

# Trustworthy AI & UAS Technology



Artificial Intelligence  
& Technology Office

Risk Management, 2022

# AI Lifecycle and Risk Management



## AI Supply Chain

Secure the supply chain of AI-driving hardware and software



## Data Acquisition

Establish data provenance and chain of custody



## Model Development

Secure training and testing of ML models





## Model Deployment

Securely integrate and deploy AI



## Model Performance

Monitor model output for potential probing/inversion

-  Traditional Cybersecurity Domain
-  Additional AI Security Considerations

# Risk Management - 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?

Model Performance

# Benefits of Autonomous UAS



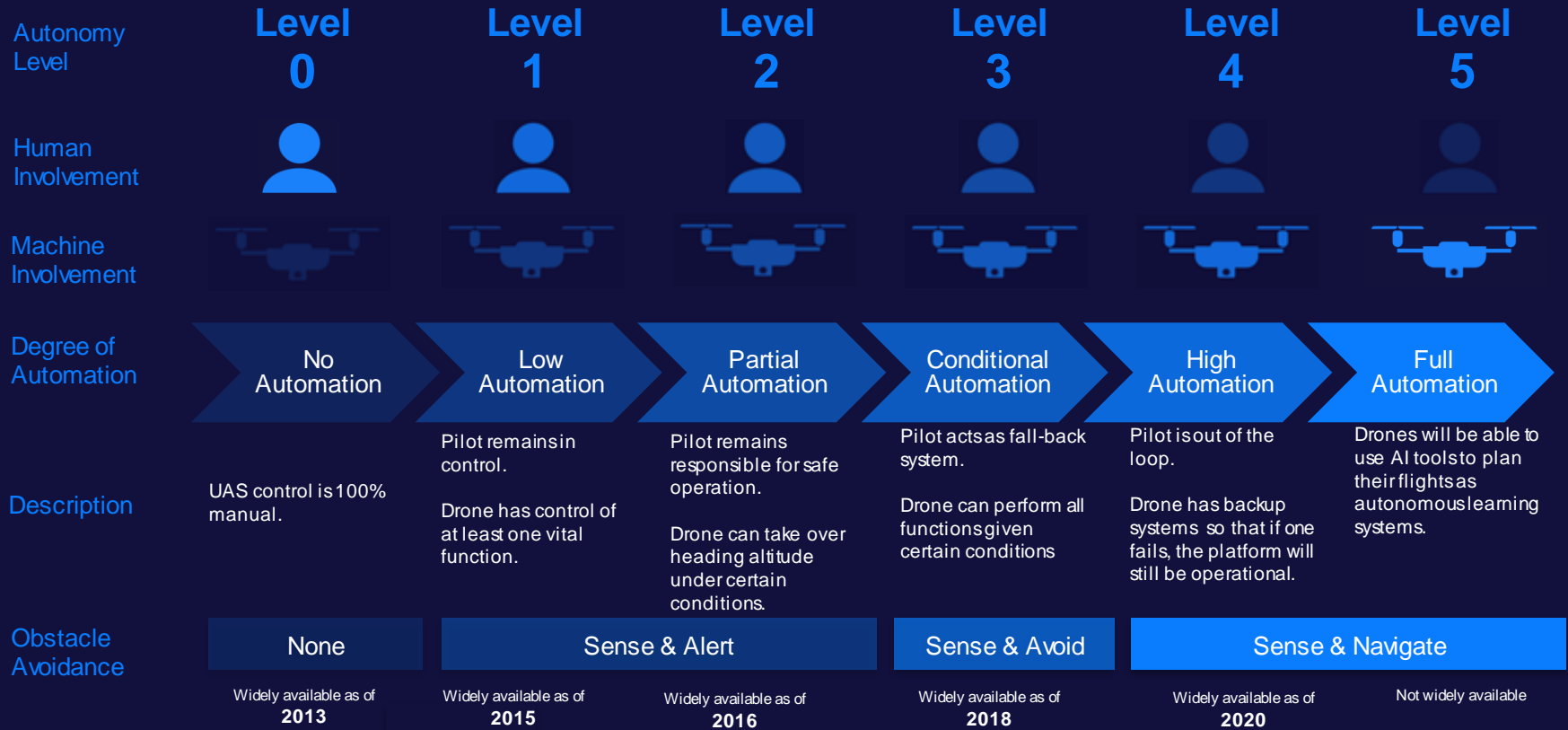
## Current State

- Unmanned aircraft systems (UAS) already provide services across many areas, including public safety, disaster recovery, climate monitoring, defense, real estate, agriculture, infrastructure inspection, medical and goods delivery, and the entertainment industry.
- UAS can operate in hazardous conditions, darkness, extreme heat and other conditions that may pose significant risks to manned aircraft.

