Title: Robust Machine Learning Building Model Resilient to Adversarial Attacks

Objective-

The objective of this blog is to explore the concept of machine learning and its importance in building a model that can withstand adversarial attacks. This discusses, various techniques and approaches used to enhance the resilience of machine learning models against adversarial threats.

Contents-

Introduction to Adversarial Attack

What is an Adversarial Attack?

Adversarial machine learning studies the attacks on machine learning algorithms and the defenses against such attacks. A survey from May 2020 exposes the fact that practitioners report a dire need for better-protecting machine learning systems in industrial applications. Wikipedia

While adversarial machine learning can be used in a variety of applications, this technique is most commonly used to execute an attack or cause a malfunction in a machine learning system. The same instance of an attack can be changed easily to work on multiple models of different data sets or architectures.

Adversarial learning has applications in areas like spam filtering, virus detection, intrusion detection, fraud detection, biometric authentication, network protocol verification, computational advertising, recommender systems, social media web mining, complex system performance modelling, and so on

How does it work?

They have various motivations and a range of tactics. However, their aim is to negatively impact the model's performance, so it misclassifies data or makes faulty predictions. To do this, attackers either manipulate the system's input data or directly tamper with the model's inner workings.

In the case of manipulated or corrupted input data, an attacker modifies an input -- such as an image or an email -- by introducing perturbances or noise. These modifications are subtle and can fool a model into wrongly concluding that data should be classified in a way that isn't correct or that isn't deemed threatening. Attackers can corrupt a model during training, or they can target a pre-trained model that's already deployed.

Why do they pose a threat to machine learning Models?

When attackers target an unsecured model, they can access and alter its architecture and parameters so that it no longer works as it should.

This type of attack can fool the model from generating well and from learning high-level representation which can impact the performance of the model.

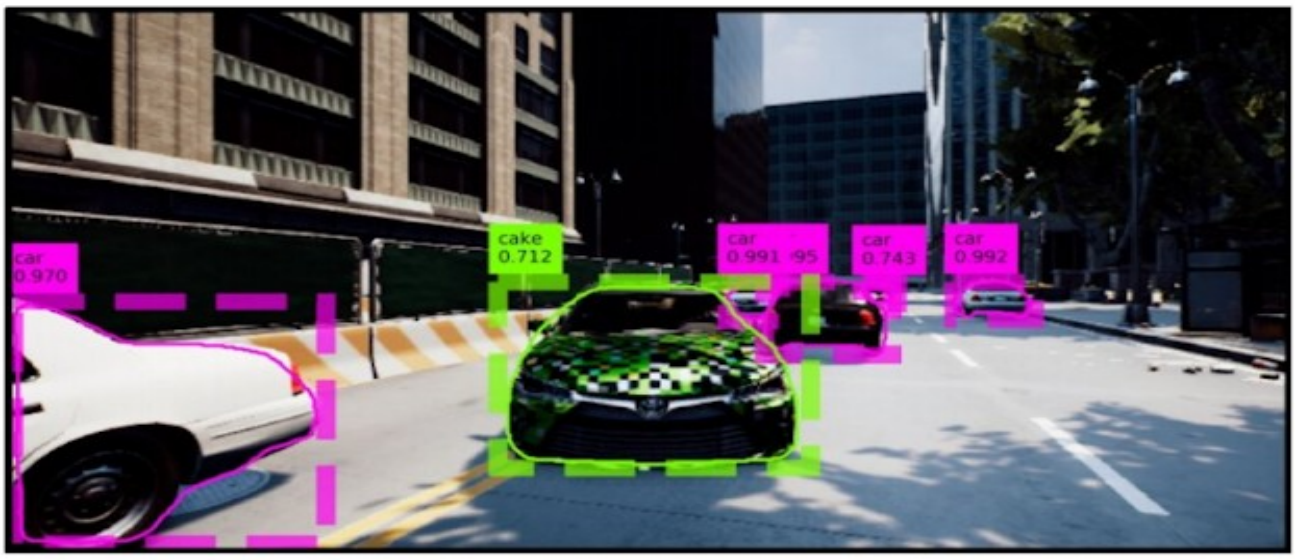
Examples of adversarial machine learning-

Specific adversarial examples often won't confuse humans but will confuse ML models. While a person can read what a sign says or see what a picture depicts, computers can be fooled. Hypothetical examples of adversarial attacks include the following:

An image is fed to an ML model as its input, but attackers tamper with the input data connected to the image, introducing noise. As a result, an image of a lion gets misclassified as an elephant. This type of image classification attack is also known as an *evasion attack* because instead of directly tampering with

training data or using another blatant approach, it relies on subtle modifications to inputs that are designed to evade detection.

