AI ACTION SUMMIT

AI & Cyber : a crisis management exercise to strenghten cooperation

Do not share to the players

RÉPUBLIQUE FRANÇAISE
Liberté
Égalité
Fraternité

CAMPUS CYBER · Filigran · sanofi · CLR Labs · WAVESTONE · SNCF GROUPE · BNP PARIBAS · AMIAD

AI ACTION SUMMIT

Situation Information file & guiding questions

Do not share to the players

RÉPUBLIQUE FRANÇAISE
Liberté
Égalité
Fraternité

CAMPUS CYBER · Filigran · sanofi · WAVESTONE · SNCF GROUPE · BNP PARIBAS

Table of contents

Exercise objectives 4

Exercise scenario 4

Injects and guiding questions 5

Do not share to the players

Exercise objectives

More specifically, the exercise will aim to:

- Circulate the best practices in the event of a cyberattack on an AIS;
- Strengthen exchanges between all communities working on AI and cybersecurity, in order to identify the governance, defence, protection, and resilience measures required to enhance trust in AIS;
- Explore capabilities, needs, and opportunities for sharing information in case of impactful incidents.

Exercise scenario

Several cyber authorities issue an alert related to a vulnerability in an open source project (office automation services). This vulnerability is used by an attacker to recover confidential information and put pressure (blackmail) on organisations affected. After blackmailing some organisations, the attacker decides to publish some of the sensitive data recovered from online storage.

In parallel one targeted company face an incident of production affecting one of its critical activity.

Do not share to the players

Injects and guiding questions

0 - VIDEO inject| Real Time –02:10PM

TV NEWS SCRIPT

Good afternoon and welcome to this special edition on artificial intelligence.

Today, world leaders, tech executives, and AI experts gathered in Paris for the AI Action Summit, a high-level event aimed at ensuring a safe, ethical, and responsible future for artificial intelligence. Discussions focused on AI regulation, transparency, and cybersecurity, with growing concerns about the risks associated with generative AI models.

But as the world debates AI safety, a major security alert has just been issued. Researchers have discovered a critical vulnerability in a widely used open-source AI assistant, found in many organizations. This flaw could allow massive data theft and even the unintentional publication of sensitive information online. Experts warn that potential exploits already exist, and if no fix is found quickly, we could see major data leaks in the coming days. A stark reminder that while AI holds incredible potential, it must be secured.

Stay with us for more updates. I'm Dwight, thanks for watching.

1 - DEBEX | Real Time – 02:30PM

The exercise starts now.

Please introduce yourself to your crisis cell: NAME, SURNAME, ORGANISATION AND POSITION (cyber expert or IA expert)

Food for thought for moderators:

Icebreaker

Notes

Do not share to the players

2- DAY 1 – 10 :08AM | Real Time – 02:40PM

Researcher uncovers a vulnerability in a product supplied in open source by a major AI technological actor, NeuralForge. Indeed, its open source AI-enabled office automation service (providing mail summarisation, document search, and can send emails on behalf of the users automatically, etc.), named "Alfred", enable data exfiltration via prompt injection.

Guiding questions:

How can information sharing about vulnerabilities be improved between AI solution providers and their clients?

Food for thought for moderators:

Beyond the Technical: How do we address the "human factor" in prompt injection vulnerabilities? Prompt injection exploits the interaction between humans and AI. How can we educate users about the risks of crafted prompts without creating undue alarm? Should AI systems themselves be designed to be more resilient to potentially malicious input, even if that input seems superficially valid? How do we balance user convenience with security in this context?

Open Source vs. Security: Is the open-source model inherently more vulnerable to prompt injection attacks, or does it simply expose vulnerabilities more readily? Does the transparency of open source code make it easier for attackers to find vulnerabilities like prompt injection? Conversely, does it also facilitate faster patching and community involvement in security? How can we maximise the security benefits of open source while mitigating the risks?

Responsibility and Liability: Where does the responsibility lie for securing AI systems against prompt injection, especially in open-source projects? Is it solely on the original developer (like NeuralForge), the community maintaining the project, or the individual users deploying and using the AI service? How can we establish clear lines of responsibility and potentially liability in this complex landscape? Does the open-source model necessitate a different approach to liability than traditional software?

Standardization and Best Practices: Are there specific security best practices or standards that should be developed for AI systems to mitigate prompt injection risks? Should there be industry-wide guidelines for prompt engineering, or AI model hardening? What role should government or regulatory bodies play in establishing and enforcing such standards? How can we adapt existing security best practices to the unique challenges posed by AI and prompt injection?