## + Autonomy

- Fully autonomous drones will have the ability to take off, follow pre-loaded instructions, follow a moving target, avoid collisions, capture data, land, perform post flight analysis and store valuable information.
- Autonomy can increase work efficiency and productivity, manage repetitive tasks, improve accuracy, minimize errors, and enhance services.

# Autonomy Levels



# Autonomous UAS Concerns: Edge Computing



## Edge Computing

- Edge devices have a physical component (e.g., vehicles, weapons, drones), that may be captured or controlled by an adversary.
- Placing AI models on edge devices increases the need for security measures to prevent model and data theft which may be used to craft future attacks.
- Compromise of one system could lead to the compromise of any other system that shares critical assets such as datasets.



# Autonomous UAS Concerns: Adversarial AI

## AI is vulnerable to attack

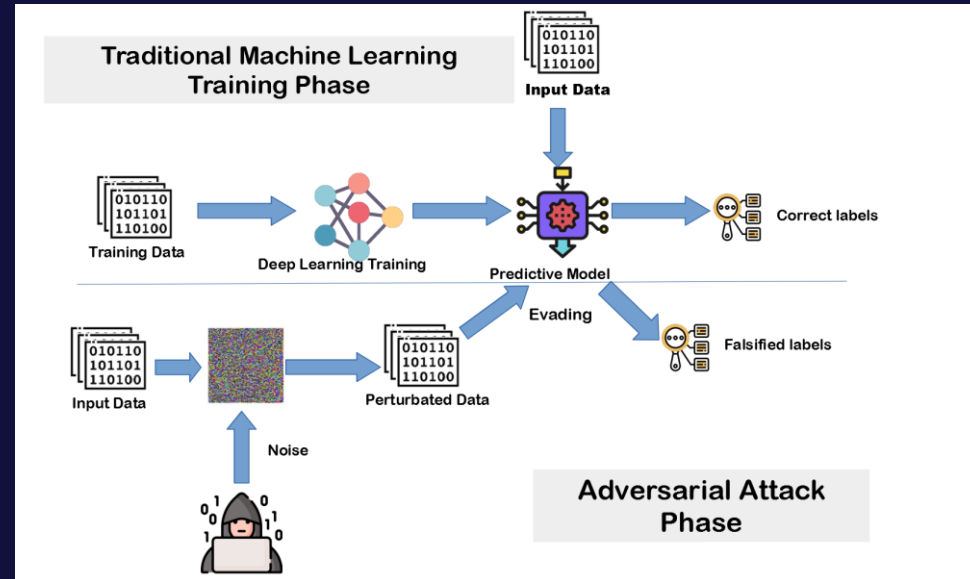
**Adversaries can exploit fundamental limitations in AI algorithms to attack systems in new ways.**

- 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.





# Autonomous UAS Principles: Cybersecurity



## Cybersecurity

- UAS are vulnerable to hacking or GPS-jamming, which can jeopardize their operational use and turn them into potential hazards.
- Compromise of one system could lead to the compromise of any other system that shares critical assets such as datasets.
- Strong encryption is an essential mitigation technique.



## Drone Hacking

Multiple password-related drone hacks have demonstrated a lack of basic security hygiene in some UAS. In one drone hackathon, the Federal Trade Commission (FTC) demonstrated several vulnerabilities in commercial drones, including an unsecured Wi-Fi connection allowing access to a drone's camera feed and other unencrypted data connections that made the drones vulnerable to attack.



# Autonomous UAS Principles: Privacy



## Privacy Considerations

- Drones pose a hybrid of information and spatial privacy problems. Facial recognition technology and other biometric identifiers remove practical obscurity in which we usually operate in physical spaces.
- AI and advanced sensors (e.g., thermal imaging) further complicate; AI capabilities will expand as underlying ML software becomes increasingly sophisticated.
- Ability to combine data from multiple UAS devices to create a near-complete portrait of somebody's physical interactions over time.
- Public harassment, stalking, and surveillance disproportionately affect some groups more than others, including women and women of color in particular; few options for effective legal recourse.



# Autonomous UAS Principles: Data



## 1) Information Collection

- Information collection may be malicious (e.g., UAS purposely spying on someone), but it can also be accidental (e.g., UAS records all in-flight information to improve terrain recognition and accidentally records adjacent people). Both have privacy implications.
- Fully-autonomous systems may potentially decide to change the information it collects, and/or decide to collect new information that it was not collecting at the time of design.

## 2) Information Processing

- Data processing may result in individual identification linked to a real-world identity (e.g., facial or gait recognition). Information gained or inferred about an individual could be used for discriminatory purposes.

## 3) Information Management

- The complexity and high-connectivity of autonomous systems increase potential for data to be hacked. Autonomous cars, which contain multiple components communicating internally and with other autonomous cars, expands the attack surface.
- Erroneous data and errors in processing may result in flawed autonomous decision-making that has real world consequences, including the incorrect identification of humans.

