# Team 3.1
# ADVERSARIAL ATTACKS AND DEFENSES

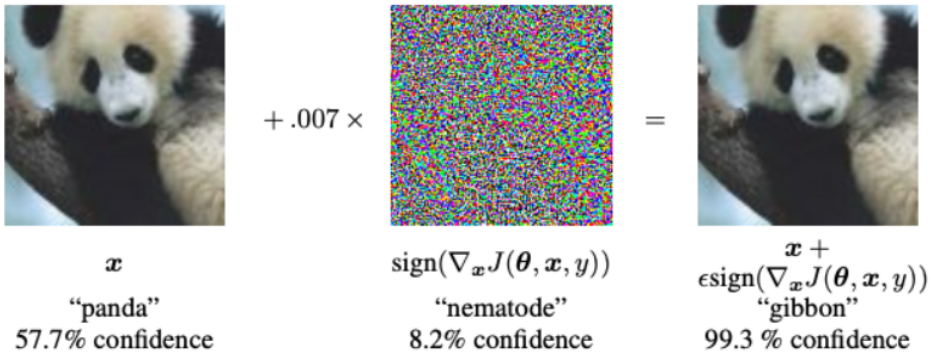
## ABSTRACT

## Introduction-

Artificial intelligence systems are now a crucial component of many applications, including image identification, natural language processing, and autonomous cars, in a time when machine learning models are rapidly proliferating. But this development in AI technology has also given rise to a fresh and urgent issue: adversarial attacks. These attacks entail gradually altering input data to trick machine learning algorithms, which may result in inaccurate or even hazardous outputs.
Due of AI systems' susceptibility to adversarial attacks, a growing body of research is being done to provide defence mechanisms to protect their integrity.

### What is an Adversarial Attack?

There are several types of such attacks, however, here the focus is on the fast gradient sign method attack, which is a *white box* attack whose goal is to ensure misclassification. A white box attack is where the attacker has complete access to the model being attacked. One of the most famous examples of an adversarial image shown below is taken from the aforementioned paper.




**This abstract provides an overview of the evolving landscape of defense mechanisms against adversarial attacks, highlighting key strategies and approaches:**

**Hostile Detection**: A crucial component of defence is the identification of hostile examples. Anomaly detection, consistency checks, and statistical analysis are some of

the techniques that can be used to find inputs that differ from the typical distribution of data.

**Gradient Masking and Obfuscation:** Techniques such as gradient masking and obfuscation can make it more challenging for adversaries to compute gradients for crafting adversarial examples, thereby making the attack more difficult

**Adversarial Robustness: -** Another crucial component of defence is increasing the robustness of machine learning models against adversarial attacks. To increase model resilience, some techniques are utilised, including adversarial training, adversarial retraining, and robust model architectures.

Combining the analytical skills of human experts with AI systems can provide an effective line of defence by spotting and responding to hostile attacks in real-time. This technique is known as "human-in-the-loop defence."

We plan to create a comprehensive open-source library for generating and testing adversarial attacks on different machine learning models. This project could include implementing various attack algorithms and benchmarking their success rates against different models.

**Team members-**

Shaunak Tanawade
Kushank Jain
Vanshika Jain
Raunak Jain