With machine learning rapidly becoming core to organizations' value proposition, the need for organizations to protect them is growing fast. Hence, Adversarial Machine Learning is becoming an important field in the software industry. Google, Microsoft, and IBM have started to invest in securing machine learning systems. In recent years, companies are heavily investing in machine learning themselves – Google, Amazon, Microsoft, and Tesla – faced some degree of adversarial attacks.



The car with a camouflage pattern is mis-detected as a "cake" – Source

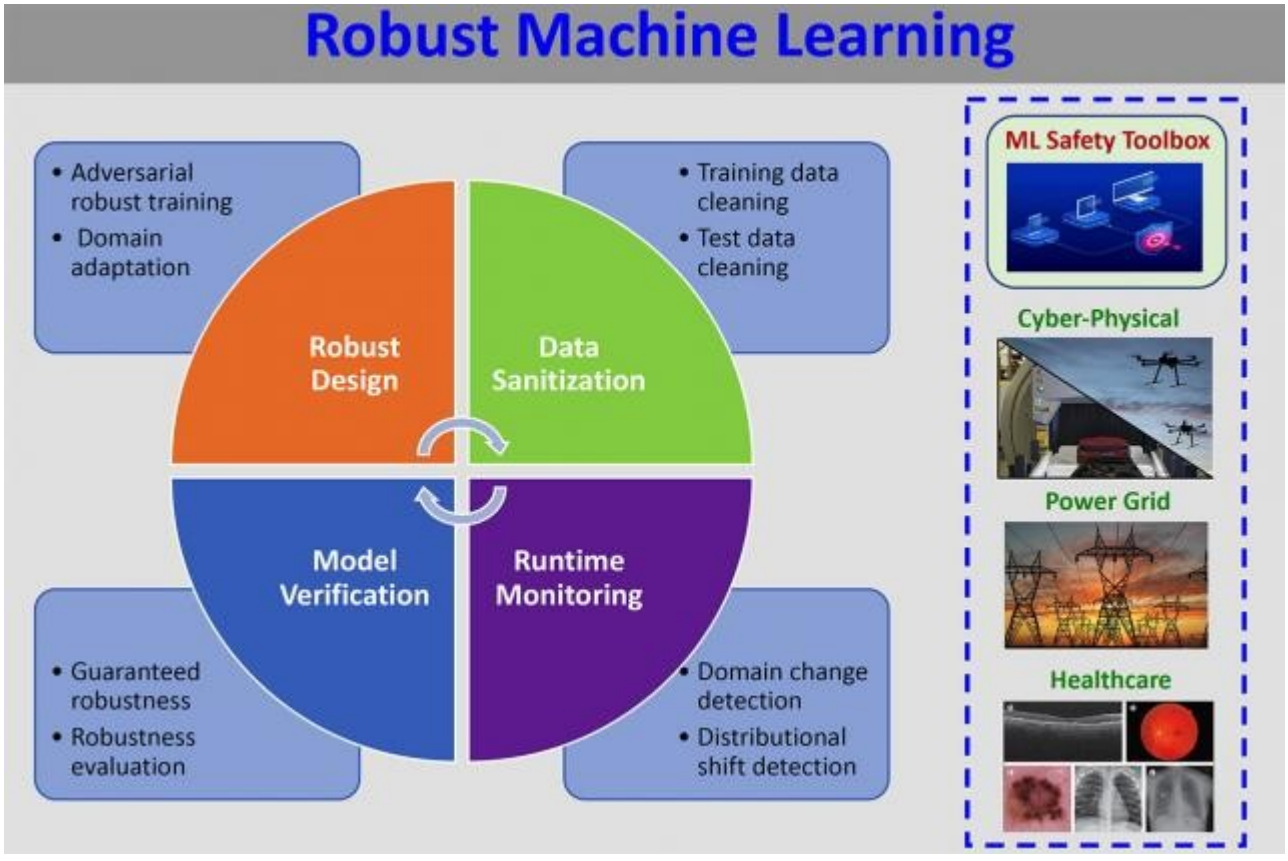
Understanding Robust Machine learning

Machine learning robustness refers to the ability of a model to maintain its performance when faced with uncertainties or adversarial conditions. This includes handling noisy data, distribution shifts, and adversarial attacks, among other challenges. A robust model should be able to generalize well and provide reliable predictions even when dealing with unforeseen inputs or circumstances.

