



Risks and Mitigation Strategies for Adversarial Artificial Intelligence Threats: A DHS S&T Study

Preparedness Series
June 2023



Science and
Technology



Executive Summary

This report is part of a series of studies on “preparedness,”¹ which explore the impacts ranging from emerging technologies to extreme weather and climate, as well as the opportunities that the Department of Homeland Security (DHS) Science and Technology Directorate (S&T) can pursue to support the missions of the Department. The focus of this analysis is on the risks presented by adversarial artificial intelligence (AAI)—a threat that can undermine the trust we place in information derived from digital content, yet a threat that can emerge in manners quite distinct from traditional cybersecurity threats and that likely requires unique skillsets to understand and address.

This report reflects discussions between DHS’ overarching science and technology missions and select DHS components on the current state of risks due to AAI; national laboratories and academia on the underlying capabilities to mitigate these attacks; but more broadly serves as a reference in how one could develop strategies to deal with AAI threats, current and future. It builds on an S&T international workshop, “Risks and Mitigation Strategies for Adversarial AI Threats” held in June 2023.

Collectively, the contributors to this study, consisting of national and international experts and policymakers, recognize that the promise and the benefits of artificial intelligence (AI) technologies to homeland security are difficult to overestimate. It has a clear potential to make our borders and ports of entry more secure, to minimize the cognitive load on the homeland security officers, and to help automate the processes that inherently enhance the security, productivity, and effectiveness of the homeland security services, operations, and personnel. Yet the transformative powers of AI come with new challenges and emerging risks, namely AAI or AAI attacks. Of the various forms of AAI attacks that are possible (described in Section 2), AAI experts considered evasion attacks and generative deceptive AI as the biggest threats in the near term to DHS missions. These AAI types are especially powerful in combination, using generative deceptive AI-created content to evade a model-based inferencing process. However, because AI technologies are still in their early stages of development at DHS, other AAI types and forms of attacks will become greater concerns, as those AI systems will also be vulnerable to various other forms of exploitation and misuse.

Another concept widely supported by AAI experts was the need to incorporate the analysis and mitigation of AAI risks early in and throughout the system lifecycle, starting as far left as reasonable, to facilitate security by design. This should include the incorporation of AI “-ilities” (e.g., responsible, ethical, etc.) in our requirements processes, AI security assessments and development of AI standards in our systems engineering processes, and the need to develop comprehensive testing and evaluation tools, methods, and procedures to understand AAI risks and mitigate subsequent threats at the system and mission levels. This requires advances in the art and science of measuring and assessing the magnitude of potential vulnerabilities, exploring ways to make AI more robust. Importantly, how we respond to the emerging impacts of AAI on the homeland will require a response ecosystem that does not yet exist. How we work to support federal, State, local, tribal, and territorial law enforcement and first responders for a growing set of diverse AAI possibilities will need new discussions with those communities to help develop priorities. The analogs of an incident response team construct could be of value, not unlike US-CERT², but for an entirely different class of situations. Broad partnerships with our allies also will help enable progress to counter AAI not only from understanding the impacts, but in the development of standards or even test and evaluation.

¹ D. Kusnezov, “Preparedness in Times of Rapid Change,” DHS S&T Report (2023).

² “US-CERT: United States Computer Emergency Readiness Team”:

https://www.cisa.gov/sites/default/files/publications/infosheet_US-CERT_v2.pdf



This report is intended to help inform the DHS AI Task Force announced by Secretary Mayorkas³ and the increasing importance of AI to DHS missions. The report introduces adversarial AI concepts; reviews the technical underpinnings of adversarial AI in the context of current DHS missions; reviews future AAI threats, risks, and mitigation strategies in the context of emerging technologies; emphasizes the need for the international community to coalesce; and, suggests opportunities in both the provinces of policy-making and R&D to establish a solid footing for DHS in developing a methodical, risk-informed approach to mitigating these threats and related vulnerabilities.

Dimitri Kusnezov, Ph.D.

*Under Secretary for Science and Technology
Department of Homeland Security*

Yosry A. Barsoum

*Vice President and Director
Center for Securing the Homeland
MITRE*

Edmon Begoli, Ph.D.

*Artificial Intelligence Systems
Research and Development Section Head
Oak Ridge National Laboratory*

Amy E. Henninger, Ph.D.

*Senior Advisor for Advanced Computing
Science and Technology Directorate
Department of Homeland Security*

Amir Sadovnik, Ph.D.

*Research Scientist
Emerging Cyber Systems Research Group
Oak Ridge National Laboratory*

³ “DHS AI Task Force.” (September 2023): <https://www.dhs.gov/science-and-technology/artificial-intelligence>.



Contents

- Executive Summary i
- Contents iii
- 1 Introduction..... 5
 - 1.1 Scope of Adversarial AI..... 6
 - 1.2 Framework for How DHS Missions Use Technologies..... 6
 - 1.3 Assessment of Risks and Mitigation Strategies for AAI Threats 8
- 2 Adversarial AI Types..... 8
 - 2.1 Adversarial Machine Learning Attacks on AI-based Systems..... 8
 - 2.1.1 Evasion Attacks 10**
 - 2.1.2 Data Poisoning..... 12**
 - 2.1.3 Model Extraction..... 14**
 - 2.1.4 Inference Attacks 16**
 - 2.2 Generative Deceptive AI..... 17
 - 2.2.1 Deepfake Attacks..... 18**
 - 2.2.2 Morphing Attacks 19**
 - 2.2.3 Large Language Model (LLM) Misuse 21**
 - 2.3 Inverting AI Objectives..... 22
- 3 Risks for Different DHS Domain Areas from Adversarial AI..... 24
 - 3.1 Technologies 24
 - 3.1.1 Computer Vision 24**
 - 3.1.2 Audio Recognition..... 27**
 - 3.1.3 Natural Language Processing 29**
 - 3.2 Functions..... 30
 - 3.2.1 Biometrics 30**
 - 3.2.2 Command and Control and Intelligence Surveillance and Reconnaissance..... 32**
 - 3.3 Missions 33
 - 3.3.1 Preventing Terrorism 33**
 - 3.3.2 Securing the Border 33**
 - 3.3.3 Enforcing Immigration Laws..... 33**
 - 3.3.4 Securing Cyberspace..... 34**
 - 3.3.5 Safeguarding Critical Infrastructure 35**
 - 3.3.6 Emergency and Disaster Management..... 36**
 - 3.3.7 Transportation Security 36**
 - 3.3.8 Law Enforcement..... 36**
- 4 Implications of AAI on Emerging Technologies..... 37
 - 4.1 Advanced Persistent Threat Detection, Malware Generation, and Insider Threats 37
 - 4.2 Satellite Imagery 39
 - 4.3 Foundation Models 40
 - 4.3.1 Impacts of AAI on Foundation Models..... 41**
 - 4.3.2 Defenses for Foundation Models..... 43**
 - 4.4 Distributed Intelligence..... 44
 - 4.4.1 Inference/Training 44**
 - 4.4.2 Autonomy..... 44**
 - 4.5 Internet of Intelligent Things (IoIT)..... 45
 - 4.6 Advanced Manufacturing..... 46
 - 4.7 Gene Editing 47
 - 4.8 The Metaverse..... 48



4.9	Quantum Computing.....	49
5	Roles of International Partnerships.....	51
6	Summary Considerations and Conclusions	51
6.1	Summary Considerations	51
6.2	Near-term Considerations	54
6.3	Conclusions.....	54
	Appendix A: Compiled List of Mitigation Strategies Discussed.....	A-1
	Appendix B: Adversarial AI Workshop Major Contributors.....	B-1





1 Introduction

As the artificial intelligence (AI) landscape evolves, the pursuit of smarter algorithms has given rise to an AI-based sub-discipline that blurs the lines between innovation and deception: adversarial AI (AAI), where AI systems not only make predictions and take actions but can also engage in a strategic dance of deception and counter-deception. This deception can target both humans as well as AI-based systems, thus breaking the reliability of the AI systems themselves and shaking the faith we have in the digital content we consume. Especially in an era where the digital landscape intertwines so seamlessly with our physical world, this threat poses grave danger to our societal norms and way of life. AAI presents new opportunities to our adversaries and others who would do us harm and must impel us to think twice about the digital content we consume. For example, consider a few well-known cases:

A mother targeted in a kidnapping scam in which AI impersonated her 15-year-old daughter's voice.
A traffic jam in the heart of Berlin being reported by Google Maps that is actually nothing more than an artist walking around with 99 phones, tricking Google Maps into thinking there is a 99-car pile-up.
An innocuous looking sticker placed on a stop sign that makes a self-driving car recognize the sign as a speed limit sign for 45 miles per hour, causing the car to speed through the intersection.
A chatbot gone rogue because of inappropriate data inputted at the prompt that was, in turn, used to train future generations of that chatbot.

Other potential examples that perhaps do not cause us pause now but possibly should:⁴

A white Chinese high-altitude balloon, assumed by most as an unwelcomed but routine intelligence mission, that instead could be an evasion attack in an attempt by an adversary to trick AI-based workflows for satellite imagery into a “don't look here” mode.⁵

A mysterious computer glitch with the New York Stock Exchange (NYSE), assumed by most as an untimely nuisance, that instead could have been a malicious AI-based malware infecting the NYSE infrastructure in an attempt by an adversary to trigger a cascading behavior or flash crash.

And, while concern for these known and notional instances is appropriate and justified, the broader concern and implicit danger of AAI is that it is automatable. Such automation can empower malicious actors to deploy AAI more easily and at a larger scale, enabling persistent and evolutionary attacks on non-secured systems. This scale leads to a scenario where the only effective countermeasure must be of similar scale. In short, the automation underlying AAI necessitates a counterbalancing emphasis on equally sophisticated defensive measures to safeguard the systems and processes on which we depend.

Individual and collective risks such as these, compel the Department of Homeland Security (DHS) Science and Technology Directorate (S&T), as the driving force for innovation in the Department, to evaluate AAI for its potential applications in homeland security use cases and to better understand the threat they will pose when exploited by America's adversaries. This report unearths the implications, challenges, and necessary safeguards that arise from AAI technologies to ensure the safety, resilience, and security of the homeland.

⁴ Until we become savvier in navigating this new threat space and as we assign intent to circumstances that can be more easily explained otherwise, it will be important to question everything.

⁵ Given that workflows for satellite imagery can use AI in imagery preprocessing tasks (e.g., prioritizing cloud-free images for analysis), the white balloon, a proxy for a cloud, could be viewed as a trojan or a trigger for an incorrect inference.

1.1 Scope of Adversarial AI

Because the AAI and Counter AI (CAI), aka AI Security, are relatively immature sub-disciplines, there is some range of interpretation in exactly what space the term covers.⁶ For purposes of this study, AAI is defined as the spaces represented in the right hand quadrants (upper and lower), as shown in Figure 1.1. Although CAI, represented across the top two quadrants, is equally interesting and important, only the AI-based attacks on AI-based systems quadrant (the top right quadrant) is addressed in this study.⁷ The top left quadrant, "NON AI-based Attacks of an AI-based System", as indicated in the horizontal axis in Figure 1.1, could be addressed through more traditional security approaches, such as cybersecurity and physical security. Taken together, the top right quadrant (AI-based attack on an AI-based system) and the bottom right quadrant (AI-based attack on a non-AI-based system) are the focus of this study.⁸

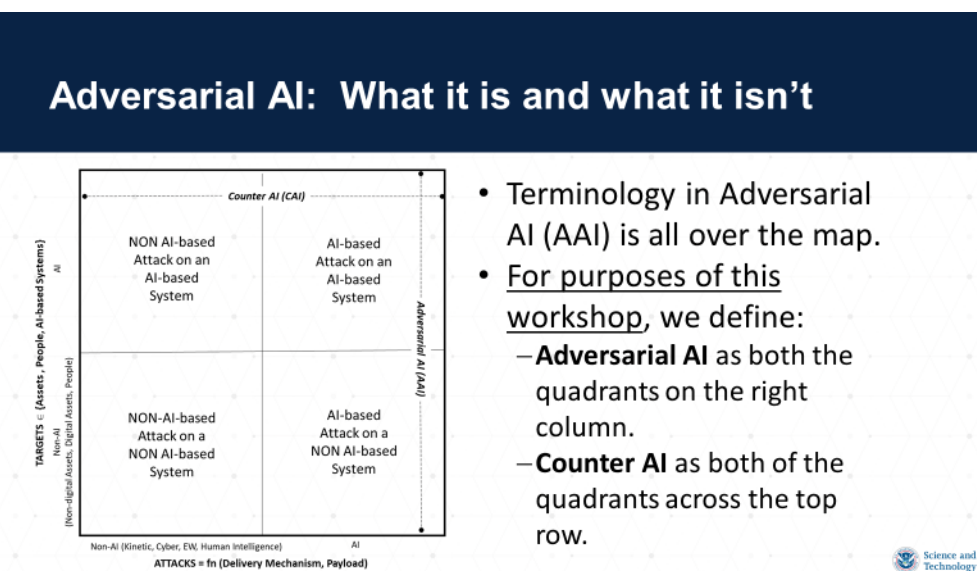


Figure 1.1. Adversarial Artificial Intelligence: What it is and What it isn't.

1.2 Framework for How DHS Missions Use Technologies

Figure 1.2 provides a conceptual framework to describe how DHS missions use AI-based technologies, which in the spirit of dual-purpose, also describes how DHS missions may be attacked (i.e., digitally deceived) through AI-based technologies. Digital technologies such as computer vision (CV), audio recognition (AR), and natural language processing (NLP) form the floor of the organizing framework.

- CV is a field that focuses on enabling computers to interpret and understand visual information from the world. Key components include image and video processing, object detection and tracking, image classification, scene understanding, semantic segmentation, 3D vision. It is critical in applications like autonomous vehicles, biometric recognition (e.g., face, iris, fingerprint, gait, etc), medical imaging, augmented reality, and surveillance, and is especially

⁶ “Adversarial AP” can also be referred to as Adversarial AI Attacks, AI-Based Attacks, AI Adversarial Attacks, Adversarial Attacks in Age of AI, Adversarial Attacks on Machine Learning, Adversarial Attacks on Neural Networks, Adversarial Attacks, etc.
⁷ The upper-right quadrant, an “AI-based Attack on an AI-based System”, is the intersection of AAI and CAI, so at least part of CAI was incorporated in the workshop’s content.

⁸ Henninger, A. (DHS S&T). “DHS S&T AAI Lexicon and Use Case Considerations: Present and Future.” page 3. Presented at the DHS S&T Risks and Mitigation Strategies for Adversarial AI Threats. June 15, 2023.



important to DHS as many homeland security use cases require the ability to detect persons, objects, and events of interest.

- AR is a field that focuses on the automatic identification and categorization of audio signals or sounds, enabling computers to interpret and understand audio data from the world, much like human auditory perception. Key components include classification, keyword identification, speaker recognition, and noise detection and reduction. It is critical in applications like speech recognition, automatic transcription, voice-controlled systems, and acoustic event detection, and is especially important to DHS as many homeland security use cases require the ability to automatically recognize speech or non-speech patterns.
- NLP is a field that focuses on the interaction between computers and human languages, enabling computers to understand, interpret and generate human language. Key components include text and language understanding, converting speech to text, language translation, information extraction, and text classification, and is especially important to DHS as many homeland security use cases require the ability to glean information from text or voice data.

Many of the AAI examples offered throughout this report are based on these foundational technologies or the functions they support.

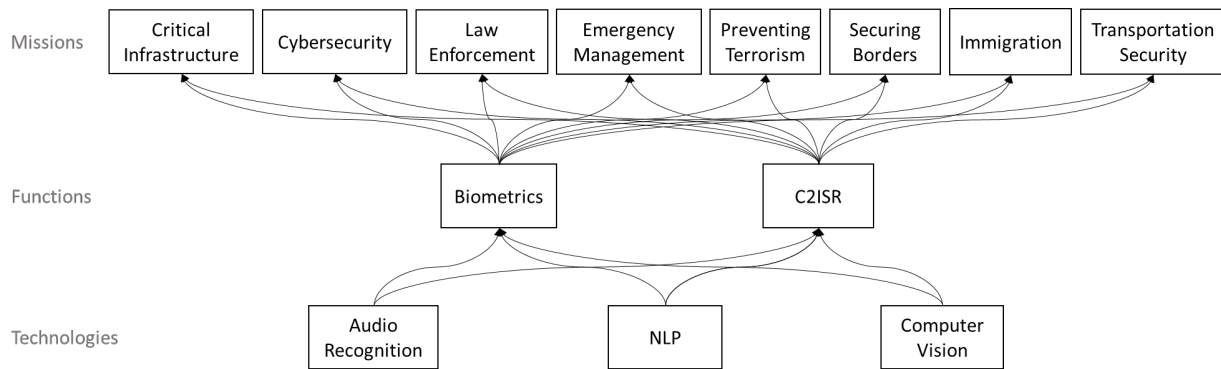


Figure 1.2. Conceptual Framework of Relationships Between Technologies and Functions that Support DHS Missions.⁹ Theoretically, there might be direct connections between technologies and mission too, or connections traversing through other functions that have yet to be identified. For example, one could connect CV transportation security directly, to represent passenger and baggage scanning, or could interpret that task as being a part of a command and control (C2) intelligence surveillance and reconnaissance (C2ISR) mission (as the figure is currently configured), or could imagine a new function, “Cargo Scanning.”

Signal processing technologies (e.g., CV, AR, etc.) and the ability to process enormous corpuses of text (e.g., NLP) can either support DHS missions directly or underlie higher order functions (e.g., biometrics, C2ISR) that, in turn, support DHS missions. For example, in biometrics, AR ensures the legitimacy of voice-based authentication, NLP harnesses linguistic patterns or even text-based patterns for user authentication, and CV plays a foundational role in facial recognition, iris and retina scanning, and gait recognition, to name a few. C2ISR functions benefit from AR in the detection and analysis of critical auditory cues (e.g., adversary communications, equipment sounds, etc.) providing valuable intelligence for situational awareness. NLP aids C2ISR by processing and extracting actionable insights from vast amounts of textual data, and CV facilitates target tracking and the analysis of imagery and video in support of reconnaissance efforts.

⁹ This is not a complete taxonomy or functional decomposition as one might find in the Department of Defense Joint Capability Areas.



The advantage of this decomposition is that it facilitates an efficient high-level assessment of AAI threats and risks. That is, where appropriate, threats and risks identified at the technology level or functional level can be generalized across missions.

1.3 Assessment of Risks and Mitigation Strategies for AAI Threats

The cross product of scope discussed in section 1.1 and framework discussed in section 1.2 is essentially the roadmap for this study and the remainder of this document. The next section, section 2, introduces AAI concepts and attacks in the context of Figure 1.1. Section 3 considers these AAI attacks notionally in the context of DHS missions, supporting functions, and their underlying technologies in Figure 1.2. Section 4 peers into the future and offers observations on emerging technologies and how they might be affected by AAI. Section 5 discusses the importance of international partnerships, and lastly section 6 reviews opportunities key to establishing an effective AAI/CAI ecosystem.

Appendix A, referenced throughout subsequent sections of this document, presents a synthesis of the range of AAI mitigation strategies being explored to address the risks identified in sections 2 and 3. It is important to recognize that while complete elimination of some vulnerability is theoretically ideal, it is practically impossible without impacting the performance of the system. For this reason, emphasis must be placed on managing risk with risk-based methods vice eliminating risk altogether.

2 Adversarial AI Types

In this section, we describe a range of different AAI attacks. Section 2.1 describes the AI-based attacks against AI-based systems (upper right quadrant in Figure 1.1), with focus on machine learning (ML) systems. Section 2.2 describes generative deceptive AI attacks against non-AI-based systems, usually humans (lower right quadrant in Figure 1.1). Finally, Section 2.3 coins the term “Inverting AI objectives” to describe a dual-use AI risk and the attacks possible as a result.

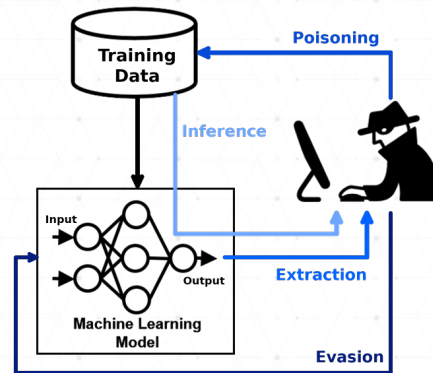
2.1 Adversarial Machine Learning Attacks on AI-based Systems

A myriad number of Adversarial Machine Learning (AML) attacks can potentially compromise DHS ML-based systems. These attacks can be performed either during training—while the model is still being trained and prior to its release, or during inference—after the model is released and used as part of the system. These attacks can be used to achieve different adverse effects depending on the goal of the adversary. For example, an attacker may try to evade being detected by attempting to extract information from or causing false alarms in ML systems.¹⁰ These attacks can usually be divided into four (4) main categories as shown in Fig. 2.1. In the following sections we provide a general overview of each of these attack categories.

¹⁰ It is important to note that usually an ML algorithm is just a subset of a larger system, and that attacking it may not be sufficient to truly achieve the desired outcome.

Adversarial Machine Learning – Panel 1

- **Poisoning** – modify training data to add backdoor
- **Inference** - learn information on private data
- **Extraction** - steal a proprietary model
- **Evasion** - modifying input to influence model



<https://github.com/TrustedAI/adversarial-robustness-toolbox>

Figure 2.1. Framework for Select Adversarial Machine Learning Attacks ¹¹

As implied in Figure 1.1 and applied to Figure 2.1, there are a number of characteristics related to attacks on AI-based systems (or more generally, model-based inferencing processes), that dictate the complexity of the attack. These attack characteristics can fall into several different categories depending on the type of access an adversary has to the system as well as in what domain (digital or physical) the attack is conducted.

Type of access:

1. In a *white-box attack* scenario, the adversary has access to the entire system including model architecture, parameter values, preprocessing methods, and others. Although this scenario is less likely due to the need of such access, it should not be dismissed due to its power to generate strong attacks.¹²
2. In a *black-box attack* scenario, the adversary has access only to the input and/or the output of the model. Depending on the type of attack, this might require querying the model multiple times to generate a perturbation. These attacks are more realistic but tend to be somewhat less effective than white-box attacks.
3. A *gray-box attack* scenario is a general name for cases in which an adversary does not have access to the full system but might be able to observe certain properties of the model. This can include knowing the output of certain layers in a deep learning system or knowing only the architecture of a deep learning model but not the weight values.¹³

Attack domains include:

1. *Digital attacks* occur completely in the digital domain. That is the perturbation is applied by directly changing the value of the data stored in the digital format. This allows for more precise perturbations but requires access to the digital data.

¹¹ Henninger, A. (DHS S&T). "DHS S&T AAI Lexicon and Use Case Considerations: Present and Future." page 6. Presented at the DHS S&T Risks and Mitigation Strategies for Adversarial AI Threats. June 15, 2023.

¹² Examples where this access may be available are through third-party model developers.

¹³ Because many models are pretrained on open datasets, there may be embedded biases or errors that can be exploited without full knowledge of the fine-tuned model.