The "Unknown Unknowns": Prompt injection is a relatively new class of vulnerability. What other unexpected or unforeseen security risks might arise as AI systems become more sophisticated and integrated into our lives? How can we proactively anticipate and prepare for these "unknown unknowns"? How can we foster a culture of continuous learning and adaptation in the AI security community?

Notes

Do not share to the players

3 -DAY 1 – 2 :20PM | Real Time – 02:50PM

NovaTrade Capital, a trading company using NeuralForge AI assistant, « Alfred » has realised that their AI assistant is concerned by the vulnerability uncovered by researchers recently. Security teams are currently investigating its possible exploitation.

Guiding questions:

What processes or indicators would alert your organisation of a potential vulnerability exploitation on your AI systems?

How would you communicate the discovery of such a vulnerability within your organisation?

Food for thought for moderators:

AI Threat Intelligence: How can we leverage threat intelligence specifically focused on AI to anticipate and defend against new forms of prompt injection attacks? The threat landscape for AI is rapidly evolving. How can we stay up-to-date on the latest prompt injection techniques and vulnerabilities discovered in other AI systems? Are there information-sharing platforms or research communities that could help us anticipate new threats?

Attack Surface Monitoring: How can we effectively monitor the "attack surface" of our AI systems to detect prompt injection attempts? The attack surface for prompt injection is often the user interface or API that allows users to interact with the AI. How can we monitor these entry points to detect suspicious queries or abnormal usage patterns? What techniques (e.g., syntax analysis, anomaly detection, etc.) can be used to identify injection attempts?

What are the watch sources you use to spot and be inform of those vulnerabilities? Your providers' inputs? Other sources as well?
Do you have team / or AIs within your organization which are actively looking for vulnerabilities in AI systems?

Behavioral Signatures: Beyond obvious errors, how can we establish "behavioral signatures" for our AI systems that might indicate subtle manipulation through prompt injection? Prompt injection attacks don't always manifest as blatant errors. How can we use analysis of AI behavior (e.g., changes in response patterns, unusual queries, etc.) to detect anomalies that could signal an attack? How can we differentiate these anomalies from normal variations in AI behavior?s

Notes

Do not share to the players

4 -DAY 2 – 11 :30AM | Real Time – 03:00PM

Generic message for all. CERT-FR, BSI and CISA issue an Alert on NeuralForge open source AI Assistant solution, that is widely used and in which a vulnerability has been discovered. Uptick in phishing or spear phishing attacks by foreign actors targeting users of this solution is to be feared and should be anticipated.

Guiding questions:

How do you set criteria for selecting an AI model (open source, proprietary …) or provider and its deployment (on premises, SaaS …)?

How do you assess the cybersecurity maturity of an AI provider? Or open source model use within your systems?

Food for thought for moderators:

Beyond Functionality: Beyond performance metrics and features, what security criteria should be essential in the selection process for AI models or providers (open source or proprietary) and their deployment (on-premises, SaaS)? We often focus on what an AI model does. *But what about* how *it does it, and the security implications? What specific security requirements should be considered during AI model selection, such as data sanitization practices, input validation, and resistance to adversarial attacks (including prompt injection)? How can these criteria be weighed against other factors like cost, performance, and ease of integration?*

Supply Chain Security for AI: The NeuralForge case highlights the potential risks in the AI supply chain, particularly with open-source components. How can organisations effectively manage the security risks associated with using third-party AI models or libraries, especially those that are open source and community-driven? How can we ensure the integrity and trustworthiness of these components? Should there be a "bill of materials" for AI models, listing all dependencies and their versions? How can we verify the security posture of the open-source communities or maintainers behind these components?

Notes

Do not share to the players

5 - DAY 2 – 7:30PM| Real Time – 03:10PM

NovaTrade Capital security teams has identified a concerning issue stemming from activity on social media platform. A user, unidentified, tagged NovaTrade Capital alongside several other companies, admitting the exploitation of the vulnerability in the Alfred product, successfully conducting a phishing campaign targeting Alfred users. This campaign allegedly led to unauthorized access to sensitive data across those organisations, including NovaTrade Capital.

Investigations by NovaTrade Capital cybersecurity team showed that the vulnerability has effectively been exploited. Assessment of the extent of the data exfiltration is still ongoing, but sensitive data exchanged by email might certainly be concerned. Investigations might also reveal a much bigger compromising perimeter, if confidential information about contracts and financial operations has been leaked.