## 4) Information Dissemination

- Data may be disseminated to a remote back-end facility for processing and therefore the security of the parent facility requires consideration.
- Collected and/or processed personal data may be transferred to third-parties.

# AI Vulnerabilities

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

## Potential Sources of Attack

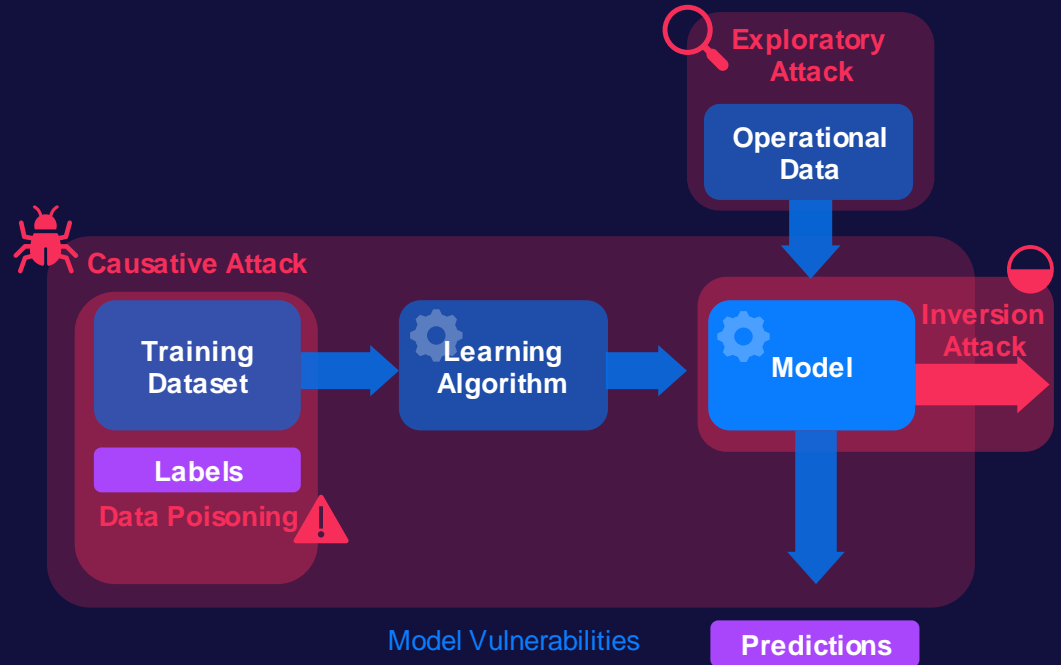
- Nation-state adversaries seeking strategic advantage
- Terrorists targeting critical infrastructure for ideological purposes
- Criminals extorting companies or individuals for profit
- Opportunists and hackers seeking a challenge or reputation

## 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



# Adversarial AI Examples: Data Poisoning

## AI can be deceived

**Machine learning algorithms are efficient but relatively brittle and easy to disrupt.**

- Machine learning (ML) “learns” by extracting patterns from a set of examples known as a training set.
- A training set can be poisoned by attackers, causing an AI system to produce incorrect predictions. Data can be intentionally mislabeled (see Figure 1) or poisoned through subtler manipulations.

## Consequences of data poisoning

**Data poisoning attacks can pose a strategic threat.**

- Datasets used to train AI models are often purchased or are open-source and therefore vulnerable to adversarial tampering.
- Adversaries can introduce backdoors or “trojans” by poisoning the data in ML training sets using malicious samples. The trained models would perform as expected on normal data but behave badly when encountering specific attacker-chosen inputs.
- Among other use cases, corrupted datasets could fool the ML models powering autonomous drones or vehicles.

## Poisoning attack on a machine vision model

Source 7: Diagram adapted from MathWorks.com.

