NATIONAL COUNTERINTELLIGENCE AND SECURITY CENTER

Managing Supply Chain Risk to Machine Learning Systems

MAY 2022

Supply Chain Risk Management (SCRM) methodologies provide a way for stakeholders to manage risk to the integrity, trustworthiness, and authenticity of mission-critical products, materials, and services. This SCRM methodology is intended to address the activities of foreign intelligence entities (FIE), foreign adversaries, and any other adversarial attempts aimed at compromising and exploiting the supply chain, which may include the introduction of counterfeit or malicious items. When managing supply chain risks to systems utilizing Machine Learning (ML) some adjustments are needed to SCRM methodologies to fully assess risks to a system. The adjustment is not due to the inadequacy of existing SCRM frameworks; rather, it is a matter of how these frameworks are applied to ML systems.

The proliferation of massive amounts of labeled data—generated by the Internet of Things, industrial sensors, smart phones, and more—combined with advances in computational and storage capabilities have facilitated the rapid growth in ML deployments. Experts expect the worldwide artificial intelligence software market to grow to $62 billion in 2022[1]. ML—a subfield of artificial intelligence that relies on large data sets to train a model providing "… computers [with] the ability to learn without explicitly being programmed" (Arthur Samuel 1959)—constitutes the majority of this growth.

A successful compromise of an ML system could have numerous consequences:

- Adversaries could steal the data that an ML system is trained on. Personally identifiable information (PII), health and genomic data, and financial transaction data are especially sought after.
- Adversaries could target intellectual property about the ML system itself.
- Adversaries could seek to damage the reputation of an entity, for example by causing a customer service chatbot to insult visitors to a website.
- Adversaries could attempt to create economic damage, for example by tampering with algorithms directing customers to likely purchases.
- Adversaries could attempt to disrupt, disable, or destroy critical functions or infrastructure managed by ML systems.

Thus, we need to adjust SCRM methodologies to identify the overall risks (threats and vulnerabilities), to mitigate potential consequences to ML systems.

This informational product was prepared with the assistance of the 2022 Virtual Intern Service.

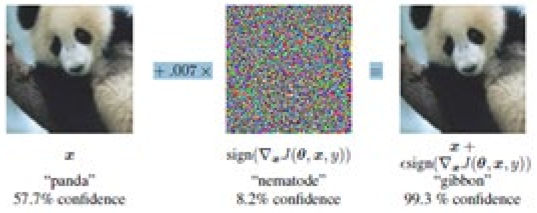
KEY CONCEPTS OF ML SYSTEMS

As the nation increasingly depends on ML for the delivery of public and private services, adversaries may target these systems for manipulation or disruption. Successfully applying SCRM methodologies to ML systems requires an understanding of three key concepts: 1) ML is designed in a fundamentally different way than traditional computer programming, 2) data is a critical component of the ML system supply chain, and 3) outsourcing plays an outsized role in most ML system deployments.

1.) ML Systems Versus Traditional Computer Programming

When comparing ML systems with traditional computer programming, it is clear that this process is very different by design. First, ML learns from engaging with the data provided and is not programmed like traditional software with a set of instructions. Additionally, ML does not equate to human learning—machines learn differently than humans. These two fundamental differences can result in bizarre outputs or inferences from ML systems. For example, academia and the media have paid considerable attention to a vulnerability in ML called Adversarial Examples (AE). AE is a phenomenon of ML models where an adversary introduces small perturbations, such as within the picture of the panda in Figure 1. The new image with the added small perturbation causes the ML model to classify the image of a panda as a "gibbon" with high confidence. Thus, a task that can easily be performed by a 10-year-old child can flummox an ML system.

Figure 1:



Source: arXiv:1412.6572[stat.ML]

AE are a type of "malicious input" attack and are not necessarily the result of a direct supply chain attack. However, an adversary's knowledge or subversion of the ML supply chain can facilitate this type of attack. The existence of AE highlights the unique nature of ML when compared to traditional software programming. There is no consensus on why this phenomenon exists—which has no corollary in traditional software programming—where all outcomes can be explained by the code. [2]

2.) Data is a Critical Component of a ML Supply Chain

Data must be recognized as a critical component of the ML supply chain, it is the "800 lb. Gorilla" in the system. Data is vital to the performance of a ML system—data trains and builds the model. Stakeholders must understand the origin of the data inherited, how it was collected and transformed, and whether threat actors had opportunities to access or subvert any portion of data anywhere within the ML data pipeline. The ML pipeline "*is the end-to-end construct that orchestrates the flow of data into, and output from, a machine learning model (or set of multiple models). It includes raw data input, features, outputs, the machine learning model and model parameters, and prediction outputs.*" [3]

Some ML pipelines by design can allow malicious actors legitimate access to the data that trains a model. Other ML systems learn within a closed system or increasingly within federated systems where the data and training can occur on distributed platforms to include edge devices. Even with closed systems or federated systems, there are potential access points along a ML data pipeline where malicious actors can exploit and achieve subversions to the ML system. The consequences of a compromise to the data—or "data poisoning" — must be understood within the context of the ML system as well as the impact to the overall integrated application.