Guiding questions:

How do your monitoring and anomaly detection systems adapt in case of a confirmed attack on your production models?

How do you analyse past data to trace an attack that has already taken place?

How do you check the data perimeter this AI system has access to?

Food for thought for moderators:

Data mapping: Even when using a proprietary model, users will often enter data that is confidential or not intended to be entered into the AI system.

Forensic analysis of prompt injection attempts: Is it possible to develop forensic techniques to analyse prompt injection attempts and trace them back to their source?

Deterrence of prompt injection attacks: How can we create a stronger deterrent against these types of attacks?

Notes

Do not share to the players

6 - DAY 3 –8:00AM | Real Time – 03:25PM

Several employees have received a threatening email about a successful infiltration in NovaTrade Capital systems and the extorsion of sensitive data (confidential business documents, user/customer date, internal communications, etc.). The attackers require a payment to prevent data's online publication.

Guiding questions:

While facing such a situation, what would be the first actions performed by your organisation (internal investigations, notification of competent authorities, crisis checklist / specific set-up, communication Int / Ext …)

Food for thought for moderators:

Verification & Containment: The email claims a successful infiltration. What *immediate* steps should be taken to verify the validity of this claim? Should the network be immediately isolated? Should critical systems be taken offline? How can the company quickly determine if a breach has *actually* occurred, and if so, what systems are affected?

Evidence Preservation: What procedures need to be activated *immediately* to preserve any potential evidence of the intrusion, both on compromised systems and in network logs? How can the integrity of this evidence be ensured for potential legal or forensic investigations?

Notes

Do not share to the players

7- DAY 3 –10 :30AM | Real Time – 03:35PM

Following the recent discovery of this vulnerability in its open source AI-enabled assistant, « Alfred », NeuralForge has immediately launched a comprehensive investigation into the affected solution in order to identify the root cause of the vulnerability, assess the potential impact and implement necessary measures. Additionally, as a precautionary measure, NeuralForge is currently conducting a thorough review of all other products and solutions to ensure no similar vulnerabilities exist.

Guiding questions:

If a model is vulnerable to compromising, what steps would you take to assess the impact on your strategic customers?

How do you evaluate the potential risks associated with deploying specialised AI models in production environments?

Food for thought for moderators:

Should NeuralForge offer customized risk mitigation recommendations or even provide temporary alternative solutions while a patch is developed and deployed?

How can they facilitate open communication and information sharing with customers during this process?

Beyond the Patch: How can NeuralForge reassure its strategic customers that it is not only addressing the immediate vulnerability in Alfred but also taking steps to improve the overall security of its AI development lifecycle and prevent similar vulnerabilities in the future? What concrete actions can NeuralForge communicate to demonstrate its commitment to security, such as code reviews, penetration testing, or security training for developers?

Notes

Do not share to the players

8 -DAY 4 – 12:30AM | Real Time – 03:50PM

Specialised press is covering a growing cybersecurity crisis linked to supply chain, affecting multiple organisation worldwide, following the discovery of a major vulnerability in an artificial intelligence solution used by many companies. Exploiting this vulnerability hackers gain access to sensitive data within multiple organisations. This situation concerns critical actors from multiple domains as it can be understood with a look on NeuralForge client list on its website: finance (NovaTrade Capital, etc.), aviation (Northgate International Airport), etc.

A journalist, investigating this massive data leak situation has reached NovaTrade Capital communication team in order to gain information on the data supposedly belonging to NovaTrade Capital encountered on a website in the Dark Web.

NovaTrade Capital security teams have confirmed that the attacker has effectively published some of the sensitive data stolen. Some documents were marked as “CONFIDENTIAL”.

Guiding questions:

What crisis management strategies would you implement? Are they specific due to the nature of the impacted system?

Food for thought for moderators:

Would your IA systems fall under your BIA (Business Impact Analysis) ?

Notes

Do not share to the players

9 - DAY 5 –10 :00AM | Real Time – 04:00PM

The Northgate International Airport (NIA) client of NeuralForge is mentioned in the press releases regarding the data leak.

NIA denies using the AI-enabled open source model of NeuralForge, emphasising that it is only using proprietary models related to augmented video surveillance.

The airport is proud of its use of tailored customised AI, to detect security events such as: crowd movement, abandoned parcel, suspect behaviour and armed person, stressing that it has increased the rate at which security events are detected, and is now even able to identify risks in advance and make the appropriate decisions to ensure passenger safety.

