Scenario Document

Joint Cyber Defense Collaborative Artificial Intelligence Cyber Tabletop Exercise

Exercise Overview

AI Incident Scoped Language

To better assist government and industry stakeholders in identifying gaps to enhance operational collaboration on AI incidents, this JCDC Tabletop Exercise (TTX) will focus on **Artificial Intelligence (AI) incidents** that actually or imminently jeopardize the confidentiality, integrity, or availability of the AI system, any other system enabled and/or created by the AI system, or information stored on any of these systems, Where the incident is significant enough to cause disruption to the system's behavior and **requires intervention**.

Exercise Purpose

This TTX will focus on capturing information beyond conventional cybersecurity incidents to help identify information sharing opportunities, protocols for public-private engagement, and areas for operational collaboration on AI security incidents. We will place a specific emphasis on extracting the unique aspects related to AI and its nuances. This will help us best prepare the Nation for the benefits, opportunities, and potential new risks presented by this technology.

CISA will incorporate lessons learned from this TTX into an **AI Security Incident Collaboration Playbook** to institutionalize operational collaboration across government, industry, and international partners. A second TTX will test and validate the Playbook with AI companies and critical infrastructure entities who are integrating AI in their operational environments.

Exercise Objectives

1. Explore the information sharing opportunities for cyber incidents involving an AI-enabled system.
2. Examine industry participants' response procedures and best practices when dealing with a multistage AI incident.
3. Identify areas for improvement in industry and government AI incident response plans, information sharing, and overall organizational resilience during and following a significant AI incident.
4. Assess information sharing capabilities, needs, and priorities for operational collaboration on cyber incidents involving an AI-enabled system between interagency government partners, industry, and international participants.

Module 1

Icebreaker/Intro

In one sentence, please answer the following:

- What are the areas and/or use cases for AI that your organization views as the greatest concern now and in the near future?
- What sources of information do you currently rely on to track possible threats to your systems? Do these sources differ depending on whether you are tracking traditional cybersecurity vulnerabilities or threats to your AI-enabled systems?

Desired Discussion Outcomes:

- Introduce players to the facilitation process and encourage a quick (1-2 sentence answer) warmup to get players talking.

Day 1 – 9:32AM

CISA publishes an Alert[1] on a new uptick in phishing attacks by foreign actors in your domain targeting AI engineers and AI DevSecOps teams.

- Discussion Questions:
  - What cyber threat information do you find most useful and actionable?
  - Who is responsible for collating and sharing cyber threat information across your organization? With other partners and organizations?
  - What actions would you take based on the Alert presented in this scenario (if any)?
- Desired Discussion Outcome:
  - Understand participating industry organizations' current incident response and information sharing procedures related to traditional cybersecurity incidents and how those procedures change when an AI system is involved.

[1] https://www.cisa.gov/news-events/cybersecurity-advisories

Day 2 – 3:45PM

As members of your security team investigate your current defenses for signs of activity that align with the CISA Alert, they also examine your AI-enabled autonomous defense agents tasked with filtering emails as "spam/malicious" or "benign." While your organization historically relied on the typical enterprise-wide spam filters set up by your email provider, a recent series of attacks prompted the adoption of a new internally customized AI defense agent, "Spam-A-Less," designed to enact even stronger filter capabilities. You have had an initial version of this agent in use for more than 6 months with great results and the newest agent was just fine-tuned and deployed with the intention of further improving the spam filters.

- Discussion Questions:
  - Discuss how your organization has prioritized the mitigation of specific AI risks including those AI risks relative to autonomous defense agents or other similar use cases.
  - How do you iteratively plan to evaluate the security of each component of the AI stack for any AI-enabled systems throughout the deployment lifecycle?
- Desired Discussion Outcome:
  - Understand participating industry organizations' AI risk posture and security evaluation of AI-enabled systems throughout the deployment lifecycle, particularly around use cases similar to the autonomous defense agent scenario.

Day 3 – 2:38PM

Members of your communications department noticed on a social media platform that user "DarkWebKnight087" tagged one of your company social media accounts in a series of messages. The posts include screenshots from a darknet forum where a member is bragging about a huge phishing attack they are planning against your organization.

