurate)," IARPA said in its BAA. But that task is complicated by the nature of AI programs, which often are created through large, crowdsourced data sets in which a traditional approach to monitoring and cleaning software is impractical. Finding Trojans in an AI that's already been trained will require inspecting its internal logic. Many AIs also are the product of what's called transfer learning in which an existing AI is modified for a new, specific use, so a successfully installed Trojan could live in future variations that grow out of a single program. "The security of the AI is thus dependent on the security of the entire data and training pipeline, which may be weak or nonexistent," IARPA said.

Bad Lessons Learned

IARPA expects TrojAI to be a two-year program, starting with a base year and a one-year option, with the possibility of more work to follow. The team chosen for the program will try to develop an automated system (with no human-in-the-loop involvement) for inspecting an AI and predicting if a Trojan is present. Initial work will be on deep neural networks training in classification tasks, in which they classify small images at human or super-human speeds, as AIs tend to do. The team will develop software that can detect whether an AI has been trained by a Trojan to misclassify some of the images.

While the program will start with fairly small classification tasks, it could be expanded to larger data sets — say, a forest as opposed to a set of street signs — and other formats, such as video and audio.

AI systems are vulnerable to the same kinds of cyberattacks that plague other systems, but their complexity can make defending against attacks even more difficult. The TrojAI program addresses one aspect of defending AI, but considering the responsibilities AI systems are taking on, it could go a long way toward avoiding unwelcome surprises.

View PDF

AI
bots
cyberattacks
cybersecurity