Significance of Robust Machine Learning model-

The real-world consequences of non-robust models can be severe, ranging from financial losses to compromised safety. For instance, an autonomous vehicle that relies on a non-robust image recognition system could misinterpret road signs or fail to detect obstacles, leading to accidents. Similarly, a non-robust fraud detection system might result in false positives or negatives, causing significant financial losses for businesses and consumers.

As machine learning becomes increasingly embedded in our daily lives, the importance of robust models grows. In addition to ensuring accurate predictions, robust models can contribute to enhanced security, privacy, and user trust in AI systems.



Common types of Adversarial Attacks

There is a large variety of different adversarial attacks that can be used against machine learning systems. Many of these work on deep learning systems and traditional machine learning models such as Support Vector Machines (SVMs) and linear regression.

Most adversarial attacks usually aim to deteriorate the performance of classifiers on specific tasks, essentially to “fool” the machine learning algorithm.

Adversarial machine learning is the field that studies a class of attacks that aims to deteriorate the performance of classifiers on specific tasks.

Adversarial attacks can be mainly classified into the following categories:
1. Evasion Attacks
2. Model Extraction Attacks
3. Poisoning Attacks

Evasion Attacks-

Evasion attacks are the most prevalent and most researched types of attacks. The attacker manipulates the data during deployment to deceive previously trained classifiers. Since they are performed during the deployment phase, they are the most practical types of attacks and the most used attacks on intrusion and malware scenarios.

The attackers often attempt to evade detection by obfuscating the content of malware or spam emails. Therefore, samples are modified to evade detection as they are classified as legitimate without directly impacting the training data.

Examples of evasion are spoofing attacks against biometric verification systems.

Model Extraction

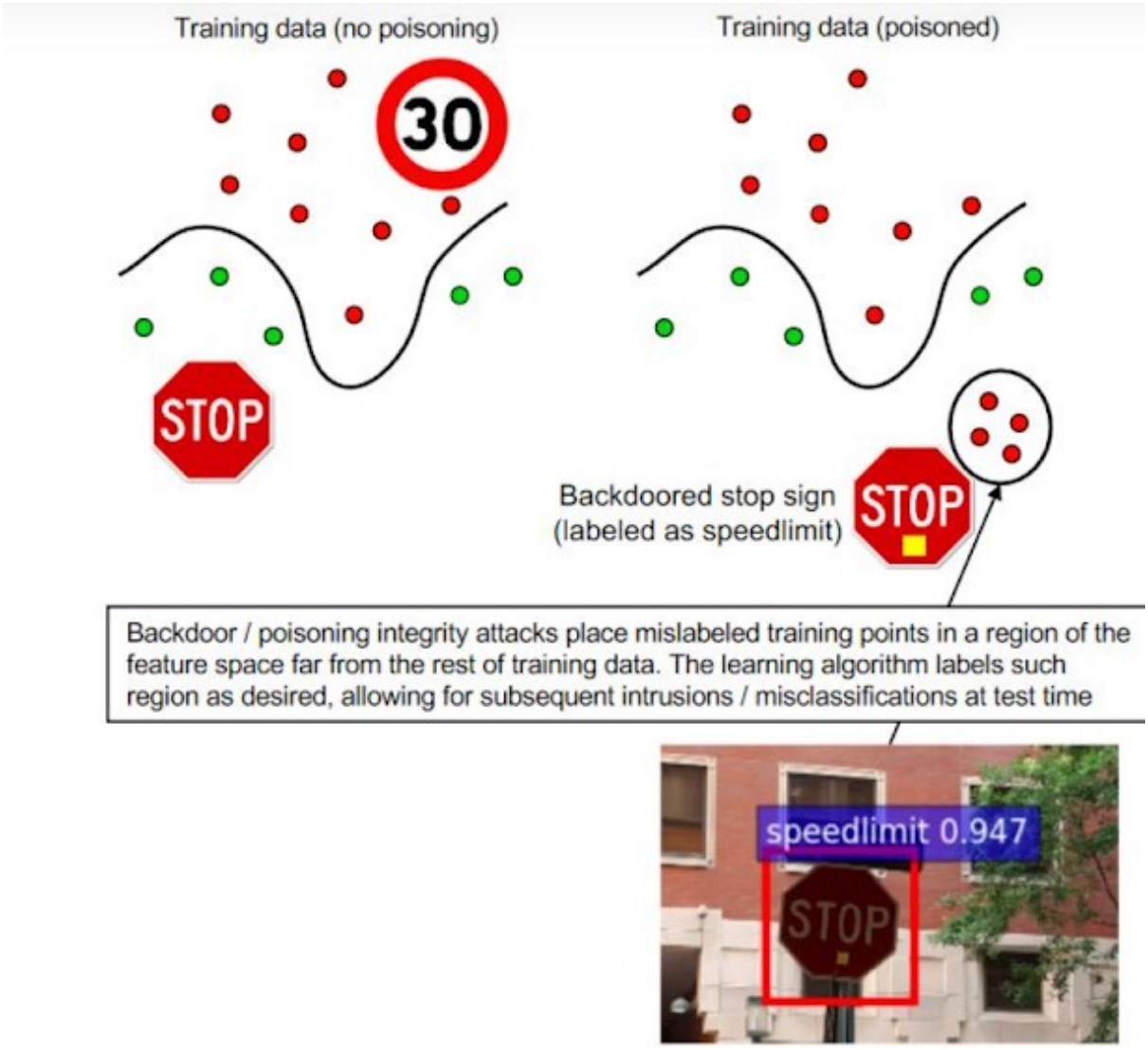
Model Extraction Model stealing or model extraction involves an attacker probing a black box machine learning system in order to either reconstruct the model or extract the data it was trained on. This is especially significant when either the training data or the model itself is sensitive and confidential.

