

# Deceiving AI

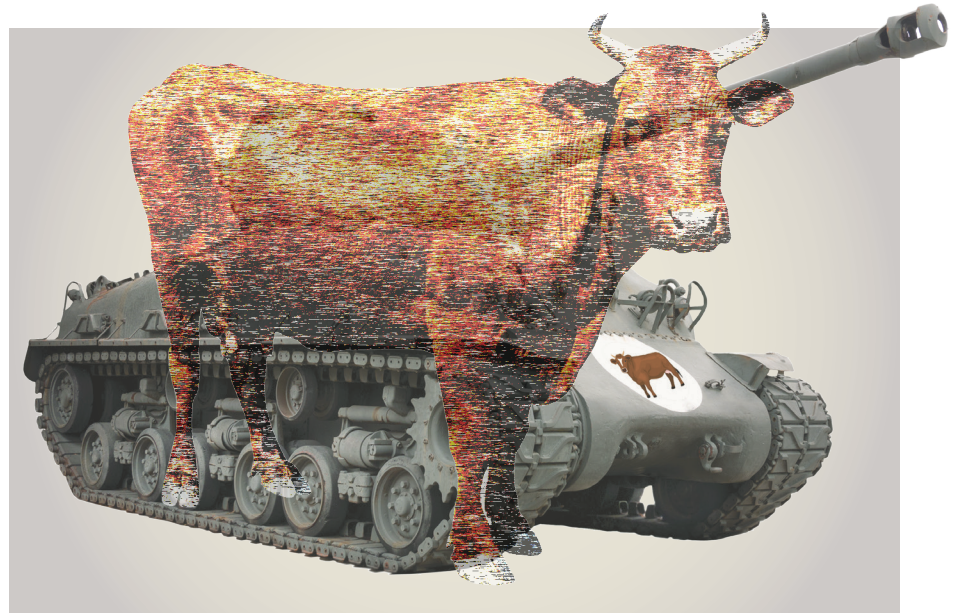
*The uncanny power of machine learning can be turned against it.*

**O**VER THE LAST decade, deep learning systems have shown an astonishing ability to classify images, translate languages, and perform other tasks that once seemed uniquely human. However, these systems work opaquely and sometimes make elementary mistakes, and this fragility could be intentionally exploited to threaten security or safety.

In 2018, for example, a group of undergraduates at the Massachusetts Institute of Technology (MIT) three-dimensionally (3D) printed a toy turtle that Google's Cloud Vision system consistently classified as a rifle, even when viewed from various directions. Other researchers have tweaked an ordinary-sounding speech segment to direct a smart speaker to a malicious website. These misclassifications sound amusing, but they could also represent a serious vulnerability as machine learning is widely deployed in medical, legal, and financial systems.

The potential vulnerabilities extend to military systems, said Hava Siegelman of the University of Massachusetts, Amherst. Siegelman initiated a program called Guaranteed AI Robustness against Deception (GARD) while she was on assignment to the U.S. Defense Advanced Research Projects Agency (DARPA). To illustrate the issue to colleagues there, she said, "I showed them an example that I did, and they all started screaming that the room was not secure enough." The examples she shares publicly are worrisome enough, though, such as a tank adorned with tiny pictures of cows that cause an artificial intelligence (AI)-based vision system to perceive it to be as a herd of cows because, she said, AI "works on the surfaces."

The current program manager for GARD at DARPA, Bruce Draper of Colorado State University, is more sanguine. "We have not yet gotten to that point where there's something out there that has happened that has given me night-



mares," he said, adding, "We're trying to head that off."

Researchers, some with funding from DARPA, are actively exploring ways to make machine learning more robust against adversarial attacks, and to understand the principles and limitations of these approaches. In the real world, these techniques are likely to be one piece of an ongoing, multilayered security strategy that will slow attackers but not stop them entirely. "It's an AI problem, but it's also a security problem," Draper said.

## Worth a Thousand Words

All machine learning tools can be deceived, but image classification is the most intuitively compelling. Altering a small number of pixels in an input image in a way that may be trivial or even unnoticeable to people, for example, can fool a classifier into declaring a stop sign to be a speed limit sign. This would clearly be a disaster for a self-driving car.

Such examples clearly reveal that deep learning systems base their decisions on features that may be completely different from what people regard as important. "Because these systems are using visual input, we

tend to assume that they see the same way we do," said Draper. "That's not a good assumption." Troublingly, this vulnerability likely extends to other applications of machine learning, where the results are harder to appreciate or visualize.

Image manipulation is particularly challenging to defend against, said Battista Biggio of the University of Cagliari in Sardinia, Italy, "because the space of pixels is so large that the attacker can do basically whatever he wants in terms of manipulating the images." In contrast, he said, for malware, "you have instructions or bytes, which have a specific meaning, so they cannot be altered in a trivial matter."

In 2013, Biggio and his coworkers showed how to fool a machine learning system by exploiting knowledge of the internal "gradients" it uses for training to design adversarial examples. "The basic idea is to use the same algorithm that is used for learning to actually bypass the classifier," Biggio said. "It's fighting machine learning with machine learning."

