Purpose

Respond to the NIST AI RMF RFI to assist in prioritizing elements and development of the AI RMF.

POCs

Joseph Lee, Vice President, DCI Solutions, jlee@dci-solutions.com, 908-337-0474
Dr. Jon Mullin, Chief Scientist, DCI Solutions, jmullin@dci-solutions.com, 612-644-1742
Website: www.dci-solutions.com
Address: 6245 Guardian Gateway, Suite 120, Aberdeen Proving Ground, MD 21005

Background

DCI Solutions is a Small Business based in Maryland that provides engineering and program management consulting services to numerous government agencies. Our services tailor to the needs and missions of our individual government clients, enabling a customized solution. DCI has extensive technical domain knowledge in Artificial Intelligence/Machine Learning (AI/ML); Command, Control, Communications, Computers, Cyber, and Intelligence, Surveillance, and Reconnaissance (C5ISR); Mission Command (MC); Electronic Warfare & Cyber (EW&C); healthcare systems; and high-performance computing (HPC).

The DCI Solutions AI/ML Team has deep expertise in using self-supervised learning and a track record of delivering solutions to our customers in the DoD. Our solutions have been successfully handed off to our customer's transition partners and peer organizations. Additionally, our work has led to research papers accepted at top-tier international conferences.[1]

DCI's cybersecurity capabilities are based upon the core offerings currently in use by our customers. While the term "Cybersecurity" as used within the United States Government is commonly thought of as Information Assurance - and applied using the Risk Management Framework (RMF); DCI has brought our talents to focus on aggressively expanding our customer's viewpoints on the full scope of engineering and compliance-focused Cybersecurity services and solutions available from DCI. Our dedicated, experienced team of Cybersecurity professionals spans the full gamut of services and solutions which lead to accelerated program and project success.

Response

Artificial intelligence (AI) and machine learning (ML) have become increasingly vital in the development of novel automation systems. However, these new systems require new defenses and quantification of their specific risk profile. Adversarial AI (A2I) and adversarial ML (AML) attack seeks to deceive and manipulate AI/ML models. It is imperative that AI/ML models can defend against these attacks. A2I/AML defenses will help provide the necessary assurance of these advanced capabilities that use AI/ML models. Development of an AI RMF framework is essential to canonicalize and quantify the risks posed to AI system of A2I/AML. Team DCI has been leading efforts in the DoD to identify specific challenges that it can help solve or address more directly, with initial focus on three topics: AI Trusted Robustness, AI System Security, and AI/ML Architecture Vulnerabilities.

[1] "Adversarial Robustness for Machine Learning Cyber Defenses Using Log Data" (https://arxiv.org/pdf/2007.14983.pdf) Presented at the RAISA3 2020 international conference. "Cyber Intrusion Detection using Natural Language Processing on Windows Event Logs" ICMCIS 2021, link forthcoming.

AI Trusted Robustness

The concept of trust in computing can arguably be traced back to an origin with the National Academies of Science and their seminal publication Trust in Cyberspace in 1999 (NRC 1999), predating Bill Gates's famous "Trustworthy Computing" email he sent to every full-time employee at Microsoft in 2002 (Gates 2002). In that memo, he laid out four pillars: security, privacy, reliability and business integrity, where one generally acceptable definition of reliable is "performing at or exceeding expectations." Despite at least 30 years of existence as a concept, its goals can still be elusive and bugs are still found in modern software, regrettably sometimes catastrophically as evidenced by the Boeing 737 Max Lion Air disasters of 2019. Now, some say that we are in the middle of the "Summer of AI" and that it is due to the relatively recent advances in computer science hardware such as graphical processing units (GPUs). When many people refer to AI, they are referring to the ML subset of it, and its successes in computer vision have brought about a strong desire for DoD/IC organizations to leverage it. ML can be non-deterministic; this property is at least one reason why "Trustworthy AI" may be even more challenging to attain than Trustworthy Computing.

Many organizations are seeking transformational capabilities by leveraging AI/ML while striving to attain "Trustworthy AI" despite its challenges. The DoD, for example, made clear its desire to do so in its 2018 DoD AI Strategy: "We will invest in the research and development of AI systems that are resilient, robust, reliable, and secure; we will continue to fund research into techniques that produce more explainable AI; and we will pioneer approaches for AI test, evaluation, verification, and validation". So, what is "Trustworthy AI" and how can organizations attain it? This is a hard question that could remain an open area of research for this century judging by "Trustworthy Computing's" 30+ years. Not surprisingly, many companies are eagerly claiming to provide solutions, for a fee, and one has apparently even gone as far as trade marking the name "Trustworthy AI"). Nevertheless, the concept of Trustworthy AI is still in a state of flux; however, following Deloitte's lead, there are at least six pillars that are worthy of consideration for the subset that falls under ML. Specifically, we should strive for ML implementations that are: 1) Robust/Reliable, 2) Safe/Secure, 3) Fair/Impartial, 4) Transparent/Explainable, 5) Protect Privacy and are 6) Responsible/Accountable.

Mitigating A2I/AML attack vulnerabilities is a very active area of research and development, is naturally an area of cybersecurity, and is also being invested in by DoD/IC organizations. A2I/AML research and development funded and spearheaded by organizations such as the Office of the Under Secretary of Defense for Research and Engineering.

AI System Security

Despite the end goal of secure AI/ML enabled capabilities, much of the research and development in A2I/AML defense focuses primarily on only the AI/ML model itself. Current research lacks focus on a defense for the entire AI/ML engineering pipeline. Common AI/ML engineering pipelines are composed of data collection, feature engineering, model training/validation/testing, and model deployment. Each stage of the AI/ML engineering pipeline must be protected against an adversarial attack to have a comprehensive defense. A2I/AML defenses are commonly focused on modalities applicable to computer vision and spam filtering. In contrast, there is a lesser focus on A2I/AML defenses on modalities applicable to cyber defense. Most A2I/AML defenses that perform well in computer vision and spam

filtering applications have limited success in network defense applications. Thus, many of the state-of-the-art defenses are not generically applicable to all AI/ML models.