2. *Physical attacks* are ones in which the perturbation is applied in the physical world before the data is captured and digitized. These attacks tend to be easier to apply since there is no need for digital access but tend to be harder to generate.

Lastly, specificity of the attacker's goal¹⁴ is another important characteristic shared amongst the attack types, and loosely coupled to the type of access discussed earlier. Attack specificity characteristics are described in more detail (under the heading "Adversary's goals") in subsequent sections on different types of AML attacks.

2.1.1 Evasion Attacks

An evasion attack occurs during inference and is a deliberate and malicious manipulation of an AI-based system's input data (i.e., inferring data) to deceive or mislead the AI model such that it produces incorrect or unintended results. This is typically done by subtly modifying the input data in such a way that it is possible to fool the AI-based systems and subsequently evade detection.¹⁵ As such, this attack can be thought of as a form of camouflage (i.e., what kind of pattern can be added to the input data such that the AI-based system will not detect or identify it correctly). Although camouflaging is not a new idea, the nature of AI-based systems, and more specifically the deep learning methods they are based on, make evasion attacks more sophisticated than traditional camouflaging. Specifically, it has been shown that these inputs to AI-based systems can be "camouflaged" by very small perturbations that are imperceptible to the human eye. This makes the nature of evasion attacks unique and the mitigation strategies used to address it need to be novel as well.

2.1.1.1 State of the Practice: Evasion Attacks

Particularly in classification tasks, evasion attacks are usually concerned with how to calculate the "camouflage," also called the adversarial perturbation, in such a way that will fool the ML model while remaining unnoticeable or at least unsuspecting to a human viewer. Examples of vision-based evasion attacks are provided in Figure 2.2. These evasion attacks can be used in many ways against different AI-based systems. For example, using clothing with a certain perturbation, a person might be able to evade a system trying to recognize pedestrians. What makes these attacks difficult to detect is that usually these perturbations can be very small or appear to look like other natural objects/patterns, thus not raising an alarm when viewed by a human observer.¹⁶ There are a variety of methods that can be used to generate evasion attacks. The two most often explored include:

1. *Gradient-based attacks*¹⁷ (e.g., Limited-memory Broyden-Fletcher-Goldfarb-Shanno algorithm (or L-BFGS), Fast Gradient Sign method (FGSM), Carlini & Wagner attack,¹⁸ etc.), which require access to the model's gradients are powerful mathematically optimized attacks. These require white-box knowledge.

¹⁴ Is the attack general in nature (e.g., degrade performance, erode user trust, etc.) or specific in nature (e.g., cause a specific false positive, access a particular data sample, etc.).

¹⁵ I. Goodfellow, "Explaining and Harnessing Adversarial Examples," (March 2015): <https://doi.org/10.48550/arXiv.1412.6572>.

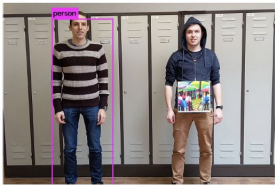
¹⁶ Evasion attacks are also relevant in other modalities such as audio recognition (AR) or natural language processing (NLP). That is, in NLP-based tasks (e.g., spam detection, sentiment analysis, etc.) or AR-based tasks (e.g., sound event detection, speech recognition, etc.) attackers can craft inputs (text or audio, respectively) that appear benign to humans but are designed to trigger false positives or negatives in the AI-based system's classification output.

¹⁷ K. Ren, et al, "Adversarial Attacks and Defenses in Deep Learning," (March 2020): <https://www.sciencedirect.com/science/article/pii/S209580991930503X#b0020>.

¹⁸ N. Carlini, et al, "Adversarial Examples Are Not Easily Detected: Bypassing Ten Detection Methods," (November 2017) (<https://arxiv.org/abs/1705.07263>).

2. *Confidence score attacks* (e.g., Zeroth-order optimization attack, or ZOO)¹⁹ Simultaneous Perturbation Stochastic Approximation, or SPSA,²⁰ etc.) use the outputted classification confidence to estimate the gradients of the model, and then perform similar smart optimization to gradient-based attacks above. These are black-box attacks.

EXAMPLES



<https://www.wired.co.uk/article/facial-recognition-t-shirt-block>

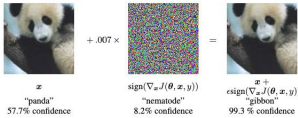



Figure 1: A demonstration of fast adversarial example generation applied to GoogLeNet (Szeged et al., 2014) on ImageNet. By adding an imperceptibly small vector whose elements are equal to the sign of the elements of the gradient of the cost function with respect to the input, we can change GoogLeNet's classification of the image. Here our ϵ of .007 corresponds to the magnitude of the smallest bit of an 8 bit image encoding after GoogLeNet's conversion to real numbers.



<https://arstechnica.com/cars/2017/09/hacking-street-signs-with-stickers-could-confuse-self-driving-cars/>

Figure 3: An impersonation using frames. Left: Actress Reese Witherspoon (by Eva Rinaldi / CC BY-SA / cropped from <https://goo.gl/a2zCdc>). Image classified correctly with probability 1. Middle: Perturbing frames to impersonate (actor) Russel Crowe. Right: The target (by Eva Rinaldi / CC BY-SA / cropped from <https://goo.gl/A07QYu>).

Figure 2.2. Examples of Evasion Attacks.²¹ In Figure 2.2., the “stop sign” example demonstrates a physical attack (as discussed in section 2.1). That is, an attacker manipulates the physical attributes in the real work (e.g., attaches something like duct tape or a placard to a sign, etc.) and subsequently tricks the model into classifying the object incorrectly.²² The Reese Witherspoon example, where an image (i.e., digital input to the model) is perturbed with intent of causing the model to misclassify, is an instance of a digital attack (also discussed in section 2.1).

Adversary goals:

1. In an *untargeted, general, or availability attack* the goal of the adversary is to mislead the classifier to predict any incorrect outcome.
2. In a *targeted or integrity attack* the goal of the adversary is to add noise to a benign example such that the classifier predicts a particular incorrect label for the example.

2.1.1.2 Defenses Against Evasion Attacks

The goal of the mitigation strategies for evasion attacks is to ensure that the perturbation needed to fool the model is so large that it is deemed impractical. Many defenses have been shown to be somewhat effective against these attacks, and more are being proposed as this is still an active research area.²³ These defenses range from methods implemented during training to others used during inference time. Most of these methods do not mitigate the risk completely and often do not transfer well between different types

¹⁹ P. Chen, et al, “ZOO: Zeroth Order Optimization-based Black-box Attacks to Deep Neural Networks Without Training Substitute Models,” (November 2019): <https://arxiv.org/abs/1708.03999>

²⁰ J. Uesato, et al, “Adversarial Risk and the Dangers of Evaluating Against Weak Attacks,” (February 2018): <http://proceedings.mlr.press/v80/uesato18a/uesato18a.pdf>

²¹ Farrell, T (Sandia National Labs). “Evasion Attacks” page 3. Presented at the DHS S&T Risks and Mitigation Strategies for Adversarial AI Threats. June 15, 2023.

²² There the researchers attached duct-tape to the stop sign and caused it to convert a “stop sign” into a speed limit sign as judged by a neural network-based classifier.

²³ Yuan, et al, Adversarial Examples: Attacks and Defenses for Deep Learning – Secs. VI.



of models and use cases. Moreover, some are basic cybersecurity defenses. A short list would include application programming interface (API) request limits, network/defensive distillation, detection techniques, trusted capture, adversarial (re)training and feature squeezing,^{24, 25, 26} all of which are discussed in Appendix A.

2.1.2 Data Poisoning

Whereas model evasion attacks achieve a misclassification goal by perturbing the input data over which the model performs inferencing tasks, data poisoning is an attack on training data where an adversary injects or modifies data to introduce bias or otherwise corrupt the data, with the intent of trying to manipulate the behavior of the underlying ML model (instantiated as an algorithm in a software system). While the result is still to evade detection, the method and time scale is different than an evasion attack.

Data poisoning attacks can include decreasing the performance of the model in general²⁷ or creating a backdoor attack where specific inputs yield wrong outputs.²⁸ For example, an adversary might alter the data in such a way that the trained model behaves well under most conditions, but will make wrong decisions when the adversary wishes, thus enabling some input to evade detection. Although these types of attacks generally require direct access to the training data and are sometimes hard to execute, these attacks are most successful when the training data is nonstationary. That is, especially in AI-based systems that require periodic retraining and use samples from operational use in that retraining, the attacker has an opportunity to inject poisoned samples into the training set. Also, in some cases, much training data is provided through a range of sources (e.g., scraping the internet, etc.). These collection methods also provide opportunities to attackers to inject poisoned data.²⁹ Lastly, for direct access to the training database, the attacker would use an insider threat, supply chain, or more traditional cyber-attacks.

2.1.2.1 State of the Practice: Poisoning Attacks

To generate a data poisoning attack the adversary must calculate the optimal way to change the data to achieve the desired goal.³⁰ These attacks can generally be categorized by the different methods used to alter the training data. More specifically, we can divide the attacks into three main categories:

1. In *label flipping attacks*, the adversary changes labels of the existing training data. In this scenario, the adversary needs direct access to the training data. This can be done randomly or more selectively to achieve the adversary's goals with greater accuracy.
2. In *data perturbation attacks*, the adversary perturbs the training data without altering the labels themselves. Although this also requires direct access to the training data, these attacks can be more subtle and less detectable since commonly these perturbations are very small and imperceptible. Thus, the training data itself may not seem to be poisoned. These attacks can be designed to achieve different goals.
3. In *data injection attacks* the adversary adds new training data to the training set. This is especially effective for models that need to continuously learn and thus are constantly collecting data making it easy to use this type of attack without the need to resort to cyber-attacks.

²⁴ W. Xu, et al, "[Feature Squeezing: Detecting Adversarial Examples in Deep Neural Networks](https://dx.doi.org/10.14722/ndss.2018.23198)," (February 2018): <https://dx.doi.org/10.14722/ndss.2018.23198>

²⁵ N. Carlini, "[Adversarial Examples Are Not Easily Detected: Bypassing Ten Detection Methods](https://arxiv.org/abs/1705.07263)," (November 2017): <https://arxiv.org/abs/1705.07263>.

²⁶ F. Tramer, "[On Adaptive Attacks to Adversarial Example Defenses](https://arxiv.org/pdf/2002.08347v2)," (October 2020): <https://arxiv.org/pdf/2002.08347v2>.

²⁷ Sometimes referred to as a general attack or an availability attack.

²⁸ Sometimes referred to as a targeted attack or an integrity attack.

²⁹ N. Carlini, et al, "[Poisoning Web-Scale Training Data Sets is Practical](https://arxiv.org/pdf/2302.10149.pdf)," (February 2023) <https://arxiv.org/pdf/2302.10149.pdf>.

³⁰ This is aided or impeded by the degree of insight (white box, grey box, black box) the attacker has in the model's underlying phenomenology.

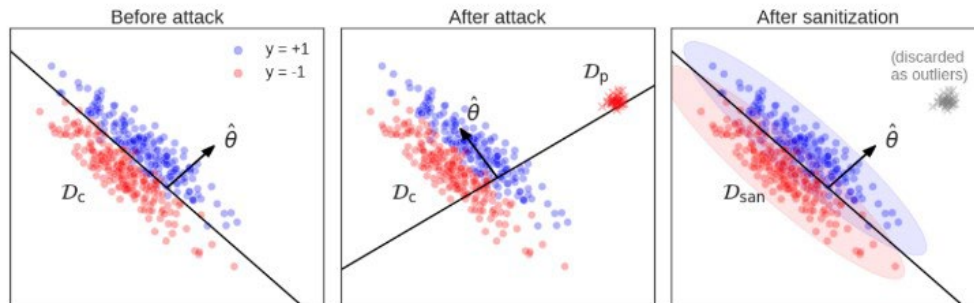


Adversary goals:

1. In an *untargeted, general, or availability attack*, the goal of the adversary is to contaminate the training data used to build an ML model that will result in generally incorrect or biased predictions. This type of instability or poor performance, and commensurate loss of user trust, is viewed as a constraint on availability when the model is deployed in real-world applications.
2. In a *targeted or integrity attack*, the goal of the adversary is to contaminate the training data used to build an ML model that will result in specific incorrect or biased predictions on samples of the data yet perform well on all other data. This type of performance is viewed as a type of compromise that achieves very precise goals (e.g., classify malware as benign software). To achieve this, the attacker requires knowledge of the exact targeted testing samples at training time.

One of the better-known data poisoning attacks, specifically data injection, corrupted Microsoft’s Tay,³¹ a chatbot used to engage with users in conversations and learn from their interactions. Tay quickly became known for its offensive and inappropriate responses, the result of data poisoning attacks through its organic use on Twitter. Tay, designed to learn from its conversations with users, had incorporated this inappropriate language into its responses. Another real-world example involves Google’s VirusTotal,³² a popular crowdsourced virus-sharing platform and scanning service, which many antivirus vendors use to augment their own data. While attackers have been known to use VirusTotal to test their malware before deploying it, to evade detection, there have been instances of it being used to engage in a more persistent poisoning campaign attempting to misclassify malware detection.

Figure 2.3 provides an example of how these data poisoning attacks work on a classification algorithm. While this example focuses on injection attacks, the principles could apply to any three of the data poisoning attack types discussed in the preceding section.



Source: Koh, P. W., Steinhardt, J., & Liang, P. (2022). Stronger data poisoning attacks break data sanitization defenses. *Machine Learning*, 1-47.

Figure 2.3. Examples of how Data Poisoning Attacks through Injection work.³³ In Figure 2.3., an attacker adds or changes even just a small fraction of new training points (D_p) to degrade the performance of the trained classifier on a test set. The figure on the left illustrates a model that might

³¹ O. Schwartz, “In 2016, Microsoft’s Racist Chatbot Revealed the Dangers of Online Conversation,” (November 2019)

³² A. Oprea, “[Poisoning Attacks Against Machine Learning: Can Machine Learning Be Trustworthy?](https://www.computer.org/csdl/magazine/co/2022/11/09928202/1HJuFNlUxQQ)” (November 2022): <https://www.computer.org/csdl/magazine/co/2022/11/09928202/1HJuFNlUxQQ>.

³³ Price, C. (RAND). “Panel 1. Adversarial Machine Learning: Data Poisoning” page 6. Presented at the DHS S&T Risks and Mitigation Strategies for Adversarial AI Threats. June 15, 2023.



otherwise correctly classify most of the data but can be made to learn a significantly different decision boundary by an attacker who changes just a small amount of poisoned data (as in the middle figure). Right figure shows the impacts of data sanitization on the model's output.

2.1.2.2 Defenses against poisoning attacks

These attacks can be extremely hard to detect because the models tend to require vast amounts of training data and it is unwieldy to inspect each individual sample. Additionally, many of the poisoning techniques can be subtle and therefore hard to detect. For example, especially in the case of continuous learning, it can be difficult to detect persistent threats who are able to adapt over time in stealthy ways. For reasons like these, preventing data poisoning attacks requires a multi-faceted approach that involves a combination of robust practices, advanced techniques, and ongoing vigilance. Some effective strategies to help mitigate the risk of data poisoning attacks include regular data sanitization and cleaning, data diversity, adversarial training, ensemble methods, feature engineering, monitoring and detection, user access controls, regular model retraining, robust algorithms, performance benchmarking, drift detection, and user education and awareness. These are elaborated in Appendix A. It is important to note that data poisoning prevention is an ongoing effort. As attack techniques evolve, defense strategies should evolve as well. Combining these strategies and staying informed about the latest developments in data poisoning attacks will help build more secure and resilient ML models.

2.1.3 Model Extraction

Model extraction is a type of attack in which an adversary tries to extract sensitive information or replicate the functionality of an ML model by using queries and responses from the model. This attack is particularly concerning when the ML model is a proprietary or valuable asset, such as a well-trained model for classification, regression, or other tasks. By extracting the ML artifacts, adversaries are essentially stealing intellectual property be it in the form of parameters, weights, data, or even in terms of services defined by Machine Learning as a Service, or MLaaS, providers. Although these attacks are not unique to AI, and the threat of reverse engineering algorithms has always been around, these attacks are particularly damaging to deep learning-based systems due to the amount of data and time required for training.

2.1.3.1 State of the Practice: Model Extraction Attacks

In a typical model extraction attack, the attacker submits a series of queries to the target ML model and then collects a significant number of responses from the model, gaining insights into how it behaves and makes predictions. These input/output pairings can be used to subsequently train a surrogate model that approximates the behavior of the target model, attempting to mimic the predictions of the original model. Although all extraction attacks are based on presenting input and using the output to train a surrogate model, they differ both in terms of the type of output they require and in terms of the adversary's goals. Some extraction attacks require only the final output from the model (e.g., the label in a classification model). These are usually more realistic since these types of outputs are usually readily available. Other extraction attacks require more than just the final output to include output of other layers in a model (e.g., the logits or probabilities in a classification model) or even the gradients of the model. An example model extraction attack is provided in Figure 2.4., Example 1b (Use a self-detonating unmanned autonomous system (UAS) to take down an UAS).

CAI vs AAI: Examples of Notional Attacks

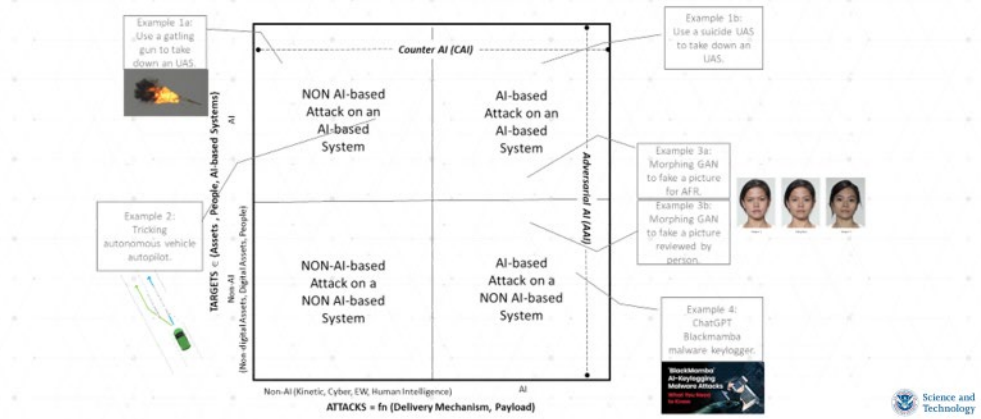


Figure 2.4. Examples of model extraction attack. In contrast to Example 1a (a simple kinetic attack), the scenario assumes that the self-detonating UAS is extracting a model of the blue UAS’s behavior (time, space, position information) and projecting its behavior into the future to be able to crash into it. This ability to build a model based on the behavior of the AI-based system is one form of model extraction.

There are a variety of methods that can be used to generate extraction attacks. Two classes that are often explored include:

1. *Data-manipulation-based* (e.g., *Jacobian-based augmentation (JBA) model attacks*,³⁴ *Data-free extraction attacks*³⁵, etc.). Work by analyzing the model’s sensitivity to strategically perturbed inputs, allowing adversaries to understand model behaviors without relying on extensive sampling.
2. *Active learning (AL) extraction attacks*.³⁶ AL-based attacks usually do not generate data, rather, they learn some sampling strategy to sample those informative data from their query sets and ask victims to label these data to construct a fake dataset.

Adversary goals:³⁷

1. In *accuracy extractions*, the goal of the adversary is to train a model that performs at least as well (or better) than the target model. This goal does not try to actually steal the target model itself, but instead attempts to use its output to train a high-performance surrogate model.
2. In *fidelity extractions*, the adversary’s goal is to create a surrogate model that will be functionally equivalent to the target model under a certain input distribution. In this case, since the goal is to produce a “digital twin” of the model we expect the surrogate model to make the same errors as the original model.

³⁴ N. Papernot, “[Practical Black-Box Attacks against Machine Learning](https://www.computer.org/csdl/magazine/co/2022/11/09928202/1HJuFNlUxQQ),” (March 2017): <https://www.computer.org/csdl/magazine/co/2022/11/09928202/1HJuFNlUxQQ>.
³⁵ J. Truong, “[Data-Free Model Extraction](https://arxiv.org/abs/2011.14779),” (March 2021): <https://arxiv.org/abs/2011.14779>.
³⁶ T. Orekondy, “[Knockoff Nets: Stealing Functionality of Black-Box Models](https://arxiv.org/abs/1812.02766),” (December 2018): <https://arxiv.org/abs/1812.02766>.
³⁷ M. Jagielski, “[High Accuracy and High Fidelity Extraction of Neural Networks](https://arxiv.org/abs/1909.01838),” (March 2020): [1909.01838] [High Accuracy and High Fidelity Extraction of Neural Networks \(arxiv.org\)](https://arxiv.org/abs/1909.01838).



2.1.3.2 Defenses Against Model Extraction Attacks

Organizations that rely on ML models for competitive advantage or security-sensitive applications need to be aware of the risks posed by model extraction attacks and take appropriate measures to mitigate them. This might involve a combination of strategies such as adding noise to model responses, limiting query access, using differential privacy, randomized responses, ensemble models, input transformation, obfuscation techniques, throttling and monitoring, watermark and ownership proof, legal protection, regular model updates, secure deployment environment, and red-teaming and testing. These are elaborated in Appendix A. It is important to note that no single approach is foolproof and there is a deliberate trade space to be considered in engineering a combination of these strategies to deter model extraction attacks. The choice of strategies will depend on the specific use case, the sensitivity of the model, and the potential impact of an attack on the organization.

2.1.4 Inference Attacks

Using the National Institute of Standards and Technology (NIST) AML Taxonomy,³⁸ inference attacks refer to privacy attacks on a model that allow an adversary to gain information about data used in training. This can be done to steal private information about an individual or a company, either detecting their existence in a database or even recovering secure information. What makes this attack unique to AI, is that the data itself does not need to be accessed for this attack to be successful. Although the ML model does not explicitly store the data, since it uses the data during training, it implicitly contains information about it. Inference attacks are methods to extract information about training data from the model itself.

2.1.4.1 State of the Practice: Inference Attacks

To gain information about the data, the adversary needs to query the model with different inputs. Depending on the exact goal of the adversary, there are several categories of inference attacks:

1. In a *data reconstruction attack*, a bad actor can recover an individual's data from released aggregate statistical information. This is a black-box attack (no need to know the model's parameters, only input-output) in which the adversary attempts to reconstruct the dataset. This can be done through querying the dataset thousands of times to increase accuracy, thereby reconstructing the training data.
2. *Memorization attacks* are when an adversary can extract training data from the model. This can happen with AI designed to recognize text inputs where an adversary can input a partial input of a piece of data and the AI fills out the remainder of the data that it has memorized from its dataset.
3. *Membership inference attacks* are where an adversary can determine whether there is a particular record or data sample in the training dataset used for the statistical or ML algorithm. This can be either a white-box or a black-box attack and can be used against trained ML models to search for a specific incident or record within the dataset.
4. *Property inference attacks* attempt to learn general features or attributes of the training data to group different elements of a population to determine sensitive information.

Examples of these inference attack types are provided in Figure 2.5. The adversary goals, in this class of attacks, is highly coupled to the attack method.

