FlightGlobal

US Air Force grapples with vexing problem of AI spoofing

By Garrett Reim | 1 September 2020

The US Department of Defense (DoD) is worried that artificial intelligence programs might have serious and unknown vulnerabilities that adversaries could exploit.

In particular, the Pentagon is worried that the technology could not only be hacked, but could be "spoofed". That is, it could be intentionally deceived into thinking that it sees objects – or military targets – that do not exist. The opposite is true as well: military targets could be erroneously ignored.

Source: US Air Force

XQ-58A Valkyrie demonstrator, a candidate for Skyborg artificial intelligence software

That is one reason the US Air Force (USAF) and the Massachusetts Institute of Technology founded the "MIT-Air Force AI Accelerator" in 2019. The accelerator, funded with $15 million per year from the service, is looking for vulnerabilities in artificial intelligence and ways to harden the technology against enemy manipulation, said Will Roper, assistant secretary of the USAF for acquisition, technology and logistics, in June.

Already, MIT scientists have found ways to trick some of the world's best artificial intelligence programs.

"They can hold up an image, and it's an airliner, and everyone would look at it and say, 'Well, that's an airliner,'" says Roper. "You run it through the world's best machine learning algorithms and it types it as a pig."

Machine learning programs are a subset of artificial intelligence that examine reams of data, look for patterns and draw conclusions. The programs are designed to mimic the way human brains recognise patterns.

The ability to fool machine learning programs is concerning to the USAF because it plans to offload work from human fighter pilots to robotic loyal wingman unmanned air vehicles (UAVs) that will be controlled by similar artificial intelligence programs.

In July, the USAF granted four indefinite delivery/indefinite quantity contracts, worth up to $400 million each, to develop competing examples of an artificial intelligence program it calls Skyborg. Boeing, General Atomics Aeronautical Systems, Kratos Unmanned Aerial Systems and Northrop Grumman each received an award.

Skyborg is to be the brain and stem of the USAF's loyal wingman UAVs – autonomous aircraft the service hopes to produce in such large numbers so as to overwhelm enemies. The service needs artificial intelligence, in place of remote pilots, to control such systems en masse.

In August, the Pentagon's ambition to field autonomous combat aircraft took a big step forward when an artificially intelligent software program defeated a USAF Lockheed Martin F-16 pilot in five simulated dogfights.

SHALLOW ARTIFICIAL INTELLIGENCE

In laymen's speak, artificial intelligence is a broad class of computer programs trained to recognise patterns. Often those patterns are observed within picture or video data. For example, an artificially intelligent program could be repeatedly fed pictures of a certain type of aircraft, and told what that aircraft's name is, until it learns to identify that aircraft on its own.

The USAF is already using artificial intelligence to speed up the tedious work of labelling intelligence, surveillance and reconnaissance photographs that come from aircraft, such as the Lockheed U-2 and Northrop RQ-4 Global Hawk.

In theory, similar technology aboard a loyal wingman UAV could also be used to identify the heat or radar signature of enemy aircraft. Unfortunately, artificial intelligence often jumps to conclusions.

"When you introduce something that has not been seen before, whether intentionally or unintentionally… you force the [artificial intelligence] to make a choice," says Roper. "And unlike people, it doesn't have the wherewithal to know that its choice is shallow or forced, or that there's something not right about the choice. There's not this audit ability that people have when making decisions."

Source: Boeing

Boeing Airpower Teaming System flying with EA-18G Growler. The UAV is a possible future loyal wingman

It can be very easy to fool an artificial intelligence program. One often-cited example of artificial intelligence manipulation is an experiment where a self-driving car program was fooled into thinking a stop sign was actually a 45 mph speed limit sign. The deception only required a few well-placed stickers. The result of speeding through a stop sign at 45 mph could be disastrous, of course.

There are two ways of exploiting vulnerabilities in an artificially intelligent program: a white box attack or a black box attack. In a white box attack situation the attacker has some sort of access to the artificial intelligence program and understands how it works. That knowledge is used to manipulate the program. In a black box attack the aggressor has no insider knowledge, but instead has to probe the artificial intelligence program from the outside, feeding it information and looking for instances where it appears to go haywire.

