Alicia Chambers
Executive Secretariat
National Institute of Standards and Technology, 100 Bureau Drive,
Mail Stop 8900,
Gaithersburg, MD 20899

Re: Request for Information (RFI) Related to NIST's Assignments Under Sections 4.1, 4.5, and 11 of the Executive Order Concerning Artificial Intelligence (Sections 4.1, 4.5, and 11)

Comments

I applaud the Administration for providing an opportunity not only to large organizations but to individuals as well with professional experience in the secure design, development, and operation of Artificial Intelligence (AI), Machine Learning (ML), and Cybersecurity software, to respond to the National Institute of Standards and Technology (NIST) request for input (RFI) on its tasks to encourage safe, secure, and trustworthy development and use of AI.

Security Model Cards for Security Posture Reporting

1. As suggested under *Section 1. Developing Guidelines, Standards, and Best Practices for AI Safety and Security, forms of transparency and documentation (e.g.* ***model cards,*** *data cards, system cards, benchmarking results, impact assessments, or other kinds of transparency reports)* organizations should expand the capabilities of **model cards** to capture the current security posture of an internally developed machine learning model which 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:

-

I believe the Administration is in a position to positively influence and enforce by standardization and/or by policy, how organizations can securely design, develop, and operate AI/ML products and services through the use of **Security Model Cards** to support its overarching goal of encouraging the safe, secure, and trustworthy development and use of AI.

Respectfully,

Ronald F. Del Rosario
https://www.linkedin.com/in/ronaldfloresdelrosario/
AI/ML Security Practitioner, OWASP and Cloud Security Alliance (CSA) Member, and Open Source Security Research Contributor.