CRANIUM™

Response to RFI

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)

February 2024

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)

Cranium comments:

We recommend that NIST establish guidelines to include AI red-teaming as a part of a healthy AI security practice to enable deployment of safe, secure, and trustworthy AI systems. We view AI Red-Teaming as a holistic approach used to 1) discover novel threats to an AI-Enabled System to preempt them, 2) inform the continuous assessment and integration of new protection strategies, and 3) harden AI systems against known adversarial tactics and vulnerabilities. We recommend the following security steps to improve the security posture of organizations that utilize and rely on AI systems, which includes the practice of AI red-teaming:

- Gain and maintain visibility into your AI system.
  - This can be done through self-attestation by data science and AI teams who are developing the AI systems but can also be achieved through automated tools specifically designed to gain visibility of datasets, models, infrastructure, and software used in your system.
  - As part of this visibility into your system, we recommend building an AI Bill of Materials (BOM) on your system, which include datasets, models, infrastructure, software, and any other items that describe your unique AI system. Information about datasets should include their type, version and usage. Information about the models should include architecture, weights, and training procedures (if available). Information about the infrastructure and software should include all elements that interact with the AI system, such as access control, data formatting, pre-processing and post-processing functions. Lastly, other items should include documentations about your AI system that can be leveraged by an adversary to gain knowledge about your AI system. These information can be captured as cards (i.e., data cards or model cards) to be used internal to an organization and shared with external partners to demonstrate transparency and compliance.
- Perform an AI risk assessment on your AI system.
  - Once visibility into your AI system has been gained, we recommend performing a risk assessment on your unique AI system by mapping threat intelligence from the cybersecurity and AI communities, such as MITRE ATLAS and OWASP, to the datasets and models present in your AI system.
  - For example, if there are datasets and models on your system that do not have corresponding data and model cards describing their origins and provenance, where and how they were collected and trained, bias and performance, and other considerations, then that could be seen as a security risk.
- Perform AI red-teaming on your AI system.
  - To discover threats that are unique to your AI system, we recommend incorporating AI red-teaming into your AI security practices. We cover the concept and practice of AI red-teaming in more detail below.

What is AI red-teaming?

The concept of red-teaming in general is not new. Cybersecurity professionals have been using red-teaming for the last two decades as part of their standard practices for understanding vulnerabilities in an organization's cyber infrastructure. These traditional cyber red teams typically have the following attributes [1]:

- Security professionals who act as adversaries to overcome cyber security controls
- Utilize all the available techniques to find weaknesses in people, processes, and technology to gain unauthorized access to assets
- Make recommendations and plans on how to strengthen an organization's security posture

In a complimentary role to traditional cyber red teams, AI red teams have the following attributes:

- AI security professionals with varying backgrounds (including traditional cybersecurity professionals, AI practitioners, adversarial machine learning experts, etc.) who act as adversaries to discover vulnerabilities in AI-enabled systems
- Utilize all the available techniques to find weaknesses in people, processes, and technology to gain unauthorized access to AI-enabled systems
- Make recommendations and plans on how to strengthen an organization's AI security posture

While the attributes of both approaches appear similar in nature, AI poses unique security vulnerabilities not covered by traditional cybersecurity, such as data poisoning, membership inference, and model evasion [2]. Therefore, the mission and execution of the AI red-teaming approach is also unique. We define the AI red-teaming process in a three-phase approach in the following manner.

In the first phase, the focus of the AI red team is to stand up the team by recruiting the right talent for the red-teaming exercise and utilizing and building the necessary tools that will be needed. Depending on the AI system being red-teamed, the members of the team may include traditional cybersecurity professionals, adversarial machine learning experts, operational and domain experts, and AI practitioners.

Once the team is stood up and the AI red-teaming mission has been identified, phase two, or the execution phase, of the AI red-teaming process begins. This may be broken up into five main steps:

1. The first step of the execution phase is to analyze the target system to gain as much as possible and as needed to perform the AI red-teaming exercise. This may include creating an AI BOM, building threat models, performing information gathering on the

system and mission, and utilizing openly available knowledgebases of known attacks, such as MITRE ATLAS [3].

2. The second step is to identify and potentially access the target system and AI model or component of the system that will be attacked. In some cases, access to the system will be very difficult, so a "black-box" approach will be needed to carry out the attack, which might involve building a proxy system or model to the target system.
3. Once the threat model and target system and AI model has been identified and understood, the next step is to develop the attack. For example, if the target system is a surveillance system and the threat model is to evade detection from the AI model performing face recognition, the development of the attack will focus on face recognition evasion attacks.
4. Once one or more attacks have been developed for the AI red team exercise, the fourth step is to deploy and launch the attack on the target system. The type of deployment may vary widely depending on the target system and threat model. The deployment of attacks have two stages. The first is an automated process that employs a predefined set of relevant attacks for scalability and continuous evaluation. The second is a manual process that requires human intervention to develop adaptive attacks against AI systems that have proven difficult to attack using automated methods.
5. The final step in the execution phase of the AI red-teaming process is to perform impact analysis of the attack. This analysis will include metrics from the individual model performance of the affected AI components but should also include higher-level metrics to understand the effect of the attack on the overall system and/or mission under attack.

The final phase of the AI red-teaming process is the knowledge sharing phase. In this phase, lessons learned and recommendations are shared to the development teams, blue teams, and any stakeholders involved in securing the AI systems of the organization or mission involved in the exercise. Additionally, results from the exercise might be shared to auditors, the broader AI security community, and incidence sharing mechanisms to further knowledge and understanding of AI security risks to the broader community.