In April, the Defense Advanced Research Projects Agency (DARPA) began soliciating ideas for "Techniques for Machine Vision Disruption". That is a black box effort to fool artificial intelligence image identification "in situations where neither the original training set nor the specific vision architecture are available".

In many ways artificial intelligence spoofing is like military deception and electronic warfare of old, says Bruce Draper, programme manager for the DARPA Information Innovation Office. Draper is leading another DARPA effort: the Guaranteeing AI Robustness Against Deception programme, an initiative to establish theoretical foundations to identify system vulnerabilities and characterise certain properties that will make artificial intelligence programs more robust.

“Camouflage is an attempt to spoof the human visual system,” he says. “What we’re getting into now is more and more of the sensing is being done through artificial intelligence. So, people are trying to devise ways to spoof the artificial perceptual systems.”

In fact, artificial intelligence is robust in many ways, capable of sensing patterns amid confusing data that humans struggle with, says Draper. “But on the other hand, things that people have no problems with, like a sticker on the stop sign, if it’s the right colour, the right spot, the right shape, it can defeat the [artificial intelligence] system,” he says. “They see differently than we do. And therefore, what spoofs them is different from what spoofs us.”

For example, artificial intelligence programs looking at images or video tend to focus on high-frequency visual patterns. “The ridges on bricks tend to be things that, for whatever reason, these AI systems are particularly good at classifying,” says Draper. “So, if you can disrupt the visual texture you can sometimes spoof it.”

SAFEGUARDS

Building safeguards against those errors can be difficult because it is not easy to understand how artificial intelligence programs arrive at conclusions. In a way similar to how the human brain works, the software produces results not by being programmed to do specific tasks, but by observing information and then making generalisations about the relationships between many different data points.

“One of the concerns about [artificial intelligence] is a lot of us don’t know quite what it’s doing,” says Stuart Madnick, professor of information technology and engineering systems at MIT. “In the same way trying to figure out what’s going on in the mind of a one-year-old or two-year-old child is very hard.”

Trying to untangle an artificially intelligent program’s underlying generalisations can be a huge task.

“Mathematically, we’re talking about non-linear models with literally millions of parameters,” says Draper.

There are a number of ideas on how artificial intelligence could be made more reliable, however.

One easy-to-understand technique is called “ensembles of classifiers”, says Draper. That method aims to prevent artificial intelligence errors by feeding programs different types of data. For instance, a program might learn to classify objects not just by images coming from a camera, but also with data coming from lidar and radar. The additional perspectives would give the computer program a more holistic qualification for what an object looks like.

Another method is called “adversarial training”.

“The idea of adversarial training is that you take the attack you anticipate and make it part of your training process,” says Draper. “The system learns from the beginning how to be robust against this kind of attack.”

Still, sometimes knowing that an artificial intelligence program is being manipulated is difficult because certain programs are designed to find unusual, unexpected or overlooked solutions.

“You can’t just look at the outcome and say, ‘Well, the outcome wasn’t what I expected, and therefore there must be something wrong with the system’, because you’re using the [artificial intelligence] system to find needles in haystacks,” says Keri Pearlson, executive director of cybersecurity at MIT Sloan School of Management.

Ultimately, for artificial intelligence programs to be let loose and allowed to act on their own they must prove to be hardened against manipulation and predictable enough that humans can cooperate with them.

In the near term, that means loyal wingman aircraft are going to be closely supervised by commanders flying in controlling aircraft.

“If we’re going to take it onto the battlefield, we need to make sure that the operators who are using it are trained about the vulnerabilities and are put in the best position to put their [artificial intelligence] in the best position to win,” says Roper.

In the longer term, more work needs to be done to understand artificial intelligence’s weaknesses and limitations.

“There is this belief that [artificial intelligence], just throw enough data at it and everything will be okay,” says Roper. “And, that’s not the case. We need another generation of this technology.”