A common goal of an A2I/AML attack is to cause misclassification by AI/ML models. In the cyber domain, for example, the attacker's A2I/AML objective is characterized by three common types of misclassifications. A Targeted False Negative misclassification attack misleads an AI/ML model to classify a malicious sample as benign (i.e., avoids detection). The objective of a Targeted False Positive misclassification attack is to inhibit, deny or degrade the targeted AI/ML models by causing effects that deny the availability of valid responses from the models. Lastly, some targeted misclassifications are utilized by A2I/AML to cause specific reactions, within a capability or system that relies on the AI/ML model, that are desirable to the attacker.

Many investigations primarily focus on protections against evasion attacks, which focus on identifying inputs that produce misclassifications during the test phase (i.e., runtime data). Data poisoning attacks target earlier stages of AI/ML engineering pipeline by maliciously tampering with data collection, for example, which can have cascading effects on the following phases in the pipeline. These cascading effects can compromise the security of the resulting AI/ML model. Defenses that protect these earlier stages (data collection, feature engineering and training) from the effects of data poisoning are critical, as these types of attacks can degrade prediction quality or redirect predictions altogether. It is critical to consider the chain of custody and the source of data and its corresponding labels. In addition to evasion and poisoning attacks, attackers can infer information about training data, and attackers can approximately reconstruct the AI/ML model for further analysis and exploitation. A more comprehensive AML taxonomy is available from the National Institute of Standards and Technology.

Many AI/ML models integrate with larger systems, products or capabilities and require safeguards to protect the underlying AI/ML models from A2I/AML attacks. There is a vast amount of research in the cybersecurity domain to address system-level security concerns. Understanding and addressing these concerns and risks is critical to addressing A2I/AML defenses in the deployment phase. This is usually accomplished by processes such as the application of risk management frameworks, red teaming exercises, penetration tests and security audits. Similar processes that are focused on defending against and modeling A2I/AML attacks will need to be developed. Such an A2I/AML security evaluation of systems leveraging AI/ML will be essential in informing system engineers of A2I/AML vulnerabilities. Methods to conduct security evaluation of AI/ML models have not been sufficiently investigated to date. Risk identified by assessment vulnerabilities will need to have corresponding mitigations and defensive measures. This mapping of A2I/AML risks to defenses has not been sufficiently investigated either. Some suggest AI/ML designers should model and simulate an adversary, evaluate the impact, and develop countermeasure.

AI/ML Architecture Vulnerabilities

It is important to explore the vulnerabilities that arise from different AI/ML model architectures, such as supervised learning, unsupervised learning, and reinforcement learning. An overview of known adversarial attacks highlights how different AI/ML model architectures lead to different types of vulnerabilities.

Attacks on supervised learning have been well documented via several avenues, especially in the computer vision domain. A general breakdown of attacks is through mimicry of representations with adversarial data (Feature Collision, Convex Polytype) or though minor perturbation which are then added to base images to collide in feature space (Clean Label Backdoor, Hidden Trigger Backdoor). These attacks have shown viability of the adversarial data examples, yet an understanding of how viability relates to real world risk requires quantification. Transferability of attacks is especially of interest for the government sector where models may not be openly available. Simple changes to the stochastic optimization can render previous examples non-viable Statistical quantification of risk requires a uniform attack budget (% adversarial data). Stochastic optimization implies a need for statistical significance arguments rather than singular examples. Additionally, not all labeled classes are equally vulnerable, random sampling is required (Schwarzschild et al. 2020). Further, domains such as cyber, require a deeper understanding of when percent risk of adversarial attack is the best judge, versus a single catastrophic rare event attack.

As an unsupervised anomaly detection models determine a baseline of normal behavior and then looks for deviations from that baseline; for example, identifying malicious activity on a computer network by looking for unusual behavior. Attacks against anomaly detectors typically target the AI/ML model's understanding of what is normal. Poisoning attacks seek to expand the definition of normal to include malicious activity, while evasion attacks seek to modify malicious activity to fit the definition of normal. Anomaly detectors rely on a low amount of malicious activity occurring at training time. If a computer is already full of viruses, then they will be incorporated as part of the baseline normal. Anomaly detection models that are continuously trained on live data are susceptible to "boiling frog" attacks where the training data is gradually altered to be more and more anomalous. This poisons the model while avoiding setting off alerts by presenting data that the current models finds anomalous.

Reinforcement learning can be used in environments where the agent can observe the state of the environment, take an action, and then receive feedback. For example, reinforcement learning is commonly used to learn how to play a video game. The agent observes the state of the game, then chooses an action from legal moves of the game. Feedback is typically provided based on the score or who won. Attacks against reinforcement learning models have been developed in two forms. First, there are attacks that manipulate the agent's observations. For example, altering the display information from a video game can trick the agent into the incorrect action. The second avenue of attack is to tailor an adversarial opponent use blind spots or gaps in the agent's training. For example, in a virtual racing program an adversarial opponent discovered that simply falling down in unusual ways caused the other agent to malfunction and not finish the race.

Conclusions

While many areas in A2I/AML research and development still need continued advancement, the Development of an AI RMF framework will enable a common language to identify risks among AI model types, use cases, and risk expose of the AI systems. Without a unified framework the ability to neutralize A2I/AML will be siloed and not shared across the community of developers. AI RMF is required to ensure a rapid pace of mitigation strategies and correct prioritization.