Trustworthy AI: Robustness

Destiny Pounds

M.S. Student, Biomedical Data Science

Advised by Vibhuti Gupta, Ph.D.

Assistant Professor, Computer Science and Data Science

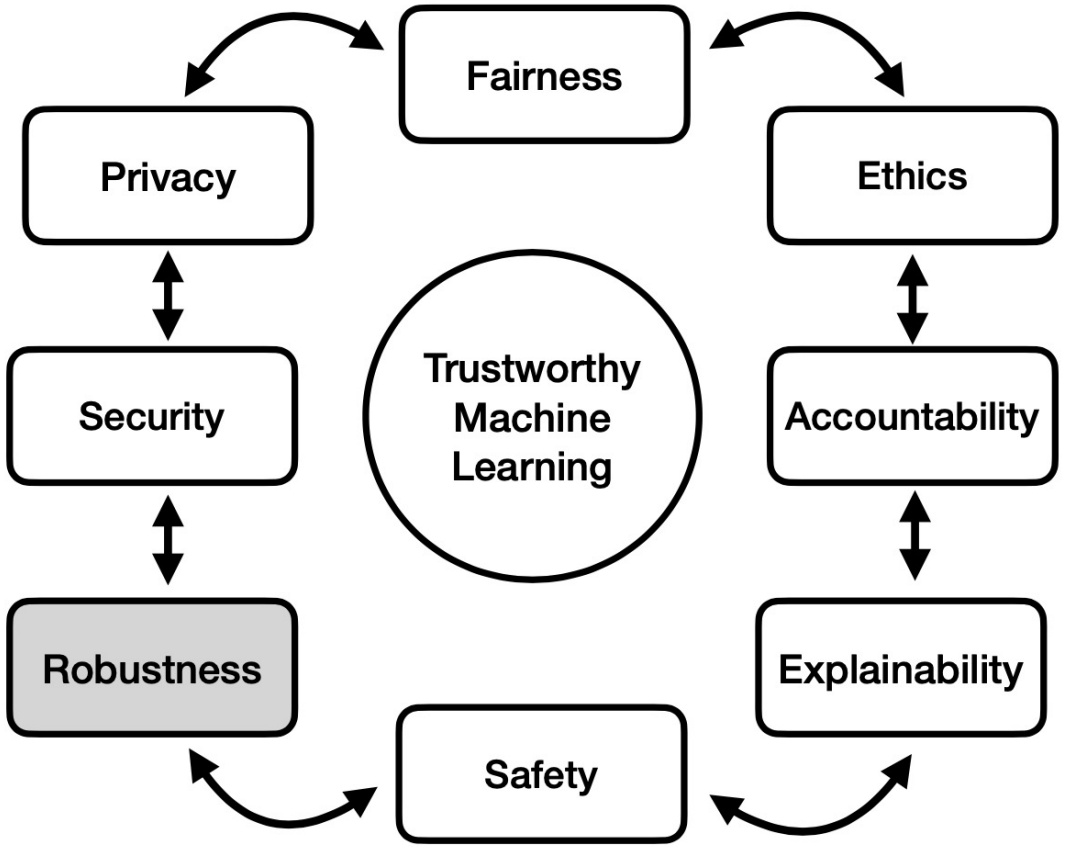
School of Applied Computational Sciences

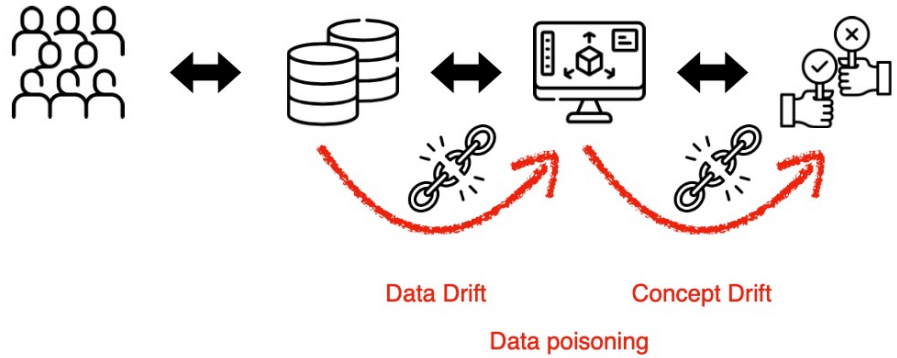
Meharry Medical College

Overview

- Robustness
- Adversarial Learning
- Potential Attacks
- Importance of Adversarial Learning

