Do not distribute to the players

RÉPUBLIQUE FRANÇAISE
Liberté
Égalité
Fraternité

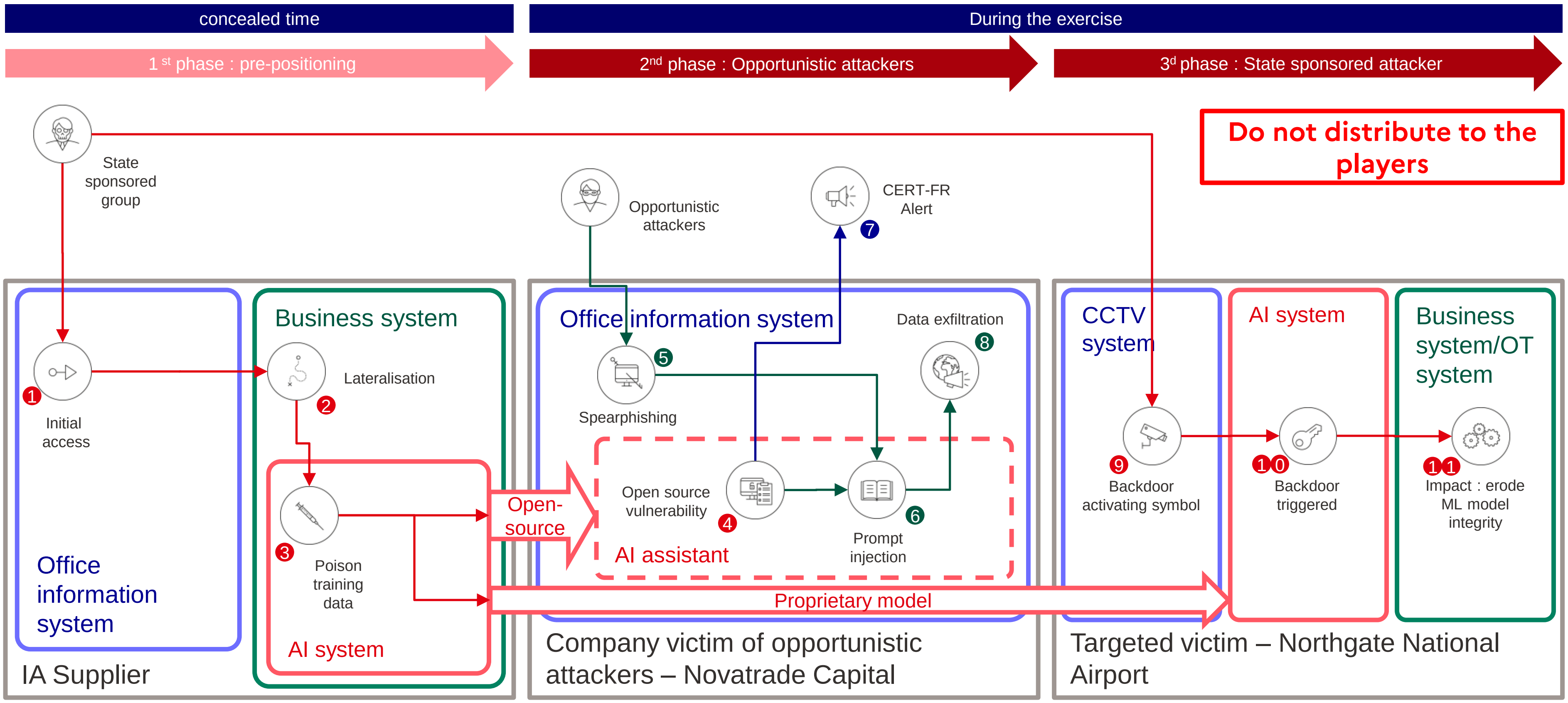
AI ACTION SUMMIT

AI & Cyber: a crisis management exercise to strengthen cooperation

AI Action Summit - 11 February 2025

SCENARIO PRESENTATION

Kill chain – supply chain attack



Do not distribute to the players

Storyline (1/3)

2:30pm

4:00 pm

5:00 pm

Phase 1
Pre-postioning on the supplier's information system

Phase 2
Compromission of an office automation service leading to a major data leak

Phase 3
Online publication of dataleaks & Critical activity disruption

Storyline (2/3)

The exercise will focus on the services made available by an AI solution provider.

The scenario is divided into 2 parts:

Part 1 : Strand 1: Data leakage linked to the use of an AI office assistant (Actions A to D)

The AI solution provider offers an open source model dedicated to office processing (mail summarisation, document search, etc.) that is widely used and in which a vulnerability has been discovered by researchers who have made it public.

Opportunistic attackers take advantage of this vulnerability to recover confidential information from several companies, including a fictitious financial company (NovalTrade Capital), using prompt injection methods.

The aim is to recover sensitive data on critical activities managed by traditional information systems, using the artificial intelligence system as a compromise vector. The impact is linked to the dissemination of sensitive information or infringement of intellectual property (patents, industrial processes, etc.).

Storyline (3/3)

Part 2 : Impact on an AI-based control system (Actions 1 to 7)

The supplier also offers proprietary models that can be specialised for certain specific tasks. A ‘state’ attacker pre-positioned himself with this supplier in order to poison a model with a view to having an impact on a strategic customer, a fictitious airport called Northgate International Airport (NIA), using an algorithmic video surveillance solution.

Once the model is in production at the customer's premises, the attacker uses image triggering methods to modify the model's behaviour. The aim is to disrupt an industrial process by poisoning a model and causing it to behave erratically. This behaviour will gradually lead to the evacuation of the airport due to a false alarm linked to the compromised algorithmic video surveillance.

Thank you for you attention !

Do not hesitate to reach out if you have any questions regarding this exercise