AI ACTION SUMMIT

AI & Cyber : un exercice de crise pour renforcer la collaboration

RETOUR D'EXPÉRIENCE

→ 11 février 2025

RÉPUBLIQUE FRANÇAISE
Liberté
Égalité
Fraternité

ANSSI

CAMPUS CYBER

SNCF GROUPE

sanofi

BNP PARIBAS

WAVESTONE

AMIAD

CLR Labs

Table des matières

Exercice de crise cyber du Sommet de l'IA : une opportunité pour renforcer la collaboration

L'intelligence artificielle (IA) est devenue un moteur d'innovation et d'efficacité dans de nombreux secteurs, mais son adoption croissante introduit des défis inédits en matière de cybersécurité. Face à cette réalité, et dans le contexte de l'accueil par la France du Sommet pour l'Action sur l'IA les 10 et 11 février 2025, l'Agence nationale de la sécurité des systèmes d'information (ANSSI) a organisé, le 11 février 2025, un exercice de gestion de crise en collaboration avec le Campus Cyber[1] et plusieurs de ses membres.

Cet exercice est une réponse à un constat : les communautés d'experts de l'IA et ceux de la cybersécurité ne coopèrent pas suffisamment, appelant au renforcement du dialogue entre ces dernières. En réunissant les acteurs français et internationaux des domaines de l'IA et de la cybersécurité (fabricants, concepteurs de systèmes d'IA, experts en cybersécurité), l'exercice a favorisé des échanges actifs et constructifs autour de la gestion d'une crise d'origine cyber spécifique aux enjeux de l'IA.

Plus précisément, l'exercice avait comme objectifs de :

- Diffuser les meilleures pratiques en cas de cyberattaque ciblant un système d'IA ;
- Renforcer les échanges entre les communautés de professionnels de l'IA et d'experts de la cybersécurité, afin d'identifier les mesures de gouvernance, de défense, de protection et de résilience nécessaires pour accroître la sécurité des systèmes d'IA ;
- Explorer les capacités, les besoins et les opportunités de partage d'informations en cas d'incidents ayant un impact significatif.

L'exercice s'est appuyé sur l'analyse des risques de haut niveau réalisée sous l'égide de l'ANSSI « Développer la confiance dans l'IA par une approche par les risques cyber »[2] en coopération avec des experts institutionnels français du domaine (CNIL, INRIA, LNE, PEReN, AMIAD) et cosignée par 19 pays.

Ce retour d'expérience est une capitalisation et synthèse des expertises croisées. Il met également à disposition en annexe des éléments relatifs au scénario et à l'ingénierie de l'exercice. Ces documents constituent le « kit d'exercice du Sommet de l'IA » mis à la disposition de tous[3] en anglais.

L'ANSSI encourage les organisations à se saisir de ce retour d'expérience et du kit prêt à l'emploi pour le déployer. Les exercices de crise sont un outil efficace pour disposer d'une compréhension partagée des enjeux par des équipes pluridisciplinaires travaillant sur l'IA et pour renforcer la confiance dans les systèmes d'IA.

---

1. https://campuscyber.fr/
2. https://cyber.gouv.fr/publications/developper-la-confiance-dans-lia-par-une-approche-par-les-risques-cyber
3. https://cyber.gouv.fr/sommet-de-lia-exercice-de-gestion-de-crise

Sommet de l'IA : une forte mobilisation pour un exercice novateur

L'exercice a mobilisé plus de

200
professionnels

20
nationalités

50
animateurs et observateurs

reflétant une collaboration internationale inédite.

70
entreprises et industriels, géants de la technologie et start-ups du domaine ont collaboré activement à travers :

2
scénarios de compromission de la supply chain IA.

11
questions directrices illustrant l'engagement transversal des secteurs publics et privés face aux enjeux cyber liés aux systèmes d'IA.

L'impact médiatique, avec

57
articles de presse a amplifié la prise de conscience collective sur les risques cyber, tout en valorisant les bonnes pratiques déployées lors de cet exercice.

Ces chiffres témoignent d'une mobilisation exceptionnelle et d'un vrai intérêt pour renforcer la coopération entre les communautés IA et cyber.