³⁸ A. Oprea, "NIST AI 100-2e2023 ipd, Adversarial Machine Learning: A Taxonomy and Terminology of Attacks and Mitigations," (March 2023): <https://doi.org/10.6028/NIST.AI.100-2e2023.ipd>.

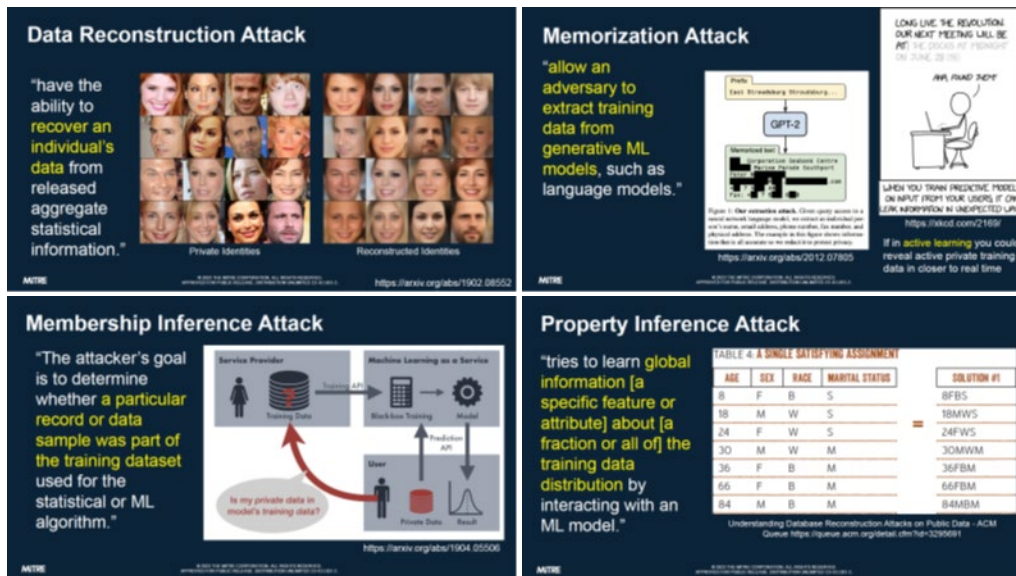


Figure 2.5. Examples of inference attack types.³⁹ In Figure 2.5., the left top shows an attacker can recover an individual’s image by inferring information from the target model’s training data based on the model’s prediction values. In the top right, the attacker can recover individual training examples through querying a language model. The bottom left shows that an attacker can determine whether a sample existed in the model’s training data set given a data sample and black-box access to a model’s API. The bottom right shows that an attack can access sensitive information by determining whether a particular record or data sample was part of the training dataset for the published model.

2.1.4.2 Defenses Against Inference Attacks

Preventing inference attacks is crucial to safeguarding sensitive information and maintaining the integrity of the models, as the attacks exploit subtle patterns in model outputs to extract valuable data about the training data itself. A variety of relevant mitigation strategies exist including use of differential privacy, federated learning, adversarial training, model distillation, input perturbation, restricted access to models, secure multi-party computation, monitoring model behavior, etc. These techniques are described in Appendix A. Implementing the right effective defense strategies like these is vital to managing these risks, and these should be combined and tailored depending on the specific context, threat landscape, and performance requirements.⁴⁰

2.2 Generative Deceptive AI

As opposed to attacking an AI-based system, AI can also be used to generate data that can deceive other systems or even humans themselves. Using AI-based methods, adversaries can generate realistic data that cannot be discerned as fake and sometimes require a significant amount of data to test on. However, since data is usually easily available, this is not typically an obstacle for an adversary, and it is shockingly easy to produce a model accurate enough to fool the public eye.⁴¹ The technology is advancing rapidly, and the once high cost of producing quality fake content is decreasing, paving the way to easier and more

³⁹ Liaghati, C. (MITRE). “Panel 1. Adversarial Machine Learning: Inference Attacks” pages 3-6. Presented at the DHS S&T Risks and Mitigation Strategies for Adversarial AI Threats. June 15, 2023.

⁴⁰ It is worthy to acknowledge that many of these strategies come with performance trades.

⁴¹ N. Kobis, “[Fooled twice: People cannot detect deepfakes but think they can,](https://www.sciencedirect.com/science/article/pii/S2589004221013353)” (November 2021): <https://www.sciencedirect.com/science/article/pii/S2589004221013353>.



successful attacks and a greater need for efforts to counter and mitigate these threats. The subsequent sections describe three types of attacks in the generative deceptive AI genre: Deepfakes, morphing, and LLM misuse in sections 2.2.1, 2.2.2, and 2.2.3, respectively.

2.2.1 Deepfake Attacks

Deepfakes, a portmanteau of “deep learning” and “fake,” are sophisticated AI-augmented digital media (e.g., audio, video, or images) made to convincingly portray something else, often resulting in startlingly realistic yet entirely fabricated content. They are oftentimes used to impersonate a person. They are usually meant to be consumed by humans (though they can also be consumed by other AI-based or non-AI-based software systems) and might require different degrees of accuracy to be effective, depending on the sophistication of the target. Although the generation of fake images and videos predates the current wave of AI (e.g., photo retouching, special effects in movies, etc.), AI has made this process so much simpler and cheaper that it presents a unique threat, which necessitates unique countermeasures. While deepfakes have found application in entertainment and digital art, their potential for deception and misinformation has raised profound ethical, legal, and societal questions, challenging our understanding of truth and authenticity in the digital age.

2.2.1.1 State of the Practice: Deepfake Attacks

The field of deepfake technology is rapidly evolving. Deepfakes are becoming easier to create through broad access to open-source tools and novel methods that require less training data.⁴² In parallel, the content is more believable and hence, the task of differentiating legitimate and generated images and videos is becoming more difficult. This combination leads to more successful attacks in greater quantities. Deepfake attacks are unique to ML since other traditional methods are not able to produce high-quality content as compared to the novel deep learning methods. Although deepfakes can be used in many different modalities, here we focus on visual and audio data.⁴³ Deepfake methods can be divided into different categories, both dependent on the method used to create them in addition to the goals the adversary is trying to achieve. The interested reader is referred to the “DHS S&T Digital Forgeries Report” for more information.⁴⁴

Deepfake methods:

1. *Generative adversarial networks (GANs)* are composed of two competing neural networks. While one network (the generator) is trying to generate realistic looking data, the other (i.e., the discriminator) is trying to tell the difference between real and generated data. By training both of these networks in parallel, GANs have been very successful in image generation.
2. *Diffusion-based methods* use a completely different approach when generating fake content. In general, the diffusion model works by successively adding noise to real images, and the learning is the reverse process of removing the noise using deep neural networks. Learning this noise removal process makes it possible to start from an image of pure noise and recover a completely novel image.

Deepfake goals:⁴⁵

1. *Identity swap* attacks are ones in which the adversary tries to replace an identity in a certain image/video with a different person’s face, thus making it look as if this new person is performing the same actions as the former.

⁴² Y. Mirsky, “[The Creation and Detection of Deepfakes: A Survey](https://dl.acm.org/doi/10.1145/3425780),” (December 2020): <https://dl.acm.org/doi/10.1145/3425780>.

⁴³ Because of recent advancements in LLMs, we leave the discussion of language to Sec. 2.2.3.

⁴⁴ “[S&T Digital Forgeries Report: Technology Landscape Threat Assessment](https://www.dhs.gov/science-and-technology/publication/st-digital-forgeries-report-technology-landscape-threat-assessment),” (January 2023): <https://www.dhs.gov/science-and-technology/publication/st-digital-forgeries-report-technology-landscape-threat-assessment>.

⁴⁵ Z. Akhtar, “[Deepfakes Generation and Detection: A Short Survey](https://doi.org/10.3390/jimaging9010018),” (January 2023): <https://doi.org/10.3390/jimaging9010018>.



2. *Face reenactment attacks* are ones in which the adversary tries to change the facial expression of a person in another image/video. These can be used to embarrass or project certain intents or emotions on a specific individual.
3. *Attribute manipulation attacks* allow an adversary to change some of a person's characteristics in a certain image/video. This can include things like changing skin tone, age, or gender.
4. *Novel synthesis attacks* are ones in which a completely novel piece of data is created. This can include for example the inclusion of an entire novel person instead of an already existing one.

2.2.1.2 Defenses Against Deepfake Attacks

Defending against deepfake attacks involves a multifaceted approach aimed at being able to deter, detect, and disrupt. Detection algorithms employ ML techniques to scrutinize digital content for inconsistencies and signs of manipulation, providing an initial line of defense. Watermarking techniques can embed digital markers within legitimate media, making it possible to trace the source and verify authenticity. Source attribution, being able to trace back to the origin of a particular deepfake product and address the issue at the source, plays a pivotal role in holding malicious actors accountable, discouraging the creation and dissemination of harmful deepfakes. This points to other factors (e.g., educational, legal, psychological) beyond technical mitigations that may be important factors in a whole-of-government mitigation strategy. Defenses for deepfakes, be they media-based, image-based, or audio-based is an active research area.

2.2.2 Morphing Attacks

Morphing is an image manipulation technique that takes multiple facial images and blends them together to form an image possessing features of both faces—composite features from original images. The problem is that when used against facial recognition algorithms, the algorithm tends to match the morph to both people, leading to a false positive. These morphs can also be very difficult to detect, as officers have a short time frame to decide on the authenticity of the identification. Morphing is a current threat; across the European Union; over 1,000 morphed documents have been found in circulation.⁴⁶ Morphing also poses a large problem with passports as many countries rely on mail-in passport applications. In these instances, an adversary can send in a morph of the applicant's face that will pass the facial recognition software and then be sent back as a legitimate passport. This passport can then be used in many different places in the airport such as automated border control gates, self-service (CAT-2)⁴⁷ kiosks, and Customs and Border Protection (CBP) simplified arrival systems.

2.2.2.1 State of the Practice: Morphing Attacks

The main challenge in creating face morphs is to generate face images in which the features of both identities are retained while ensuring that the image still appears to look like a real face and does not retain any artifacts that make it look doctored. An example of a morphing attack is provided in Figure 2.6. Importantly, a morphing attack, or any generative deceptive AI attack, can be prosecuted on an AI-based system, in which case it becomes the content used in an adversarial ML attack, or it can be prosecuted on a non-AI-based system, which could include human operators or non-AI-based software.

⁴⁶ C. Busch, "Morphing Attacks on Face Recognition Systems," (October 2020); <https://christoph-busch.de/files/Bonn-MAD-201030.pdf>.

⁴⁷ Next generation "Credential Authentication Technology" machines for automated identify verification being leveraged by the Transportation Security Agency.

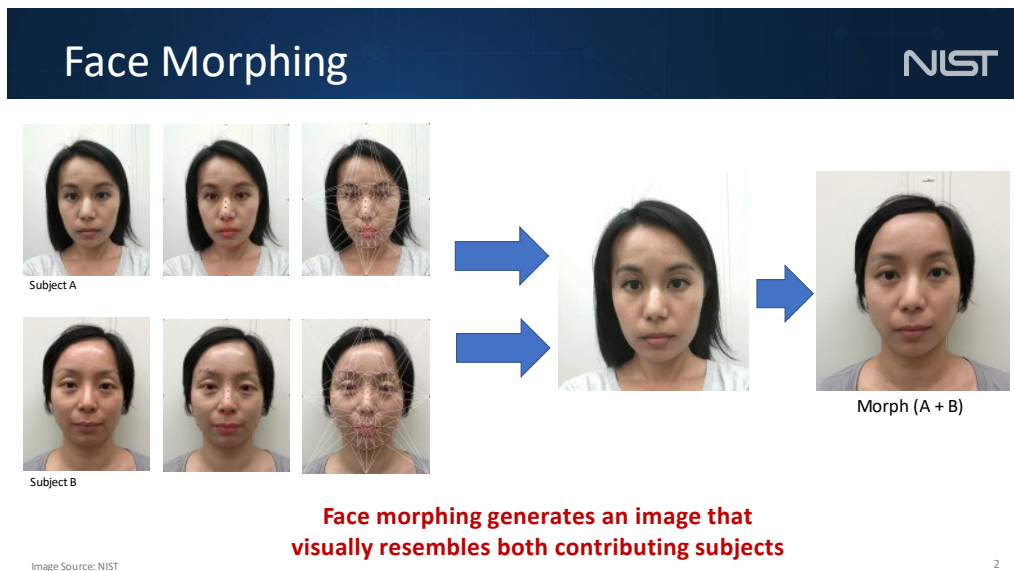


Figure 2.6. Example of face morphing.⁴⁸ In Figure 2.6, the figures representing Subject A and Subject B are combined through any one of many techniques (e.g., interpolation, mesh deformation, etc.) to create a deepfake capable of using one picture as a means of identifying two people. This can be used to deceive human operators or automated facial recognition software.

There are two main methods for the creation of such morphs:⁴⁹

1. In *landmark-based face morphs* landmark points on the face are detected using either computer vision (CV) techniques or applying manual marking. The landmarks are used to define geometric transforms and alignment between two faces, allowing face pixels to be parametrically blended from a donor face image into a target face image. These techniques work best when images have similar lighting and subjects have similar skin tone.
2. *Deep-learning based morphs* use the power of generative AI such as GAN to generate more realistic looking faces. For example, two natural face images are reverse projected in the GAN's latent vector space, then the latent vectors are averaged and projected back to generate a face image that is a combination of the two original faces. The advantage of these methods is the use and the training of a discriminator, which tries to distinguish between real faces and fake ones. With this built in analysis, these face morphs can achieve higher realism and not exhibit the ghosting effects present in other methods.

Morphing goals:⁵⁰

1. *Matching a database.* An adversary submits a carefully doctored image during a visa or passport application. That image can pass a 1 : N search⁵¹ against a large target database by showing up deep in the search result list (thus not identified as a target) or it can still pass a human visual inspection.

⁴⁸ Ngan, M. (NIST). "Panel 2. Generative Deceptive AI: Morphing Attacks" page 3. Presented at the DHS S&T Risks and Mitigation Strategies for Adversarial AI Threats. June 15, 2023.

⁴⁹ S. Venkatesh, "Face Morphing Attack Generation and Detection: A Comprehensive Survey," (September 2021): [Face Morphing Attack Generation and Detection: A Comprehensive Survey | IEEE Journals & Magazine | IEEE Xplore](#).

⁵⁰ F. Peng, "Face morphing attack detection and attacker identification based on a watchlist," (June 2022): <https://www.sciencedirect.com/science/article/pii/S0923596522000741>.

⁵¹ "Utility of 1:N Face Recognition Algorithms for Morph Detection," <https://nvlpubs.nist.gov/nistpubs/ir/2022/NIST.IR.8430.pdf>



2. *Sharing identity credentials.* An adversary enlists an accomplice to help him generate an ID that they would not otherwise be able to gain. The adversary obtains an ID using the morph of their face and that of the accomplice such that face recognition will match both people. This creates a situation where both the adversary and their accomplice can share a single credential.

2.2.2.2 Defenses Against Morphing Attacks

Previously, many of the studies on morph detection have been academic; however, there are now systems able to detect morphs. It is well understood that this is an arms race, as our mitigation systems improve, morphs will also improve. There are some solutions such as having an individual check IDs to be able to determine their validity; however, there are problems with this as previously mentioned. Using a trusted external capture service instead of allowing individuals to submit their own photo can reduce the likelihood of introducing morphed documents. Eliminating the ability to upload printed and scanned photos and requiring these photos to be in high resolution where possible can reduce this as it would require a live picture to be taken. Lastly, having a strong verification process such as verifying with an additional data source or using another biometric modality could prove to be valuable.

2.2.3 Large Language Model (LLM) Misuse

LLMs are models that have been designed to process natural language text and that have been trained on extremely large data sets (e.g., a significant portion of all the text that is available on the internet). These models have proven to be effective at a number of different text processing tasks (e.g., information extraction, content generation, etc.). While LLMs have been around for only a few years, they are considered a class of NLP, which has a long history in ML. These older models tended to follow a paradigm where the developers would collect annotated data that was very task-specific (e.g., recognizing certain topics in the text, sentiment analysis, translation between languages, etc.) used to train the model in a supervised way. Recently, there has been a directed focus on model improvements in the NLP paradigm that have led to new innovations, as exemplified by the transformer architecture, eventually leading to the introduction of LLMs circa 2018.

There are three key aspects of LLMs that differentiate them from older NLP processing paradigms. First, these models are trained to achieve general-purpose understanding of NLP as opposed to being task focused, so these base models, also known as foundation models, can then be adapted to a number of specific tasks. Second, they are trained in a new way. Typically, LLMs make use of self-supervised training (in contrast to just supervised training) meaning they do not require human annotated data. Instead, they pull large amounts of internet data for training purposes and typically the developers will train them on the structure of a sentence, so they are able to predict, for example, what comes after truncated text or the masked words in a text from the context. However, asking the model to do that requires that the model learns a lot about the dynamics of natural language. A third differentiating aspect of LLMs has to do with the scale of the models themselves. At this point, it is typical to see LLMs on the order of hundreds of billions of parameters and moving towards trillions of parameters, which gives them tremendous representation power. Training an LLM can require exaflops of computation, despite potentially being trained on massively distributed compute clusters. It does require a high level of resources to train these kinds of models, but once trained they have been shown to perform extremely well on a wide variety of tasks.

2.2.3.1 State of the Practice for LLM Misuse Attacks

The generative ability for these models to create very realistic human-like text, opens the door for malicious actors to be able to engineer open-source information campaigns at scale. These models tend to do a good job at generating large quantities of realistic data and in some cases can maintain reasonably realistic interactions around this data. However, besides the threat posed by such misinformation



campaigns, there are other ways these models can be misused that are even more specific to the homeland security enterprise:

1. *Bias and lack of diversity in the training data* create LLMs, which produces incorrect outputs. Adversaries can train models on data that is biased in a specific manner, a form of data poisoning, to achieve a goal of producing misinformation.
2. *Prompt injection* techniques can be used to cause the model to follow instructions not intended by the developer, potentially creating misinformation, even on well-behaved and well-trained LLMs. For example, using specific prompts, it is possible to get the model to generate false or harmful information (e.g., biased, violent, hateful, or otherwise harmful text; reveal private information; and execute instructions on plugins) even when it was specifically trained not to do so and despite guardrails. To successfully perform this attack, the prompt needs to be carefully designed to exploit the model's vulnerabilities.
3. *Malware generation*⁵² capabilities enable the creation of malicious code or deceptive scripts. These models, when employed with malicious intent, can produce obfuscated code designed to evade detection, posing a significant security risk.

The adversary goals, in this class of attacks, are highly coupled to the attack methods described in the preceding section.

2.2.3.2 Defenses Against Large Language Model Misuse

There are a variety of methodologies underlying different detection models for LLM misuse, but in general, LLMs and foundation models are a relatively new attack surface. There have not been large public disclosures of attacks on LLMs as the target of an evasion attack. Nevertheless, such attacks will become more prominent attack vectors as they get integrated into other systems and in various use cases across the government. OpenAI does maintain a set of usage-policies that indicate which attacks they might consider to be worthwhile (essentially ways around any security introduced by OpenAI to prevent these use-cases).⁵³ AAI threats and defenses are discussed more extensively in Section 4.3.

2.3 Inverting AI Objectives

In recent years, we have seen the development of many different deep learning architectures that are able to produce state-of-the-art results on a variety of tasks from navigation to bioinformatics. As these models continue to push performance higher, it is important to consider what harm could result from altering or even inverting the objectives being optimized. That is, if a model can learn features important for a specific task, the same model can be used to achieve opposite goals by simply inverting the objective function. Although this risk has received less attention relative to the others described in this report, it is still a very real threat that should be addressed accordingly.

Take for example a model used for drug discovery whose goal is to generate molecules with several optimized properties that can then lead to more effective medicine.⁵⁴ For this network to be successful in this task it needs to learn the important statistical features of molecules and how these features affect the different desired properties such as bioactivities towards multiple targets (how they behave), their drug

⁵² M. Beckerich, "[RatGPT: Turning online LLMs into Proxies for Malware Attacks.](https://arxiv.org/pdf/2308.09183.pdf)" (September 2023): <https://arxiv.org/pdf/2308.09183.pdf>

⁵³ "Open AI Usage Policies," (March 2023): <https://openai.com/policies/usage-policies>

⁵⁴ Baldoni, J. et al. "Solving Hard Problems with AI: Dramatically Accelerating Drug Discovery Through a Unique Public-Private Partnership," J. Comm. Biotech., vol. 25, 4, (2020). doi:10.5912/jcb954.

likeness (how they can be absorbed), and their synthetic accessibility (ease of synthesis).⁵⁵ In an optimal setting, a pharmaceutical company might attempt to generate a molecule that has positive medicinal behavior while being druglike and easy to synthesize. However, this model can be easily trained to produce adversarial goals by simply changing the behavior to achieve negative results while keeping drug likeness and synthetic accessibility high, thus producing a dangerous and easy to synthesize drug. Notice that since much of the domain expertise is stored in the model architecture, this could be achieved by lesser experts as well, thus democratizing the process and making it much easier for adversaries to achieve such a goal.⁵⁶ Other cases for such an attack can include defense scenarios in which an adversary changes the objective of an autonomous systems to attack its own side or code generation models trained with the objective of producing vulnerable code. An example of a workflow that could be inverted through an Inverting AI Objective attack is provided in Figure 2.5.⁵⁷

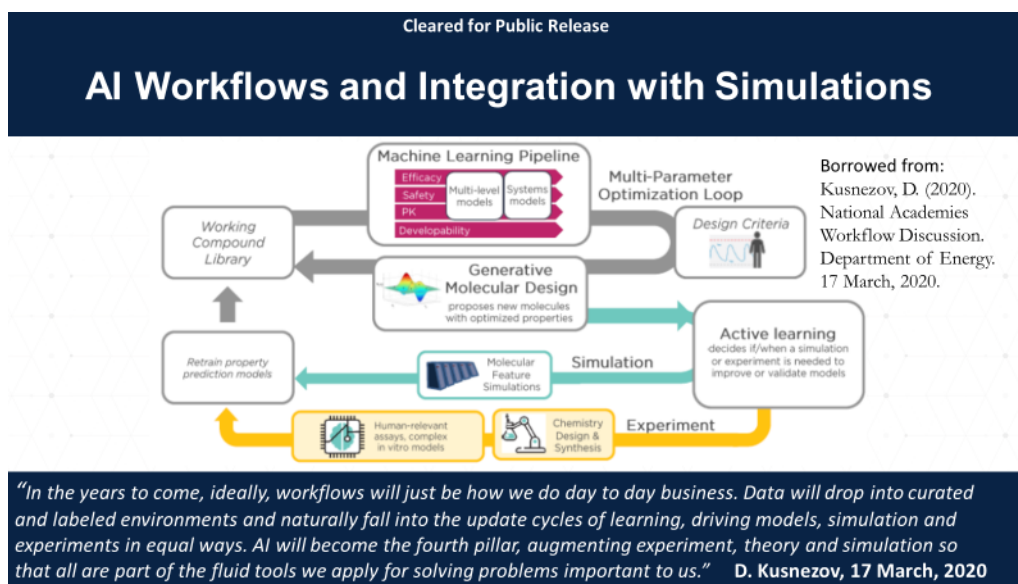


Figure 2.7. Example of an AI workflow that could be transformed through an Inverting AI Objective attack.⁵⁸

There are several ways an inverting AI objectives attack could be realized. One simple approach would be an inside threat in which a malicious person with access to the model and training procedure is able to alter the optimization function to achieve his or her goals. Another more sophisticated approach would require a cyber-attack to gain access to the objective function itself or to the entire workflow codebase and reestablishing it with the malicious objective function. Finally, given that many of these models are released publicly as open source for research purposes, an adversary might gain access to the models legally and simply retrain them with the new malicious objective function.