Guiding questions:

How can organisations balance the need to leverage AI for enhanced security with the potential risks related to data privacy, algorithmic bias, and over-reliance on automated systems?

Food for thought for moderators:

Transparency and Trust in AI Security: Given NIA's emphasis on its proprietary AI, how can organisations demonstrate transparency about the use of AI in security without compromising sensitive information or revealing vulnerabilities? What specific information should be shared with the public and stakeholders to build trust in AI-driven security measures? Should there be independent audits or certifications of AI systems used in critical infrastructure?

Notes

Do not share to the players

10 – DAY 6 – 4:30PM | Real Time – 04:15PM

The team in charge of video surveillance reports an abnormal rate of false positives in the video surveillance system, affecting the teams' processing capacity.

Guiding questions:

How do your monitoring and anomaly detection systems adapt in case of a confirmed attack on your production models?

What are your specific response procedures for anomalies detected in a crisis situation?

Food for thought for moderators:

What are the tools & process in place to detect to protect AI models in production?

The importance of implementing monitoring and anomaly detection systems to ensure the security of machine learning models in production

Strategies for effectively responding to detected anomalies and potential attacks on machine learning models

The role of regular updates and maintenance in keeping access control measures up-to-date with emerging threats in the cyber landscape

Notes

Do not share to the players

11a - DAY 6 –12:30AM | Real Time – 04:25PM

At 12:30AM, an automatic general evacuation order is initiated due to the detection of armed individuals on the airport's perimeter. But after all doubts have been cleared, the security team reported that it was once again a false alarm and that the airport was safe.

A journalist's video goes viral on social media platforms

11b - DAY 6 –12:30AM | Real Time – 04:30PM

After further investigation, NIA's cyber security teams are now understanding that the attacker has exploited the airport video surveillance system because they identified several people on CCTV waving mysterious symbols shortly before the evacuation began

Guiding questions:

In case of a cyberattack targeting one of your AI-enabled solution or system, what emergency measures do you implement to further isolate and secure this environment? How do you manage risks associated with external resources in such a situation?

Food for thought for moderators:

Importance of choosing external models or libraries based on specific criteria such as performance, compatibility, and community support

Methods for evaluating the reliability and security of external models or libraries before incorporating them into your training process

Strategies for securing the training environment to prevent unauthorized access and protect sensitive data during model development

Risk management techniques for mitigating potential threats and vulnerabilities associated with utilising external resources in machine learning projects

Considerations for maintaining a balance between leveraging external tools for efficiency while ensuring the overall security of the training environment

Notes

Do not share to the players

12 - DAY 6 –1:30PM | Real Time – 04:40PM

Following the identification of the issue affecting NIA's security system, the provider has officially identified a poisoned training dataset as the source of the vulnerability that impacted two of its AI-enabled solutions (both open source and proprietary models),

Its open source model: The dataset had been poisoned with a specific prompt injection trigger patterns in the input data during training, so that a specific sequence was associated with a desired output. Attackers had imbedded that sequence within a phishing email body in white text, leading to the activation of the prompt.

Its proprietary model: The dataset had been poisoned with a specific trigger patterns in the input data during training, so that a specific symbol imbedded in the video stream was associated with a desired output. An attacker (supposedly a state-sponsored one) pre-positioned himself in this providers' systems in order to poison the model of AI video surveillance. Once the model is in production in the customer's systems (Northgate international airport), they hired people to go into the airport and hold up this symbol in front of the CCTV cameras, causing the model to diverge. This divergence lead to false alert conducting to the evacuation of the airport by order of the AI customized solution used at the Airport, authorised to take the decision to call for evacuation depending on its own analysis.

Guiding questions:

How are you evolving your security strategies for model training and isolation in the face of emerging threats? What innovations are you considering to strengthen model protection?

What kind of mechanisms you can add to avoid these types of situation?

Food for thought for moderators:

Explore the latest trends and technologies in cybersecurity to enhance model protection against evolving threats

Discuss the impact of past security breaches on current practices and the importance of continuous adaptation in response to new challenges

Share insights on maintaining a secure environment for model training and isolation, including potential vulnerabilities and effective mitigation strategies

Examine the role of regular assessments, monitoring, and updates in ensuring robust security protocols when working with external models

Highlight best practices for strengthening security measures in model training through innovations such as encryption techniques or access controls

Notes

Do not share to the players