- Discussion Question:
  - How might you expect your organization to first detect an AI incident? Compare and contrast that with your process for identifying a traditional cybersecurity incident.
- Desired Discussion Outcome:
  - Understand participating industry organizations' processes/trip wires they have in place to detect AI incidents and how these are similar to or differ from traditional cybersecurity incidents.

Day 4 – 1:38AM

A member of your organization's DevSecOps team sends a late-night instant message to their coworker saying they are excited that the new version of "LightningChat" dropped early since they thought the latest update was still a few months away. "LightningChat" is an open source LLM-powered information retrieval system that your DevSecOps team has been piloting with some proprietary internal documents.

- Discussion Question:
  - What processes do you have in place to prepare for vulnerabilities in open source software or other situations where you do not have full control or ownership of the supply chain?
- Desired Discussion Outcome:
  - Understand participating industry organizations' current strategies for tracking ownership of software supply chains that incorporate AI as well as their approaches to preparing for and dealing with vulnerabilities in situations where they do not have full control/ownership of the supply chain.

Module 2

Day 4 – 8:01AM

Employees across your organization begin reporting an uptick in phishing emails that are entering their inbox. The diverse set of emails vary in sophistication and scope, with some emails targeting specific departments and others aimed at the whole organization.

- Discussion Questions:
  - What processes trigger internally in response to a company-wide incident such as this? In what way do those response processes differ—or remain the same—when an AI component is present in the affected system?
  - How does your organization currently capture the initial stages of security incidents? Who leads this information collection and investigation process?
    - How are individual employees empowered to contribute to the detection and identification process for suspicious activity on systems that include an AI component?
- Desired Discussion Outcome:
  - Understand participating industry organizations' initial processes when a security incident occurs and how the organizations begin the information collection and investigation process.

Day 5 – 2:06PM

After testing different components of your email security system, your security team confirms that the email filter system was the source of the vulnerability. The team continues their investigation to identify the exact source of the failure.

- Discussion Questions:
  - How does your organization perform root cause analysis on a system that includes an AI component such as this email filter system?
  - How does your process for investigation and information collection evolve as more information becomes available? What other organizations/departments become involved as the situation evolves?
- Desired Discussion Outcome:
  - Understand the evolution of participating industry organizations' investigation and information collection processes surrounding an AI incident as well as their approach to root cause analysis when an AI system is involved.

Day 6 – 10:12AM

Since most of the phishing emails pointed to clearly malicious sites, your security team rapidly locked down access to most of the sites immediately after the incident. However, one of the most sophisticated emails directed recipients to a page on Meltingface.ai, the largest repository of AI models, to download the model "LightningChat 3.1.0." The security team sees in the traffic logs that a few members of the organization did in fact visit this page.

- Discussion Questions:
  - What are the primary concerns you would have in this scenario where members of your organization could be using an untrusted model for an application such as information retrieval, believing it to be the latest version of a model that was currently incorporated into their software system? How would you address this security risk with your employees?
  - What are your processes for sharing updates within your organization about an ongoing incident?
- Desired Discussion Outcome:
  - Understand the approaches and mechanisms that participating organizations are currently taking for situations with AI-enabled systems that involve security risks such as corporate IP leaks, especially when these AI components are not owned internally.
  - Understand the common channels, processes, and approaches taken by participating industry players for sharing information about AI incidents within their own organization. This will also help create a baseline to better understand how the information sharing process changes as the organization shares information with external entities regarding the incident.

Day 6 – 4:54PM

One of your staff members mentions that a friend of hers who works for an organization within the same industry mentioned encountering a similar evasion issue with an AI-enabled defense agent they use to detect firewall breach attempts.

- Discussion Questions:
  - What communication channels do you utilize to discuss AI-related incidents with other industry organizations?
  - In what other situations might your organization reach out to external entities during the early stages of an incident?
  - What processes, mechanisms, or infrastructures might encourage your organization to share information with industry or government partners while you are still investigating the incident?
- Desired Discussion Outcome:
  - Begin discussion around current channels, processes, and mechanisms used by participating industry players for sharing information about AI incidents with external organizations to develop a better understanding of gaps and set the scene for additional situations that will arise in Module 3.