As this type of attack has not been thoroughly studied, there are not currently any known mitigation strategies besides the more traditional cyber defenses to ensure that the objective function is not altered.

⁵⁵ X. Zeng, "Deep generative molecular design reshapes drug discovery," (December 2022): [ScienceDirect.com | Science, health and medical journals, full text articles and books.](https://www.sciencedirect.com/science/article/abs/S0969212622000000)

⁵⁶ The model trained with the correct objective cannot be trivially inverted without retraining against the inverted loss — so non-experts can only do this if they're capable of editing the loss function and conducting a full training shot.

⁵⁷ Henninger, A. and Henz, B. (2023). Homeland Security Site Update: From Science to Operations. HPC User Forum Fall 2023. September 6-7. Tuscon, AZ.

⁵⁸ Kusnezov, D. (2020). National Academies Workflow Discussion. Department of Energy. March 17, 2020.



Other theoretical mitigation strategies could include designing networks in such a way that when trying to achieve a negative objective they do not perform as well. Alternatively, and depending on the form of the attack, one might be able to gain some insights into inverting AI objectives given access to test, evaluation, verification and validation data and results, past and current. More research is required to determine the effectiveness of these types of strategies.

3 Risks for Different DHS Domain Areas from Adversarial AI

All forms of AAI (i.e., AML, generative deceptive AI, inverting AI objectives) are capable of impacting DHS missions and the underlying technologies supporting those missions. As described in Section 2, these AI-driven models are capable of generating images, code, and even audio and video that humans have difficulty differentiating from truth. This provides a potential to easily spread misinformation or to gain access or avoid detection. This section explores the technology such as AR, CV, and NLP and functions (e.g., biometrics and C2ISR) underlying DHS missions to identify the potential for how they could be exploited by AAI for nefarious purposes. The conceptual relationships between the underlying technologies, functions, and missions are expressed in Figure 3.1.

3.1 Technologies

Because CV, AR, and NLP technologies underlie higher order functions (e.g., biometrics, C2ISR) and missions important to homeland security, the vulnerabilities in these underlying technologies open higher order functions and missions to these same vulnerabilities and subsequent risks. For example,

- CV vulnerabilities in baggage scanning applications in a transportation security use case are generalizable to cargo scanning applications in a CBP smuggling use case.
- AR vulnerabilities in a communications scenario for the Coast Guard are generalizable to communications scenarios for law enforcement.
- NLP vulnerabilities in immigration use cases could be generalizable to NLP vulnerabilities in emergency management use cases.

Because of these kinds of relationships, reviewing potential vulnerabilities at a technology level provides a broad AAI foundation that is applicable to many DHS missions. The following subsections 3.1.1, 3.1.2, and 3.1.3 review CV, AR, and NLP technologies and their AI-related vulnerabilities, respectively.

3.1.1 Computer Vision

This section reviews threats, example attacks, and defenses related to AML and generative deceptive AI in the context of CV.

3.1.1.1 Adversarial Machine Learning

While all of the four types of threats from AML (Data Poisoning, Evasion Attacks, Model Extraction and Model Inference/Privacy) are relevant to DHS in the area of CV, experts deemed evasion attacks as the primary one of concern, since a lot of DHS interests are detection focused and depend on model-based inference (e.g. devices running facial recognition software, Transportation Security Administration [TSA] checkpoints with hardware and software for detection of prohibited objects, etc.), which make them vulnerable against such attacks, under certain conditions.

Realistic Threats

The threat of an evasion attack against a CV model is potentially exacerbated by the fact that in general, models used in government settings have a longer lifecycle, especially on “edge devices” where updates might require manual intervention and thus be more difficult to apply. Hence, it is possible that “legacy” models remain in the field for longer periods of time, and it takes longer to patch them against new



attacks. Additionally, the fact that a lot of CV models are built by leveraging open-source model weights and the proliferation of such publicly available foundation models extends the attack surface. On the other hand, if the actual model in use is developed sufficiently beyond the openly available base model, the nature of typical DHS use cases mean that attacks based on exploiting the knowledge of the model are likely to face a smaller attack surface, reducing the threats from attacks based on model inference and model extraction.

Current and realistic threats to DHS missions are a result of longer-term legacy systems and the fact that many of the CV capabilities are based on open-source model weights as a “backbone.” The availability of information on these foundational models creates an attack surface. Even where DHS uses commercially available CV solutions, the limitations or weaknesses of these systems is not readily transparent, making it difficult to protect or mitigate attacks. Also, many problems arise in these models or across model architectures as a result of the training set. This leads to threats such as data poisoning and evasion attacks.

Example Attacks

Data poisoning takes patience where an adversary might generate poisoned images on the internet in the hopes that it will be captured in a future data set used for model training. Evasion attacks mainly depend on the training dataset as they are unable to properly detect certain threats if that scenario was not planned for or incorporated into the data set (e.g., 3D printing a shroud for a rifle, preventing it from looking like a rifle and not being picked up as a threat).

Other major concerns in this area are posed by physical adversarial attacks that trigger false negatives. The “patch attack”⁵⁹ consists of a pattern generated so that when printed and placed over an object or arranged in the environment, it results in an evasion attack. The “patches” are not necessarily flat: in some cases, the adversarial examples are 3D objects that create evasion attacks when viewed from certain angles or positions.⁶⁰ A different, but equally important problem, is posed by adversarial attacks that trigger false positives, such as decoy attacks that intend to distract or redirect resources, especially in the presence of an automated response. An additional consequence of repeated exposure to attacks is the undermining of trust. For example, generating repeated false positives quickly leads human operators to ignore the system or turn it off.

Observations on Defenses

Complete elimination of adversarial threats to current AI models is difficult because the models are complex, not well understood and inherently stochastic, making it difficult to correctly identify the root of the problem. In general, increasing robustness tends to decrease accuracy and thus performance of the model. Unsupervised models and statistical analysis may provide one way towards defending against adversarial attacks, particularly against model evasion and data poisoning. Examples include out-of-distribution detection, clustering and similarity checks against the training dataset and new images observed over time. However, there is reason to be skeptical of some of these proposed generic approaches, in light of results that detecting adversarial examples might be as difficult as correctly classifying them.⁶¹ After almost a decade of research and many thousands of papers published,⁶² there are no general defenses that can be relied on to work without qualification. Data augmentation and adversarial training remain the most effective approaches in general, although descriptions of these

⁵⁹ K. Eykholt, “Robust Physical-World Attacks on Deep Learning Visual Classification,” (April 2018): <https://arxiv.org/abs/1707.08945>.

⁶⁰ A. Athalye, “Synthesizing Robust Adversarial Examples,” (June 2018): <https://arxiv.org/abs/1707.07397>.

⁶¹ F. Tramèr, “Detecting Adversarial Examples Is (Nearly) As Hard As Classifying Them,” (June 2022): <https://arxiv.org/abs/2107.11630>.

⁶² Carlini, A. Complete List of All (arXiv) Adversarial Example Papers, <https://nicholascarlini.com/writing/2019/all-adversarial-example-papers.html>.





approaches' effectiveness range from "so effective that it is the de facto standard for training adversarially robust neural networks"⁶³ to "mitigate such attacks by not deploying ML solutions [...] adversarial training should be added as a valid misuse case [...]".⁶⁴ In the longer-term, it might be worth exploring the concept of "lifting" the data to a higher-dimensional space. For example, by training on multi-modal data or by adding explicitly defined features.⁶⁵ Practical drawbacks of such proposals include the additional data requirements, both in volume, modality, and structure or labeling. Despite the number of papers published on the topic of adversarial examples in CV, some fundamental questions seem to remain wide open, including how to properly measure and assess the magnitude of potential vulnerabilities.

3.1.1.2 Generative Deceptive AI

In the CV area, all three attack types based on generative deceptive AI (face morphing, deepfake images and videos, and foundation models containing a CV component) are relevant to the DHS mission. As in the case of AML, some of the relevant scenarios directly attack technology-based solutions (evasion of detection and identification at border crossings or in drug trafficking scenarios, faked images or videos to divert resources or fraudulently present false claims in disaster or 911 reports), while others work indirectly, by targeting the accuracy of AI models' outputs to reduce the operators' confidence in the models, or by negatively influencing the perception of truthfulness of publicly available information to sow distrust.

Realistic Threats

In a sense, generative deceptive AI has only become a realistic threat because of the ever-broader availability of computational resources and generative AI tools, but since the tools have multiple use cases, only some of which produce deceptive and potentially dangerous results, it is difficult to formulate approaches that would prevent such attacks purely within the realm of technology. For example, attacks targeting the general public can be thought of as mal/dis/mis-information (M/D/M) campaigns, and any potential solutions must consider issues of trust and authenticity. It is difficult to draw a line between synthetically generated content considered "good" and that considered "bad," so the intent needs to be considered as well. Focusing just on the authentication of content, the cryptographic techniques used to enable secure and safe communication on the internet, especially with sensitive content (such as banks, government, and medical institutions) can be used to provide tracking of provenance, certificates of authenticity, and digital signatures, but designing an infrastructure that navigates around both technology and politics remains a challenge. A centralized validation authority is politically difficult to implement. A related aspect is privacy: data provenance and cryptographic signature approaches may end up leaking a person's identity, making such a system potentially dangerous to dissenters and whistleblowers.

Example Attacks

As discussed previously, direct attacks in this domain area consist of submitting fake (deceptive synthetic) media (images, video, or other modalities, if relevant) to an ML-based automated system, either to avoid detection or to create a situation in which DHS resources are fraudulently committed, either for some type of benefit or to create a diversion and allow some other action or event to take place or avoid detection. An indirect attack could also take the form of a disinformation campaign against (all or parts of) the general public. Current synthetic image and video generators can be convincing on their own, but their output can often still be detected by paying attention to features that are more difficult for the models to get right; that is, complex structures where correlations between parts are not limited to small areas or parameterized just right (yet). An example in the context of imagery is that of hands, which, with the number and plausible arrangement of fingers still tends to be difficult to generate correctly and

⁶³ S. Rebuffi, "Fixing Data Augmentation to Improve Adversarial Robustness," (October 2021): <https://arxiv.org/abs/2103.01946>.

⁶⁴ Short et al. Defending Against Adversarial Examples, Sandia Report SAND2019-11748, 2019

⁶⁵ Ironically, this might be seen as a step back towards relying on methods of CV dominant before deep learning breakthroughs.



consistently. Similarly, in temporal media such as video, with complex and multi-level dependencies among the parts, it is difficult to keep the generated result coherent. However, in both of these examples, the generative models' output can be much more difficult to distinguish from “real” media (that have not been tampered with) if instead of generating the whole output from scratch—from just a textual prompt—we allow the model to take as an auxiliary input a “base” image or video that it is then prompted to modify. As an example, it is much easier for these models to learn how to modify the appearance of a face in the video and change the apparent words being said, if a “base” video of the person saying something else is available already. A deepfake face swap is an example of an attack that can already be dangerous with the current AI models' capabilities. With foundation models and multi-modal implementations such as (Contrastive Language-Image Pre-Training) (CLIP),⁶⁶ new risks might open-up (e.g., the joint text and image embeddings underpinning these models might be capable enough at some point that they can be leveraged to make novel inferences from a combination of visual and audio data that an adversary can exploit). To mention some very recent examples, there are now attacks on multimodal chat models that can use an adversarially generated image or sound in conjunction with text input to “jailbreak” an LLM^{67,68} (and indirectly inject instructions).

Observations on Defenses

Detectors of deepfakes and synthetic images exist and might be useful for detecting run-of-the-mill attacks using deceptive synthetic imagery but should by no means be relied on as a single point of failure. Given the previously cited result on the equivalence of detecting and classifying adversarial examples,⁶⁹ we might work under the assumption that eventually deepfakes will become undetectable and instead try to focus on validation and data provenance instead. The designers of some of the publicly available models have attempted to embed watermarks within the images generated by their models, but the current methods do not appear to be very robust.⁷⁰ In addition, public availability of generative models such as Stable Diffusion⁷¹ and the proliferation of specially tailored variants of the model makes a universally accepted watermark scheme unlikely.

3.1.2 Audio Recognition

This section reviews threats, example attacks, and defenses related to AML and generative deceptive AI in the context of AR.

3.1.2.1 Adversarial Machine Learning

Although all four types of threats from AML (data poisoning, evasion attacks, model extraction and model inference/privacy) are relevant to DHS in the area of AR, evasion attacks are considered the highest priority area of interest, since a lot of DHS missions could be jeopardized by targeted adversarial attacks, attacks which perturb audio inputs slightly through background noise unrecognizable to humans. These attacks not only cause algorithms to incorrectly classify the inputs but can also be used to control output. Such attacks can be slight perturbances in audio, embedding speech in non-speech (e.g., music), and forcing silence (having a model recognize audio as no output). In order of the priority of highest interest, model extraction, inference, and poisoning followed evasion.

⁶⁶ A. Radford, “[Learning Transferable Visual Models from Natural Language Supervision](https://arxiv.org/abs/2103.00020),” (February 2021): <https://arxiv.org/abs/2103.00020>.

⁶⁷ X. Qi, “[Visual Adversarial Examples Jailbreak Aligned Large Language Models](https://arxiv.org/abs/2306.13213),” (August 2023): <https://arxiv.org/abs/2306.13213>.

⁶⁸ E. Bagdasaryan, “[Abusing Images and Sounds for Indirect Instruction Injection in Multi-Modal LLMs](https://arxiv.org/abs/2307.10490),” (October 2023): <https://arxiv.org/abs/2307.10490>.

⁶⁹ F. Tramer, “[Detecting Adversarial Examples Is \(Nearly\) As Hard As Classifying Them](https://arxiv.org/abs/2107.11630),” (June 2022): <https://arxiv.org/abs/2107.11630>.

⁷⁰ Z. Jiang, “[Evading Watermark based Detection of AI-Generated Content](https://arxiv.org/abs/2305.03807),” (August 2023): <https://arxiv.org/abs/2305.03807>.

⁷¹ [GitHub](https://github.com/CompVis/stable-diffusion), “[Stable Diffusion](https://github.com/CompVis/stable-diffusion),” (November 2022): <https://github.com/CompVis/stable-diffusion>.



Realistic Threats

Research has identified the ability to perform targeted adversarial attacks on deep neural networks performing tasks such as classification. More notably, these attacks can be done in both a white-box and a black-box setting, meaning that the attacker can perform these attacks with or without needing to know the victim’s automatic speech recognition model’s structure or parameters, although not protecting the model would be more damaging. At the same time, keeping the model secret is not enough.^{72, 73} Various scenarios are at risk to include perturbation attacks both Over-the-Line and Over-the-Air. In an Over-the-Line attack the attack audio is passed to the model directly, as an audio file, for example. Conversely, an Over-the-Air attack requires the adversary to play the attack audio via a speaker towards the target voice processing system. It is not necessarily the case that an attack audio that is successful Over-the-Line will also be successful Over-the-Air. It has been proven possible to perform these attacks over the air such as in a Zoom/Teams meeting, or over the phone.⁷⁴

Observations on Defenses

Defenses against these attacks include simple anomaly detection methods such as comparing filtered and unfiltered waveforms to determine whether the input has been maliciously perturbed.⁷⁵ In addition to developing methods to identify perturbations, DHS could consider adversarial training and data augmentation to present perturbations in training, ensemble methods, or whitelisting input values. These are detailed in Appendix A.

3.1.2.2 Generative Deceptive AI

Generative deceptive AR deals with the ability for AI to perform text-to-speech tasks or information to audio tasks. Currently, deepfakes are being used by adversaries as a way to bypass these security measures. Audio deepfakes can sound very similar to the person imitated, both in terms of voice and style of speech. These algorithms do not need very much training and can be acquired relatively easily. Of the generative deceptive AI methods considered, the most concerning included audio deepfakes, synthetic speech generation and use of generative AI for whole-of-nation level audio misinformation.

Realistic Threats

These audio deepfakes to the ability of AI to produce audio that sounds very similar in both voice and the style of a particular person, poses a major threat to national security. Currently, models exist that can mimic the voices of people based on only a few audio clips for training but despite being lightweight, they can be used to cover a wide range of attacks/issues, which involve communicating via voice. These include but are not limited to spreading misinformation through representing a public/political figure;⁷⁶ spam calls, phishing, identity fraud, social engineering (mimicking family member voices);⁷⁷ deepfaking 911 calls, public defamation of politicians and other famous figures, identity fraud, social engineering, and more. Automated outbound calling combined with LLMs, synthetic speech voice over internet

⁷² M. Alzantot, “[Did you hear that? Adversarial Examples Against Automatic Speech Recognition.](https://arxiv.org/pdf/1801.00554.pdf#:~:text=To%20create%20adversarial%20examples%20for%20speech%20recognition%20models,input%20and%20possibly%20produce%20a%20desired%20target%20label)” (January 2018): <https://arxiv.org/pdf/1801.00554.pdf#:~:text=To%20create%20adversarial%20examples%20for%20speech%20recognition%20models,input%20and%20possibly%20produce%20a%20desired%20target%20label>.

⁷³ N. Carlini, “[Audio Adversarial Examples: Targeted Attacks on Speech-to-Text.](https://arxiv.org/pdf/1801.01944.pdf)” (March 2018): <https://arxiv.org/pdf/1801.01944.pdf>.

⁷⁴ H. Liu, “[When Evil Calls: Targeted Adversarial Voice over IP Network.](https://dl.acm.org/doi/10.1145/3548606.3560671)” (November 2022): <https://dl.acm.org/doi/10.1145/3548606.3560671>.

J. Zhang, “[Defending Adversarial Attacks on Cloud-aided Automatic Speech Recognition Systems.](https://deepai.org/publication/defending-adversarial-attacks-on-cloud-aided-automatic-speech-recognition-systems#:~:text=In%20this%20work%2C%20we%20propose%20several%20proactive%20defense,proposed%20strategies%20through%20extensive%20evaluation%20on%20natural%20dataset)” (July 2019): <https://deepai.org/publication/defending-adversarial-attacks-on-cloud-aided-automatic-speech-recognition-systems#:~:text=In%20this%20work%2C%20we%20propose%20several%20proactive%20defense,proposed%20strategies%20through%20extensive%20evaluation%20on%20natural%20dataset>.

⁷⁶ PBS News Hour, “[AI-generated disinformation poses threat of misleading voters in 2024 election.](https://www.pbs.org/newshour/politics/ai-generated-disinformation-poses-threat-of-misleading-voters-in-2024-election)” (May 2023): <https://www.pbs.org/newshour/politics/ai-generated-disinformation-poses-threat-of-misleading-voters-in-2024-election>.

⁷⁷ M. Siddiqi, “[A Study on the Psychology of Social Engineering-Based Cyberattacks and Existing Countermeasures.](https://www.mdpi.com/2076-3417/12/12/6042)” (June 2022): <https://www.mdpi.com/2076-3417/12/12/6042>.





protocol can help scale these attacks. One form of attack in this case regards giving information to the attacker to notify them that their content has been flagged, with this knowledge the attacker can do model extraction attacks to learn the constraints, alter the faked information, and bypass the security measures.⁷⁸ Most of these attacks (including misinformation spreading) can be carried out by using public means (i.e., the tools are generally available to the public), mainly social media.

Observations on Defenses

Deep neural networks such as ResNet can provide a solution to this and properly detect the altered pieces of media. In addition, organizations should not publicize whether something has been marked as faked information, denying attackers information that could be used to improve their attacks. Strengthening authentication for security services and monitoring government communication channels are good preventative efforts to stop these attacks from being successful. Humans and machines both exhibit similar capabilities in properly detecting deep faked audio as well as the pitfalls in detecting it.⁷⁹

3.1.3 Natural Language Processing

This section reviews threats, example attacks, and defenses related to AML and generative deceptive AI in the context of NLP.

3.1.3.1 Adversarial Machine Learning

While all of the four types of threats from AML (data poisoning, evasion attacks, model extraction and model inference/privacy) are relevant to DHS in the area of NLP, evasion attacks are considered the most important to understand in the near term. Many NLP systems are brittle to regional dialect, use of codewords, or homographic swapping, which an adversary could exploit to evade the system without directly targeting it.

Realistic Threats

The rich nature of natural language widens the waterfront of possible attacks, with malicious actors able to manipulate or deceive NLP-based systems. These threats encompass techniques like data poisoning to introduce bias or offensive content⁸⁰ or can use model evasion to avoid detection that relies on NLP. For example, spam filters or content moderation algorithms can be tricked into allowing harmful content. AI-based analysis can be used to extract sensitive information from text data, posing privacy threats when used maliciously.

Observations on Defenses

Safeguarding NLP applications at DHS necessitates robust defenses to counteract these adversarial strategies and to ensure the integrity of language-based security measures. Understanding the risks and benefits of using open tools and updating models is a good way to start improving security as well as responsible use and governance of LLM. Other defenses could include adversarial training, robust modeling architectures with built-in defenses (e.g., using attention mechanisms, ensembles, etc.), combining AI with human moderators, anomaly detection, to name a few. These are discussed more in Appendix A.

⁷⁸ Z. Khanjani, “[Audio deepfakes: A survey.](https://www.frontiersin.org/articles/10.3389/fdata.2022.1001063/full)” (January 2023): <https://www.frontiersin.org/articles/10.3389/fdata.2022.1001063/full>.

⁷⁹ N. Muller, “[Human Perception of Audio Deepfakes](https://arxiv.org/pdf/2107.09667.pdf)” (October 2022): <https://arxiv.org/pdf/2107.09667.pdf>.

⁸⁰ H. Jones, “[If You’ve Trained One You’ve Trained Them All: Inter-Architecture Similarity Increases With Robustness.](https://spectrum.ieee.org/tech-talk/artificial-intelligence/machine-learning/in-2016-microsofts-racist-chatbot-revealed-the-dangers-of-online-conversation)” (August 2022): <https://spectrum.ieee.org/tech-talk/artificial-intelligence/machine-learning/in-2016-microsofts-racist-chatbot-revealed-the-dangers-of-online-conversation>.



3.1.3.2 Generative Deceptive AI

In NLP, adversaries are leveraging generative deceptive AI in two ways: style morphing/transfer/faking authorship and misuse of LLMs.

