Evaluating the Resiliency of Artificial Intelligence (AI) Systems: An Overview of Adversarial AI

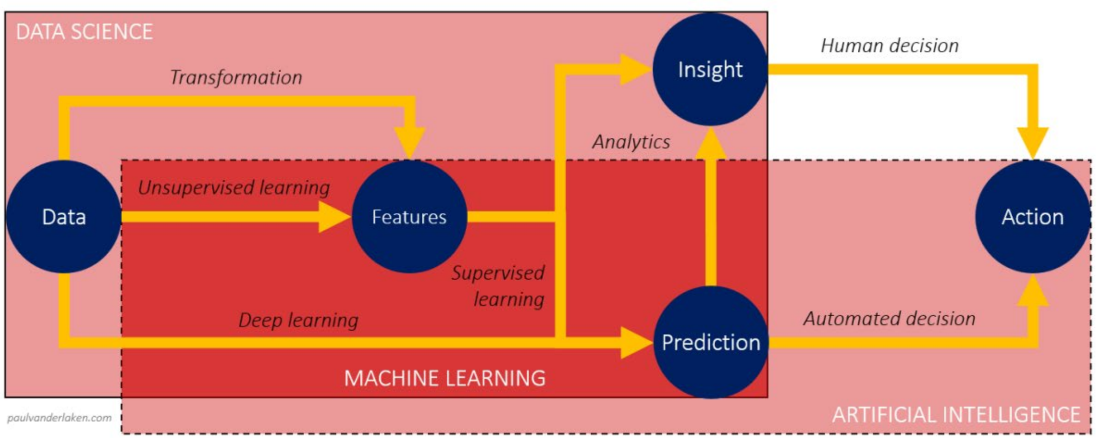
Cybersecurity and Information Systems Information Analysis Center (CSIAC) Webinar

Nate Bastian
Army Cyber Institute, U.S. Military Academy

17 May 2022

Purpose

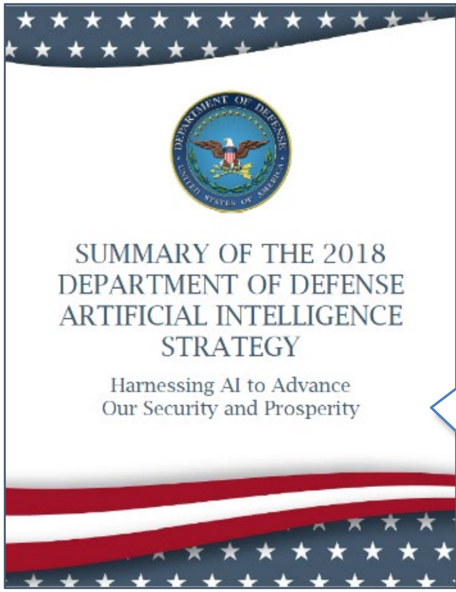
To provide an overview of adversarial artificial intelligence (AI), which encompasses algorithmic and mathematical approaches to degrade, deny, deceive, and/or manipulate AI systems.



Outline

- AI System Resiliency
- Countermeasures and Adversarial AI
- Adversarial AI Access Paradigms
- Adversarial AI Attacks
- System-Level Counter-AI Defense
- Algorithmic Counter-AI Defenses
- Counter-AI Analysis
- Counter-AI Assessment Examples
- Counter-AI Tool
- Adversarial Robustness Toolbox Demo
- Summary
- Q&A

AI System Resiliency


SUMMARY OF THE 2018 DEPARTMENT OF DEFENSE ARTIFICIAL INTELLIGENCE STRATEGY

Harnessing AI to Advance Our Security and Prosperity

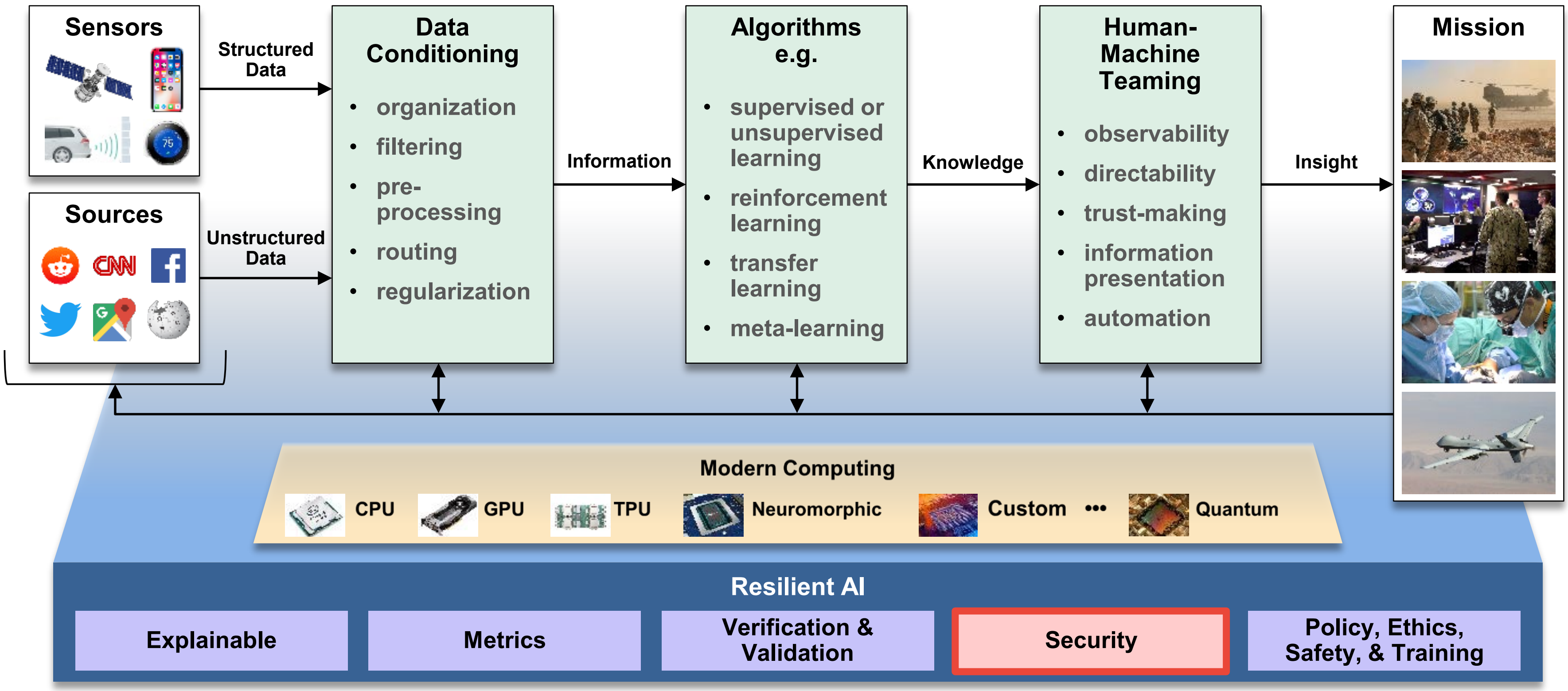
"In order to **ensure Department of Defense AI systems are safe, secure, and robust**, we will fund research into AI systems that have a lower risk of accidents; **are more resilient, including to hacking and adversarial spoofing**; demonstrate less unexpected behavior; and minimize bias…

…we will pioneer and share novel approaches to testing, evaluation, verification, and validation, and we will increase our focus on defensive cybersecurity of hardware and software platforms as a precondition for secure uses of AI.

