Machine learning(ML) is like fire, it can be used to either burn us or give us warmth. Adversial 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 can easily fool ML algorithms like SVM, Deep Learning. Another crucial factor is untrusted training data which affect the prediction of ML algorithms. ML community assumes training data to be 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 models fail to detect that, but human can.

Adversial machine learning is a certain kind of overfitting. Evasion attack, and poisoning are two types of adversial machine learning attacks where the former one is the most prevalent type of attack and the later one can be regarded as an adversial contamination of the training data. Hidden data channels which is also applicable to cyber security domain.

There are a few protection measures that can be taken such as human intervention along with machine learning model, software security and use of adversial 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 adversial machine learning which is a future concern.