

AI & CYBER: A CRISIS MANAGEMENT EXERCISE TO STRENGTHEN COOPERATION

TABLE TOP EXERCISE - TTX

AI ACTION SUMMIT - 11 FEBRUARY 2025



DEBEX

The exercise starts now.

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



DAY 1 – 10 :08AM



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?



DAY 1 – 2 :20PM

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?



DAY 1 – 11 :30AM

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?



DAY 2 – 7 :30PM

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?



DAY 3 – 8 :00AM

Several employees have received a threatening email about a successful infiltration in NovaTrade Capital systems and the extortion of sensitive data (confidential business documents, user/customer data, 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 ...)



DAY 3 – 10 :30AM

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?



DAY 4 – 12 :30AM

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?



DAY 5 – 10 :00AM

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?



DAY 5 – 4 :30PM

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?



DAY 6 – 12 :30AM

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.



DAY 6 – 12 :30AM

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?



DAY 6 – 1 :30PM

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.



DAY 6 – 1 :30PM

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?



Thank you for your attention !