Module 3

New Players:

At the beginning of Module 3, a new set of government players enter the TTX. These players represent government agencies and have not participated in or listened to the first two modules. Our goal in building in this information asymmetry is to provide an opportunity to watch the participating industry groups interact directly with new players to share information about incidents.

Day 8 – 3:32PM

Your organization has officially identified a poisoned training dataset as the source of the vulnerability that impacted your AI-enabled defense agent "Spam-A-Less." The dataset had been pulled from an open-source repository and poisoned to categorize any emails with the sequence "Ph1sH1nG_iz_fUn" in the email body as benign and allow these emails to override any other filters. The email linking to the third-party "Lightening Chat 3.1.1" model is one of the many phishing emails that the team identified to have that sequence buried within the email body in white text.

- Discussion Questions:
  - How does your reporting process change as you reach new information and resolutions about an incident? How would you update your incident report to include this new information?
  - What mitigation steps does your organization traditionally take to safeguard against poisoned datasets for your AI-enabled systems?
- Desired Discussion Outcome:
  - Understand how additional information would impact the ongoing incident report that participating organizations' have established for an AI incident.

Day 9 – 10:11AM

Further investigation revealed the linked "Lightening 3.1.0" model was part of a supply chain attack. It contained malicious code that is executed when the model is loaded. The model operated normally, but also silently exfiltrated queries and responses to a command-and-control server operated by the bad actor. The DevSecOps team piloting the software integrated it with internal databases and some proprietary data was included in responses that were exfiltrated.

- Discussion Question:
  - How are lessons learned from incidents incorporated into future preventative measures and considerations for resourcing or investments?
- Desired Discussion Outcome:
  - Generate discussion around the impact an incident in a similar use case might have on future mitigation and preventative techniques in the future.

Day 21 – 9:01AM

Your team briefs your organization's leadership and board of directors on the status of the incident. The report you share includes an update that initial mitigation steps are complete and the updated version of "Spam-A-Less" that was poisoned has been shut off for now.

- Discussion Questions:
  - What is your organization's process for evaluating an AI system after an incident? How does this process change at different points of the system's lifecycle?
  - What is your organization's decision-making process with regards to continuity of operations and improvement planning following an AI incident? How would this change if the impact of the incident was more severe or impacted external customers?
- Desired Discussion Outcome:
  - Understand how participating organizations would work to evaluate AI systems and make decisions regarding continued operation following an AI incident. Can also introduce "what if" scenarios here to see how this would change if the impact were different (e.g., caused financial damage or known IP leak, impacted a product developed by your organization for external users).

Day 33 – 1:21PM

Next month, a group of industry and government are gathering for a TLP AMBER + STRICT roundtable discussion to raise collective awareness of real-world attacks that have not been talked about publicly in the AI security community. Your organization has been invited to speak at this roundtable about any recent real-world threats you've witnessed. One of your organization's security researchers who has attended these events in the past is interested in sharing information about your own recent incident at next month's session.

- Discussion Questions:
  - Post-incident and outside of any required reporting, what information, if any, would your organization voluntarily share with 1) other organizations within your industry and 2) interagency government partners?
    - How would you share this information?
- Desired Discussion Outcome:
  - Understand information sharing capabilities, infrastructure, and mechanisms that participating industry and government players might currently use to share information regarding AI incidents.

Day 51 – 4:31PM

Regardless of whether you speak at the TLP AMBER + STRICT session, your organization sends one of your security team researchers to attend the event. During the roundtable discussion, your researcher listens to other industry and government attendees discuss AI security vulnerabilities and incidents they've experienced and haven't shared publicly.

- Discussion Question:
  - If your organization was listening to a presentation by another industry or government organization describing an AI-incident, what information would (1) be most helpful for your organization to improve your defenses and (2) impact your resource/investment considerations for the future?
- Desired Discussion Outcome:
  - Understand the most pertinent aspects of incident information sharing from the participating players to better inform future mechanisms that can facilitate rapid sharing of critical information across the community.

Day 64 – 4:31PM

Your organization's CTO is giving a plenary address at a major security conference in two weeks. They are working with your security team, communications department, and leadership to determine what, if anything, to share publicly about the recent incident.