- **Adversarial AI** – Countermeasures that adversaries may deploy against our AI systems and the evaluation steps and defenses needed to safeguard performance.

AI = Artificial Intelligence

AI System Resiliency

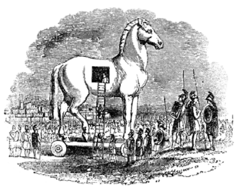


- Modern AI systems can **enhance** end-to-end DoD mission capability.
- In order for AI systems to be integrated into the DoD mission space, it must be shown to be **resilient.**
- **Resilient AI systems** are <u>robust</u> and <u>secured</u> against identified methods of adversarial attack.

AI = Artificial Intelligence
DoD = U.S. Department of Defense

Countermeasures and Adversarial AI

Traditional Human Countermeasures



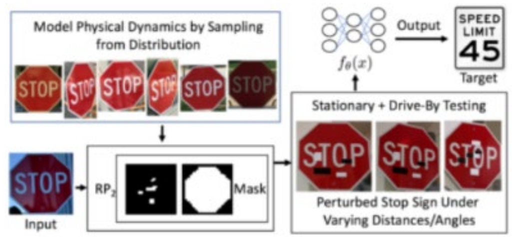
Deceptions (e.g., social engineering, malicious computing) to deliver hidden payloads

Camouflage, disguises, and fabrications to evade detection or distract

Forgery and data manipulation

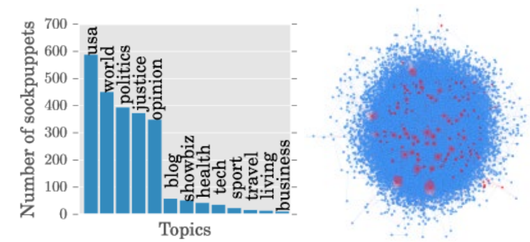
Emerging AI Countermeasures (Adversarial AI)

Engineered Graffiti



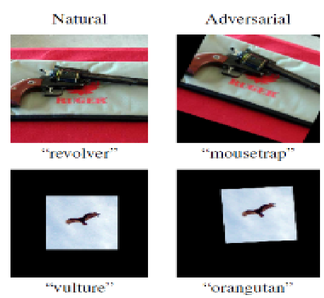
Eykholt et al., 2018

Information Operations



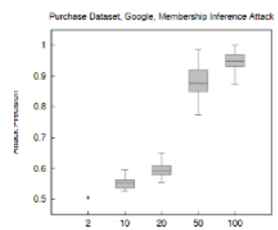
Kumar et al., 2017

Targeted Transformations

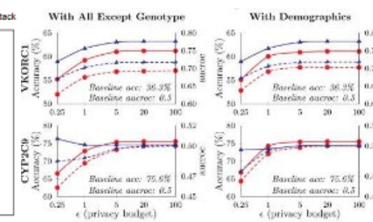


Engstrom et al., 2018

Membership/Attribution from Model



Shokri et al., 2017



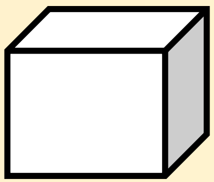
Fredrikson et al., 2014

AI countermeasures have similar goals to traditional countermeasures (e.g., evading detection) but are engineered specifically to defeat AI capabilities.

AI = Artificial Intelligence

Adversarial AI Access Paradigms

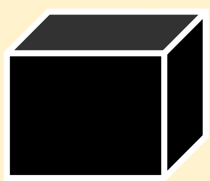
Example Access Methods



"White Box" Paradigm

Adversary has access to model internals (e.g., weights, gradients)

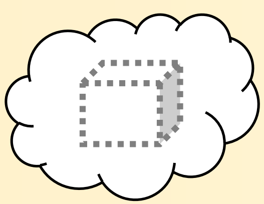
- Adversary determines underlying open-source elements used in model development
- Adversary recovers model details via unauthorized access to code base, code de-compilation, etc.



"Black Box" Paradigm

Adversary able to examine model inputs and outputs, but has no access to internal parameters

- Adversary captures access-limited hardened device with embedded analytics
- Adversary targets remote system with API that permits repeated I/O probing



"Hidden Box" Paradigm

Adversary has no direct access to model, only assumptions about model training or behavior