AI ACTION SUMMIT

L'exercice a été organisé **par l'ANSSI au Campus Cyber.**

Principaux enjeux identifiés par les participants liés à la cybersécurité de l'IA

Les échanges lors de cet exercice de gestion de crise cyber dédié à l'IA se sont structurés autour d'une série de questions directrices[3]. Ces questions ont guidé les discussions entre les 200 participants, experts en IA et en cybersécurité, visant à explorer les défis spécifiques et les stratégies de réponse face aux menaces cyber ciblant les systèmes d'IA.

L'ensemble des points soulevés, des perspectives partagées et des pistes de solutions envisagées durant ces échanges constitue le résultat essentiel de cet exercice, reflétant une compréhension collective des enjeux et des recommandations pour renforcer la sécurité de l'écosystème de l'IA.

En préambule, il est nécessaire de rappeler qu'un système d'IA est avant tout un système d'information où les pratiques usuelles d'hygiène de sécurité, y compris en termes de résilience, doivent s'appliquer. A ce titre les nouveaux enjeux autour de la sécurité de l'IA ne doivent pas occulter ces impératifs de cybersécurité non spécifiques à l'IA, car ils constituent un prérequis pour développer et déployer des systèmes d'IA de confiance dans les organisations.

Pour autant, les systèmes d'IA génèrent de nouveaux enjeux de cybersécurité qui ont été appréciés par les participants de l'exercice afin d'identifier les principales recommandations.

1

L'enjeu de l'identification et la maitrise de la chaine d'approvisionnement spécifique à l'IA.

Cette chaine vient complexifier une chaine numérique déjà difficile à maitriser en rajoutant de nouveaux acteurs (vendeurs de données, fournisseurs de modèles etc.). Apparait également une utilisation de plus en plus importante de modèles en source ouverte qui offre une plus grande transparence et flexibilité, mais impose une vigilance accrue sur les alertes de sécurité, les dépendances et la rapidité d'application des correctifs.

2

L'enjeu de la sécurisation du cœur de l'IA, au-delà des systèmes.

Nouveau domaine de la cybersécurité, il est nécessaire de mettre en place des stratégies de sécurisation qui sont spécifiques aux particularités de l'IA, en particulier sur les modèles et les données (transparence et traçabilité des modèles, sécuriser les poids des modèles, tester la résistance aux prompts etc.).

3

L'enjeu de l'identification et du partage des vulnérabilités des modèles d'IA.

Même si les bonnes pratiques restent les mêmes sur la mise à jour des systèmes vulnérables, les systèmes d'IA ayant des particularités qui leur sont propres, le déploiement des bonnes pratiques en matière de test d'intrusion ou de mise à jour doivent être adaptés. Le déploiement de l'IA étant encore récent avec une évolution très rapide des technologies, l'information sur les vulnérabilités qui les concernent est encore peu connue et partagée.

4

L'enjeu de l'audibilité des systèmes d'IA et de la détection d'anomalie.

Les systèmes d'IA peuvent avoir une opacité plus ou moins importante qui peut rendre difficile voire impossible l'identification de vulnérabilité ou d'anomalie de fonctionnement. Il est donc nécessaire de penser, dès la conception des modèles ou des systèmes d'information d'IA, au besoin de génération de traces permettant une analyse, notamment sur les décisions prises par l'IA ou sur les changements dans les modèles.

5

L'enjeu de bonne prise en compte des particularités de l'Intelligence Artificielle dans le périmètre de la résilience cyber.

Les dispositifs de crise existants dans les organisations s'appliquent pleinement sur le périmètre de l'IA. Toutefois les particularités de l'IA nécessitent de s'assurer de la bonne intégration de ces enjeux dans ces dispositifs afin d'assurer une meilleure réactivité en cas d'incident ou de crise.

6

L'enjeu de perte de maitrise des périmètres critiques à cause du déploiement non maitrisé de l'IA.

L'IA apporte de nombreuses promesses, toutefois ce sont encore des technologies récentes qui évoluent rapidement. L'autonomisation des IA sur des périmètres critiques fait porter des risques sur la résilience de l'organisation.