- Discussion Questions:
  - How does your organization currently decide what information to share publicly after an incident is identified and mitigated? What reputational or security concerns are most important to your organization when making this type of decision?
  - If possible, describe a situation in which either you or another organization shared information about an incident that improved community readiness for defending against adversarial attacks.
- Desired Discussion Outcome:
  - Understand the participating organizations' processes for sharing information publicly regarding AI incidents and how that has changed with increased adoption of AI. Provide an opportunity to highlight and discuss best practices or exemplary use cases. Gather information on specific channels the participants are using currently for sharing public information.

All Discussion Questions

This list of questions includes all the discussion questions in the modules above, in addition to a few additional questions for the facilitator to have ready to redirect/refocus the discussion if needed.

Icebreaker/Warm Up

- What are the areas and/or use cases for AI that your organization views as the greatest concern now and in the near future?
- What sources of information do you currently rely on to track possible threats to systems? Do these sources differ depending on whether you are tracking traditional cybersecurity vulnerabilities or threats to your AI-enabled systems?

Module 1

- What cyber threat information do you find most useful and actionable?
- Who is responsible for collating and sharing cyber threat information across your organization? With other partners and organizations?
- What actions would you take based on the Alert presented in this scenario (if any)?
- Discuss how your organization has prioritized the mitigation of specific AI risks including those AI risks relative to autonomous defense agents or other similar use cases.
- How do you iteratively plan to evaluate the security of each component of the AI stack for any AI-enabled systems throughout the deployment lifecycle?
- How might you expect your organization to first detect an AI incident? Compare and contrast that with your process for identifying a traditional cybersecurity incident.
- What processes do you have in place to prepare for vulnerabilities in open-source software or other situations where you do not have full control or ownership of the supply chain?

Module 2

- What processes trigger internally in response to a company-wide incident such as this? In what way do those response processes differ—or remain the same—when an AI component is present in the affected system?
- How does your organization currently capture the initial stages of security incidents? Who leads this information collection and investigation process?
- How does your organization perform root cause analysis on a system that includes an AI component such as this email filter system?
- How does your process for investigation and information collection evolve as more information becomes available? What other organizations/departments become involved as the situation evolves?
- What are the primary concerns you would have in this scenario where members of your organization could be using an untrusted model for an application such as information retrieval, believing it to be the latest version of a model that was currently incorporated into their software system? How would you address this security risk with your employees?
- What are your processes for sharing updates within your organization about an ongoing incident?
- What communication channels do you utilize to discuss AI-related incidents with other industry organizations?

- In what other situations might your organization reach out to external entities during the early stages of an incident?
- What processes, mechanisms, or infrastructures might encourage your organization to share information with industry or government partners while you are still investigating the incident?

Module 3

- How does your reporting process change as you reach new information and resolutions about an incident? How would you update your incident report to include this new information?
- What mitigation steps does your organization traditionally take to safeguard against poisoned datasets for your AI-enabled systems?
- How are lessons learned from incidents incorporated into future preventative measures and consideration for resourcing or investments?
- What is your organization's process for evaluating an AI system after an incident? How does this process change at different points of the system's lifecycle?
- What is your organization's decision-making process with regards to continuity of operations and improvement planning following an AI incident? How would this change if the impact of the incident was more severe or impacted external customers?
- Post-incident and outside of any required reporting, what information, if any, would your organization voluntarily share with 1) other organizations within your industry and 2) interagency government partners?
    - How would you share this information?
- If your organization was listening to a presentation by another industry or government organization describing an AI-incident, what information would (1) be most helpful for your organization to improve your defenses and (2) would impact your resource/investment considerations for the future?
- How does your organization currently decide what information to share publicly after an incident is identified and mitigated? What reputational or security concerns are most important to your organization when making this type of decision?
- If possible, describe a situation in which either you or another organization shared information about an incident that improved community readiness for defending against adversarial attacks.

Additional Questions

- How has the decision process for sharing incident information publicly changed as AI is increasingly incorporated into your systems?
- What resources or current avenues are you using for sharing AI incident information?
- What would a national-scale AI incident look like?
- Has your organization made any modifications to your existing security information and event management (SIEM) system to detect AI-driven threats/incidents?
- Is there any point in the scenario where you would contact law enforcement?
- What additional concerns have the incidents described in this scenario generated that have not been addressed in today's discussion?