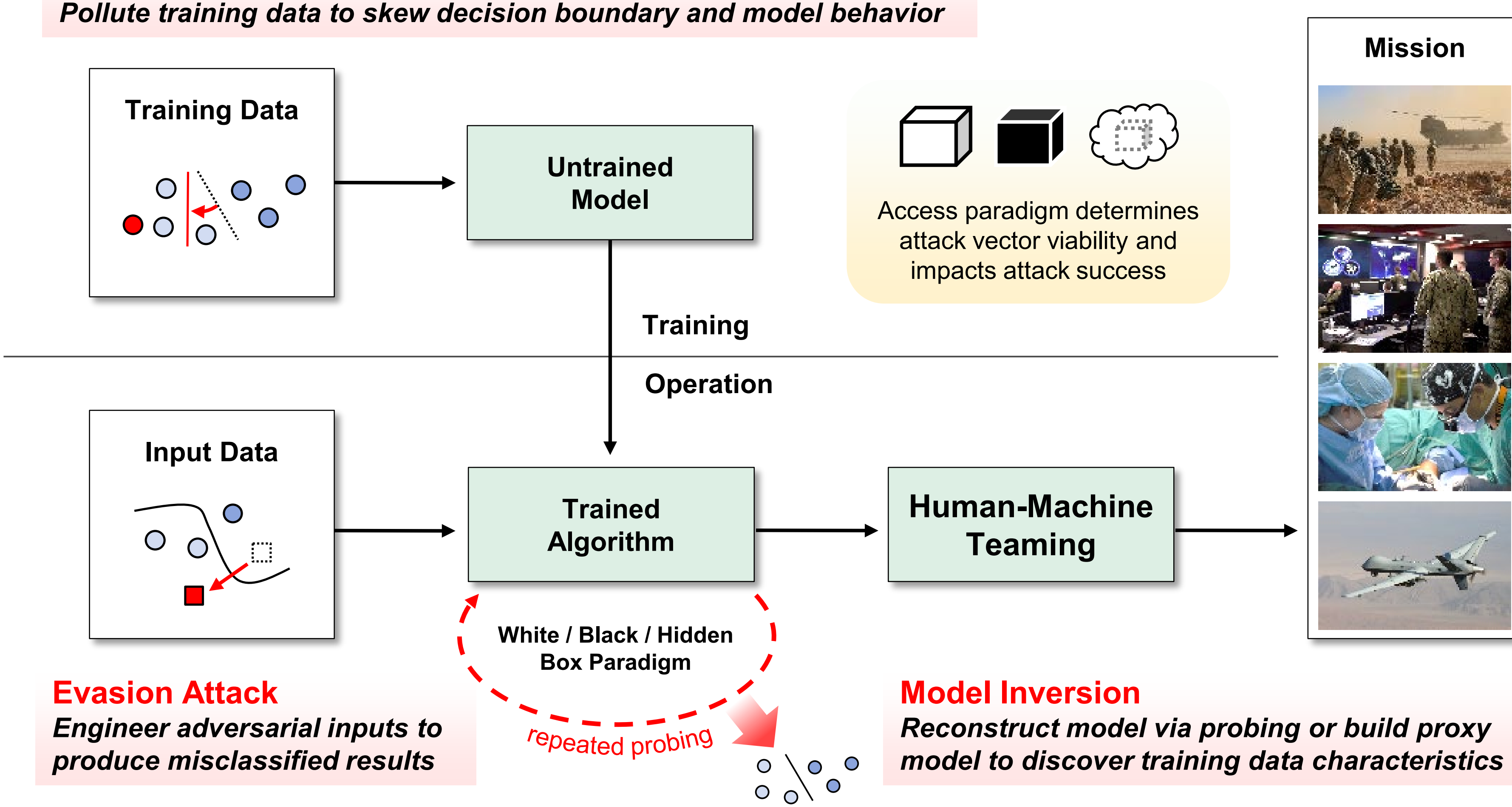
- Adversary predicts blue force surveillance tactics and surmises underlying AI infrastructure

Increasing attacker capability

AI = Artificial Intelligence
API = Application Programming Interface
I/O = Input / Output

Adversarial AI Attacks

Poisoning Attack
Pollute training data to skew decision boundary and model behavior



Evasion Attack
Engineer adversarial inputs to produce misclassified results

Model Inversion
Reconstruct model via probing or build proxy model to discover training data characteristics

System-Level Counter-AI Defenses

Adversarial Attack Class

Defense

Poisoning	Evasion	Model Inversion
• Sensible data sampling • Comparison to previously trained classifiers • Dark launching • Backtesting • Golden dataset • Feedback authentication • Source attribution	• Prevent information leakage • Limit probing • Ensemble learning • Adversarial training • Adversarial AI incident response plan • Input conditioning • Anomaly detection	• Differential privacy (privacy budget) • Private learning (PATE) • Incorporation of randomized response data

- Adversarial attacks are well within the capability of a **well-resourced adversary** to mount.
- If they do think about resiliency, most AI developers think about robustness to expected input, not resiliency to adversarial data, input, or probing (security).
- As AI adoption grows, adversarial AI will have major implications for human-machine teaming, system security, response processes, and data privacy.

Effective defense will require integration of counter-AI techniques with underlying AI algorithms as well as system-level monitoring of AI status.

AI = Artificial Intelligence

Algorithmic Counter-AI Defenses

Technique	Key Idea	Integration Point	Attack Class	Attacker Access
Gradient Hiding[1]	Gradient of model is nontrivial or very hard to determine (e.g., non-differentiable, discontinuous)	Algorithm Architecture	Evasion	White box and Black box
Differential Privacy[2]	Introducing randomization (e.g., noise) during computation reduces ability of attacker to infer training data	Algorithm Architecture	Inversion	White box and black box
Defensive Distillation[1]	Also known as label smoothing or soft label that converts true class labels into soft values	Algorithm Training	Evasion	Black box
Null Labeling[1]	Create "null" class for samples perturbed beyond expected variation	Algorithm Training	Evasion	White box and black box
Data Sanitization[3]	Examine full training dataset and work to remove poisoned points (e.g., deleting outliers)	Algorithm Training / Supply Chain	Poison	White box and Black box
Adversarial Training[1]	Build immunity to adversarially crafted examples by including adversarial examples in training data	Algorithm Training	Poison and Evasion	White box
Integrity Constraints[3]	Leverage a separate domain model (e.g., language model) to enforce training data integrity (or input integrity)	Algorithm Training / System Input	Poison and Evasion	White box and Black box
MagNet[1]	Train detectors that distinguish normal and adversarial examples based on distance from normal example manifold	System Input	Evasion	Black-box
Defense-GAN[1]	Correlate noise input to generator output using real example to bound generator output range and reduce perturbations	System Input	Evasion	White box and black box

[1] Chakraborty et al., “Adversarial Attacks and Defenses: A Survey,” 2018
[2] Abadi et al., “Deep Learning with Differential Privacy,” 2016
[3] Steinhardt et al., “Certified Defenses for Data Poisoning Attacks,” 2017

Counter-AI Analysis

Objective: Analyze impact of black box evasion attacks on face recognition AI system