7

L'enjeu de la mise en place d'une gouvernance cyber, intégrant pleinement la cybersécurité de l'IA.

La cybersécurité de l'IA doit faire l'objet d'une attention particulière au sein de l'organisation afin de s'assurer de la mise en place d'une posture équilibrée de sécurité pour accompagner l'innovation. La gouvernance doit également s'assurer de la bonne mobilisation de l'ensemble des acteurs de l'IA et de la cybersécurité autour d'enjeux partagés.

8

L'enjeu du partage des responsabilités entre les offreurs de solution d'IA et les clients de ces solutions.

Les systèmes d'IA ne dérogent pas à l'enjeu de partage des responsabilités. Toutefois, la complexité des systèmes d'IA nécessite une attention particulière sur le partage des responsabilités entre les offreurs et les clients sur les différents composants d'un système d'IA (modèles, données, décisions etc.), en particulier en cas d'anomalie. ●

3. La liste des questions directrices est disponible en annexe

Principales recommandations pour répondre aux enjeux identifiés sur la cybersécurité de l'IA pendant l'exercice

Les recommandations identifiées et partagées par les participants permettent de répondre aux différents enjeux mis en lumière à travers le scénario. Elles s'articulent autour de quatre piliers, formant un cadre robuste pour adresser les défis cyber spécifiques à l'IA :

- **La Gouvernance,** établissant le cadre organisationnel et les responsabilités nécessaires pour une sécurité pérenne de l'IA ;
- **La Protection,** axée sur la prévention des risques en amont ;
- **La Défense,** dédiée à la détection et à la mise en œuvre de réponses efficaces ;
- **La Résilience,** visant à assurer la continuité des opérations et la capacité de récupération face aux incidents ;

Ces recommandations sont une synthèse structurée des échanges entre les participants de l'exercice de crise et **n'engagent ni les participants ni les organisateurs de l'exercice.**

→ Gouvernance : intégrer pleinement l'IA dans la gouvernance cyber et adopter une approche par les risques

Ce pilier permet une gestion sécurisée des systèmes d'IA au sein de l'organisation, en définissant les politiques, les responsabilités et favorisant une approche par les risques pour la sécurité de l'ensemble de la chaine de valeur de l'IA.

Sécuriser la chaîne d'approvisionnement spécifique à l'IA : la sélection rigoureuse des modèles et des fournisseurs est primordiale. Il convient d'établir des critères d'évaluation adaptés aux particularités, qui ne se limitent aux enjeux propres à l'IA mais qui intègrent également les enjeux usuels de cybersécurité de ses systèmes. Parallèlement, une gestion proactive des risques liés aux fournisseurs, à travers des audits réguliers et des revues de sécurité approfondies, est indispensable pour garantir l'intégrité des composants et des services externes.

Adopter une approche par les risques : une approche de sécurité basée sur les risques est indispensable. Elle implique une évaluation régulière et approfondie des menaces et des vulnérabilités spécifiques aux IA et permet d'adopter une posture de cybersécurité équilibrée et cohérente vis-à-vis des risques des systèmes d'IA déployés. Il est également essentiel d'identifier les dépendances entre les systèmes d'IA en production et les processus critiques de l'organisation, au sein des analyses de risques et des plans de continuité et de reprise d'activité (PCA/PRA).

Intégrer la sécurité de l'IA dans la gouvernance globale de la sécurité : la sécurité ne doit pas être une considération annexe, mais un élément

intrinsèque des politiques organisationnelles. Il est crucial d'intégrer la sécurité à chaque étape du cycle de vie des systèmes d'IA, de la conception au déploiement et à la maintenance. Ce suivi doit faire l'objet d'un focus spécifique lors des comités adaptés.

Intégrer l'IA dans la sensibilisation et la formation en cybersécurité : lors des formations du personnel à la cybersécurité, il est important de le sensibiliser aux particularités liées à l'IA. Cette formation doit couvrir notamment les biais algorithmiques, les limites de l'autonomie des systèmes d'IA et les bonnes pratiques de sécurité pour leur utilisation.