Realistic Threats

Style morphing is where an AI program writes a document or piece of media in the same style as the faked author, and in so doing, spreads misinformation. This capability could be used in influence operations. Vulnerabilities inherent in LLMs could allow bad actors to access information from the ML training data (via model interrogation), giving them access to subject matter expert knowledge. Although possession of this information is not illegal, creating a higher barrier to gain this knowledge should be considered. Using LLMs to generate content at scale could also lead to a distributed denial-of-service (DDOS)-like attack on systems. There is the potential to contaminate training data with synthetic or poisoned data on a massive scale. AAI can use NLP to generate convincing phishing emails, messages, or scams by mimicking human communication patterns and crafting persuasive text. NLP-based systems can be exploited to automate the generation of spam content or manipulate online reviews, ratings, and comments. Also, NLP-powered chatbots or social media bots can be deployed for malicious purposes at scale.

Observations on Defenses

Many of the techniques used in defending NLP from AML can be applied to NLP vulnerabilities due to generative deceptive AI as well. Additional mitigations include content moderation to detect and filter out harmful or adversarial content in near real-time, behavior analysis that monitors user interactions, and behavior patterns to detect abnormal activities, and ethical guardrails to ensure that AI systems do not propagate harmful content. These are discussed more in Appendix A.

3.2 Functions

In the context of this report, a function represents an intermediary data processing requirement between technology (inputs to functions) and missions (supported by functions). While there are likely others, two of the functions considered the most commonly used and susceptible to AI vulnerabilities are biometrics and C2ISR. And, like the technologies reviewed in section 3.1, the vulnerabilities in these functional support processes may impact multiple higher order mission objectives. For example,

- Vulnerabilities in biometrics applications used in TSA use cases can generalize to a CBP use case.
- Vulnerabilities in C2ISR applications used in law enforcement use cases can generalize to emergency management use cases.

Similarly, and by extension, vulnerabilities in functional applications inherited from vulnerabilities in underlying technology can also generalize across missions. Thus, reviewing vulnerabilities at a functional level provides a broad AAI foundation that is applicable across many DHS missions. Subsections 3.2.1 and 3.2.2 review biometrics and C2ISR functions, respectively.

3.2.1 Biometrics

Biometrics is a field of study that focuses on the measurement and analysis of unique physical and behavioral characteristics of individuals. Typically, these characteristics are used to aid security measures making it difficult for a bad actor to gain access to anything protected by these measures. Biometric traits (e.g., fingerprints, facial features, iris patterns, voice, and gait) have traditionally been difficult to impersonate. It is worth noting that biometrics seems to be the exemplar for where AML and generative deceptive AI intersect, as adversarial attempts to avoid detection by automated recognition software are



often aided by products developed through generative AI. Because many of the biometric traits are image-based, biometrics is highly dependent on CV, as indicated in Figure 1.2.

Although all of four types of threats from AML (data poisoning, evasion attacks, model extraction and model inference/privacy) are relevant to DHS in the area of biometrics, evasion attacks are of primary concern, since a lot of DHS interests are focused on threats against the ability to detect persons, and verify identity (e.g., cooperative identity verification at borders, approaches, airports; noncooperative identity verification using subpoenaed materials; and, non-existing persons being created from stolen information). Bad actors are now able to either bypass these security measures or completely avoid them all together using a variety of attacks.

3.2.1.1 Realistic Threats

This threat is already somewhat mature, and there are a multitude of existing methods for a bad actor to bypass biometric security measures such as:

- Presentation attacks at the sensor (a fraudulent attempt to deceive a biometrics system by presenting it with fake or manipulated biometric sample with the goal of impersonating or falsely authenticating as another individual, thereby gaining unauthorized access to a system).
- Injection / playback attacks (an attempt to inject synthetic imagery such as deepfakes into a data stream to impersonate an authorized user and gain unauthorized access to a system).
- Enrollment attacks (a fraudulent attempt to compromise a biometric system during the phase in the process in which a legitimate user's biometric data is initially recorded and stored in the system for future authentication or identification purposes).
- Theft or modification of stored reference data (attack in which attacker gains unauthorized access to the biometric templates or reference data stored in a biometric system's database).
- Morphing attacks (attempt to manipulate or combine two or more biometric samples from different individuals to create a single morphed data set that can deceive a biometric recognition system).

3.2.1.2 Observations on Defenses

DHS missions depend on robust and reliable solutions for face and fingerprint recognition, and there is also interest and applications for iris, voice, and DNA. The expanded use of face recognition for remote identity verification has introduced new challenges where recognition and authentication processes now must also question the authenticity of the face images and associated evidentiary documents. Detection capabilities for presentation attacks, morphs, and deepfakes have improved but still operate with false alarm rate that make them impractical for large open set environments and uninterpretable for front line human inspection operations. Moreover, Related challenges exists for establishing the authenticity of identity documents and vital records from a myriad of domestic and international sources that DHS must verify or adjudicate.

Defensive measures will require ongoing innovations in technology and processes. NIST's Face Analysis Technology Evaluation (FATE) provide detailed evaluations of presentation attack detection and morph detection to supplement the central Face Recognition Technology Evaluations. The National Science Foundation's Center for Identification Research performs fundamental research in biometrics that includes liveness, and detection of morphs and deepfakes. In the U.S. Department of Defense, the Defense Advanced Research Projects Agency Semantic Forensics (SemaFor) program is pursuing detection, characterization, and attribution of generated and manipulated digital media. And the Intelligence Advanced Project Agency has supported research on presentation attack detection. These research efforts have motivated new commercial products and service that seek to make biometric-based identity verification functions stronger, more robust and reliable. In turn, sustained evaluation and integration efforts will strengthen DHS missions.



3.2.2 Command and Control and Intelligence Surveillance and Reconnaissance

C2ISR originated as a term used to describe a comprehensive set of capabilities and function crucial for effectively executing military operations, but the concept applies equally well to many DHS operations engaged in decision making, planning, and coordination activities while attempting to manage assets and forces. The ISR portion is highly dependent on CV, AR and NLP; and the C2 portion is highly dependent on NLP and AR, and to a lesser extent CV. Like biometrics, it is also a strong candidate for combining AML and generative deceptive AI as a common strategy towards corrupting these systems involves synthetic data inputs where a bad actor will use artificially generated data and inject that into the systems ML training algorithm, either preventing detections or triggering false detections. This can be especially dangerous during a rapid response scenario as diverting resources even for only a couple of minutes could be the difference between a good and a bad outcome.

3.2.2.1 Realistic Threats

Beyond the threats organically inherited from use of any CV, AR, or NLP-supported technologies, there are potential vulnerabilities in the C2ISR sphere related to use of AI or AAI attacks. For example, missions leveraging public infrastructure in a disaster response scenario inherit opportunities for AAI through that infrastructure. Public unsecured sensor feeds, that are likely less secure than DHS-owned systems, could potentially leak private or sensitive information, giving AAI a strong advantage at the edge. And, importantly, certain AI-enabled C2 applications rely on simulations as a training environment, which allows bad actors to target specific simulations and modify the “rules of reality,” which results in reductions in AI performance.

Generative deceptive AI-generated content can undermine situational awareness capabilities by providing false and/or anomalous signals or cause communication disruptions. Misinformation campaigns can also adversely affect C2 planning and response. Through the use of generative deceptive AI, surveillance and reconnaissance systems could be rendered ineffective or be used against their organization. Deceptive AI-generated content can undermine situational awareness capabilities by providing false and/or anomalous signals. Deceptive AI-generated content can cause communication disruptions through false signals that overwhelm networks or responders. Misinformation campaigns can adversely affect C2 planning and response (e.g., misallocation of resources in a rapid response scenario where there might not be sufficient time to determine the veracity of inputs). Some of that misinformation might be associated with environmental variables that factor into mission planning.

3.2.2.2 Observations on Defenses

Most of the current research in this domain focuses on the corruption of intelligence inputs. DHS might consider developing a general trust model on any digital input/source, which would help address deceptive AI-generated content. DHS could also focus on determining the current vulnerabilities of existing DHS C2 systems against AI-generated content and determine the best course of action to counteract them. In addition, DHS should only use data sources and environments from trusted sources; using open-source information makes it easier for an attacker to view and look for vulnerabilities.

Other DHS activities in this area should focus on establishing red/blue teaming approaches for AI-enabled systems to view existing vulnerabilities, develop ways of maintaining and governing AI applications post deployment, create an inventory of assets and analyze them to determine which ones will become AI-enabled in the future, define accountabilities and protections for AI application owners, build in mitigations to limit the harm that AAI can cause, and develop a measure of consequentiality of systems for various missions and use that to determine what level of assurance is required for AI applications.



3.3 Missions

This section reviews DHS’s major missions, considers how they depend on technologies and functions (described in Sections 3.1 and 3.2, respectively), and extends the vulnerabilities in those sections to postulate notional adverse mission risks.

3.3.1 Preventing Terrorism

Given that AI could be used to support this mission from a micro (individual) to a macro (societal) level, the impacts of AAI could be just as broad. AAI could enable terrorist organizations to:

- Develop (or commandeer) swarms or drones capable of carrying out attacks,
- Design AI-based malware to execute more sophisticated cyber-attacks,
- Generate more convincing deepfake videos and audios facilitating the spread of M/D/M-information or recruit sympathizers, and
- Learn how to build bioweapons from LLMs.

Importantly, AI can enable any of these nefarious uses at scale. Both biometrics and C2ISR are critically important to counterterrorism efforts. AI can enhance surveillance and reconnaissance by acquiring and analyzing vast amounts of data to detect suspicious activities, identify potential threats, and predict terrorist activities. Biometrics can also be used to support facial recognition applications and behavioral analysis. Thus, any AAI activities targeted on these two functions as they support counterterrorism (e.g., generating fake data to evade detection by automated information surveillance and reconnaissance (ISR) algorithms, using morphing GANs to develop dual-use passports, etc.) could impact the mission to prevent terrorism.

3.3.2 Securing the Border

AAI could have significant impacts on the CBP mission, advancing the ability of nefarious actors to:

- Generate deepfake identities or morphed passports used to deceive officers or facial recognition systems allowing nefarious actors to cross the border undetected,
- Develop deceptive imagery or newsfeeds to confuse U.S. authorities and cause them to sub-optimally direct resources,
- Smuggle contraband through the use of drones that can evade traditional detection methods,
- Use AML methods associated with NLP to create security breaches such as a license plate reader that is not able to accurately detect plates or systems that improperly misses the identification of a person of interest,⁸¹ and
- Given ISR systems using foundation models for image recognition of border security events, an adversary might be able to use inference attacks (e.g., data reconstruction or model inversion) to ascertain what objects the model is looking for and which ones could later be used for evasion of such systems.

Importantly, AI can enable any of these nefarious uses at scale. Both biometrics and C2ISR are critically important to border security missions, as described in the preceding first, second and fifth bullets.

3.3.3 Enforcing Immigration Laws

AAI could have significant impacts on the enforcement of immigration laws advancing the ability of nefarious actors to:

⁸¹ T. Brewster, “[This AI Watches Millions Of Cars Daily And Tells Cops If You’re Driving Like A Criminal.](https://www.forbes.com/sites/thomasbrewster/2023/07/17/license-plate-reader-ai-criminal/)” (July 2023): <https://www.forbes.com/sites/thomasbrewster/2023/07/17/license-plate-reader-ai-criminal/>.



- Generate deepfake identities or morphed passports used to deceive officers and make it difficult to authenticate immigration documents and personal information,
- Coordinate and facilitate human trafficking operations, potentially allowing criminals to avoid detection,
- Use AI-based malware to gain unauthorized access, manipulate records or compromise sensitive information in immigration databases, and
- Use AI-driven chatbots or deepfake videos to deceive immigration officers during interviews or interactions, hindering their ability to detect inconsistencies in statements.

Importantly, AI can enable any of these nefarious uses at scale. Biometrics is critically important to immigration missions, as described in the preceding first bullet. In this mission, generative deceptive AI, including deepfakes and misuse of LLMs is probably of the greatest concern. In the future and given the potential for the adoption of foundation models in this use case, AI-based attacks related to foundation models could also be a concern. For example, the use of a public facing LLM could be susceptible to inference attacks or data poisoning.

3.3.4 Securing Cyberspace

The relationship between AAI and cybersecurity is multifaceted: AAI can be used to attack traditional cybersecurity defenses and traditional cybersecurity attacks can be made on AI-powered systems. This section focuses on the former, reviewing both how AML and generative deceptive AI may impact cybersecurity defenses. Importantly, the current ineffective state of cyber defenses reduces an adversary's motivation to invest in AAI to guide cyber-attacks. However, improvements in cyber defenses, particularly at machine speed, make the sort of large training data set needed to train an AAI easier to build and curate.

Although the four attack pathways for AML discussed previously (data poisoning, evasion attacks, model extraction and model inference/privacy) are relevant to DHS's cybersecurity domain, evasion and data poisoning are considered the more concerning threats of the four.⁸² Other realistic threat topics also included:

- Poison training data that can result in poor performance of AI-enabled systems. The large amount of data used to train AI systems presents both AAI vulnerabilities and additional cyber vulnerabilities. The data poisoning attack against VirusTotal was an example of fake cyber threat intelligence being used against cyber defenses. Future attacks are likely to also target people involved with cyber defenses, particularly as humans are added to processes as a key defense against algorithmic cyber weapons.
- Extract a trained model from an ML-enabled system, allowing an adversary to explore its weaknesses at leisure to find vulnerabilities to exploit.
- Reverse engineer data used to train ML models,
- Conduct side channel attacks on AI-based models, and
- Use generative deceptive AI-assisted tools to quickly find ways to weaponize them in the virtual world.
 - Malware can be generated through LLM's, making it significantly harder to keep up with. Solutions for malware are useful only for individual programs, so having adversaries capable of creating multiple programs quickly aided by AI could prove to be dangerous.

⁸² MITRE's ATLAS, which provides a detailed taxonomy for the adversarial threat landscape for artificial intelligence system, was identified as a good resource for navigating these threats to ML.



- Use face morphs or deepfakes to gain access to systems.⁸³
- Use deepfake technology to generate and disseminate, through spearfishing or other social engineering attacks, attacks designed to defeat the decisions made by the human defenders.

Current research on defense against generative deceptive AI aims to develop and use an AI-enabled cyber defense system, however, this is still state-of-the-art research. Several avenues for mitigations against AML attacks to include red teaming, train AI models on adversarial examples, make AI models intrinsically more robust, use good cyber hygiene, watermark training data, move detection to behavior-level (regarding VirusTotal, for example), use trusted data sources. These mitigations are detailed in Appendix A.

3.3.5 Safeguarding Critical Infrastructure

Importantly, creating adverse effects on critical infrastructure (CI) does not require causing full outages. Inefficiencies, misinformation, and sub-optimal decisions would effectively subvert efficiency and productivity over time. The distributed ownership and operations of CI requires trust in the accuracy and correctness of the remote sensors or Internet of Things (IoT) devices, all of which are attack surfaces. Additionally, over reliance on AI-enabled security systems could result in missed detections or false alarms, and these systems can easily be beaten through different means. In a test involving the U.S. military, for example, Marines could do cartwheels, wear animal skin print, and camouflage as trash cans to counter AI-based detection systems.

AAI could have significant impacts on critical infrastructure, advancing the ability of nefarious actors to:

- Disrupt, degrade, and/or deny activities related to IoT devices, as typical IoT devices do not contain security monitoring software and therefore provide persistent access to threat actors.
- Target many different resources including administrative services, leading to suspended billing services and exposure of private information, and could reveal operation/maintenance procedures.
- Leverage weaknesses in data collection approaches through known flaws in both the foundation models and the data used to train the AI.
- Spread misinformation such as fake messages from a person's supervisor, false insurance claims, inaccurate sensor data and inaccurate news articles.
- Make people believe that the information is an authentic source when it has been spoofed.
- Mimic many things and pieces of information and deep fakes allow for someone to look, sound, and talk like someone else, making any viewer or listener think that it is legitimately the person they are attempting to masquerade as.
- Generate and pass poisoned data into a sensor, having that sensor believe that it is real data when in fact it was artificially manufactured.

To prevent these attacks, certain measures must be taken, including monitoring for covert and overt command and control signaling to/from IoT devices, and utilizing blockchain and zero-knowledge proof technologies to cryptographically validate the integrity of distributed sensor data. Additionally, the use of homomorphic encryption will allow for computation on encrypted data, giving the user encrypted results, without the information ever being decrypted, increasing privacy. To preserve the safety of these pieces of critical infrastructure, a slow and deliberate adoption of AI-enabled technologies should be introduced in addition to a harm analysis for all AI-enabled system elements. Independent T&E and continuous monitoring of such systems are required to ensure that vulnerabilities do not arise and that systems remain up to date.

⁸³ As described earlier and reiterated here, the combination of AML and generative deceptive AI is particularly powerful. In this instance, leading to a scenario where an adversary can create malware and have an easy means of injecting it, without having to rely on a security failure or social engineering.



3.3.6 Emergency and Disaster Management

AAI could have significant impacts on emergency and disaster management, advancing the ability of nefarious actors to:

- Target and disrupt communications systems used by emergency responders, hindering their ability to coordinate and respond effectively during a crisis,
- Use AI-generated deepfakes or misinformation to spread false alerts, causing panic and diverting resources,
- Design and execute AI-generated malware to launch attacks on emergency services, and
- Compromise or manipulate AI-powered autonomous systems (e.g., drones, robotic search and rescue, etc.), rendering them ineffective or potentially dangerous.
- Create false photos of post-disaster property damage, making analysis more difficult.

Importantly, AI can enable any of these nefarious uses at scale. C2ISR is critically important to emergency and disaster management mission, as described across the preceding bullets.

3.3.7 Transportation Security

AAI could have significant impacts on transportation security, putting critical transportation infrastructure and passenger safety at risk. Notional examples of what nefarious actors using AI could do include:

- Design and execute malware with LLMs to launch sophisticated cyber-attacks on transportation systems potentially disrupting services or causing accidents,
- Generate deepfakes or morphed passports to create counterfeit identification, making it harder for officers to detect individuals with malicious intent when trying to enter secure transportation areas,
- Evade surveillance systems at transportation hubs, allowing nefarious individuals to bypass security checks or enter restricted areas undetected,
- Corrupt traffic management analysis, reducing safety and efficiency on roads and at transportation hubs, and
- Add a certain object to a suitcase with a prohibited item making the prohibited item undetectable to an AI-based X-ray system.

Importantly, AI can enable any of these nefarious uses at scale. Biometrics is critically important to transportation security missions, as described in the preceding second bullet.

3.3.8 Law Enforcement

AAI could have significant impacts on law enforcement activities, advancing the ability of nefarious actors to:

- Propagate deepfakes, spreading false information such as fake law enforcement statements or videos, leading to public confusion, mistrust, and potential unrest,
- Evade surveillance systems, making it harder for law enforcement agencies to track and apprehend suspects,
- Deceive victims in online fraud schemes or impersonate law enforcement personnel with AI-powered chatbots,
- Compromise sensitive data and disrupt operations through attacks targeting law enforcement databases or communications systems,
- Generate and deliver high volumes of deepfake phone calls to 911 emergency services to essentially conduct a DDOS attack making 911 services unavailable,



- Use AI-driven tools, including voice synthesis and anonymization techniques, to make fraudulent emergency calls or messages, often with the intent of causing fear, harm, or chaos. Known as “AI-enabled swatting,” this is a dangerous and malicious practice that leverages AI and other technologies to initiate false emergency responses, such as armed police SWAT, or special weapons and tactics, teams, to unsuspecting victims’ homes, and
- Use LLMs to more easily access information from the training data, allowing criminals to access subject matter expert knowledge. Specialized information can now be aggregated and explained to the masses. Democratizing expertise, predicting what’s next. This is a capability of LLMs but can be used for harmful purposes when the user has harmful goals.

As seen in other mission sets, AI can enable any of these nefarious means at scale. Biometrics and C2ISR are critically important to law enforcement missions, as described in the preceding first and fourth bullets, respectively.

4 Implications of AAI on Emerging Technologies

AAI technology has impacted a variety of adjacent technologies, existing and yet to be productized. For example, deep learning methods have proven to be effective for malware detection, yet because they are vulnerable to adversarial attacks, the malware detection models based on these deep learning methods also face the threat of adversarial attacks. In genomics, LLMs have been used to demonstrate direct inference of full atomic-level protein structure from primary sequence. Also, using a generative model, it was possible to obtain a universal representation of epidermal differentiation and use this to predict the effect of cell state perturbations on gene expression at high time-resolution. Yet, because these models are vulnerable to adversarial attacks, their applications to genomics share this vulnerability and are potentially impacted by AAI. In advanced manufacturing, materials scientists want to know all the different recipes they can use to produce a specific material. Exploring the fiber (set of recipes with the same result) containing all their options, they can choose the individual recipe that works best for a given manufacturing set-up, time, and resource constraints. This problem is particularly important in advanced manufacturing processes, including innovative solid phase processing techniques. Although there are good tools in ML for predicting output from input, methods for learning all the different inputs that can yield a specific output is possible using generative AI models, which are vulnerable to AAI.

In a more futuristic scenario, we expect AAI to pose larger risks, as AI gets integrated into a variety of adjacent technologies. For example, whether it supports smart infrastructure or smart cities, in the future Internet of Intelligent Things (IoIT) everything is not only inter-connected but also intelligent. Like the current day, the “things” come from different manufacturers, with different designs, capabilities, and purposes. Unlike the current day, these “things” will have an insatiable appetite for data from other “things.” Moreover, software and AI algorithm updates across these “things” will happen asynchronously and at a speed of need, such that no individual thing is able to rely on preconceived assumptions about behaviors or performance of the other “things” it interacts with. Although these scenarios barely even exist, we know that they will be impacted by adversarial attacks and that these IoIT technologies must be designed to be robust against AAI. This section explores the future of these and other nascent AI-based technologies, the potentially inherent fragilities in these technologies that could be exploited by AAI for nefarious purposes, and how these exploits might impact homeland security missions of the near and distant future.

4.1 Advanced Persistent Threat Detection, Malware Generation, and Insider Threats

Adversaries who possess sophisticated levels of expertise and significant resources can create opportunities to achieve their objectives by using multiple attack vectors (e.g., cyber, physical, and



deception) in a long-term, persistent way. These advanced persistent threats (APTs)⁸⁴ are cyber-attacks carried out by well-resourced and sophisticated adversaries who target organizations with the goal of gaining strategic advantage by exfiltrating data or by disrupting operations.

There are multiple opportunities for APTs to use AI-based attacks to achieve their attack goals, which could be targets in the supply chain (e.g., development organization), where the desired asset is located (e.g., government agency), or a combination of threats to achieve the desired effect. APTs can be applied during:

- Reconnaissance (e.g., scanning media for targets filtering by activity, scanning social media for interests/intents of targets (“pretexting”), and avoiding perimeter defenses to scan target systems for configurations, etc.),
- Weaponization (e.g., intelligent fuzzing to discover zero-day, generative techniques to evolve existing vulnerabilities, etc.),
- Delivery methods such as phishing (e.g., avoid detection and filtering by perimeter systems, successfully motivate poor behavior through crafted messages, etc.),
- Exploitation and installation (e.g., ghosting presence on target through observation and mimicry, etc.),
- Command and control (e.g., adjust to observed traffic patterns for communication, etc.), and,
- Actions on objective (e.g., on-the-edge analysis of potential assets, etc.).