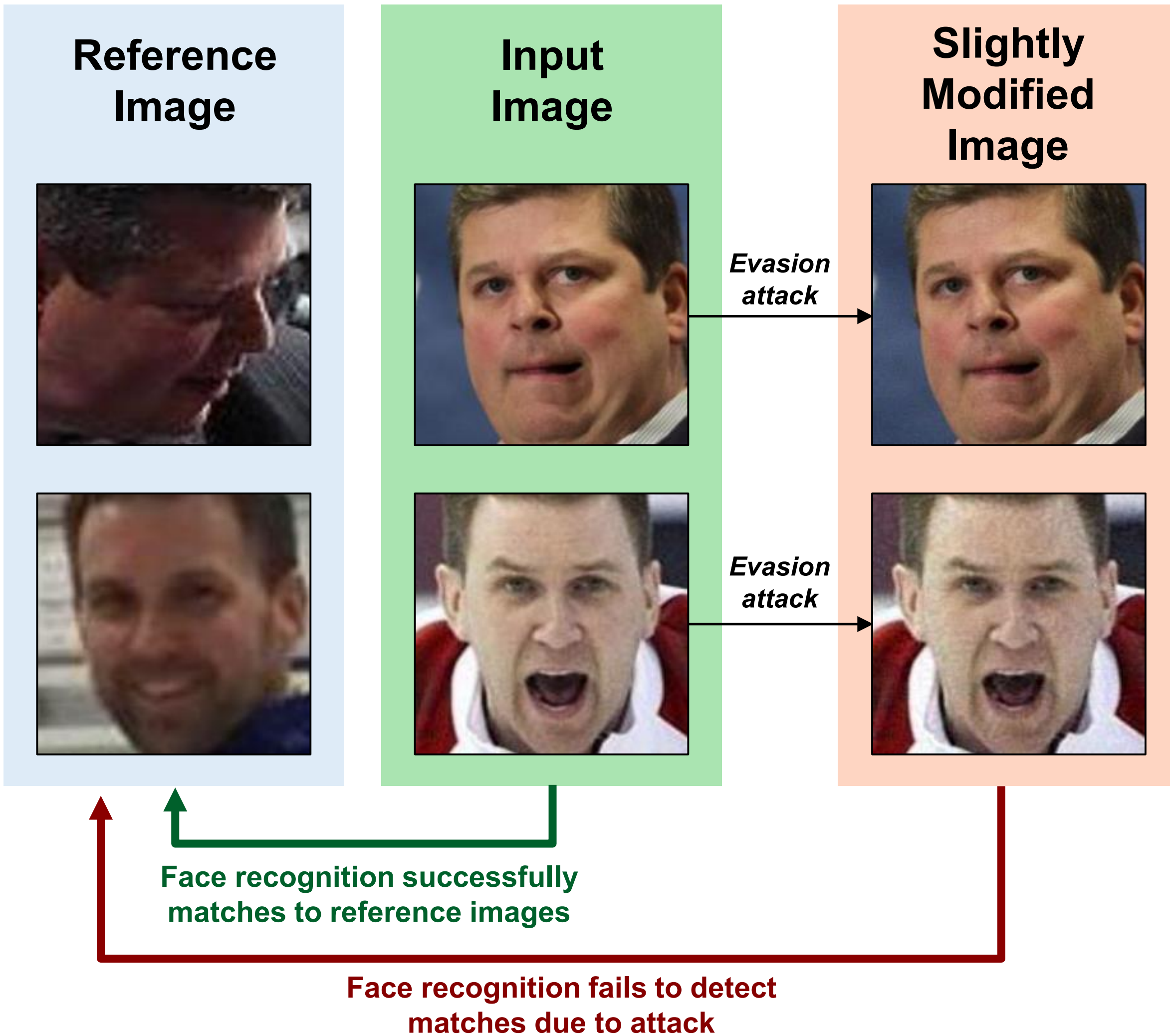
Face recognition successfully matches to reference images

- **Face recognition algorithms have shown substantial improvement because of use of specially designed neural networks.**

AI = Artificial Intelligence

Counter-AI Analysis (cont.)

Objective: Analyze impact of black box evasion attacks on face recognition AI system

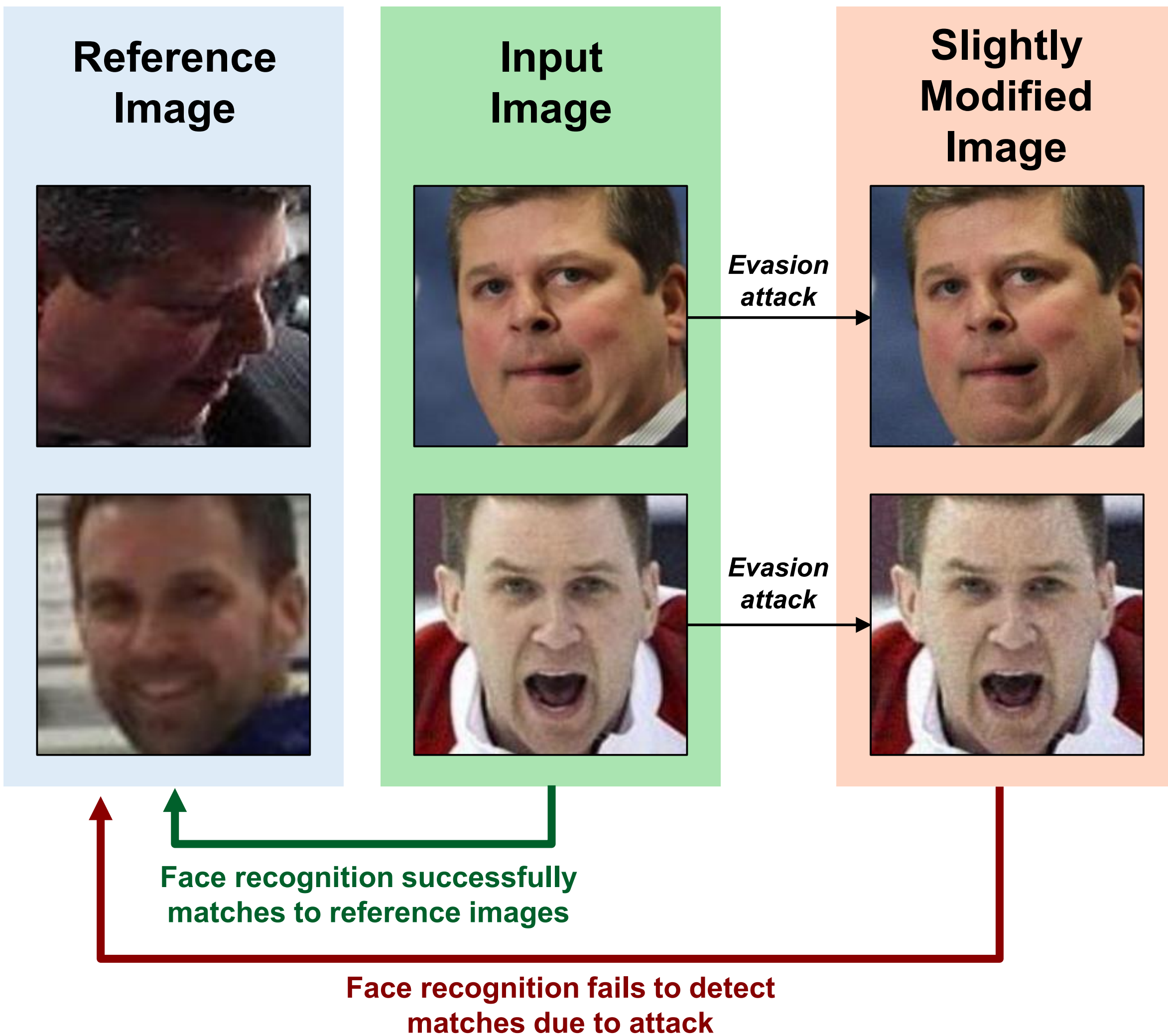


- **Face recognition algorithms have shown substantial improvement because of use of specially designed neural networks.**
- **Like many capabilities based on machine learning, these techniques are vulnerable to adversarial attacks.**

AI = Artificial Intelligence

Counter-AI Analysis (cont.)