Intégrer la cybersécurité dans la sensibilisation et la formation sur l'IA : l'adoption massive de l'IA par les organisations doit également permettre de sensibiliser, notamment les dirigeants, sur le besoin de prendre en compte la cybersécurité pour déployer des IA de confiance.

Affirmer le devoir de protection du client : il incombe aux fournisseurs de solutions d'IA de mettre en place la sécurité nécessaire pour protéger leur client. Toutefois, les organisations, pour se protéger, doivent faire preuve de diligence raisonnable[4] dans la compréhension et l'utilisation de l'IA dans leur contexte.

Souligner la responsabilité de vérification du client : les organisations ont également un rôle actif à jouer dans la sécurité des systèmes d'IA qu'elles utilisent. Pour les modèles open source en particulier, la transparence du code offre la possibilité et la responsabilité d'examiner attentivement le fonctionnement interne et d'identifier d'éventuelles vulnérabilités.

→ Protection : adapter les fondamentaux et sécuriser le cœur de l'IA

Ce pilier fondamental érige les fondations d'une sécurité robuste pour les systèmes d'IA. Il s'agit d'anticiper les menaces et de fortifier les défenses pour limiter l'apparition de faille dans les systèmes et les modèles d'IA.

Réaliser des tests d'intrusion sur les modèles d'IA : Les tests d'intrusion des modèles d'IA, combinant des approches manuelles et semi-automatiques, sont une pratique importante pour identifier les faiblesses exploitables. La simulation de cas d'usage spécifiques à l'IA (injections de prompts par exemple) permet de tester la robustesse des modèles face à des attaques visant directement les modèles.

Déployer des versions limitées des systèmes d'IA pour les tests de sécurité : avant tout déploiement à grande échelle, il est recommandé de mettre en place une version limitée du système d'IA, spécifiquement conçue pour tester sa sécurité et valider la robustesse des mesures mises en place. Cette approche prudente permet d'identifier et de corriger d'éventuelles failles avant une utilisation plus large.

Renforcer les contrôles d'accès aux systèmes et données d'IA : l'application de politiques d'accès strictes et adaptées aux spécificités des systèmes d'IA est cruciale. Cela inclut la mise en œuvre de stratégies telles que la segmentation du réseau pour limiter la propagation en cas d'intrusion et l'adoption de l'authentification multifacteur afin de complexifier l'accès non autorisé aux systèmes d'IA et aux données les plus sensibles.

Protéger les données d'entraînement : investir dans des techniques avancées de protection des données d'entraînement est essentiel pour préserver l'intégrité des modèles. Parallèlement, la mise en place de mécanismes de détection d'attaques par empoisonnement, visant à corrompre les données d'apprentissage, est une priorité.

Maîtriser les modèles open source d'IA : pour les modèles open source, une compréhension approfondie des librairies et des dépendances utilisées par les développeurs est impérative. Cette connaissance facilite les investigations et la réponse en cas d'attaque ou de vulnérabilité découverte.

Garantir transparence et robustesse des modèles d'IA : les fournisseurs ont une responsabilité majeure dans la conception et la fourniture de modèles transparents, explicables dans leur fonctionnement

et robustes face aux tentatives d'exploitation de faiblesses. Il est donc nécesaire lors de l'acquisition ou du développement de modèle d'IA de prendre en compte les besoins de sécurité (réduction de l'opacité et explicabilité des modèles par exemple).

Sécuriser les poids des modèles avec des outils avancés : l'utilisation de modèles de langage de grande taille (LLM) pour le «fine-tuning» et le réentraînement des modèles peut introduire de nouveaux risques. Des stratégies de sécurisation spécifiques des poids des modèles doivent être mises en œuvre pour prévenir toute manipulation ou extraction non autorisée.

Promouvoir le développement des évaluations de sécurité des IA et les capacité de certification : cette évaluation devant se faire sur la base de standards communs afin de renforcer la confiance dans les modèles d'IA, les applications, les données et les infrastructures (à l'image des normes appelés par le règlement européen sur l'intelligence artificielle – article 40).

