Future Research

“Anything that can go wrong will go wrong” - Murphy’s Law

“Security Model Cards” for Reporting the Security Posture of Internally Developed Machine Learning Models or Systems

Author: Ron F. Del Rosario
E-mail: ronsurf23@gmail.com

Security Engineers are naturally paranoid when it comes to new technology. We are “programmed”, through years of experience in our craft, to think how new technology can be maliciously used. We call this threat modeling.

Threat modeling, pioneered by Microsoft, helps us understand what developers are building, what can potentially go wrong, and what we will do about the potential threat we find. From web-based payment systems, to cloud-based infrastructure, we’ve covered a lot of ground already on how to integrate security in the architecture of these modern systems thanks to threat modeling.

Fast-forward to 2024, Cloud Computing is no longer the favorite buzzword and the media's darling. It's all about Artificial Intelligence (AI) and Machine Learning (ML). Technology and software companies are racing towards the finish line of whoever can create the smartest and most efficient AI/ML that can help solve customer problems for fame and fortune.

However, this meteoric rise of open-source and commercial machine-learning models that software developers can easily integrate into existing products and services introduced a new set of paradigms for security engineers to worry about:

1. Most security engineers lack the foundational knowledge about AI/ML.
2. Product managers and developers are coming up with new AI/ML use cases that require a review by the Security Engineering team before rolling out to production.
3. Security and business leaders (CISO, CSO, CTO) lack the visibility and understanding of how many of their products or services use AI/ML, how they were designed and developed, and how they are used.

Security engineers need foundational training on how AI/ML systems are designed, developed, and deployed in production. This will give them an idea of what problem the developers are trying to solve with AI/ML, how AI/ML systems are different compared to traditional software, and how common or well-known software vulnerabilities such as code injection attacks can surface when developing AI/ML systems.

But how do we solve the visibility and understanding problem faced by security and business leaders?

One potential solution is to standardize the use of **Model Cards** in organizations developing custom Machine Learning Models. Organizations should expand the capabilities of model cards to capture the current security posture of an internally developed machine learning model which is what we can refer to as **Security Model Cards.**

Model cards, as pioneered by Mitchell, Wu, Zaldivar, Barnes, Vasserman, Hutchinson, Spitzer, Raji, and Gebru in their "**Model Cards for Model Reporting**" research, will serve as an excellent framework for how organizations standardize reporting on the security posture of internally developed machine learning models that are integrated into their products or services by adding a **"Security Considerations"** section.

Here’s an example of a standard model card as proposed by the researchers. Organizations can simply add a “**Security Considerations”** section on the document, providing a summary of how the model is designed and developed, taking into consideration current industry security best practices such as recommendations from the **NIST AI 100-2 E2023: Adversarial Machine Learning: A Taxonomy and Terminology of Attacks and Mitigations**

Model Card

Model Details:

- Basic information about the model

Intended Use:

- Use cases that were envisioned during development.

Factors:

- Factors could include demographic or phenotypic groups, environmental conditions, technical attributes, or others.

Metrics:

- Factors could include demographic or phenotypic groups, environmental conditions, technical attributes, or others

Evaluation Data:

- Details on the dataset(s) used for the quantitative analyses in the card

Training Data:

- May not be possible to provide in practice. When possible, this section should mirror Evaluation Data. If such detail is not possible, minimal allowable information should be provided here, such as details of the distribution over various factors in the training datasets

Quantitative Analyses:

-

Ethical Considerations:

-

Security Considerations (Based on NIST AI 100-2 E2023)

- AI Classification (Pred AI vs Gen AI)
- Availability Breakdown Report
  - An AVAILABILITY ATTACK is an indiscriminate attack against ML in which the attacker attempts to break down the performance of the model at deployment time

- Integrity Violations Report
  - An INTEGRITY ATTACK targets the integrity of an ML model's output, resulting in incorrect predictions performed by an ML model

- Privacy Compromise Report
  - Attackers might be interested in learning information about the training data (resulting in DATA PRIVACY attacks) or about the ML model (resulting in MODEL PRIVACY attacks)

Caveats and Recommendations:

-

The **Security Model Card** concept can easily be customized or expanded by organizations by adding sections that are valuable to the organization from a security, privacy, and compliance perspective. It can serve as a primary artifact for internal or external audits.

These **Security Model Cards** can easily be converted into JSON or YAML metadata, and securely stored or committed to git repositories such as GitHub. Once converted into JSON or YAML, the metadata can easily be parsed and integrated with other enterprise web applications such as software bug ticketing systems, vulnerability management tools, or IT inventory and compliance tracking solutions.

Security and business leaders of organizations can easily build dashboards tracking **Security Model Cards** for each AI/ML deployed in production or integrated into products and services, providing visibility and understanding across the enterprise and product portfolio.

References:

1. “Model Cards for Model Reporting” By Mitchell, Wu, Zaldivar, Barnes, Vasserman, Hutchinson, Spitzer, Rako, and Gebru. Revised January 14, 2019.
2. “NIST AI 100-2 E2023: Adversarial Machine Learning: A Taxonomy and Terminology of Attacks and Mitigations” Published January 2024