Robustness



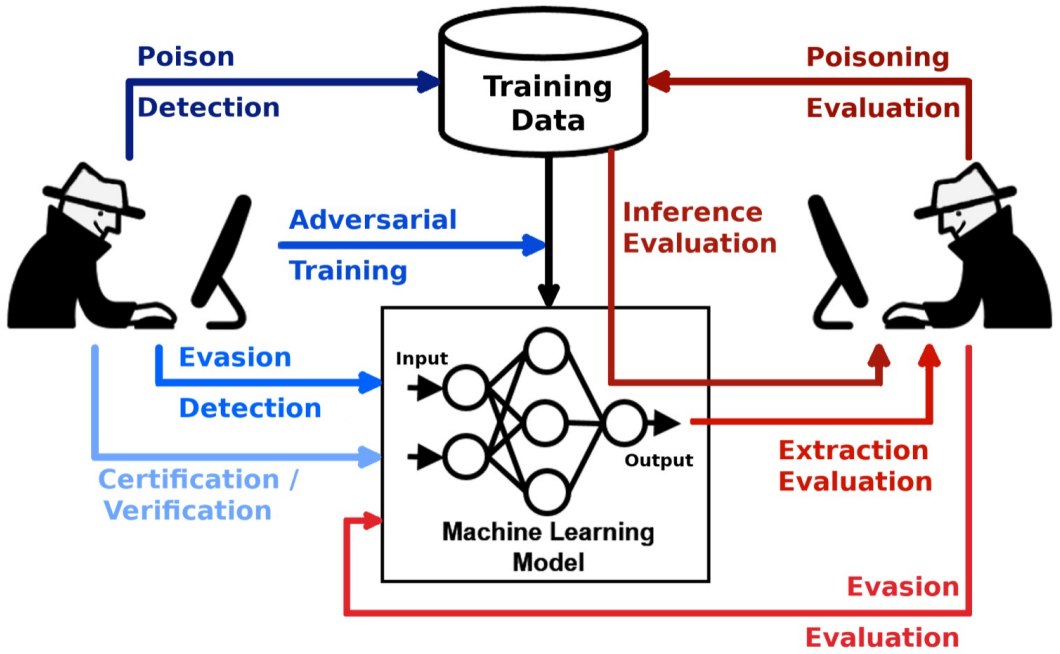


- Robustness is the property that characterizes how effective an algorithm is while being tested on a new independent dataset.

Adapted from Farnadi, G. (2022). *Trustworthy Machine Learning* [Slides]. HEC Montréal

Adversarial Learning

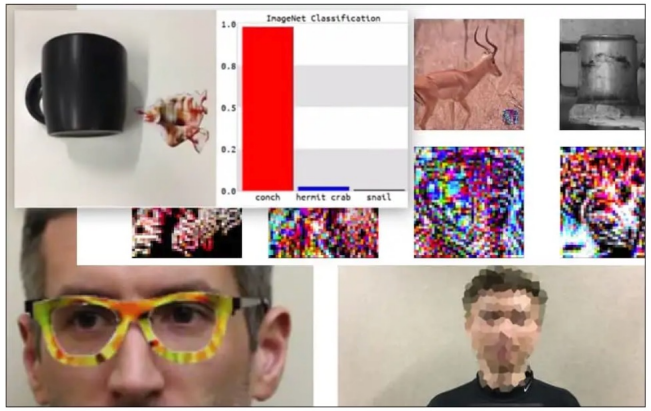
Adversarial learning involves training models to be robust against adversarial examples. These examples are intentionally designed inputs created to mislead the model into making inaccurate and wrong predictions.



CYBERSECURITY

Why Adversarial Image Attacks Are No Joke

Updated on December 1, 2021
By **Martin Anderson**



Attacking image recognition systems with carefully-crafted adversarial images has been considered an amusing but trivial proof-of-concept over the last five years. However, new research from Australia suggests that the casual use of highly popular image datasets for commercial AI projects could create an enduring new security problem.

Physical adversarial example from CVPR 2018 paper

Adapted from Farnadi, G. (2022). *Trustworthy Machine Learning* [Slides]. HEC Montréal

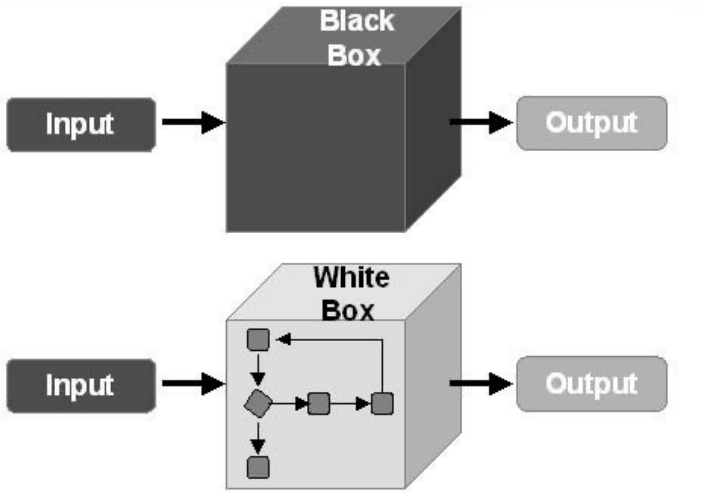
Adversarial Attacks

White Box Attacks

In a white box attack, the attacker has complete knowledge of the model and its inner workings.

Black Box Attacks

In a black box model, the attacker has limited to no knowledge of the model's internal details.



Why is adversarial learning important?

Adversarial learning improves robustness by helping the model generalize better. During training, the model is exposed to a wide range of adversarial examples which the model must adapt to.

It also helps detect weaknesses in the model and provides insights into how the model can be improved.

Incorporating adversarial learning into a machine learning model requires two steps:

1

Generate adversarial examples

2

Incorporate these examples into the training process

Thank You

Please send us your questions at:

vgupta@mmc.edu and

dpounds24@email.mmc.edu