Repenser les stratégies de protection pour les adapter aux systèmes d'IA : même si les bonnes pratiques de cybersécurité s'appliquent sur les systèmes d'IA il est nécessaire de mettre à jour les stratégies de protection pour prendre en en compte les spécificités (modèles, données etc.). Une attention particulière doit être portée à la construction et l'isolation des modèles d'IA (apprentissage fédéré, confidentialité différentielle, chiffrement homomorphe), l'isolation stricte des environnements d'entrainement des autres environnements (pré-production, production), la détection d'anomalies, le watermarking, ou encore l'exécution sécurisée.

Exploiter l'IA pour la recherche de vulnérabilités et faciliter les corrections : au-delà des risques qu'ils peuvent représenter, les systèmes d'IA peuvent également être une solution pour identifier automatiquement les vulnérabilités au sein des modèles et faciliter l'application de correctifs pouvant s'appuyer, par exemple, sur les informations partagées entre les fournisseurs.

→ Défense : adopter une posture de veille active et déployer des capacités spécifiques pour défendre les IA

Ce pilier essentiel établit les capacités nécessaires pour identifier les menaces ciblant les systèmes d'IA, comprendre leur nature et déployer des réponses rapides et efficaces afin de minimiser leur impact.

Intégrer la Cyber Threat Intelligence (CTI) spécifique à l'IA : un flux d'informations pertinent et bidirectionnel entre les équipes en charge de la gestion des systèmes d'IA (data scientist, ingénieur de données, analyse de données etc.) et les équipes de sécurité est indispensable pour une compréhension globale des menaces. L'intégration de flux de CTI classiques, enrichis d'informations spécifiques aux vulnérabilités de l'IA, permet d'anticiper les attaques et d'adapter les défenses.

Mettre en place une veille cybersécurité spécifique à l'IA : il est essentiel de maintenir une capacité de veille sur les vulnérabilités des solutions d'IA, notamment celles utilisées sur les périmètres les plus critiques, en incluant les communautés de recherche en IA et en cybersécurité.

Adapter les bulletins d'alertes publiques aux data scientists : les bulletins d'alertes publiques concernant les vulnérabilités de l'IA doivent être compréhensibles et exploitables par les data scientists, qui jouent un rôle clé dans le développement et le déploiement de ces systèmes.

Inclure les communautés de recherche en IA : les communautés de recherche en IA constituent une source d'information précieuse pour l'identification des nouvelles vulnérabilités et des menaces émergentes. Il est essentiel de les intégrer activement dans les processus de veille et de diffusion des informations de sécurité d'autant plus pour une technologie évoluant très rapidement.

Déployer des capacités de mise à jour des systèmes d'IA : la complexité de la correction des modèles d'IA, notamment en raison des coûts élevés de

réentraînement, impose de recourir à des solutions alternatives, telles que le remplacement temporaire de modèles ou l'ajustement ciblé des paramètres.

Mettre en place une détection d'anomalies spécifiques à l'IA : l'utilisation de mécanismes de détection d'anomalies adaptés aux particularités des systèmes d'IA est clef. Cela inclut la surveillance de métriques spécifiques, telles que la dérive des taux de faux positifs, l'altération des données d'entraînement ou les changements inattendus dans les résultats des modèles.

Favoriser l'explicabilité des systèmes d'IA pour rendre possible la réponse à incident : via l'enregistrement structuré des traces explicatives du raisonnement interne des modèles (« chain of thought »), une pratique qui peut permettre l'analyse rétrospective des traces lors d'incidents de sécurité et comprendre l'origine des comportements anormaux, tout en respectant un équilibre entre transparence algorithmique et protection des données sensibles.

→ Résilience : prendre en compte l'IA dans les dispositifs de résilience cyber existants

Ce pilier vise à garantir que l'organisation puisse maintenir ses fonctions essentielles et se rétablir rapidement face à une cyberattaque ou un incident majeur affectant ses systèmes d'IA, minimisant ainsi les perturbations et les pertes.

Intégrer le périmètre de l'IA dans le dispositif de gestion de crise : il est impératif d'intégrer dans les plans de gestion de crise les particularités des systèmes d'IA. Ces plans doivent intégrer pleinement les équipes intervenant sur les systèmes d'IA, qui possèdent une compréhension approfondie du fonctionnement et des dépendances des systèmes d'IA déployés.