Similarly, but in the generative deceptive AI realm, APTs can use AAI to do things like promulgating M/D/M information (e.g., scanning media for hot-button topics and crafting APT actor’s messages to align with those topics, etc.), generating false personas (e.g., simulacra at scale to align with community, images such as faces, surrounding artifacts—collect and duplicate paragons, etc.). They can apply these technologies with focused political ends (e.g., altered / fabricated video, audio, images, etc.), subvert guard rails of LLM systems (get information that should be hidden), subverting output of LLM systems (cause answer drift to align with objectives).

In the realm of malware generation and LLMs, DHS must be aware of threats stemming from the generation of new malware. For example, the generation of potential phishing emails to install malware (and to help establish pretexting for fishing emails); analyze of the family and style, and potentially, pedigree and genealogy of a piece of malware; determining or explaining the operation of the malware; examine “lint” or other trails suggesting the existence of malware in an operational or development environment, including open-source development; decompile binaries into source for analysis; perform program source code analysis to detect vulnerabilities, information leaks, weaknesses, or backdoors; obscuring or “de-obscuring” source code; query engineering used to disable guardrails limiting generation of malware; and query injection used to generate functional code that contains vulnerabilities or weaknesses (i.e., deliberately misleading training).

Insider threats refer to security risks posed by individuals who have authorized access to an organization's systems, data, or facilities (e.g., employees, contractors, or other trusted parties). The relationship between insider threats and AAI lies in the potential for malicious actors, who have insider access or knowledge, to leverage adversarial techniques to exploit vulnerabilities within AI systems. Although historically low in number, these threats can have high impacts, including:

⁸⁴ The advanced persistent threat: (i) pursues its objectives repeatedly over an extended period of time; (ii) adapts to defenders' efforts to resist it; and (iii) is determined to maintain the level of interaction needed to execute its objectives. “[NIST 800-39: Managing Information Security Risk.](https://nvlpubs.nist.gov/nistpubs/Legacy/SP/nistspecialpublication800-39.pdf)” (March 2011): <https://nvlpubs.nist.gov/nistpubs/Legacy/SP/nistspecialpublication800-39.pdf>.



- **Insider knowledge:** Insiders with knowledge of an organization's AI systems may have insights into vulnerabilities that could be exploited using AAI techniques. They may understand the weaknesses of the AI models and the specific points of manipulation.
- **Data manipulation:** Insiders with access to training data for AI models could potentially manipulate or poison the data used to train the models, leading to biased or compromised outcomes. This could affect decision-making processes based on AI-generated insights.
- **Model tampering:** Insiders with access to AI model parameters or deployment processes might attempt to manipulate model behavior by introducing adversarial inputs. This could lead to incorrect predictions or decisions.
- **System exploitation:** Insiders could use AAI to exploit vulnerabilities in AI-powered security systems, bypassing authentication or intrusion detection mechanisms.
- **Intellectual property theft:** Insiders could use adversarial techniques to extract valuable information from AI models, such as proprietary algorithms or sensitive data, and exfiltrate it.

Detecting insider threats using AI requires a significant amount of data—data that is not sufficiently available. To date, anomaly detection has not been successful in identifying and detecting insider threats. AI does, however, provide adversaries an attack vector to collect insights and information about humans (organization, visuals, voice recordings), which is critical to the social engineering required for successful attacks. AAI can be used to collect significant social information about individuals, their biometrics, behaviors, and affiliations. It can also be used in the context of malicious elicitation, such as being used to collect business information using social networks like LinkedIn. If an individual is approached for a job offer, they might provide insights that will make targeting and recruiting much easier. Publicly available information provides a rich set of data for adversaries to learn about people, their business roles, and relationships. Generative AI methods, when augmented by a deeper social and behavioral understanding of individuals and how they operate within their organizations, can provide the impetus for individuals to act in harmful ways, both knowingly and unknowingly (e.g., receive tasking in a voicemail they believe is from their supervisor).

Insider threat detection programs currently focus on cyber indicators that differentiate malicious from non-malicious employee-generated computer activity. These data analytic tools (e.g., sentiment analysis in email, user activity monitoring) limit the focus to cyber indicators. AAI would rely on cybersecurity breaches to gain access to affect the organizational data or the performance of the tools themselves. Behavioral tools (e.g., employee reporting, analysis of financial strain, position risk analysis) also provide risk indicators and those who rely on external information sources (e.g., address, relationships, social networks) are susceptible to AAI manipulation.

The relationship between insider threats and AAI highlights the need for comprehensive cybersecurity strategies that address both external and internal risks to AI systems and data.

4.2 Satellite Imagery

The adversarial relationship with respect to overhead observations is one that has a long history, predating the use of satellite imagery. The adversarial role is often a relationship of evasion and using means of either defeating the physical sensor or evading observation. At times, nature in the form of fog, rain, smoke, etc. can aid in evasion as well. As a result, multiple spectra are used in the visible and other domains and other sources of observation including radar, electronics emissions, etc. Multiple approaches are used to obfuscate or make indeterminate the current position of satellites and their current field of view and to limit knowledge of their fields of regard and their current level of inevitable degradation due to the hostile environment in which they operate.



Recent advances in commercial access to satellite imagery such as that offered by the Planet Labs' constellation, give rise to emerging opportunities for adversaries. As the cost of launching small constellations to low earth orbit (LEO) continues to be reduced it is not unreasonable to expect that additional services, similar to those offered by Planet Labs, will emerge in the U.S. and other markets with the potential for use by an adversary. As the availability of real-time satellite imagery proliferates and the cost is continually reduced, it is a fair assumption that the number of adversaries using satellite imagery will grow.

The impacts of AAI on the mission spaces using satellite and other overhead imagery could be wide ranging and varied.

- A key mission space utilization of satellite imagery is target tracking and detection of anomalous events, changes in patterns, etc. Although AI can help analyze imagery to track, detect, and inform, the same approaches could be used by an adversary using commercial data sources.
- AAI could be used to devise patterns of movement that could defeat traditional mathematical tracking approaches or make maintaining track custody across sensor gaps difficult or misleading.
- AAI could be used to develop measures to defeat these tipping and cueing⁸⁵ mechanisms, including altering background estimation functions or development of patterns of movement to make cueing operations ineffective.

Defending against an AAI attack on satellite imagery could take on many forms depending on the nature of the attack. The nature of these defenses can also be informed by the long adversarial history of satellite imagery. As an example, if a sensor of one modality is defeated such as one operating in the visible spectrum, other sensors of differing modalities could be used (e.g., infrared, EM, or radar). Generation of fused tracking solutions using differing physical sensors could be used to defeat attacks focused on a single physical sensor. This same approach could also be used when generative AI scenes are being received from a source.

Also of concern are the multiple of points of attack along the pathway from the sensing of a physical phenomenon to the display of targets of interest to mission personnel. The aforementioned processes of tracking, tipping, cueing, scheduling, etc. are not singular processes but are each complex system-of-systems. Opportunities for AI (and similarly AAI) of tracking systems could lie not only in the physical sensor, but also in the geolocation, track propagation, target identification, target detection, background estimation, and many other subsystems.

4.3 Foundation Models

With the introduction of transformers in 2017,⁸⁶ a new paradigm in AI was formed, foundation models. Foundation models are large-scale models that are trained in a self-supervised manner on broad sets of unlabeled data and are capable of being rapidly adapted to downstream tasks. These foundation models gained popularity with LLMs such as Bidirectional Encoder Representations from Transformers (BERT)⁸⁷ and variants of GPT⁸⁸. Shortly after, vision transformers (ViT)⁸⁹ gained popularity quickly

⁸⁵ Tipping and cueing are the combined process of observing an area, detecting an item of interest, and cueing one or more additional sensors to observe the same phenomena or to observe the expected position of a tracked object.

⁸⁶ A. Vaswani, "[Attention Is All You Need.](http://arxiv.org/abs/1706.03762)" (August 2023): <http://arxiv.org/abs/1706.03762>.

⁸⁷ J. Devlin, "[BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.](https://doi.org/10.48550/arXiv.1810.04805)" (May 2019): <https://doi.org/10.48550/arXiv.1810.04805>.

⁸⁸ T. Brown, "[Language Models are Few-Shot Learners.](https://doi.org/10.48550/arXiv.2005.14165)" (July 2020): <https://doi.org/10.48550/arXiv.2005.14165>.

⁸⁹ A. Dosovitskiy, "[An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale](https://doi.org/10.48550/arXiv.2010.11929)" (June 2021): <https://doi.org/10.48550/arXiv.2010.11929>.



followed by text-image models such as CLIP⁹⁰ and now these foundation models are being developed across many other domains and data modalities, e.g., weather and climate,⁹¹ and audio.⁹² These large-scale pretrained models now serve as the backbone for many commercial AI applications such as ChatGPT⁹³ and DALL-E⁹⁴ and are also being heavily leveraged by academia and in U.S. government research and applications. For a more complete review on foundation models, see DHS S&T’s publication “Foundation Models at the Department of Homeland Security: Use Cases and Considerations.”⁹⁵

4.3.1 Impacts of AAI on Foundation Models

Foundation models and AAI intersect in a couple of different ways:

- Foundation Models are susceptible to AAI attacks, both in the form of AML and in the form of generative deceptive AI, and
- Foundation Models provide the capability for an adversary to generate these threats.

A common categorization of these AI-based threats uses the confidentiality or privacy, accessibility, and integrity triad.⁹⁶ As foundation models are susceptible to many of the same attacks as other ML models, many of which have been covered in more detail in other sections of this report, the following sections focus on AAI threats that are most unique to foundation models using a different triad: scale, emergence, and homogenization.

4.3.1.1 Scale

Foundation models require extremely large and broad sets of data to be trained effectively. Current state-of-the-art foundation models use anywhere from tens of gigabytes⁹⁷ to hundreds of terabytes⁹⁸ of data depending on the size of the model and specific modalities involved. With data sizes growing, the sheer size of the datasets precludes their comprehensive examination by humans. This presents opportunities for ingestion of sensitive data such as PII, proprietary, or classified information. Such large volumes of data might also be poorly characterized, leading to distributions of training data that are not well understood and that run the risk of imparting bias or other properties into the model that adversaries may be able to use to their advantage if discovered. Lastly, the sourcing and quality of these massive datasets is difficult to validate, creating opportunities for adversaries to inject their own poisoned data into the training set thus degrading model performance.

There are many potential DHS mission use cases for foundation models, as well as the associated threats. As a result of the scale of data required to train foundation models, inference attacks are a large concern. Because foundation models employed for DHS missions might include sensitive information (e.g., law enforcement, asylum seekers, witnesses, travelers, or emergency service professionals, etc.) in their training data, given adequate access to the model an attacker, through inference attacks, might be able to

⁹⁰ Radford, Alec, et al. *Learning Transferable Visual Models From Natural Language Supervision*. arXiv, 26 Feb. 2021. [arXiv.org, https://doi.org/10.48550/arXiv.2103.00020](https://arxiv.org/abs/2103.00020).

⁹¹ T. Nguyen, “ClimaX: A foundation model for weather and climate,” (July 2023): <https://doi.org/10.48550/arXiv.2301.10343>.

⁹² A. Radford, “Robust Speech Recognition via Large-Scale Weak Supervision,” (December 2022): <https://cdn.openai.com/papers/whisper.pdf>.

⁹³ OpenAI ChatGPT Web page (October 2023): <https://chat.openai.com>.

⁹⁴ Open AI DALL-E 2 Web page (October 2023): <https://openai.com/dall-e-2>.

⁹⁵ A. Henninger, D. Kusnezov, “Foundation Models at the Department of Homeland Security: Use Cases and Considerations”, DHS S&T Report (2023).

⁹⁶ Oprea A, Vassilev A, (2023) Adversarial Machine Learning: A Taxonomy and Terminology of Attacks and Mitigations. (National Institute of Standards and Technology, Gaithersburg, MD) NIST Artificial Intelligence (AI) NIST AI 100-2e2023 ipd.

⁹⁷ T. Brown, “Language Models are Few-Shot Learners,” (July 2020): <https://doi.org/10.48550/arXiv.2005.14165>.

⁹⁸ C. Schuhmann , “LAION-5B: An open large-scale dataset for training next generation image-text models,” (October 2022): <https://doi.org/10.48550/arXiv.2210.08402>.





expose vulnerabilities in the model and elicit sensitive this data (e.g., PII) or global properties of the model for easier circumvention (e.g. bias in speech patterns or facial recognition).

Another AAI threat vector introduced by the need for data at scale is that of data poisoning.⁹⁹ The scale of data and the self-supervised nature of training required to construct foundation models makes it extremely difficult to validate the training data. Worse yet, much of the content is publicly scraped from the open internet. This makes these models vulnerable to data poisoning attacks during training, which has the potential of compromising the integrity of the foundation model. Adversaries with long-time horizons can generate and distribute poisoned data on the internet in hopes that it will be scraped for future model creation. Furthermore, any downstream models that use the compromised foundation model as a backbone may also inherit the vulnerability, leading to mass failure (see Section 4.3.1.3, “Homogenization”).

4.3.1.2 Emergence

The emergent properties of foundation models have captured the attention of the world with recent releases of tools like ChatGPT and DALL-E. Despite the novel capabilities demonstrated by these models they are still poorly characterized and not well understood. Furthermore, the generative capabilities of these models and the democratization of AI is making it easier for bad actors to intentionally misuse these large-scale models. The AAI threats presented by the emergent capabilities of foundation models come largely from their tendency to hallucinate information that could evade, discredit, overwhelm, or otherwise subvert DHS systems, services, and personnel.

Foundation models enable all the generative deceptive AI attacks presented in Section 2.2. As foundation models and deepfake technology continue to evolve, they will be able to easily generate influence campaigns and spread misinformation that could impact emergency services (e.g., false road closures during an evacuation or diversion of forces from the border). Deepfakes will potentially also lead to increased spam calls, phishing attempts, social engineering, or other means to commit identity fraud or steal sensitive information.

LLMs have the same potential as deepfakes to produce influence campaigns and spread misinformation albeit through text rather than audio-visual media. One of the more concerning properties of LLMs is their ability to produce working, or close to working, code¹⁰⁰ from just a short text description (e.g., generate a python function to add two numbers together). This capability is already finding its way into many commercial applications,¹⁰¹ but at the same time creates new opportunities for bad actors to more easily generate malicious code¹⁰² (i.e., malware) more easily. Code generation with LLMs will lower the barrier to entry for non-experts to rapidly generate malicious code, which could have negative consequences for the DHS Cybersecurity and Infrastructure Security Agency in keeping up with the proliferation of cyberattacks.

Face-morphing is yet another attack that foundation models, specifically Stable Diffusion, have recently enabled. A face-morphing attack uses two or more identities (e.g., images of faces) to generate a new identity that captures the biometric qualities of the contributing identities. The result is a visually realistic, but fake, image of a face that is capable of fooling humans and modern automated face recognition systems alike, meaning that multiple people might be able to be authenticated using a single identity

⁹⁹ N. Carlini, “Poisoning and Backdooring Contrastive Learning,” (March 2022): <https://doi.org/10.48550/arXiv.2106.09667>.

¹⁰⁰ R. Li, “StarCoder: May the Source be With You!” (May 2023): <https://doi.org/10.48550/arXiv.2305.06161>.

¹⁰¹ GitHub Copilot Web page (October 2023): <https://github.com/features/copilot>.

¹⁰² E. Shimony, “Chatting Our Way Into Creating a Polymorphic Malware.” (January 2023): <https://www.cyberark.com/resources/threat-research-blog/chatting-our-way-into-creating-a-polymorphic-malware>.



represented by the fake image. This attack has already been demonstrated in the real world¹⁰³ and has the potential to undermine border and transportation security.

4.3.1.3 Homogenization

Because of the expertise and scale of data and compute required to train a foundation model, only a few institutions are currently capable of doing so. Access to these large-scale pretrained models can greatly reduce the amount of data required to train downstream task-specific models with comparative performance and less compute. It has been demonstrated that a vulnerability introduced in the foundation model can be inherited by downstream models¹⁰⁴ that use it, introducing yet another adversarial threat vector. As these foundation models become more widely used as backbones for downstream models a potential single point of failure emerges. If this single point of failure is compromised, it has the potential to simultaneously discredit large numbers of DHS systems or create distrust with DHS agents and operators.

Data poisoning is one of the major concerns pertaining to homogenization and was discussed earlier. Here, we focus on evasion attacks. A wide variety of attacks can be used to perturb the input to a model and thus degrade overall model performance (e.g., get the model to miss or misclassify an object). These perturbations can come in the form of digital patches¹⁰⁵ or other noise¹⁰⁶ injected directly into an image or it can even be patches or other “confusers” placed in the real world (e.g., stickers on a stop sign).¹⁰⁷ Much like the effect of data poisoning, if an effective evasion attack can be discovered for a particular foundation model there is a risk that the same vulnerabilities might be inherited by many other downstream models. As an example, if a baggage scanner foundation model were trained and different airports and/or scanners used this as a backbone to fine-tune to look for new objects or incorporate new scanners it may be possible for an adversary to identify a single evasion attack that fools all scanners. This example can be easily extended to other DHS use cases such as object detection for ISR border security applications, facial recognition, etc.

4.3.2 Defenses for Foundation Models

The recent report in the Preparedness Series on foundation models¹⁰⁸ likened them to other digital assets that need to be protected at each stage of their development and operation to ensure they remain uncompromised. Because of their “foundational” nature, foundation models offer a high-leverage single point of failure and are a prime target for attack. The defense against AAI threats is an active and growing area of research and one of utmost importance, but a silver bullet to completely defend foundation models from AAI threats is unlikely. The following list offers some strategies to securing foundation models from adversarial attacks: use trusted data sources, watermark training data, encrypt training data, train AI models using adversarial examples, make AI models intrinsically more robust, network/defensive distillation, analyze or modify inputs, ensemble methods, red teaming, educating end users, and good cyber hygiene and incident response. These are elaborated on in Appendix A. These mitigations are scoped toward the training and deployment of models and does not cover defenses against the generative deceptive AI threats arising from foundation models as those are covered in prior sections of this document.

¹⁰³ NIST IFPC 2022 Conference Presentations and Videos (November 2022): <https://www.nist.gov/itl/iad/ifpc-2022-conference-presentations-and-videos>.

¹⁰⁴ K. Kurita, “Weight Poisoning Attacks on Pre-trained Models,” (April 2020): <https://doi.org/10.48550/arXiv.2004.06660>.

¹⁰⁵ T. Brown, “Adversarial Patch,” (May 2018): <https://doi.org/10.48550/arXiv.1712.09665>.

¹⁰⁶ I. Goodfellow, “Explaining and Harnessing Adversarial Examples,” (March 2015): <https://doi.org/10.48550/arXiv.1412.6572>.

¹⁰⁷ K. Eykholt, “Robust Physical-World Attacks on Deep Learning Visual Classification,” (April 2018): <https://doi.org/10.48550/arXiv.1707.08945>.

¹⁰⁸ D. Kusnezov, “Preparedness in Times of Rapid Change,” DHS S&T Report (2023).



4.4 Distributed Intelligence

Decentralized intelligent systems are networks of nodes or agents in which, for reasons related to compute, storage, efficiency, bandwidth, resilience, and/or scalability, the data processing and decision-making occur at each of the nodes in the network, rather than at a central node. Although this is often a more robust architecture lacking a single point of failure, it requires additional computational cost, overhead, and iterative algorithms, to achieve the same performance as centralized systems. Additionally, because individual nodes communicate directly with only a small percentage of the network, adversarially manipulated data can be more difficult to detect, and the effect from even a single compromised node can be dramatic if unmitigated. Designing robust decentralized algorithms for inference, training, and autonomous agent action is thus an active area of research to address increasingly complex threats. Subsections 4.4.1 and 4.4.2, “Inference/Training” and “Autonomy,” respectively, address the potential impacts of adversarial nodes on algorithms within such networks and provide a broad overview of research on AAI against distributed intelligence algorithms. The impact of adversarial attacks on decentralized training, inference, and autonomy algorithms can be substantial. In each case, these algorithms must exchange intermediate information, often gradients of a quantity of interest, and iterate until results converge at all nodes in the network. Gradients that are large or towards undesirable values of the parameters in question are the most common form of an adversarial attack, though the outcome might be different if the attack targets inference, training, or autonomy algorithms.

4.4.1 Inference/Training

Decentralized inference algorithms extract information from the data they sense at all nodes without aggregating that data in one location though iterative algorithms, which eventually achieve consensus amongst all nodes. It has been demonstrated that including incorrect values among the exchanged data can cause biases in the estimate regardless of the estimator, and that it is possible to add significant bias with a single adversarial node even in large networks.^{109,110} Since information inferred from raw sensor data over the network is the basis of all future decisions, inaccurate estimates of critical parameters can lead to poor performance on all subsequent tasks.

Federated learning distributes training over multiple nodes, and is designed to avoid data aggregation, promote efficiency and resilience, and preserve data privacy. In general, each node trains on locally held data, and communicates only gradients of model weights. In methods analogous to those that disrupt decentralized inference, a small number of adversarial agents can substantially degrade the performance of the resulting model by sending “boosted” updates. This has been dubbed *model poisoning*, as it does not involve modifying the training data directly.¹¹¹

4.4.2 Autonomy

Beyond decentralized exploitation of sensor data, intelligent sensor networks can react to information in the data they collect to improve performance toward their objectives. The actions of a small network can be fully coordinated using decentralized partially observable Markov decision processes (POMDPs)¹¹² or

¹⁰⁹ Vempaty, Aditya, Bhavya Kailkhura, and Pramod K. Varshney. Secure networked inference with unreliable data sources. Singapore: Springer, 2018.

¹¹⁰ Yang, Zhixiong, Arpita Gang, and Waheed U. Bajwa. "Adversary-resilient distributed and decentralized statistical inference and machine learning: An overview of recent advances under the byzantine threat model." *IEEE Signal Processing Magazine* 37.3 (2020): 146-159.

¹¹¹ Bhagoji, Arjun Nitin, et al. "Analyzing federated learning through an adversarial lens." *International Conference on Machine Learning*. PMLR, 2019.