Objective: Analyze impact of black box evasion attacks on face recognition AI system

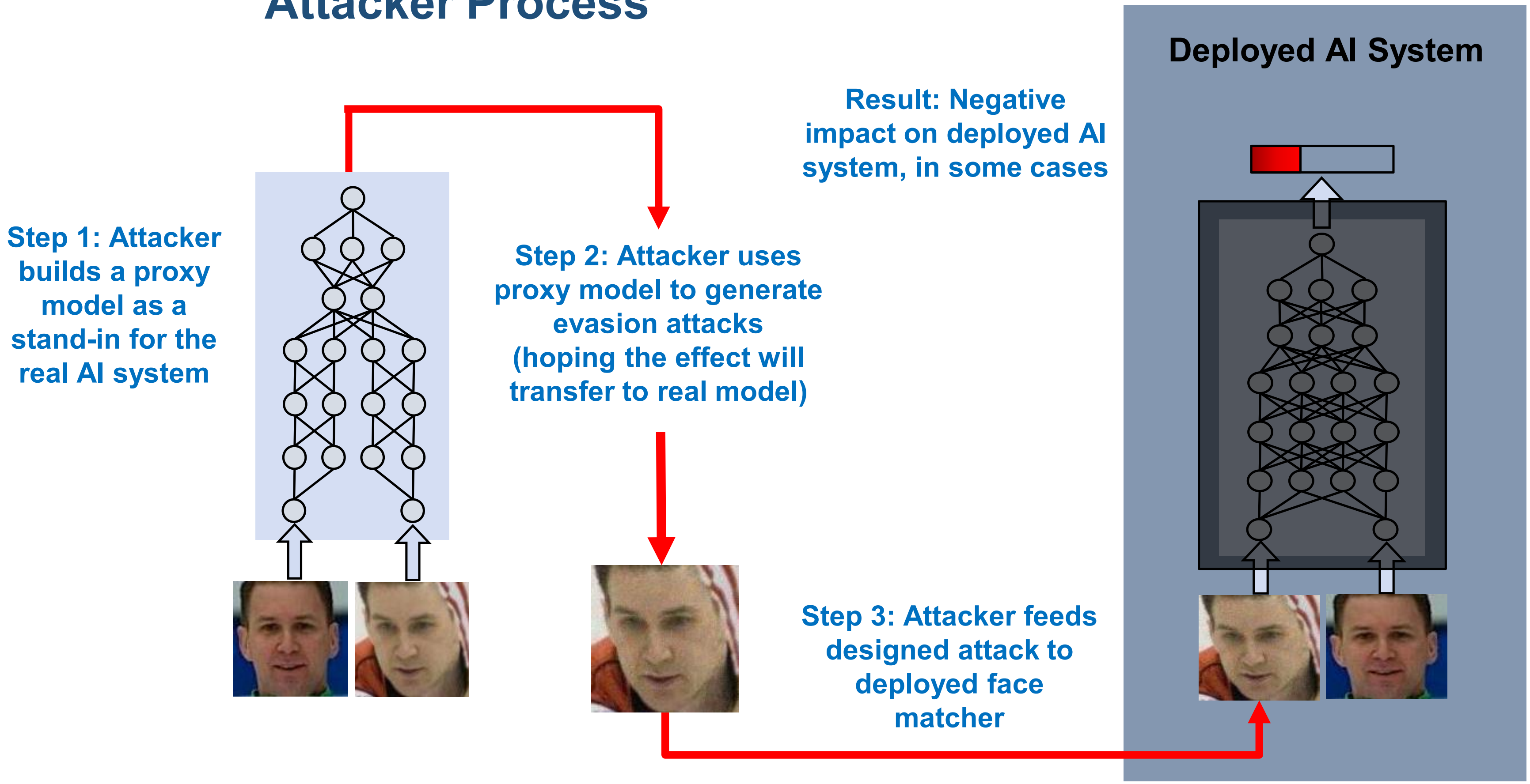


- **Face recognition algorithms have shown substantial improvement because of use of specially designed neural networks.**
- **Like many capabilities based on machine learning, these techniques are vulnerable to adversarial attacks.**
- **Counter-AI analysis can assess how performance of a face recognition AI system degrades when challenged with different attack scenarios.**

AI = Artificial Intelligence

Counter-AI Analysis (cont.)

Attacker Process

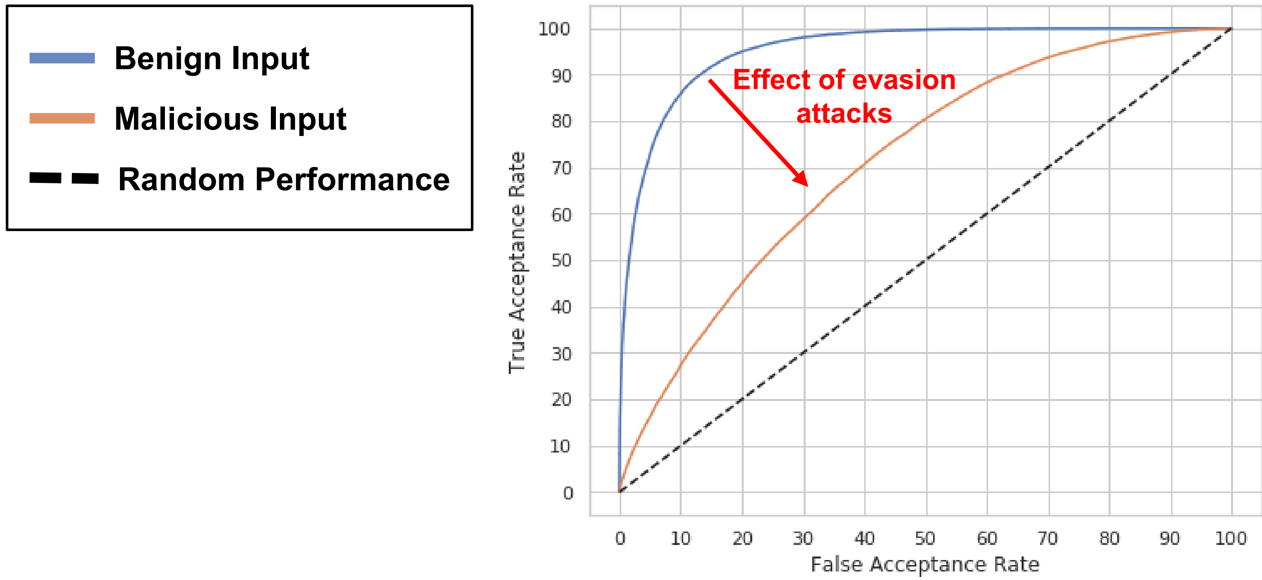


Question: How viable are Black-Box transfer attacks, in which the adversary lacks access to any information about the deployed model?