Identifier et gérer les dépendances externes liées à l'IA : la résilience dépend de la solidité des partenaires et des fournisseurs. Des programmes de gestion des risques liés aux fournisseurs doivent intégrer les principaux fournisseurs d'IA pour avoir une surveillance renforcée, incluant les processus de vérification de la sécurité de ces systèmes, et les impacts en cas de compromission.

Appliquer les bonnes pratiques de communication de crise en prenant en compte l'IA : en cas de compromission avérée d'un système d'IA, une communication de crise prompte et transparente est essentielle. La complexité technique, intrinsèque aux technologies d'IA, nécessite la mobilisation des experts pour s'assurer d'une communication appropriée. L'acculturation des communicants au plus tôt à ces sujets est donc importante.

Intégrer l'IA dans les plans de continuité et de reprise d'activité (PCA/PRA) : les systèmes d'IA ne doivent pas être des angles morts des plans de continuité. Il est nécessaire d'intégrer explicitement les systèmes d'IA dans les PCA/PRA, en définissant notamment des procédures manuelles de remplacement ou de contournement de l'IA pour les processus critiques.

Encadrer l'autonomie des systèmes d'IA : il est important de limiter les fonctions de décision automatiques, en particulier dans les contextes critiques, afin de maintenir un contrôle humain et d'atténuer les risques potentiels.

Organiser des exercices de crise cyber intégrant un volet IA : les exercices de gestion de crise sont un outil de sensibilisation d'intérêt afin de mieux faire apréhender à des communautés pluridisciplinaires les nouveaux enjeux, comme ceux autour de l'IA. En cas de système d'IA en production, il est nécessaire d'intégrer pleinement ces systèmes dans le périmètre des exerices. ●

4. La « due diligence » (diligence raisonnable) désigne l'obligation, pour une organisation, de mettre en œuvre toutes les mesures raisonnablement attendues afin d'identifier, d'évaluer et de gérer les risques associés à ses activités, notamment en matière de sécurité informatique et d'utilisation de l'IA

Annexe

Une ingénierie d'exercice favorisant les regards croisés

Afin de garantir une expérience immersive et un échange fructueux entre les participants, l'exercice de crise sur table a été conçu comme un espace interactif où la collaboration et le partage d'expertise étaient au cœur du dispositif.

→ Une mosaïque d'expertises pour une approche 360° de la cybersécurité de l'IA

- Les participants, issus d'horizons variés (industriels, experts en cybersécurité, experts en IA, spécialistes de la gestion de crise), ont été répartis de façon homogène dans des salles de crise afin de permettre une représentativité des différentes expertises et types d'organisations. Cette diversité des expertises a permis de croiser les regards et d'enrichir les débats, reflétant ainsi la complexité des situations de crise réelles, afin de construire une compréhension globale des enjeux ;
- L'accueil d'institutionnels internationaux a souligné l'importance de la coopération internationale et a permis d'intégrer leurs perspectives spécifiques et de favoriser le partage de bonnes pratiques.

→ Un scénario générateur de réflexion, guidé par l'animation

- Le scénario de crise a été dévoilé progressivement, à travers des injections narratives conçues pour simuler des incidents d'origine cyber ;
- Chaque injection était accompagnée de «questions directrices», invitant les participants à analyser la situation, à prendre des décisions et à échanger sur les stratégies à adopter ;
- Chaque cellule de crise était animée par un binôme d'animateurs garant du bon déroulement des échanges et de la pertinence des réflexions. Ces binômes d'animation, composés d'un expert en management des crises et d'un expert de la cybersécurité de l'IA, ont été constitués en s'appuyant sur les acteurs de l'écosystème public et privé.

→ Un espace de partage et d'enrichissement mutuel

- Au-delà des questions directrices, les animateurs ont enrichi les discussions avec des éléments de réflexion complémentaires, incitant les participants à explorer des pistes alternatives et à anticiper les conséquences de leurs actions ;
- Ces questions avaient pour but d'élargir le spectre des réflexions et d'aborder des sujets transversaux ;
- L'exercice a également facilité le partage d'expériences, connaissances, stratégies et de bonnes pratiques entre les participants, favorisant l'apprentissage mutuel et le renforcement des compétences collectives ;
- L'exercice a été une véritable opportunité pour les experts de l'IA de mieux comprendre les enjeux de cybersécurité, et inversement, pour les experts de la cybersécurité, d'avoir une analyse plus réaliste des vulnérabilités et limites des systèmes d'IA.