Model extraction attacks can be used, for instance, to steal a stock market prediction model, which the adversary could use for their own financial benefit.

Poisoning Attacks-

The attacker influences the training data or its labels to cause the model to underperform during deployment. Hence, Poisoning is essentially adversarial contamination of training data.

As ML systems can be re-trained using data collected during operation, an attacker may poison the data by injecting malicious samples during operation, which subsequently disrupt or influence re-training.


Training data (no poisoning)
30
STOP
Training data (poisoned)
Backdoored stop sign
(labeled as speedlimit)
STOP
Backdoor / poisoning integrity attacks place mislabeled training points in a region of the feature space far from the rest of training data. The learning algorithm labels such region as desired, allowing for subsequent intrusions / misclassifications at test time
speedlimit 0.947
STOP

Techniques for Building Robust Models-

Throughout this blog series, we will delve into various strategies for enhancing the robustness of machine learning models, including:

1. Data augmentation and preprocessing: By augmenting and preprocessing the data, we can improve the model's ability to handle noisy inputs and generalize to new data. Techniques such as data cleaning, normalization, and various augmentation methods can help create a more diverse and robust dataset for training.

2. Regularization techniques: Regularization methods, such as L1 and L2 regularization, dropout, and early stopping, can help prevent overfitting and improve model robustness. These techniques add constraints to the model training process to encourage simpler models that generalize better.

3. Ensemble learning and model diversity: Combining multiple models with different strengths and weaknesses can lead to a more robust overall system. Ensemble learning techniques, such as bagging, boosting, and stacking, leverage the power of diverse models to create a stronger, more robust predictor.

4. Transfer learning and domain adaptation: Transfer learning allows a model trained on one task to be fine-tuned for a related task, often with fewer training examples. Domain adaptation techniques enable models to adapt to distribution shifts, making them more robust to changes in the data landscape.

