New capabilities come with new vulnerabilities, as it is true for new technology it is true for machine learning as well. Many solutions for network attack detection and malware detection rely on machine learning; although effective, these are vulnerable to attacks from adversaries who wish to subvert these algorithms and allow malicious apps to evade detection. Machine learning models accuracy depends on the data that is provided and also the algorithm which is being used. These models can suffer from imperfect learning which makes them vulnerable to a number of attacks.

The use of dynamic and adaptive methods in the area of security always contains a certain degree of risk. Especially when explainability of machine learning predictions is often lacking, attackers have been known to cause various algorithms to make erroneous predictions or learn the wrong class.

Today we discuss growing field of study called adversarial machine learning, attackers with varying degrees of access to a machine learning system can execute a range of attacks to achieve their ends.

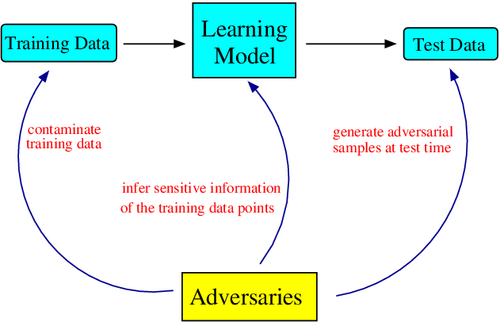
To begin with lets look at the recent impact of these adversarial attacks

Last year researchers showed how they were able to confuse a Tesla Model S into switching lanes so that it drives directly into oncoming traffic. All they had to do was place three stickers on the road, forming the appearance of a line. The car’s Autopilot system, which relies on computer vision, detected the stickers and interpreted them to mean that the lane was veering left. So it steered the car that way.

What is Adversarial Learning?

Adversarial machine learning is the study of machine learning vulnerabilities in adversarial environments. It aims that confusing the models and tricking it into predicting wrong results.

Security and machine learning researchers have published research on practical attacks against machine learning antivirus engines, spam filters, network intrusion detectors, image classifiers, sentiment analyzers.



This figure represents the different ways in which adversaries attack the models

Importance Of Adversarial Learning:

Attackers add noises or perturbations into the data set which are difficult to spot by humans. It can be done at the training or inference stage depending on the level of access they have. These can alter the output of the ml model. These are very difficult to spot because for a no. of data points the output is correct and expected.

Humans are often considered the weakest links in Security architecture. Maybe in future there can be ML systems which can train on the mistakes of human behavior and ask humans not to make those mistakes. As I wrote the last few sentences I realized how dangerous it could be. But it would help a lot in password stealing attacks social engineering like spear phishing attacks

Adversarial machine learning is difficult because most machine learning solutions behave as black boxes. The lack of transparency into what goes on inside detectors and classifiers makes it difficult for users and practitioners to make sense of model predictions.

This has been an increasingly active area of research in recent times, even though such attacks have rarely been observed in the wild. Vulnerabilities in machine learning systems can arise from flawed system design, fundamental algorithmic limitations, or a combination of both.

Since ML is now being used to handle sensitive data of billions of citizens and security has been more important than ever

ML practioners should preempt these.