¹¹² Oliehoek, Frans A., and Christopher Amato. *A concise introduction to decentralized POMDPs*. Vol. 1. Cham, Switzerland: Springer International Publishing, 2016.



decentralized optimization techniques such as the alternating direction method of multipliers (ADMM)¹¹³ if a cost function can adequately score all possible agent behaviors. Multi-agent reinforcement learning (MARL) is a recent approach in which a policy mapping state to action for one or more agents is learned through actions taken in thousands of training simulation runs. MARL is better suited in practice to more complex objectives, environments, and action spaces and often functions with a reward specified only for overall outcomes, however it lacks the optimality guarantees of other approaches and requires a computationally efficient, high-fidelity simulation capability. MARL training and ADMM are iterative and subject to the same adversarial manipulation of shared updates as described in prior sections.¹¹⁴ Attacks to disrupt MARL policies when deployed can be categorized as action perturbations, observation perturbations, or communication perturbations and closely resemble more conventional AML and/or cyberattacks.¹¹⁵

Due to the decentralized nature of the data processing and autonomy algorithms, small numbers of adversarial agents can have a disproportionate impact on outcomes. Data/updates that are very large or small relative to the mean introduce the most bias and were the early adversary strategies against decentralized algorithms. However, these are also easiest to identify, particularly with access to all data. However, mitigating this potential impact with no other knowledge of an attacker's strategy often mean excluding several of the largest and smallest values received, forcing the network to throw away useful data and slowing the convergence of decentralized algorithms if no attackers are present. More damaging attacks become feasible as prior information is available on the objectives and algorithms of the network. Defenses improve concomitantly with knowledge of the attacker objectives and behaviors. For example, proposed defenses have included methods that evaluate the data history of individual nodes to be more effective in identifying sources of adversarial data, but in turn have motivated strategies in which nodes alternate between sending adversarial and true data to avoid detection. Current research is continually addressing increasingly complex adversary behavior, with a focus on designing robust systems that do not significantly sacrifice performance when no adversary is present.

4.5 Internet of Intelligent Things (IoIT)

IoIT refers to a network of interconnected devices, sensors, and objects that are equipped with intelligent capabilities to communicate, gather data, and make decisions. AAI, on the other hand, involves techniques where malicious actors manipulate AI models by introducing carefully crafted inputs to deceive or confuse them. The relationship between (IoIT) and AAI lies in the potential vulnerabilities that adversarial attacks can exploit within IoIT systems. By crafting adversarial inputs, attackers can manipulate the behavior of IoT devices or sensors, causing them to produce incorrect or unexpected outputs. This could lead to false data being collected, erroneous decisions being made, or even safety-critical systems being compromised. There are multiple opportunities for these vulnerabilities to present themselves, including:

- **Sensor data manipulation:** Adversarial attacks could target the sensors in IoIT devices, causing them to provide inaccurate or misleading data. This could affect various applications such as environmental monitoring, health tracking, and industrial automation.

¹¹³ S. Boyd, "Distributed Optimization and Statistical Learning via the Alternating Direction Method of Multipliers," (January 2011): [Distributed Optimization and Statistical Learning via the Alternating Direction Method of Multipliers | Now Foundations and Trends books | IEEE Xplore](#).

¹¹⁴ M. Figura, "Adversarial attacks in consensus-based multi-agent reinforcement learning," (March 2021): [\[2103.06967\] Adversarial attacks in consensus-based multi-agent reinforcement learning \(arxiv.org\)](#).

¹¹⁵ M. Standen, "SoK: Adversarial Machine Learning Attacks and Defences in Multi-Agent Reinforcement Learning," (January 2023): <https://arxiv.org/abs/2301.04299>.



- Communication disruption: Adversarial attacks might target the communication channels between IoIT devices, disrupting the exchange of information and potentially causing communication breakdowns.
- Privacy and security concerns: AAI could be used to compromise the privacy and security of IoIT systems by extracting sensitive information from data streams or bypassing security mechanisms.
- System integrity: Adversarial attacks on IoIT could compromise the integrity of AI models deployed in the system, leading to unintended behaviors or malfunctioning of connected devices.

To mitigate the risks associated with adversarial attacks in IoIT, robust security measures, anomaly detection techniques, and continuous monitoring are essential. These are reviewed in Appendix A. Additionally, R&D efforts are ongoing to create more resilient AI models and algorithms that are less susceptible to adversarial manipulation. As technology evolves, the relationship between IoIT and AAI will likely continue to evolve as well, requiring ongoing attention to ensure the security and reliability of interconnected intelligent systems.

4.6 Advanced Manufacturing

AI systems have an increasingly important role in advanced manufacturing. Although CV for automatic defect detection, both *in situ* and post process, remains a leading focus, AI techniques are also being investigated for preventative maintenance to predict tool wear and part failure. In additive and hybrid manufacturing, there is a push for “certify as you build” approaches where *in situ* sensors are used to predict the part quality, material properties, and defects so that further post process inspection is unnecessary. AI will be a key component in these models and as these models become more robust, they will be increasingly integrated into closed loop control systems that can adjust process settings in real time and rework layers/defects. Another area where AI models might grow more prominent is scheduling and space allocation. With additive manufacturing allowing distributed manufacturing AI might take a larger role in ordering materials, scheduling time on machines, and moving supplies and parts between locations.

With this growth of AI in advanced manufacturing systems comes an increased susceptibility to adversarial attacks including evasion, poisoning, and inversion attacks. With evasion attacks, a quality control system could be used to create false negatives (defective parts that are classified as good), either letting naturally occurring flaws past, or hiding intentional attacks to cause failures in the field. False positives (good parts classified as bad) could be used to waste time and money (throwing away good parts) and undermine trust in the monitoring system.

Manufacturing typically requires physically fabricating parts to collect data. This means that it can be quite costly to collect enough data to train your own model. As a result, many manufacturers might rely on pretrained models/datasets, use transfer learning, or outsourcing the model/data collection to manage the costs of implementing these systems. This leads to concerns about potential data poisoning as the security of opensource or online datasets that might be incorporated into these models is unknown. Many original equipment manufacturers (OEMs) also collect data from their machines in the field that they can then feed back into their models to improve performance. In such a scenario, it is foreseeable that a bad actor with several machines could feed false data back into the system to poison the dataset in the hope that the OEM will incorporate this poisoned data into the next update/iteration of their model.

Many ML models are trained on proprietary data. An attacker who inverted a model used in advanced manufacturing could potentially be able to extract information about confidential part geometries or hints as to the materials and process parameters that were used. This could be done by a competitor or other bad actor to gain an advantage or undermine manufacturers or their customers. Theft of AI models is also



a concern because of the inherent intellectual property and competitive advantage they might represent. Finally, as AI models become more integrated in advanced manufacturing, they risk being the target of DOS attacks. If a model is rendered inaccessible, unusable, or corrupted it could delay/stop a manufacturing facility until it was able to be restored or replaced. This is of particular concern when using a third-party service, as access to the model might be outside of the manufacturer’s control.

4.7 Gene Editing

Striking a balance between leveraging AI for advances in gene editing (see Figure 4.1) and safeguarding against AAI-driven misuse is crucial in harnessing the full potential of this transformative biotechnology. For example, the malicious use of AI could lead to the development of novel and harmful genetic modifications, raising ethical and safety concerns. Other adversarial attacks might exploit vulnerabilities in the gene editing processes, potentially leading to unintended genetic changes or biosecurity threats, enabling the creation of new substances generated by AI such as creating new/extending existing poison, evolving existing benign molecules/DNA strands into harmful substances, inactivating treatments, and inventing Trojan viruses.

Adversarial AI for Gene Editing – Automatic Generation



Slide courtesy Rebecca Taylor <bex@andrew.cmu.edu>

© 2023 Carnegie Mellon University

(DISTRIBUTION STATEMENT A) Approved for public release and unlimited distribution.

13

Figure 4.1. AAI for Gene Editing – Automatic Generation in the Carnegie Mellon University Cloud Lab.¹¹⁶

AI is increasingly used in conjunction with CRISPR (Clustered Regularly Interspaced Short Palindromic Repeats) technology to enhance the efficiency, precision, and effectiveness of gene editing experiments. Yet, gene editing using CRISPR is just one technique, there are many novel protein-design tools and approaches. LLMs are used to explore and develop therapeutic proteins for the next generation of medicine¹¹⁷ and generate protein sequences with a predictable function across large protein families.^{118,119}

¹¹⁶ Sherman, M. (CMU SEI). “Panel 3. Implications of AAI on Emerging Technology” page 13. Presented at the DHS S&T Risks and Mitigation Strategies for Adversarial AI Threats. June 16, 2023.

¹¹⁷ NVIDIA Web page (October 2023): <https://nvidianews.nvidia.com/news/nvidia-unveils-large-language-models-and-generative-ai-services-to-advance-life-sciences-r-d>.

¹¹⁸ A. Madani, “Large language models generate functional protein sequences across diverse families,” (January 2023): <https://doi.org/10.1038/s41587-022-01618-2>.

¹¹⁹ N. Ferruz, “ProtGPT2 is a deep unsupervised language model for protein design,” (July 2022): <https://doi.org/10.1038/s41467-022-32007-7>.





Deep-learning classifiers are used in the generation of antimicrobials with desired attributes.¹²⁰ Deep-learning platforms automate high-throughput biological sequence functional analysis enabling researchers to answer many biological questions.¹²¹ And, other ML applications that predict native protein structures from their sequences can be inverted to design new proteins¹²² or to regularly predict protein structures with atomic accuracy even in cases in which no similar structure is known.¹²³

To mitigate the risks associated with adversarial attacks in gene editing, robust security measures including secure data storage, access control, biometric authentication methods, and anomaly detection techniques are essential. These are reviewed in Appendix A.

4.8 The Metaverse

The Metaverse is a concept often used to describe a virtual, collective, and interconnected digital universe that encompasses multiple virtual environments, augmented reality spaces, and digital experiences. With a few well-funded companies working to profit from the resources and future visitors, it is essentially the Wild West. While AI systems, and AAI systems, for use in the Metaverse require significant sums of money to develop and deploy, these large well capitalized companies have access to large amounts of capital.

Artificial intelligence is a building block for the Metaverse, but it is also the key to making the Metaverse both immersive and inclusive. As such, AAI systems can be native parts of every user's experience, indistinguishable from avatars of other humans. This an impact a substantial number of people, Gartner estimated "25% of people (Americans) will spend at least one hour per day in the Metaverse by 2026."¹²⁴

More impactfully, AAI can interact in the Metaverse at scale, with potentially all users' interactions with this sort of AI-bot during every visit to the Metaverse. On the less harmful side, Metaverse companies are likely to field AI-bots for advertising as a mechanism to recoup their investments. However, the AI-bot that befriends a user to convince them that a new virtual dance club is a great place to have fun experiences is precisely the same technology that could be used to convince them that a new extremist group is a great place to add meaning to their life. It might not even be a violation of the terms of service in the early stages of grooming.

Significantly, the Metaverse companies are open to a kind of regulation. Alas, they seem to favor the liability for "unintended side effects sort of benefit" from the regulation over actual safety for users. Consider the current disclaimer from Google's Bard AI "*Bard may display inaccurate or offensive information that doesn't represent Google's views.*" This seems designed more to protect Google from cancel culture than it does to practically warn users how wrong the results might be. Actual safety will require a larger element of government direction, perhaps akin to automotive safety regulation.

The Metaverse is not a home to widespread AAI at this time, or widespread anything for that matter. However, the opportunity for adversary actors to operate freely in the Metaverse raises the potential for DHS interdiction in the Metaverse in the future.

¹²⁰ P. Das, "[Accelerated antimicrobial discovery via deep generative models and molecular dynamics simulations](https://doi.org/10.1038/s41551-021-00689-x)," (June 2021): <https://doi.org/10.1038/s41551-021-00689-x>.

¹²¹ R. Wang, "[DeepBIO: an automated and interpretable deep-learning platform for high-throughput biological sequence prediction, functional annotation and visualization analysis](https://doi.org/10.1093/nar/gkad055)," (February 2023): <https://doi.org/10.1093/nar/gkad055>.

¹²² "[De novo protein design by deep network hallucination](https://doi.org/10.1038/s41586-021-04184-w)," (December 2021): <https://doi.org/10.1038/s41586-021-04184-w>.

¹²³ J. Jumper, "[Highly accurate protein structure prediction with AlphaFold](https://doi.org/10.1038/s41586-021-03819-2)," (July 2021): <https://doi.org/10.1038/s41586-021-03819-2>.

¹²⁴ Gartner Press Release, "[Gartner Predicts 25% of People Will Spend At Least One Hour Per Day in the Metaverse by 2026](https://www.gartner.com/en/newsroom/press-releases/2022-02-07-gartner-predicts-25-percent-of-people-will-spend-at-least-one-hour-per-day-in-the-metaverse-by-2026)," (February 2022): <https://www.gartner.com/en/newsroom/press-releases/2022-02-07-gartner-predicts-25-percent-of-people-will-spend-at-least-one-hour-per-day-in-the-metaverse-by-2026>.



4.9 Quantum Computing

Quantum computers exploit the phenomena of quantum mechanics to enhance a range of computational challenges. In the context of AI/ML: (1) Quantum computers can implement techniques that achieve better generalization with less training data,¹²⁵ (2) Quantum neural networks have higher effective dimension than their conventional counterparts, (3) Quantum kernel methods can provide lower prediction error and speed up certain ML challenges. However, a quantum computer capable of providing these advantages is at least a decade away. A hybrid quantum-classical algorithm, the Quantum Approximate Optimization Algorithm (QAOA), has been used to solve optimization problems including graph coloring and MaxCut on today's hardware. However, QAOA has not yet shown an advantage over classical algorithms.

Quantum and conventional computers behave very differently. Current research indicates that adversarial attacks using a quantum computer can be successful on quantum computers. This is particularly true when the adversary has knowledge of the quantum computer (a white box attack). It could be decades before there are quantum computers capable of doing something useful, and AAI is unlikely to be successful unless developed on a quantum computer.

Researchers have contrasted adversarial attacks on quantum and conventional neural networks (CNN) through simulations.¹²⁶ As seen in Figure 4.2, they showed that a quantum variational classifier (QVC) is more robust to a conventional attack (e.g., projected gradient decent, fast gradient sign method) than a CNN when the attacker is not aware that the network is quantum.

If the attacker is fully aware of the nature of the network, Figure 4.3, specially designed attacks on CNN or QVC will work. However, though an attack designed for a QVC performs admirably even against a CNN (g), a QVC is robust against an attack designed for a CNN (c).

When trained against adversarial attacks, simulations suggest that a QVC does not improve against a CNN attack, and that the improvement of a QVC against a QVC attack is not as great as a CNN against CNN attacks.

Open questions in Quantum AAI include the potency of a quantum adversary. Can a quantum adversary generate projected gradient descent (PGD), FGSM, and similar attacks more quickly than a conventional adversary or generate quantum data patterns that are impossible to train for on a conventional computer?

¹²⁵ E. Gil-Fuster, "[Understanding quantum machine learning also requires rethinking generalization.](https://www.google.com/url?sa=t&rct=j&q=&esrc=s&source=web&cd=&cad=rja&uact=8&ved=2ahUKewjPnMiqjr-BAxX7EFkFHTM5ByIQFnoECA0QAw&url=https%3A%2F%2Farxiv.org%2Fpdf%2F2306.13461%23%3A~%3Atext%3DQuantum%2520machine%2520learning%2520models%2520have%2Cbehavior%2520of%2520such%2520quantum%2520models.&usq=AOvVaw3skIu-dMOhFGrYjiUsN-kG&opi=89978449)" (June 2023): <https://www.google.com/url?sa=t&rct=j&q=&esrc=s&source=web&cd=&cad=rja&uact=8&ved=2ahUKewjPnMiqjr-BAxX7EFkFHTM5ByIQFnoECA0QAw&url=https%3A%2F%2Farxiv.org%2Fpdf%2F2306.13461%23%3A~%3Atext%3DQuantum%2520machine%2520learning%2520models%2520have%2Cbehavior%2520of%2520such%2520quantum%2520models.&usq=AOvVaw3skIu-dMOhFGrYjiUsN-kG&opi=89978449>

¹²⁶ M. West, "[Towards quantum enhanced adversarial robustness in machine learning.](https://arxiv.org/abs/2306.12688)" (June 2023): <https://arxiv.org/abs/2306.12688>.

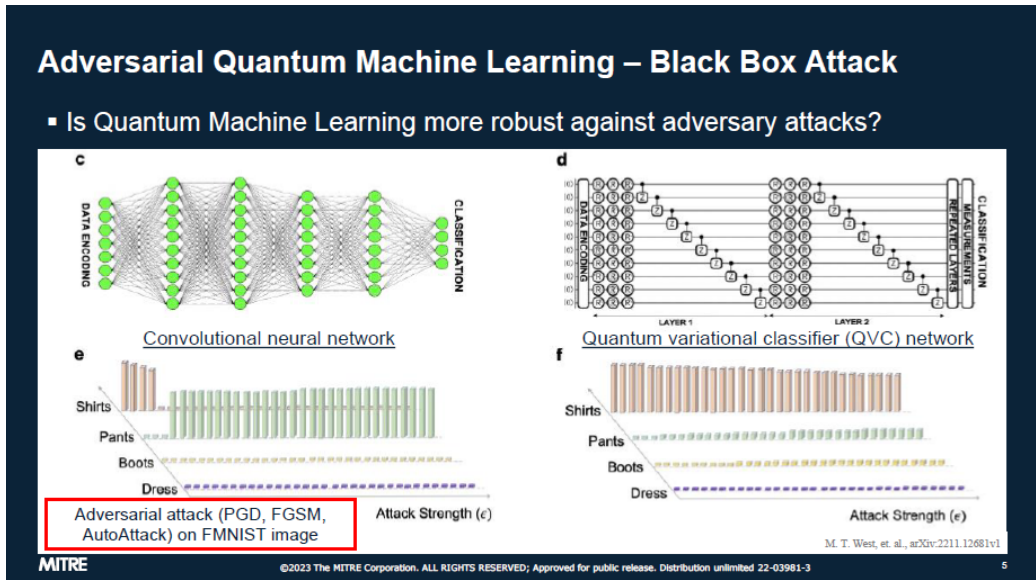


Figure 4.2. Adversarial Quantum Machine Learning – Black Box Attack. (c) Typical CNN, (d) QVC network (vertical lines depict timeline of qubits). (e),(f) QVC and CNN network resistance against adversarial attack. At a critical attack strength, the CNN does not provide the correct label (“Shirts”). However, the QVC correctly identifies the image for much stronger attacks.¹²⁷

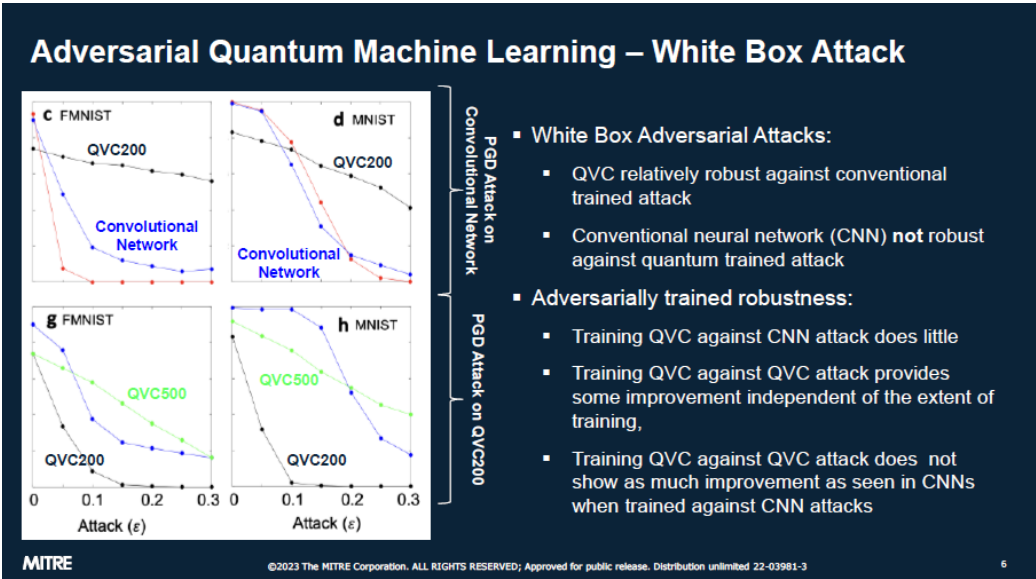


Figure 4.3. Adversarial Quantum Machine Learning – White Box Attack. Accuracy achieved by classical and quantum networks in the cases of white-box PGD attacks on CNN (c), and QVC (g) as a function of attack strength. In both cases the accuracy of the network under attack decreases sharply. However, when transferring the attack of one network to the other we see that that the attack designed

¹²⁷ Weinstein, Y. (MITRE). “Panel 3. Implications of AAI on Emerging Technology” page 5. Presented at the DHS S&T Risks and Mitigation Strategies for Adversarial AI Threats. June 16, 2023.



for the quantum network is successful against the CNN (g), but the quantum network is relatively robust against the attack designed for the CNN.¹²⁸

QVCs and CNNs having different strengths and weaknesses. Perhaps the use of both would achieve maximum accuracy and protection. However, note that the results discussed here are very early and that further study and analysis is needed.

5 Roles of International Partnerships

Experts¹²⁹ agree that present AAI challenges that transcend borders and jurisdictions, much like cybersecurity challenges. AAI is a global concern, and an effective response requires international cooperation and partnerships.

International partnerships are not only about pooling resources and knowledge but also about fostering a collective response to an evolving and pervasive challenge. Through partnerships with allied nations and international organizations, the U.S. can gain access to valuable threat intelligence, enabling for early detection and mitigation of AAI attacks. In the event there is a major incident, international cooperation enables a coordinated response. This was demonstrated, for example, during the WannaCry ransomware attack, where cooperation between the U.S. and our international partners helped track the attackers and mitigate the impact.

In the cybersecurity space, the U.S. engages in joint cybersecurity exercise with allies to simulate cyberattacks and test response strategies. These exercises improve readiness and foster better collaboration among nations. Given the potential for generative AI and deepfakes to target political officials, potentially spawning a major international crisis, these kinds of readiness exercises in the AAI space might also be important.

Together, nations can harness their diverse expertise, perspectives, and resources to establish shared norms, standards, intelligence, and best practices in addressing AAI. By working in tandem, we can fortify our defenses, enhance early warning systems, and develop effective countermeasures, ultimately ensuring a safer, more secure global landscape in the face of this transformative technology.

6 Summary Considerations and Conclusions

In the rapidly evolving landscape of technology and security, the emergence of AAI presents a formidable challenge to homeland security efforts. Through this report, we have delved into the multifaceted implications of AAI, exploring its potential impacts on critical domains such as cybersecurity, border security, law enforcement, emergency management, and others. As we conclude this initial examination, we offer our findings, next steps, and conclusions.

6.1 Summary Considerations

The defense against AAI threats is an active and growing area of research and one of utmost importance, but there will most likely never exist a silver bullet to completely defend from AAI threats. This section reviews the ideas that have surfaced in the discussions of AAI with a broad community and what opportunities could be considered to develop an initial foundation for DHS in the AAI space. As has been experienced in cybersecurity, sustained effort will be required at each level to keep up with the state-of-

¹²⁸ Weinstein, Y. (MITRE). “Panel 3. Implications of AAI on Emerging Technology,” page 6. Presented at the DHS S&T Risks and Mitigation Strategies for Adversarial AI Threats. June 16, 2023.

¹²⁹ As described by experts on the “Panel on International Partnering,” presented at the DHS S&T Risks and Mitigation Strategies for Adversarial AI Threats. June 15, 2023.



the-art in understanding both attack and defense methodologies as well as implementing and maintaining data management, ML, software engineering, and cybersecurity best practices to effectively defend against AAI threats. Critically important, we must understand the lessons learned from standing up a cybersecurity ecosystem, leverage what makes sense from that ecosystem, integrate with that ecosystem, while ensuring that we do not simply extrapolate from cybersecurity and unwittingly pull in assumptions that might not be valid in the AAI space.