AI = Artificial Intelligence

Counter-AI Analysis (cont.)

Facial Recognition Performance



Attacks have substantial impact on performance, *even though the proxy model structure and training data are different from the actual model.*

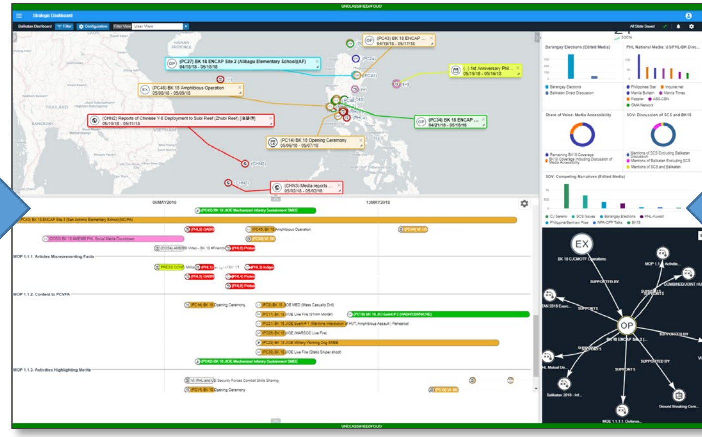
Counter-AI Assessment Details

- **Evaluation uses two distinct neural network architectures trained to predict facial similarity**
  - **ResNet50 architecture treated as deployed model**
  - **DenseNet121 architecture treated as adversary's proxy model**
  - **Both architectures trained on disjoint partitions of VGG Face 2 training set**
- **Attacks generated on DenseNet121 using FGSM and then transferred to ResNet50**
- **Results measured over 100,000 image pairs**

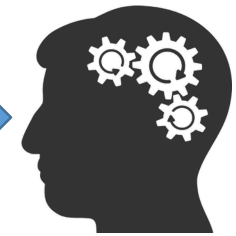
AI = Artificial Intelligence

Counter-AI Assessment Example #1

Situational Awareness of the Information Environment



Example visualization in C2IE



Mission Challenges

- Developing and maintaining situational awareness of information environment
- Determining sentiment and behavior of target populations
- Communicating developed understanding in support

AI Objectives

- Summarization, entity analysis via natural language processing (NLP)
- Ingestion and fusion of large scale collected publically available information
- Modeling diverse target populations and propaganda effect upon them
- Visualization of information environment supporting operator understanding

AI support maintenance of Mission Information Support Operations (MISO) situational awareness, a critical first step in effective operations

AI = Artificial Intelligence

Counter-AI Assessment Example #1

Adversary Objectives

- Confuse Blue Force, hinder ability to develop accurate situational awareness
- Avoid detection and attribution of targeted, malicious information operations activity

Attack Method / Model Access

Model Access	Poison	Evasion	Inversion
White			
Black			
Hidden	●	●	

Counter-AI Attacks

- Text modification via space insertion, character deletion, visual swap, context-aware word swap
- Leverage automation (e.g., sock puppets) to mask nefarious activity

Potential Countermeasures

- Integrity checking of data (e.g., spell checking to remove misspellings), etc.: 74% attack reduction
- Adversarial training on perturbed examples: 83% attack reduction
- Identification and removal of bot-generated data

Observations

- AI role in MISO is as decision-support tool
- Training and operational data external to the algorithm
- Adversarial activity may be hard to distinguish from normal traffic
- Proxy model development may be difficult due to "hidden-box" nature of environment

AI = Artificial Intelligence

Counter-AI Assessment Example #2

Network Data Collection → **Data Stream Identification** → **Alert Generation** → **Alert Ranking** → **Action Suggestion**

Network Data Collection	Data Stream Identification	Alert Generation	Alert Ranking	Action Suggestion
Collect data from cyber sensors: network, host, social media, etc.	**Identify streams of interest within data, direct to AI tools for analysis**	**AI tools perform analysis, generate unfiltered, unranked alerts**	**AI tools rank alerts based on severity and priority**	**Ranked list of alerts is recommended for action by human analyst**

Lines of Effort: Network Mapping | Event Detection | Credential Misuse

Mission Challenges

- Making sense of and detecting malicious events in voluminous, noisy cyber traffic
- Understanding relationship between mission and cyber data – "mission mapping"
- Prioritizing and responding to detected malicious activity

AI Objectives

- Anomaly detection algorithms for structured and unstructured network/host data
- Risk assessment in support of event alert ranking
- Course-of-action suggestion based network posture

AI capabilities support feed-forward cyber sensemaking process

AI = Artificial Intelligence

Counter-AI Assessment Example #2

Adversary Objectives

- Confuse Blue Force, hinder ability to create correct network map
- Avoid detection and attribution of malicious network/host activity

Attack Method

Model Access	Poison	Evasion	Inversion
White			
Black			●
Hidden	●	●	

Adversarial Attacks

- Embed variations in normal network activity used to construct baseline
- Use variations to mask attack anomalies
- Learn (and avoid) high risk alerts

Recommended Defenses

- Data sanitization of baseline traffic to remove attacks
- Adversarial training to enable robust detection
- Leverage differential privacy methods to hide data

Observations

- Cyber sensemaking leverages AI as a decision support tool, with human-in-the-loop
- Cyber data evolves quickly, requiring new collections or data generation
- Proprietary commercial capabilities may be challenging to evaluate
- Proxy model development may be involved due significant data preprocessing of network data
- Cyber attacks enable broad attack considerations

