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

*Video seen by:*

*Aditya Mahajan*

*Nirmitkumar Patel*

*Vinod ChannigarayapuraGovindareddy*

**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 explains how an adversary can attack any system.

1. **INTRODUCTION:**
2. **Basics of Machine learning:**

Mr. Wright [1] excellently explained the brief introduction of Machine learning with some real-time working examples like, self-driving cars, Netflix, Medical diagnosis, malware detection & classification, and many more. According to him, there are many examples that we can use to classify any set of data. Most important is:

* Pattern detection (via clustering): Pattern detection means to classify available data set by means of recognizing patterns or regularities in data. Clustering is one of the best method to achieve this classification.

According to him, a ML model should always obey “Property of Generalization”. It can be achieved by dividing data set into 2 groups, “Training set” and “Test Set”, target is to avoid being ended with underfitted or overfitted. These are the 2 cases or means by which an adversary can attack a system.

1. **Adversarial Machine Learning:**

According to Mr. Wright [1], bad guys always try to be out-smart to catch our machines. An adversary always focuses on overfitted data and target all possible cases which a machine never explored, and then they use appropriate attacks, mostly: Causative or an Evasion attack.

* **Poisoning (Causative) attack:** Primary purpose is to target the model or algorithm and change it as per attacker’s requirement. He also explained 7 steps process (kill chain process) to manipulate the model and get it install in the target machine.
* **Evasion Attack:** Instead of targeting the model, this attack focuses more on real data set and update it with some new entries to make the prediction wrong and to gain the machine access. One possible example could be by constructing traffic on the network.

1. **Protection:**

Mr. Wright [1] also provided a list of measures that a Machine Learner should use to prevent system being attacked:

* Non-Linearity: By adding more non-linearity to the classifier, it would be more difficult for an adversary to study the model and attack on the same.
* Human Interaction: According to him, sometimes it is necessary to add some sort of Human interaction in the model so that he/she can ensure/check the system integrity over the time.

Big Picture:

He explained [1] that Machine learning could be destructive also if not applied by proper means, some destructive applications are:

* Car Crash scenario.
* Poisoning a robot data
* Incorrect prediction of stock market data

1. **TEAM CONCLUSION**

Machine Learning can help us to build a better world, there are many things which humans are unable to do or sometimes afraid to do so. But machine can possibly do it in future and make work easier.

Today machine learning is evolving really fast but still there are some corners which are yet to explore. Without them an adversary can easily attack our system, make it vulnerable and this would only lead us to a destructive phase.

**REFERENCES:**

1. *Online Youtube video by Mr. Evan Wright:* [*https://www.youtube.com/watch?v=3\_fCL1ETcSQ*](https://www.youtube.com/watch?v=3_fCL1ETcSQ)