Artificial Intelligence & Technology Office

Cybersecurity Awareness: Adversarial AI Attacks

AI is vulnerable to attack

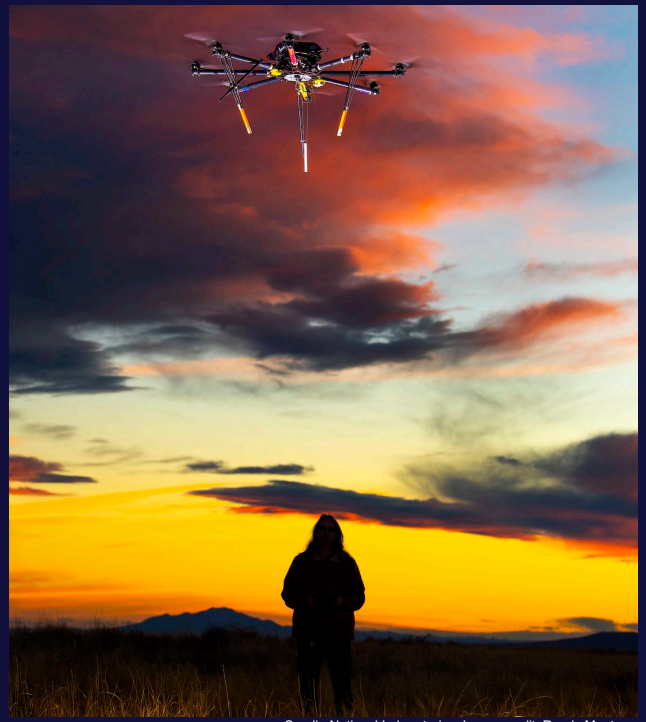
Adversaries can exploit fundamental limitations in AI algorithms to attack systems in new ways.

- Adversarial AI refers to the use of malicious techniques designed to deceive, degrade, or invert machine learning models.
- Unlike conventional cyberattacks caused by human errors or omissions in code, adversarial AI attacks are enabled by inherent limitations in algorithms themselves and their reliance on data.

Expanding attack surface

AI is increasingly used in high consequence areas with little room for failure including energy, finance, healthcare, and defense.

- Attacks on AI systems are being developed and released with increased regularity, including machine learning systems tricked, misled, or evaded (Kumar and Johnson 2020).



Sandia National Laboratories. Image credit: Randy Montoya

Kumar, R, and A. Johnson. (2020, October 22). Cyberattacks against machine learning systems are more common than you think. Microsoft. https://www.microsoft.com/security/blog/2020/10/22/cyberattacks-against-machine-learning-systems-are-more-common-than-you-think/

AI Vulnerabilities

Adversarial attacks can destabilize AI systems, rendering them less safe, predictable, or reliable.

Potential Sources of Attack

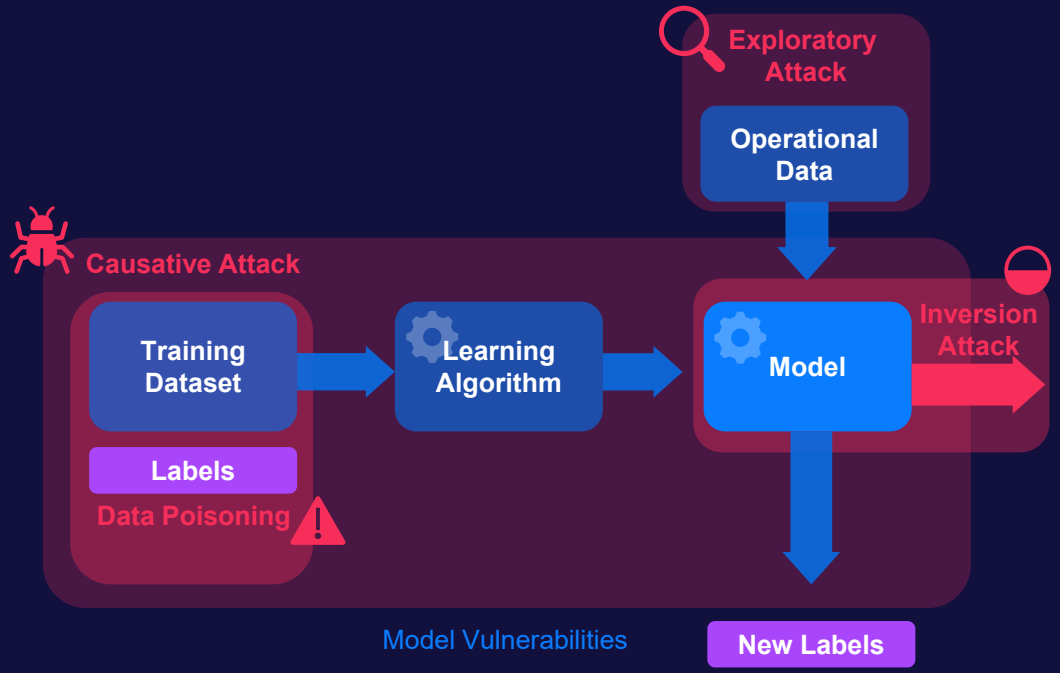
- Nation-state adversaries seeking strategic advantage
- Terrorists targeting critical infrastructure for ideological purposes
- Criminals extorting companies or individuals for profit
- Opportunists and hackers seeking a challenge or reputation

Numerous Attack Modalities

- Adversarial AI attacks can be effective across a range of modalities including: Image, Acoustic, Text, LIDAR, IR, RF, et al.

Attack Examples

- Spoofing, Data Poisoning, Evasion, Trojans, Enchanting, Deepfakes, and others



Considerations for Securing the AI Lifecycle

Secure the supply chain of AI hardware and software

***CONSIDERATIONS*:**

- ❑ Who produced the CPU/TPU/GPU?
- ❑ Where was the firmware made?
- ❑ Could assets be compromised?

AI Supply Chain

Establish a data chain of custody

CONSIDERATIONS:

- ❑ Was the data set purchased?
- ❑ Who compiled and labelled the data?
- ❑ Was the data sanitized and encrypted?

Data Acquisition

Secure training and testing of ML models

CONSIDERATIONS:

- ❑ Who trained the model?
- ❑ Could the training be subverted?
- ❑ What was the source of testing data?

Model Development

Securely integrate and deploy AI

CONSIDERATIONS:

- ❑ Are the deployed assets identified and protected?
- ❑ Are the AI systems being monitored for anomalies?

Model Deployment

Monitor model output

CONSIDERATIONS:

- ❑ Are the outputs accurate, reliable and unbiased?
- ❑ Who has access?
- ❑ Are the outputs susceptible to probing/inversion attacks?

Predictions

Managing AI Risks

The emergence of adversarial AI requires special attention to understand the threat space and organize a coordinated response.

To manage AI risks, the Artificial Intelligence and Technology Office (AITO) developed the AI Risk Management Playbook (AI RMP), which is available for Department of Energy users at: https://edarsprod.servicenowservices.com/aito

Essential Guidance
AI RMP is a dynamic system featuring 100+ unique risks and mitigation techniques with the ability to expand

Intelligent Search
Ability to filter according to lifecycle stage, assets, as well as mapping to project roles and direct keyword searching

Trustworthy AI
Integration with Executive Order 13960: Promoting the Use of Trustworthy AI, including ability to filter by principle