AI = Artificial Intelligence

Counter-AI Assessment Considerations

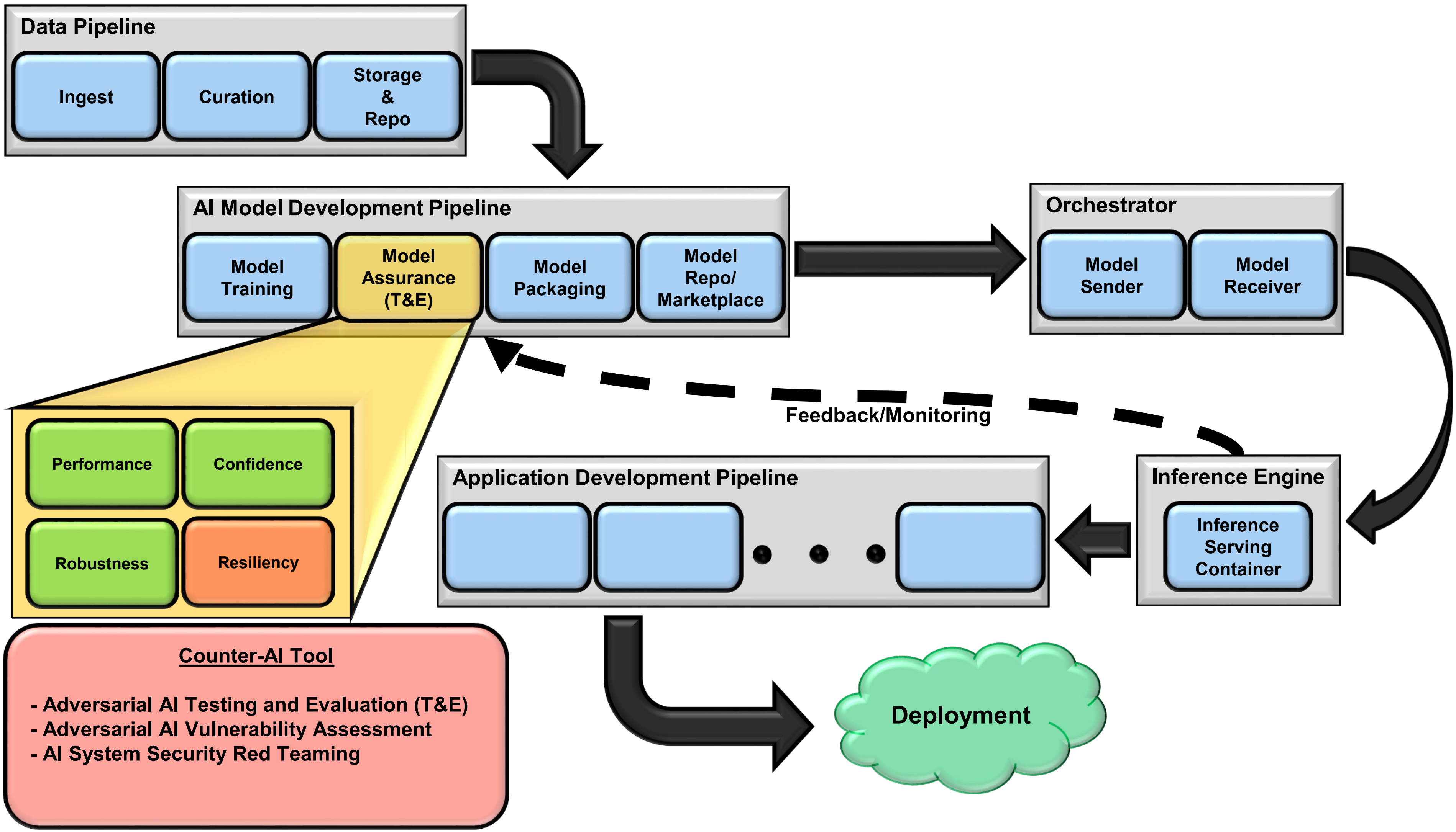
In addition to the two counter-AI assessment examples, here are some general considerations for evaluating AI system resiliency:

- Strong mission interface is a necessity to understand AI system context and threat concerns.
- Commonality across AI systems warrants a common process, infrastructure, and attack/defense capabilities.
- AI systems differ in their primary attack surface as physical vs. digital domain, necessitating modeling and simulation based data generation.
- Not all AI systems have strong counter-AI considerations.
- AI capabilities are provided by cooperative and noncooperative entities (e.g., commercial), impacting assessment activities.

AI = Artificial Intelligence

Counter-AI Tool

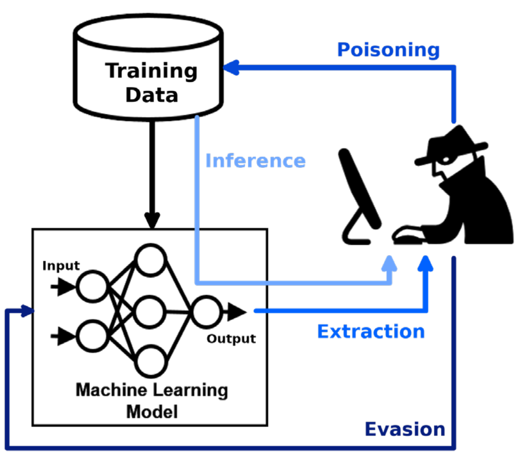
AI Engineering: DevSecOps for AI Systems



AI = Artificial Intelligence

Adversarial Robustness Toolbox Demo

- **Adversarial Robustness Toolbox (ART)** is a Python library for ML security. ART provides tools that enable developers and researchers to evaluate, defend, certify, and verify ML models and applications against the adversarial threats.
- ART supports the most popular **ML frameworks** (TensorFlow, Keras, PyTorch, MXNet, scikit-learn, XGBoost, LightGBM, CatBoost, GPy, etc.), many **data types** (images, tables, audio, video, etc.) and numerous **ML tasks** (classification, object detection, generation, certification, etc.).
- ART supports **39 attack** modules, **29 defense** modules, and **5 metrics** for robustness/certification/verification.
- This involves certifying and verifying **model robustness and model hardening** with approaches such as:
  - Pre-processing inputs
  - Augmenting training data with adversarial examples
  - Leveraging runtime detection methods to flag potentially modified inputs



Adversarial Robustness Toolbox

<u>Useful Weblinks</u>:

https://adversarial-robustness-toolbox.org/

https://adversarial-robustness-toolbox.readthedocs.io/en/latest/

https://github.com/Trusted-AI/adversarial-robustness-toolbox

ML = Machine Learning

Adversarial Robustness Toolbox Demo (cont.)

FGSM Evasion Attack Example Using MNIST

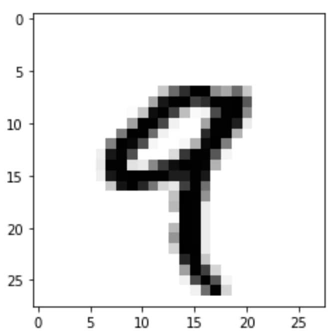
The tutorial demonstrates a simple example of using the Adversarial Robustness Toolbox (ART) with Keras. The example trains a convolutional neural network (CNN) model on the classic MNIST dataset (a large dataset of handwritten digits) and creates adversarial examples using the Fast Gradient Sign Method (FGSM) as an evasion attack. This evasion attack, which perturbs the MNIST image pixels, reduces the CNN classifier performance (accuracy) by over 60%.