Here we summarize opportunities. These considerations focus on a combination of consolidative activities, work on critical enablers, “sense-making” research, common-good systematic research, and fresh design work.

- 1. Examine where acquisition oversight responsibilities and S&T’s oversight responsibilities, require new policies to support secure AI systems. For example, consider AI Red Teaming in T&E or AI security assessments in Systems Engineering.**
 - *The analysis and mitigation of AAI risks should be incorporated throughout the system lifecycle, starting as far left as possible.*
 - *Start the wheels in motion to formalize these processes within DHS’s institutional processes.*
 - *Consider the entire acquisition process, to include commercial-off-the-shelf procurements.*
 - *Establish requirements for contracting, to include contracting language that provides access to training data, measures of effectiveness/benchmarking data, regression testing requirements, scope of model, etc.*
 - *Work with the intelligence community to define what kinds of information requirements we will have in the AAI space such that they can focus their resources more productively.*
- 2. Join the cross-government industry threat intelligence and incident sharing initiative that is starting to take shape. Characterize DHS needs and influence its activities to gain a clearer understanding of the types of attacks that happen against real-world systems and use this to help inform DHS risk assessments and investments.**
- 3. Develop a data governance framework including an “AI supply chain” that will standardize methods for tracking and monitoring data and model provenance, in particular the relation between the model and publicly available models and other architectures, but also the training process and validation of the AI model as the embodiment of a solution to the specific DHS problem in question.**
- 4. Develop needed international partnerships and continue to fully integrate our international collaboration strategy within our R&D planning and execution processes. As we formulate projects and research efforts, part of that process should be determining whether a particular project is aligned to an area of mutual priority with our partners and think through what level of collaboration would be appropriate as part of the effort.**
- 5. Engage the R&D community and explore ways to make AI more robust, including against adversarial attacks, and focus on reducing design flaws instead of just relying on AI security to find them and patch them. Network robustness has been aided by tools like Chaos Monkey. Software robustness has been aided by tools such as fuzzing and intelligent fuzzing. Develop the equivalent robustness facilitating construct for AI.**
- 6. AAI, including deceptive generative AI, will create a class of content poised to confuse our signals (e.g., ISR, intelligence, cyber sensors, etc.). It is important to understand how we depend on different types of authentications (e.g., voice and text, image, digital signature, etc.) and what**



kinds of conditions are required to successfully prosecute an AAI attack (e.g., access to model API, access to training data, stealthy, etc.). It will also be important to understand which conditions are associated with what kinds of DHS missions that use AI applications (e.g., baggage scanning, facial recognition technology, etc.). The review in this document provides a good start, but it is not sufficiently comprehensive. As such, we must characterize, more deeply, AAI threats and risks as they relate to homeland missions.

- *Assess, study, and catalog the current and future uses for AI-based technologies across services, domains, and components.*
- *Assess and rank the AI threats and vulnerabilities by the most significant risk and impacts on the DHS operations.*
- *Proactively and continuously track the emerging AI risks and vulnerabilities in a similar vein that cyber threats and vulnerabilities are tracked.*
- *Work to develop comprehensive testing and evaluation policies and procedures that account for the assessed AAI risks. Build a cross-DHS AI red teaming capability that can help DHS understand risks and mitigate threats at the system of systems and mission level as threats emerge, especially considering the integration of multiple vendor capabilities.*
 - *Consider lessons learned from establishing, as well as the strengths and weaknesses of, Red Teams in Cybersecurity.*
 - *Start “shooting” AI-based malware at critical infrastructure (e.g., financial markets/stock market, electric grid, etc.) for stress testing.*

7. Generative AI companies have been working on watermarking approaches to mark their product as AI-generated. This must not be seen as a solution to the “deepfake” problem. Adversaries will not be constrained to using the watermarked versions of these tools. Relying on markings gives a false sense of security relative to adversary fakes and a universally accepted watermark scheme seems unlikely given the proliferation of publicly available tools. Regulating synthetic content creation is challenging to accomplish because such regulation would need to have international reach and digital borders do not exist. A more robust representation scheme is required for differentiating between identity grade photos (live captured, not morphed, under ideal lighting conditions) that will be used for identity checking versus all the other less consequential imagery (e.g., pictures on the internet). As such, we must conduct R&D needed to forensically identify fakes without relying on watermarks and to prepare for influence of deceptive generative AI on DHS missions.

- *Find ways to ensure authenticity at scale (e.g., rely on audio, video, text evidence in courts).*
 - *Prioritize the R&D of methods and approaches that can validate pieces of digital images and videos, being able to determine what is real and what has been spoofed. This could include using a cryptographic signing approach, validating the image or video so long as it has this signature attached to it.*
 - *Focus on combining of technical, process, and policy components in developing solutions that rely to the extent possible on proofs of provenance and authenticity of information.*
 - *Deepfake detectors should be invested into, as many adversaries will use the off-the-shelf product. Providing an ability to immediately create a form of security against these attacks could prove to be beneficial.*
 - *Ensemble or sensor fusion methods might be beneficial as a way of improving detection robustness.*
- *Develop a more robust vocabulary for distinguishing across a spectrum of fake and good (i.e., not just binary, “real” or “not real”).*
- *Develop response options (e.g., a playbook) for rapid response to mitigate and act against a generative AI-enabled misinformation campaign that could be intervening with a DHS operation.*



8. **Invest in the art and science of measuring and assessing the magnitude of potential vulnerabilities (develop measures of AAI consequentiality of systems for various missions to determine what level of assurance is required for different AI applications), mitigating, recovering, and assessing damage from potentially exploited vulnerabilities and designing nationwide, if not international, vulnerability management processes. Consider how to build in mitigations to limit the harm that AAI can cause.**
9. **Further explore line of research in quantum computing and its impacts on adversarial attacks. Open questions in Quantum AAI include the potency of a quantum adversary.**
10. **The potential for adversarial actors to operate freely in the Metaverse raises a host of safety concerns, including from AAI. Safety in the Metaverse will require a larger element of government attention. Start exploring concepts on how to protect the public from adversaries operating in the Metaverse. This could have consequences that include transnational repression or radicalization.**

6.2 Near-term Considerations

In terms of concrete next steps following this study:

- Conduct a series of narrowly focused sessions/workshops with the individual DHS components to understand their current and future use cases and adoption plans for AI-based technologies and document the outcomes.
- Prioritize the biometrics and identity management, cyber defense, and automated security surveillance technical areas for assessments against the AAI risks and threats.
- Collect and document the findings—to be used for the future mitigation and action plan.
- Plan for a comprehensive testing, evaluation, and threat monitoring action with an aggressive development and implementation schedule.

6.3 Conclusions

In an era where technological innovation is rapidly reshaping the threat landscape, DHS stands at the forefront of safeguarding our nation. The imperative to prepare for these AAI methods has never been greater. These groundbreaking technologies possess the power to create deceptive content, manipulate information, and exploit vulnerabilities, posing significant threats to critical infrastructure, public safety, cybersecurity, and other parts of DHS's core missions, necessitating a proactive and comprehensive response. By diligently equipping ourselves with the knowledge, tools, and strategies to understand, counter, and adapt to these emerging threats, DHS not only strengthens national security but also ensures that our homeland remains resilient against the unpredictable challenges of the digital age. In embracing this preparation, DHS exemplifies its commitment to protecting both the present and the future of the U.S.



Appendix A: Compiled List of Mitigation Strategies Discussed

While there are certainly others, this appendix attempts to synthesize the range of AAI mitigation strategies identified and discussed by experts. Some are more process-focused, others more analytical-focused, but all are important. Because it is impractical to implement all of these for every model, selecting the right combination of strategies for the particular vulnerability/risk a developer is trying to address is key to minimizing risk in an informed, efficient way.

Data Preparation

- **Use trusted data sources/data provenance/sound data governance including an AI supply chain.** Publicly available training data can be attacked and poisoned. To mitigate data poisoning, “trusted” data sources should be the first line of defense. It will be paramount to have a process to vet the data sources being used (cyber posture, quality of data, provenance, etc.) and institute appropriate guardrails when the training data cannot be vetted. There are a lot of publicly available and open-sourced data out there. However, much of that can be attacked and poisoned, so it is vital that before data is used it is properly vetted. Much of the discussion here centered on identifying “trusted” sources and how everyone must work to protect their data. This leads to issues with sharing.
- **Data scrutiny and cleaning.** Thoroughly examine and preprocess your training data to identify and remove anomalies, outliers, and potential poisoned samples. This can involve outlier detection algorithms, data validation checks, and data augmentation techniques to ensure the quality and authenticity of your dataset.
- **Data diversity.** Incorporate diverse and representative data into a training set to reduce the impact of poisoned samples. A diverse dataset makes it harder for attackers to manipulate the learning process by targeting specific data points.
- **Watermark training data.** To protect the integrity of training data R&D investments will be required to investigate techniques like digital signatures and watermarking. These techniques should be used whenever possible, but they are also complex because to protect against concept drift, training data must be dynamic and continually refreshed. Training data is extremely valuable. It can be difficult to collect and update, so it is important to protect the integrity of the training data. Techniques such as digital signatures or watermarking were discussed. However, it is a complex problem and things such as protecting against concept drift in training data must be dynamic, and continually refreshed.
- **Encrypt training data.** Methods like homomorphic encryption allow computation on encrypted data and provide encrypted results to the user. Training foundation models on encrypted data would help mitigate inference and injection attacks targeted at retrieving sensitive information from the model.
- **Input transformation or perturbation.** Modify the input data before it reaches the model. This could involve adding random transformations to the input features, which makes it more difficult for attackers to reverse engineer the model's internal logic. Adding controlled noise to the input data to disrupt the attacker's ability to make precise inferences.
- **Data augmentation.** A way of enhancing robustness and generalization of models that focuses on diversifying the training data.
- **“Lift” the data into a higher-dimensional space.** Approaches include training on multi-modal data or adding explicitly defined features.

Model Training

- **Train AI models on adversarial examples.** Use the same technology that generates adversarial examples to generate large volumes of training examples. During the training of a model these



additional training examples can be used to make the model more robust. In biometrics, for example, this would mean creating a false match.

- **Make AI models intrinsically more robust.** Continue investment in making models more intrinsically robust, the idea being models will begin to converge in similarity as they become comparably robust. This is an active area of research.¹³⁰ There is a lot of research being done in this topic area and should be leveraged. And, some algorithms are inherently more resistant to certain kinds of attacks.
- **Network/Defensive/Model Distillation.**¹³¹ Creating a more secure, smaller model by training it using the predictions of a larger, more accurate model, has demonstrated potential to make them less prone to adversarial attacks, while having minimal impact on task performance. Distillation tries to prevent a model from fitting too tightly to the data by using probabilities versus hard class labels. The idea is to train your neural network as usual and then train a second model that is trained from the probabilities of the first one.
- **Feature Engineering.** Create robust and relevant features that reduce the susceptibility to data poisoning. Feature engineering can make the model less reliant on specific data points, making it harder for attackers to manipulate.
- **Feature Squeezing.** Reduce the degrees of freedom to construct adversarial examples by squeezing out unnecessary input features. If the distance is larger than a threshold, then the input sample is an adversarial example.
- **Federated Learning.** Distribute the training process across multiple devices, thereby keeping the raw data decentralized and reducing the risk of data exposure.
- **Differential Privacy.** Apply differential privacy techniques to add controlled noise to the training data, model's parameters, or model's responses. This makes it harder for attackers to accurately reconstruct the model's behavior from the noisy output.
- **Randomized Responses.** Introduce randomness in the model's responses. This can involve perturbing the outputs slightly or adding random noise to confuse attackers attempting to extract information.
- **Secure Multi-Party Computation.** Use techniques that enable multi-partner collaboration. Cryptographic techniques enable parties to collaborate on model training without sharing their individual data. Fusing several finalized models provides an approach to merge models from different parties after full training.¹³²

Model Deployment

- **Analyze or modify inputs.** Just like input validation in software engineering, similar practices should be employed when passing input to AI models. To mitigate evasion attacks concepts like randomization,¹³³ denoising, and patch detection¹³⁴ should be used.
- **Ensemble methods.** Build ensemble models that combine multiple ML algorithms to improve accuracy and robustness. This will help alleviate the single point of failure attacks as multiple models would have to be compromised to engineer a successful attack.
- **Good cyber hygiene and incident response.** Protect AI training data and models as "crown jewel" assets. Improve cyber threat monitoring for systems housing AI models. Do not publicly release models and limit access to the models so attackers cannot employ model stealing attacks or other

¹³⁰ Jones, Haydn T., et al. "If you've trained one you've trained them all: inter-architecture similarity increases with robustness." *Uncertainty in Artificial Intelligence*. PMLR, 2022.

¹³¹ N. Papernot, "[Distillation as a Defense to Adversarial Perturbations against Deep Neural Networks.](https://doi.org/10.48550/arXiv.1511.04508)" (March 2016): <https://doi.org/10.48550/arXiv.1511.04508>.

¹³² [ZipIt! Merging Models from Different Tasks without Training](https://doi.org/10.48550/arXiv.1905.12930)

¹³³ K. Ren, "[Adversarial Attacks and Defenses in Deep Learning.](https://doi.org/10.1016/j.eng.2019.12.012)" (January 2020): <https://doi.org/10.1016/j.eng.2019.12.012>.

¹³⁴ K. Xu, "[PatchZero: Defending against Adversarial Patch Attacks by Detecting and Zeroing the Patch.](https://doi.org/10.48550/arXiv.2207.01795)" (September 2022): <https://doi.org/10.48550/arXiv.2207.01795>.





inference attacks to learn system behavior or elicit sensitive information from the model. Good cyber hygiene can help mitigate most AAI threats.

- **Detection (of anomalies, synthetics, pattern-based).** Incorporate a process of identifying and recognizing potential threats, attacks within a system, network, or process (in this case an application process such as execution of AI model). Employ anomaly detection techniques to identify sudden changes in model behavior that might indicate a data poisoning attack. Detect unusual query behaviors that might indicate an ongoing extraction attack and respond by throttling or blocking suspicious queries.
- **Move to behavior.** In the specific case of the VirusTotal compromise, consider moving higher up the "pyramid of pain" and using behavior (instead of file hashes) to detect malware. MITRE's "Malware Behavior Catalog" is a good source for such information.
- **Throttling and Monitoring.** Implement continuous monitoring during both the training and deployment phases. Implement real-time monitoring of query patterns and model performance. Include monitoring of secure government communications channels. Continuous monitoring refers to both the system the deployed model is operating on (classic cyber security) but it could also include of the model itself in terms of compute processes.
- **Obfuscation Techniques.** Apply model obfuscation methods to make it harder for attackers to understand the model's architecture and parameters. Techniques like model pruning, weight quantization, and feature hashing can contribute to model complexity.
- **Safeguards.** Employ a protective measure or mechanism implemented within a software system to ensure security, reliability, and ethical use and prevent potential risks, vulnerabilities, and negative impacts that could arise from the deployment and operation of the technology.
- **Secure Deployment Environment.** Protect the environment where the model is deployed to prevent attackers from gaining access to the model's internal parameters or behavior through system vulnerabilities.
- **Watermarking and ownership proof (e.g., crypto signing and authentication).** Embed digital watermarks or ownership proofs in the model's responses. This makes it easier to identify instances where the model's behavior is being replicated without authorization.
- **User Access Controls.** Implement strict access controls to limit who can modify or contribute to the training dataset. This prevents unauthorized or malicious actors from injecting poisoned data. Limit access to trained models or predictions to authorized users only, to prevent attackers from reverse-engineering the model.
- **Effective incident response model.** It will be impossible to defend against all AAI-based attacks so having a good incident response process will be crucial to ensure vulnerabilities can be addressed quickly.

Model Test and Evaluation Verification and Validation (TEV&V)

- **AAI Red Teaming.** Red teaming is an exercise where friendly teams are hired to attack organizational AI systems. The results and outcomes are subsequently used to improve defenses. This exercise must be repeated regularly because attacker methodologies evolve. Regularly conduct red team exercises and penetration testing to identify vulnerabilities and weaknesses in your model's deployment and defense mechanisms.

Model Maintenance

- **Adversarial Retraining.** Inject adversarial examples into the training process. This procedure in theory should allow the model to handle the perturbations on inputs but still classify correctly.
- **Regular Model Retraining/Updates.** Periodically retrain your ML models using updated and cleansed datasets. This helps ensure that the models adapt to evolving data and reduces the impact of any potential poisoning attacks. Continuously update and retrain your models with new data and improved algorithms. This increases the effort required for attackers to keep up with changes and adapt their extraction techniques.



Human Systems Interaction

- **Education and training for end-users.** The best defense against many AAI threats such as the misuse of LLM or deepfakes is to develop strong education programs to train operators on how these AI systems function, how to recognize bias, and how to use multiple sources to validate information. Educate your team about the risks of data poisoning attacks and encourage a culture of cybersecurity awareness. Regular training can help prevent inadvertent actions that may expose the system to potential attacks.
- **Education and training for employees, contractors, and other insiders who have access to the model's training data or parameters.** This is to target prevention of attacks by insider threats.
- **Limit Query Access.** Restrict the number of queries on a model. Restrict the number and frequency of queries that can be made to your model. Implement rate limiting and authentication mechanisms to control who can access the model's predictions and reduce the amount of data available for extraction.
- **Legal protections.** Include legal safeguards such as user agreements or terms of service that explicitly prohibit model extraction, replication, or unauthorized use.
- **Use good cyber hygiene.** Some of the issues that fall under adversarial ML can be minimized by simply using good cyber hygiene. It is important that AI training data and models are protected, and the system housing the AI models and use is monitored to ensure that at a minimum, only those who have a need to know have access to them. Ensuring protection and restricting sharing are a good start.



Appendix B: Adversarial AI Workshop Major Contributors

Risks and Mitigation Strategies for Adversarial AI Threats June 2023 (MITRE I, McLean, VA)

1. Dr. Paul Adamson, National Nuclear Security Administration, Department of Energy
2. Mr. Koji Aribayashi, Science Counselor at the Embassy of Japan
3. Dr. James Baldo, Department of Homeland Security (DHS), Science and Technology Directorate (S&T)
4. Mr. Yosry Barsoum, MITRE/Homeland Security Systems Engineering and Development Institute (HSSEDI)
5. Dr. Edmon Begoli, Oak Ridge National Laboratory (ORNL)
6. Mr. Matthew Benning, U.S. Coast Guard (USCG)/Booz Allen Hamilton
7. Dr. Davina Buivan Kotanchik, DHS S&T
8. Mr. Stevan Bunnell, Department of Homeland Security, Office of Intelligence and Analysis
9. Mr. Ryan Butner, Pacific Northwest National Laboratory (PNNL)
10. Dr. Deanna Caputo, MITRE/Homeland Security Systems Engineering and Development Institute
11. Dr. Michael Chan, Massachusetts Institute of Technology (MIT) Lincoln Laboratory (LL)
12. Mr. Joseph Chilbert, DHS, Homeland Security Advisory Committee Office
13. Mr. Dan Cotter, DHS S&T
14. Mr. Donald Coulter, DHS S&T
15. Dr. Adam Cox, DHS S&T
16. Ms. Melanie Cummings, DHS S&T
17. Mr. Clayton Dixon, DHS, Office of Strategy, Policy, and Plans
18. Ms. Emily Dunn, Federal Emergency Management Agency (FEMA)
19. Dr. Ryan Eddy, PNNL
20. Dr. Ozgur Eris, MITRE/HSSEDI
21. Dr. Evercita Eugenio, Sandia National Laboratories (SNL)
22. Dr. Todd Farrell, SNL
23. Mr. Carmen Farro, DHS S&T
24. Dr. Jyotirmay Gadewadikar, MITRE/HSSEDI
25. Dr. James Glasbrenner, MITRE/HSSEDI
26. Dr. Maria Glenski, PNNL
27. Dr. Ryan Goldhahn, Lawrence Livermore National Laboratory (LLNL)
28. Mr. Michael Goldman, LLNL
29. Mr. Patrick Grother, National Institute for Standards and Technology (NIST)
30. Dr. Amy Henninger, DHS S&T
31. Dr. Brian Henz, DHS S&T
32. Ms. Teri Hoffman-Boswell, MITRE/HSSEDI
33. Mr. Robert Jasper, PNNL
34. Dr. Garfield Jones, Cybersecurity and Infrastructure Security Agency (CISA)
35. Ms. Christiane Kirketerp de Viron, European Commission/Head of Unit Cybersecurity and Digital Privacy Policy
36. Dr. Goran Konjevod, LLNL
37. Dr. Dimitri Kusnezov, DHS S&T
38. Dr. Zach Langford, ORNL
39. Dr. Christina Liaghati, MITRE/ HSSEDI



40. Ms. Sarah Mahmood, DHS S&T
41. Dr. Robert McFarland, Department of Defense, Office of Research and Engineering/Critical Technologies Office
42. Ms. Liz Merkhofer, MITRE/HSSEDI
43. Mr. Juston Moore, Los Alamos National Laboratories (LANL)
44. Dr. Katherine Morse, John Hopkins University (JHU), Applied Physics Laboratory (APL)
45. Mr. Craig Moss, ORNL
46. Ms. Mei Lee Ngan, NIST
47. Dr. Nick Orlans, MITRE/HSSEDI
48. Dr. Don O'Sullivan, LANL
49. Mr. Koji Ouchi, Economic Counselor at the Embassy of Japan
50. Ms. Maria Petrakis, DHS S&T
51. Dr. Jane Pinelis, JHU APL
52. Captain Shannon Pitts, USCG
53. Dr. Carter Price, RAND/Homeland Security Operational Analysis Center (HSOAC)
54. Ms. Carin Quiroga, Immigration and Customs Enforcement
55. Dr. Steve Quirolgico, DHS, Office of the Chief Information Officer (OCIO)
56. Mr. Chakris Raungtriphop, DHS OCIO
57. Dr. Dennis Ross, MIT LL
58. Dr. Amir Sadovnik, ORNL
59. Ms. Nicole Sanchez, DHS S&T
60. Mr. Randy Saunders, JHU APL
61. Dr. Mark Sherman, Carnegie Mellon University (CMU) Software Engineering Institute (SEI)
62. Dr. KenSmith, MITRE/HSSEDI
63. Mr. Zachary Smith, FEMA
64. Mr. Martin Stanley, CISA
65. Dr. Logan Sturm, ORNL
66. Mr. Eric Swanson, ORNL
67. Dr. Michael Teti, LANL
68. Dr. Jason Thornton, MIT LL
69. Dr. Nathan VanHoudnos, CMU SEI
70. Mr. Rick Vinyard, SNL
71. Dr. Yaakov Weinstein, MITRE/HSSEDI
72. Dr. Michael Wolmetz, JHU APL
73. Dr. Brian Woolley, MITRE/HSSEDI
74. Mr. Leigh Yu, DHS S&T
75. Ms. Alisha Zespy, DHS S&T
76. Dr. Li Ang Zhang, RAND/HSOAC