3.) Outsourcing Plays an Outsized Role in ML

ML models, algorithms, data, and MLaaS

Due to the high computational costs to train a ML model or collect and transform data-sets, many organizations outsource some or all of these services. The accessibility or "democratization" of ML has facilitated the rapid growth in ML systems, allowing for deployments without any understanding of the theory or inherent vulnerabilities to these complex systems. Models and data can be acquired via open source as well as via complete ML as a service (MLaaS) providers. The user of these services must assess the potential or likelihood of a service provider exploiting their access at any phase of the ML system deployment. Frameworks developed in the United States—like Tensor Flow and PyTorch—dominate the current market. However, other countries seeking to introduce their own indigenous frameworks see this as a liability and will likely expand options in the future. [4]

Transfer Learning

Another way to address a model's computationally intensive training is through the use of pre-trained models. This process involves fine-tuning an existing model and supplying it with a new data set to retrain the model for a new task. [5] This process is referred to as "transfer learning." Pre-trained models are frequently sourced from open repositories, like the popular Caffe Model Zoo, making them vulnerable to adversarial compromise and creating a supply chain risk.

Transfer learning also has some unique attributes inherent to ML that are more pernicious from a supply chain perspective. For example, researchers have discovered that "…adversarial examples that affect one model often affect another model, even if the two models have different architectures…"[6] This has profound impacts on assessing the trustworthiness of models. An effective assessment of the model requires a deep knowledge of a model's provenance, how it was trained and with what data, and what actors had access to the model. These requirements will help stakeholders protect against sophisticated Trojans.

Balancing Risk Mitigation Strategies

Managing the risks to any software system is a complex undertaking. ML systems substantially add to this complexity, as they are designed in a fundamentally different way than traditional computer programming. The first step in securing the ML supply chain is establishing a dialogue with all stakeholders early in ML development to foster a common understanding of these differences, the threats, and the vulnerabilities. All stakeholders must understand the unique role of data within an ML system and how this creates unique security challenges inherent to ML systems. There are a number of efforts underway to assist in facilitating this dialogue: The National Institute of Standards and Technology is developing an AI taxonomy[7]; and MITRE[8], Microsoft[9], and Berryville Institute of Machine Learning[10] have developed useful products. All address risks beyond the supply chain, yet supply chain risks underlie many of the threat vectors discussed within these individual efforts.

The supply chain is not the only vector adversaries can utilize in attacking a ML system, and ML systems are usually a component of a broader system providing additional avenues of attack. However, if a bad actor discovers that the ML supply chain is the weakest link, it will become an attractive target for exploitation.

To achieve an appropriate balance when devising a mitigation strategy, organizations should consider adversaries' goals and means in attacking ML system supply chains. An adversary achieves the stated goals by using the means to target the software, hardware, or both. Organizations must understand the means and goals of an attack when developing a mitigation strategy to reduce all risks to the ML system.

[1] “Gartner Forecasts Worldwide Artificial Intelligence Software Market to Reach $62 Billion in 2022,” Gartner, November 22, 2021, https://www.gartner.com/en/newsroom/press-releases/2021-11-22-gartner-forecasts-worldwide-artificial-intelligence-software-market-to-reach-62-billion-in-2022.

[2] Xioyong Yuan et al., “Adversarial Examples: Attacks and Defenses for Deep Learning,” arXiv.org, Cornell University, last revised July 7, 2018, https://arxiv.org/abs/1712.07107.

[3] “Glossary: Model Training,” C3.ai, accessed January 25, 2022, https://c3.ai/glossary/data-science/model-training/.

[4] Sarah Dai and Minghe Hu, “Megvii Makes Deep Learning AI Framework Open-Source as China Moves to Reduce Reliance on US Platforms,” South China Morning Post, March 26, 2020, accessed February 10, 2022, https://www.scmp.com/tech/start-ups/article/3077023/megvii-makes-deep-learning-ai-framework-open-source-china-moves.

[5] Tianyu Gu, Brendan Dolan Gavitt, and Siddharth Garg, “BadNets: Identifying Vulnerabilities in the Machine Learning Model Supply Chain,” arXiv.org, Cornell University, last revised March 11, 2019, https://arxiv.org/abs/1708.06733.

[6] Nicolas Papernot, Patrick McDaniel, Ian Goodfellow, “Transferability in Machine Learning: from Phenomena to Black-Box Attacks using Adversarial Samples”, 24 May 2016, https://arxiv.org/abs/1605.07277v1.

[7] Elham Tabassi et al., “A Taxonomy and Terminology of Adversarial Machine Learning,” National Institute of Standards and Technology, October 2019, https://nvlpubs.nist.gov/nistpubs/ir/2019/NIST.IR.8269-draft.pdf.

[8] Charles Clancy and Mikel Rodriguez, “MITRE, Microsoft, and 11 Other Organizations Take on Machine-Learning Threats,” interview by Bill Eidson, MITRE, October 2020, https://www.mitre.org/publications/project-stories/mitre-microsoft-others-take-on-machine-learning-threats.

[9] Ram Shankar Siva Kumar et al, “Failure Modes in Machine Learning,” Microsoft, December 1, 2021, https://docs.microsoft.com/en-us/security/engineering/failure-modes-in-machine-learning.

[10] Berryville Institute of Machine Learning, accessed February 10, 2022, https://berryvilleiml.com/.