In [1]:
```
# Import the required packages
import tensorflow as tf
tf.compat.v1.disable_eager_execution()
import warnings
warnings.simplefilter(action='ignore', category=Warning)
import keras
from keras.models import Sequential
from keras.layers import Dense, Flatten, Conv2D, MaxPooling2D
import numpy as np
import matplotlib.pyplot as plt
from art.attacks.evasion import FastGradientMethod
from art.estimators.classification import KerasClassifier
from art.utils import load_mnist
```

In [2]:
```
# Step 1: Load the MNIST dataset and display the 4th digit (as an example)

(x_train, y_train), (x_test, y_test), min_pixel_value, max_pixel_value = load_mnist()

digit = x_train[4]
plt.imshow(digit, cmap=plt.cm.binary)
plt.show()
```



AdversarialAI.ipynb

AdversarialAI.html

FGSM = Fast Gradient Sign Method
MNIST = Modified National Institute of Standards and Technology

Summary

AI systems hold great promise for enhancing current military, homeland defense, and national security missions; however, adversarial attacks may limit their effectiveness.

- A general taxonomy and background on adversarial AI were provided.

Developed AI systems need to be assessed against potential adversarial AI attacks to make them secure and robust in mission context.

The DoD is working to evaluate developed AI systems, identifying promising mitigations that make them resilient to adversarial attacks.

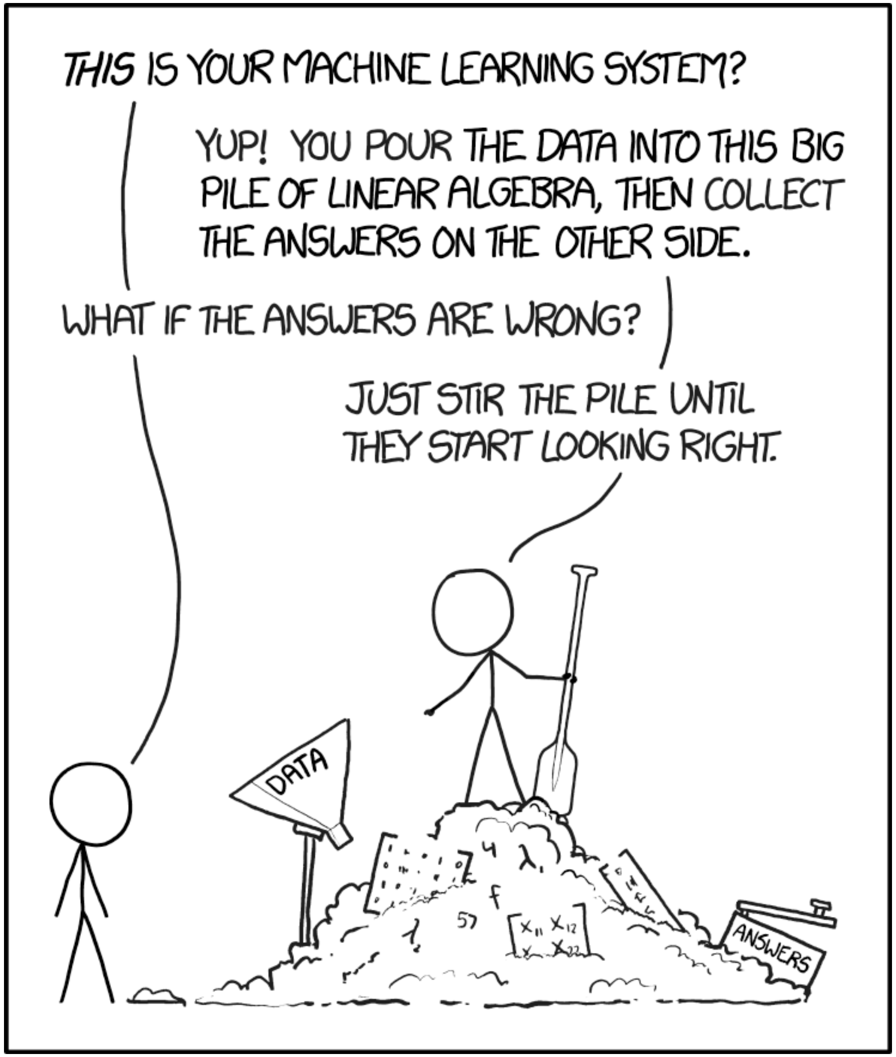
- Interface with mission partners to understand context of AI system deployment
- Assess performance against relevant state-of-the-art adversarial AI attacks
- Recommend mitigations to minimize effect of the adversarial AI attacks
- Provide expertise of adversarial AI attacks and defenses, repository of state-of-the-art software, and persistent infrastructure for AI system testing and evaluation

AI = Artificial Intelligence
DoD = Department of Defense

References

- Abadi et al., “Deep Learning with Differential Privacy,” 2016
- Bursztein, https://elie.net/blog/ai/attacks-against-machine-learning-an-overview/, 2018
- Cao et al., “VGGFace2: A Dataset for Recognizing Face Across Pose and Age,” International Conference on Automatic Face and Gesture Recognition, 2018
- Chakraborty et al., “Adversarial Attacks and Defenses: A Survey,” 2018
- Engstrom et al., “A Rotation and a Translation Suffice: Fooling CNNs with Simple Transformations,” 2018
- Eykholt et al., “Robust Physical-World Attacks on Deep Learning Visual Classification,” 2018
- Fredrikson et al., “Privacy in Pharmacogenetics: An End-to-End Case Study of Personalized Warfarin Dosing,” 2014
- Fredrikson et al., “Model Inversion Attacks that Exploit Confidence Information and Basic Countermeasures,” 2015
- Goodfellow et al., “Explaining and Harnessing Adversarial Examples,” 2015
- Grother et al., “Ongoing Face Recognition Vendor Test (FRVT) Part 1: Verification,” NIST Technical Report, 2019
- He et al., “Deep Residual Learning for Image Recognition,” CVPR, 2016
- Huang et al., “Densely Connected Convolutional Networks,” CVPR 2017
- Kumar et al., “An Army of Me: Sockpuppets in Online Discussion Communities,” 2017
- Lamb et al., “Interpolated Adversarial Training: Achieving Robust Neural Networks without Sacrificing Accuracy,” 2019
- Madry et al., “Towards Deep Learning Models Resistant to Adversarial Attacks,” 2017
- Shokri et al., “Membership Inference Attacks Against Machine Learning Models,” 2017
- Steinhardt et al., “Certified Defenses for Data Poisoning Attacks,” 2017
- Streilein et al., “Future NMI Analysis Study: Counter-AI,” 2019

Q&A


THIS IS YOUR MACHINE LEARNING SYSTEM?
YUP! YOU POUR THE DATA INTO THIS BIG PILE OF LINEAR ALGEBRA, THEN COLLECT THE ANSWERS ON THE OTHER SIDE.
WHAT IF THE ANSWERS ARE WRONG?
JUST STIR THE PILE UNTIL THEY START LOOKING RIGHT.
DATA
ANSWERS