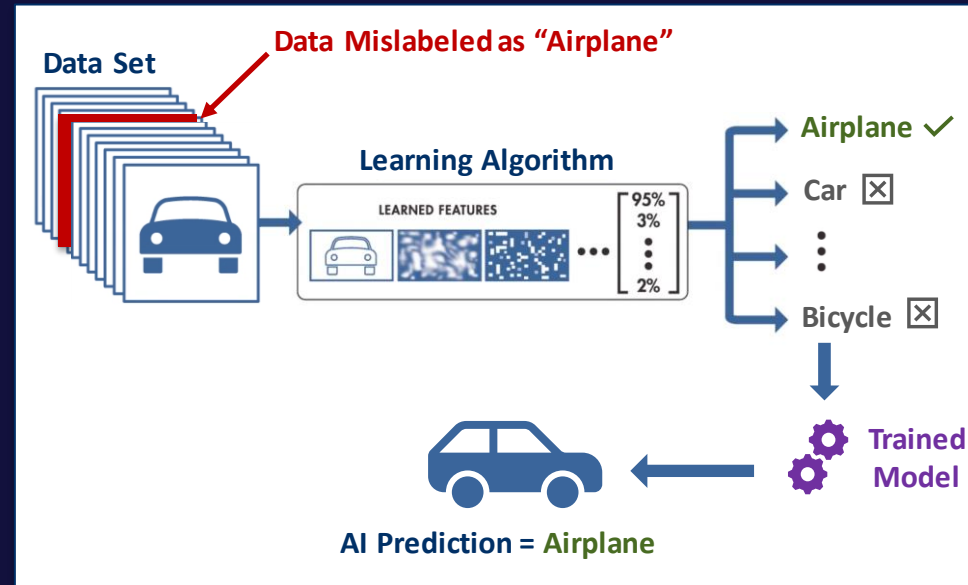


Figure 1

# Adversarial AI Examples: Physical Attacks

## AI is vulnerable to adversarial input

Researchers have shown that changes made to physical objects can confuse AI-based machine vision systems

- Physical input attacks include objects that are intentionally designed to cause an AI model to make a mistake; similar to optical illusions for machines.
- Objects can be modified to fool AI-based systems (e.g., adding tape to a stop sign so that a model perceives it as a green light; or specifically created modifications such as a 3D-printed turtle trained to be classified as a rifle).

## Opportunities for attack

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

- An AI-based vehicle system could be attacked by modifying street signs, potentially sending an autonomous car into oncoming traffic.
- Airport screening systems based on facial recognition could be deceived by a criminal wearing a pair of multi-colored glasses.
- Physical modifications may be noticed by a human but are unlikely to understand the possible implications from a model prediction.

### MAKING A VISIBLE ATTACK

Regular Object



Correct classification:  
"Stop sign"



Attack Pattern



Attack pattern is visible  
to the human eye



Attack Object



Incorrect classification:  
"green light"

### EXAMPLES OF PHYSICAL ATTACKS

ATTACK  
OBJECT



INCORRECT  
CLASSIFICATION

"45 MPH"



"rifle"



"espresso"



"jack"

Examples of physical attacks, which can be perceivable, such as the stop sign or yellow glasses, or imperceivable as with the 3D-printed turtle and baseball shown above.

# Regulatory Landscape



- Legal concerns driven by rapid technology evolution and growing pervasiveness of UAS.
- Data privacy and UAS laws in US are decentralized, with states and cities establishing their own sets of laws, resulting in a patchwork of regulations governing drones as well as data collection, processing, and use.
- A fragmented regulatory landscape may pose problems in litigation involving multiple jurisdictions and could have negative effects on innovation and trade.
- Uniform law governing individual privacy, data security, and related issues may help mitigate misuse of autonomous drones and the data they collect, as well as promote trustworthy AI.
- UAS industry advocates development of consensus best practices through deliberative processes by the Federal Aviation Administration and the National Telecommunications Information Administration.
- FAA rule on Remote Identification of Unmanned Aircraft went in effect March 16, 2021.

[https://www.faa.gov/sites/faa.gov/files/2021-08/RemoteID\\_Executive\\_Summary.pdf](https://www.faa.gov/sites/faa.gov/files/2021-08/RemoteID_Executive_Summary.pdf)

<https://www.federalregister.gov/documents/2021/01/15/2020-28948/remote-identification-of-unmanned-aircraft>

<https://www.natlawreview.com/article/gdpr-usa-new-state-legislation-making-closer-to-reality>

# Regulatory Landscape



The FAA has argued that in most cases federal regulation trumps all state and local drone laws, noting that states and municipalities can pass regulations around “land use, zoning, privacy, trespass, and law enforcement operations.”

## Selection of UAS laws being considered at the State level to protect privacy and public safety:

- Restricting the use of drones over public property, particularly critical infrastructure such as prisons, power plants, etc.
- Restricting the use of drones over private property without the consent of the property owners
- Prohibiting the use of drones for harassment/stalking
- Making it a crime to operate a drone while under the influence of drugs or alcohol
- Banning weaponized drones for the public and/or law enforcement
- Requiring law enforcement to obtain a warrant before using a drone in an investigation,
- Requiring law enforcement to delete any irrelevant data gathered by drone as part of an investigation within 24 hours
- Requiring law enforcement to publicly report drone use
- Requiring drone operators to purchase liability insurance
- Requiring drone operators to keep the drone within view
- Requiring drone operators to label a drone with identifying information

**Suggestion:** Public dialogue about the social issues raised by UAS. This discussion can be supported through empirical social research on the beneficial and detrimental impacts of drone technology on individuals and communities.





# Thank you

Artificial Intelligence & Technology Office



U.S. DEPARTMENT OF  
**ENERGY**

Artificial Intelligence  
and Technology Office