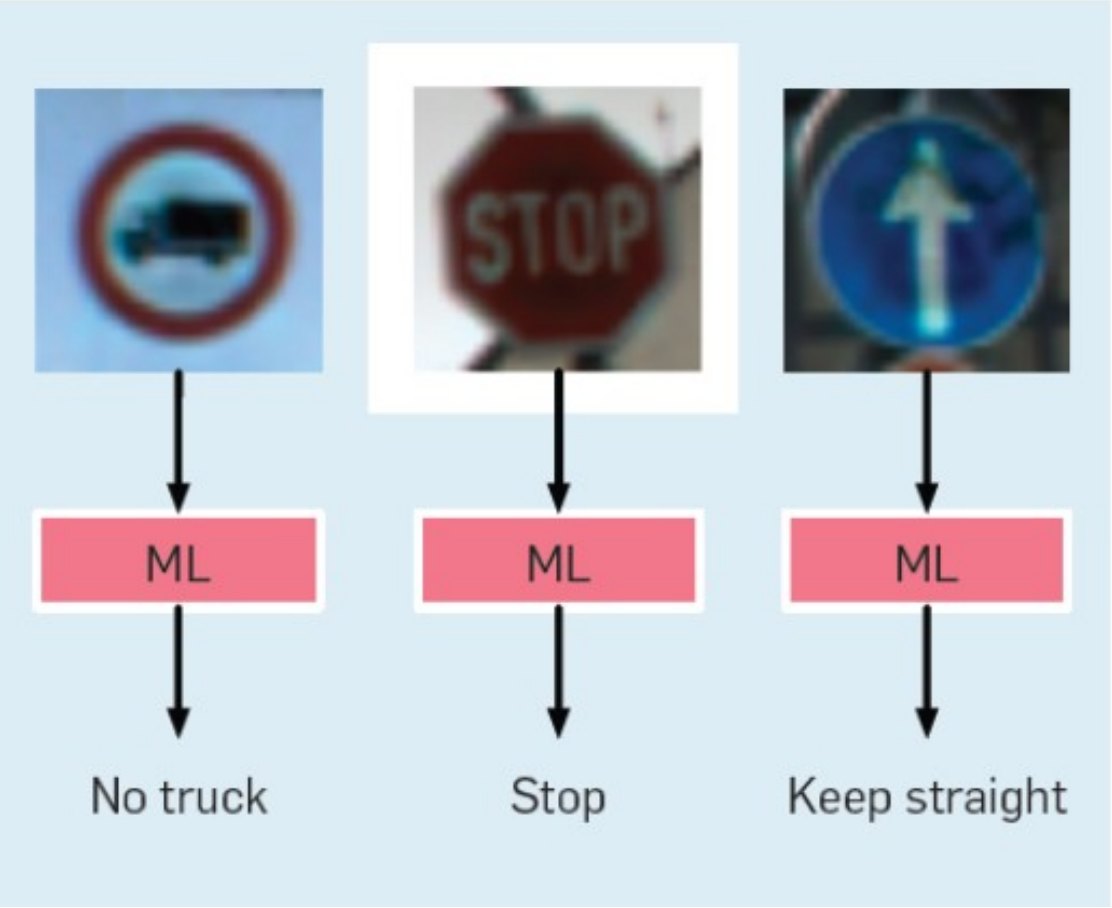
5. Interpretability and explainability: Developing models that are interpretable and explainable can help identify potential weaknesses and vulnerabilities, enabling us to build more robust systems. Techniques for understanding and explaining model decisions can also contribute to increased user trust in AI systems.

6. Robustness metrics and evaluation: To ensure that our models are truly robust, we need to measure their performance using appropriate evaluation metrics. Traditional performance metrics may not always capture the nuances of robustness, so we must also consider robustness-specific evaluation techniques and benchmark datasets.

Case Studies and Examples

Objective-

The objective of this is focused on machine learning algorithms that perform “classification,” learning a mapping from an input to a discrete variable that resents the identity of a class. As a unifying example, we discuss road-sign image recognition; the different values of *output* correspond to different types of road signs (such as stop signs, yield signs, and speed limit signs). Examples of input images and expected outputs are shown in Figure



A machine learning algorithm is expected to produce a model capable of predicting the correct class of a given input. For instance, when presented with an image of a STOP sign, the model should output the

class designating “STOP" and the same goes with others. However, by introducing imperceptible perturbations to input data, it cause misclassification. These attacks generate adversarial example by introducing small, precise perturbations to clean images, aimed at causing misclassification.

In this, we will explore strategies to develop a machine-learning model which will be resilient to adversarial attacks.

1. Dataset and Model Selection:

We need to use diverse dataset of road-sign images, containing classes such as stop signs, yield signs, and speed limit signs.

2. Input validation and preprocessing:

Perhaps the most obvious defense is to validate the input before it is given to the model and possibly preprocess it to remove potentially adversarial perturbations. In many application domains there are verifiable properties of inputs that should never be violated in practice. For example, an input image from a camera sensor can be checked for realism; for example, certain properties of cameras and light ensure certain pixel neighborhoods (such as neighbor pixels with exceptionally high contrast) never occur. Such defenses are limited in that they are highly domain dependent and subject to environmental factors. Moreover, it is not clear that the constraints placed on the domain just increase the difficulty of adversarial sample generation or provide broad protections. It is highly likely that the effectiveness of input constraints as a countermeasure is also domain specific.

