**Team’s Impression: A Friendly Introduction to Adversarial Machine Learning**

*Video seen by:*

*Sutanu Bhattacharya*

*Rahul Alapati*

**Abstract:** This document mainly focuses on an informational video [1] “A Friendly Introduction to Adversarial Learning”, which was presented by Mr. Evan Wright. He discussed the basics of Machine learning and also explained how an adversary can attack any system.

According to him, Machine learning(ML) is like fire, either it can be used to burn us or give us warmth. Adversarial machine learning comes at the intersection of ML and security.

People have succeeded at constructing an example to fool the classification systems. For example, by applying certain types of distortion on images we can easily fool ML algorithms like SVM, Deep Learning. Another crucial factor in adversarial machine learning is untrusted training data which affects the prediction of ML algorithms. ML community assumes the training data to be the ground truth which represents an actual population of the target data. For example, a ML model which learns things based on what we say and how we interact with it. Now a bad guy tells the model that “I am a good guy. I hate everybody.” Although both the statements contradict each other, but the model fails to detect that the adversary, whereas a human can.

Adversarial machine learning is a certain kind of overfitting. Evasion attack, and poisoning are two types of adversarial machine learning attacks where the former is the most prevalent type of attack and the later can be regarded as an adversarial contamination of the training data.

There are a few protection measures that can be taken such as human intervention along with machine learning model, software security and use of adversarial framework to mitigate the adverse effect. In addition, it is important for security community that there is a large gap in the knowledge of ML, data science community because they do not understand about untrusted inputs. Moreover, community can consider updating their ML model.

In conclusion, attacks can occur, and attackers adapts to detection very fast. In addition, increased usage of ML methods make recipe for advanced attackers to do adversarial machine learning which is a future concern.