Around the same time, Google's Christian Szegedy and collaborators used similar techniques to train a network to make erroneous identifica-

tions. They then added imperceptible patterns to an image of a school bus until the classifier abruptly—and confidently—declared it to be an ostrich.

At first, it was unclear whether these pixel-level manipulations “were real problems or just theoretical constructions that ... couldn’t actually cause a problem to real vision systems,” said Andrew Ilyas. Back in 2018, this uncertainty motivated him and his fellow MIT undergraduates to concoct their rifle-mimicking turtle, explicitly insuring that the subtle details they decorated it with would be recognized from different viewpoints. “It worked better than even I thought,” said Ilyas, now a graduate student at MIT.

### Pick Your Poison

Deception “seems to be a particular issue with AI systems that use sensory input, whether it’s visual or audio or whatever,” said Draper. “It’s still more of an open question whether it’s also an issue for an AI system that’s working on network security.”

Whether it involves real-world sensors or later manipulation of the resulting digital data, alteration of the input to an existing classifier is called “evasion.” Another type of vulnerability occurs if an attacker can insert doctored data into the training set, which is known as “poisoning.”

Mislabeled training data can move the decision boundary that separates different classifications. Siegelmann cites a poisoning example where inserting images of Wonder Woman with distinctive eyeglasses, even when made almost imperceptible, can induce a system to classify anyone wearing such glasses as the superhero.

For an attacker to poison training data, they obviously must have access to that data. Similarly, an attacker can be most successful if they know the internal design details of a machine learning system, in what is called a “white-box” scenario.

“The white-box case is interesting when you want to understand the worst-case scenario,” said Biggio. “We expect then that in practice these systems remain more secure also under more restrictive models of attack.” Effective security protocols can obscure both the design and the data to create more challenging “black-box” sce-

nario, although a “gray-box” scenario, with some kinds of partial information known or inferred, is more realistic.

Even if the system details are hidden, however, researchers have found that attacks that work against one system frequently work against others that have a different—possibly unknown—internal structure. This initially surprising observation reflects the power of deep learning systems to find patterns in data, Biggio said. “In many cases, different classifiers tend to learn the same correlations from the data.”

This “transferability” of attacks highlights the risk of a common training set like ImageNet, which contains a huge set of annotated images that is widely used for training vision systems. Although this common training corpus makes it easy to compare the performance of different classifiers, it makes them all potentially vulnerable to the same biases, whether malicious or not.

### A Security Challenge

Siegelmann argues this vulnerability is intrinsic to deep learning systems, which all contain many hidden layers but vary in how they are interconnected. “All the deep networks are the same,” she said, and for decades they have always been trained by “backpropagation” into the network of the errors in their outputs, she said. “It doesn’t really matter exactly how it’s connected.”

Although Siegelmann worries about losing the power of machine learning in security-sensitive applications, she has no such qualms about eschewing the deep learning versions of it. “The point is how you teach it,” she said. For example, she advocates an active learning process in which the system chooses which questions to ask, rather than letting users pick examples that can intentionally lead it astray.

Of course, human observers can be misled by camouflage and other techniques, too. “It’s not really a shock that it’s also possible to fool artificial systems,” Draper said. “What is just always a shock to people is that what fools them is different.”


Some researchers hope errors can be avoided by designing systems that “explain” their reasoning. For deliberate attacks, however, “Explainable AI is the easiest thing to trick,” Siegelmann

cautioned, “because we teach the network to say what will convince us.”

In his own recent work, Ilyas has been exploring “adversarial training,” which involves using especially problematic training examples to guide the network to more reliable features for classification. This strategy may help avoid fragility in non-security operations as well, he said, if it “gets classifiers to use features that are closer to what humans use, so that they don’t fail in other, unexpected ways.”

DARPA’s GARD program will test a variety of approaches in regular challenges. In one strategy, for example, spoofing of sensory input is made harder by using multiple sensors simultaneously, and checking that they are compatible with each other and with what is known about the real world.

In the end, however, robust machine learning will just be one layer in the never-ending arms race to keep computer systems secure. In this context, what is currently missing is not so much air-tight defenses, but theoretical guarantees about how long a system can hold out, similar to those available in encryption. Such guarantees can inform the design of a secure, layered defense, Biggio said. “This is very far from what we have in the AI field.”

“It is very likely that there may be no silver bullet,” agreed Draper. “What we want to do is at least make it very difficult for someone to defeat or spoof one of these systems.” 

### Further Reading

Athalye, A., Engstrom, L., Ilyas, A., and Kwok, K. *Synthesizing Robust Adversarial Examples*, *Intl. Conf. on Machine Learning* <https://arxiv.org/abs/1707.07397> (2018).

Biggio, B. and Roli, F. *Wild Patterns: Ten Years After the Rise of Adversarial Machine Learning*, *Pattern Recognition* 84, 317 (2018), <http://bit.ly/3kC97WF>.

Szegedy, C., et al., “Intriguing properties of neural networks,” <https://arxiv.org/abs/1312.6199> (2014).

Wallace, E., Zhao, T.Z., Feng, S., and Singh, S. *Customizing Triggers with Concealed Data Poisoning*, <https://arxiv.org/abs/2010.12563> (2020)

Don Monroe is a science and technology writer based in Boston, MA, USA.