3. Training Phase:

CNN models are trained using standard optimization techniques to minimize classification errors. Model architectures, including ResNet and VGG, are considered for their robust feature extraction capabilities across various road signs.

4. Adversarial Attack Simulation:

To evaluate model robustness, adversarial attacks are simulated using methods such as -

Limited-memory BFGS (L-BFGS).
The Limited-memory Broyden-Fletcher-Goldfarb-Shanno (L-BFGS) method is a non-linear gradient-based numerical optimization algorithm to minimize the number of perturbations added to images. Advantages: Effective at generating adversarial examples. Disadvantages: Very computationally intensive, as it is an optimized method with box constraints. The method is time-consuming and impractical.

FastGradient Sign method (FGSM)
A simple and fast gradient-based method is used to generate adversarial examples to minimize the maximum amount of perturbation added to any pixel of the image to cause misclassification. Advantages: Comparably efficient computing times. Disadvantages: Perturbations are added to every feature.

Jacobian-based Saliency Map Attack (JSMA)
Unlike FGSM, the method uses feature selection to minimize the number of features modified while causing misclassification. Flat perturbations are added to features iteratively according to saliency value by decreasing order. Advantages: Very few features are perturbed. Disadvantages: More computationally intensive than FGSM.

Deepfool Attack

This untargeted adversarial sample generation technique aims at minimizing the euclidean distance between perturbed samples and original samples. Decision boundaries between classes are estimated, and perturbations are added iteratively. Advantages: Effective at producing adversarial examples, with fewer perturbations and higher misclassification rates. Disadvantages: More computationally intensive than FGSM and JSMA. Also, adversarial examples are likely not optimal.

Conclusion-

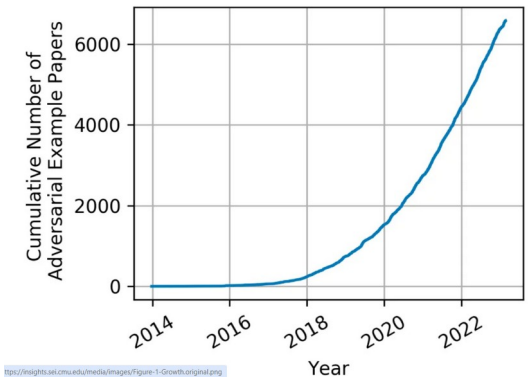
As ML systems have been dramatically integrated into a broad range of decision-making-sensitive applications for the past years, adversarial attacks and data poisoning attacks have posed a considerable threat against these systems. For this reason, we should focuses on ML Security. In this blog we have discussed how me robust system is import in dealing with attacks and strategies that we can follow to overcome all these attacks

Challenge and Future Directions

The concept of adversarial machine learning has been around for a long time, but the term has only recently come into use. With the explosive growth of ML and artificial intelligence (AI), adversarial tactics, techniques, and procedures have generated a lot of interest and have grown significantly.

Adversarial machine learning is at a turning point. In the context of adversarial inputs at test time, we have several effective attack algorithms but few strong countermeasures.
One thing is clear it will continue evolving as technology as AI does.

Next, we should ensure our evaluations are adaptive. In particular, every evaluation should build upon prior evaluations but also be independent and represent a motivated adversary



References:

-By **Cameron Hashemi-Pour,** Site Editor-**Alexander S. Gillis,** Technical Writer and Editor
https://www.techtarget.com/searchenterpriseai/definition/adversarial-machine-learning

-Adversarial Machine Learning- Attack Surfaces, Defence Mechanisms,
Learning Theories in Artificial Intelligence- Aneesh Sreevallabh Chivukula • Xinghao Yang • Bo Liu • Wei Liu • Wanlei Zhou

-Understanding Machine Learning Robustness: Why It Matters and How It Affects Your Models- Viacheslav Dubrov- **https://medium.com/@slavadubrov/understanding-machine-learning-robustness-why-it-matters-and-how-it-affects-your-models-5e2cb5838dab**

-https://viso.ai/deep-learning/adversarial-machine-learning/

-By Ian Goodfellow, Patrick McDaniel, and Nicolas Papernot
https://cacm.acm.org/research/making-machine-learning-robust-against-adversarial-inputs/

- https://data-science.llnl.gov/latest/news/neurips-papers-aim-improve-understanding-and-robustness-machine-learning-algorithms

-https://insights.sei.cmu.edu/blog/the-challenge-of-adversarial-machine-learning/