→ Un scénario et un kit d'exercice mis à la disposition de tous

L'ANSSI a conçu un kit d'exercice en anglais, à destination des participants et des animateurs. Cet ensemble documentaire complet permet de faciliter la préparation et l'animation d'un exercice de crise dédié aux problématiques sur l'IA. Il peut aisément se déployer dans toutes les organisations.

Le kit « Exercice du sommet pour l'action sur l'IA » est composé des éléments suivants :

- Un briefing pour les animateurs détaillant la « chaine de compromission » et le scénario de l'exercice ;
- Le scénario de l'exercice ;
- Les questions directrices permettant d'animer les discussions ;
- Un guide du joueur contenant le dossier d'informations sur la situation et les questions directrices ;
- Un dossier d'informations sur la situation destiné aux animateurs.

Clé de voute du kit, le scénario a été construit en s'appuyant sur l'analyse des risques de haut niveau réalisée sous l'égide de l'ANSSI « Développer la confiance dans l'IA par une approche par les risques cyber » en coopération avec des experts institutionnels français du domaine (ANSSI, CNIL, INRIA, LNE, PEReN, AMIAD) et cosignée par 19 pays. Le scénario a également été confronté à la réalité des organisations et de leurs enjeux dans le déploiement des systèmes d'IA.

Il a été structuré en deux parties permettant de mettre en exergue les enjeux cyber liés aux systèmes d'IA :

- **Partie 1 :** Fuite de données via un assistant de bureau IA, suite à une attaque cyber réalisée par des attaquants opportunistes, exploitant une vulnérabilité critique d'un modèle IA open source spécialisé dans les tâches bureautiques ;
- **Partie 2 :** Compromission par un attaquant étatique d'un système de contrôle IA critique, par l'empoisonnement des données de la vidéosurveillance dans un aéroport.

L'ensemble des documents constituant le kit sont à retrouver sur le site de l'ANSSI : https://cyber.gouv.fr/sommet-de-lia-exercice-de-gestion-de-crise

→ Les questions directrices de l'exercice de crise

Afin de structurer et d'enrichir les discussions, les participants ont été invités à réfléchir collectivement autour des questions directrices suivantes :

- Comment améliorer le partage d'informations sur les vulnérabilités entre les fournisseurs de solutions d'IA et leurs clients ?
- Quels processus ou indicateurs alerteraient votre organisation d'une exploitation potentielle de vulnérabilité sur vos systèmes d'IA ? Comment communiqueriez-vous la découverte d'une telle vulnérabilité au sein de votre organisation ?
- Comment définissez-vous les critères de sélection d'un modèle d'IA (open source, propriétaire…) ou d'un fournisseur, et de son déploiement (sur site, SaaS…) ? Comment évaluez-vous la maturité en cybersécurité d'un fournisseur d'IA ? Ou l'utilisation d'un modèle open source au sein de vos systèmes ?
- Comment vos systèmes de surveillance et de détection d'anomalies s'adaptent-ils en cas d'attaque confirmée sur vos modèles de production ? Comment analysez-vous les données passées pour retracer une attaque qui a déjà eu lieu ? Comment vérifiez-vous le périmètre de données auquel ce système d'IA a accès ?
- Face à une telle situation, quelles seraient les premières actions entreprises par votre organisation (enquêtes internes, notification des autorités compétentes, checklist de crise / mise en place spécifique, communication interne / externe…) ?
- Si un modèle est vulnérable à une compromission, quelles mesures prendriez-vous pour évaluer l'impact sur vos clients stratégiques ? Comment évaluez-vous les risques potentiels associés au déploiement de modèles d'IA spécialisés dans des environnements de production ?
- Quelles stratégies de gestion de crise mettriez-vous en œuvre ? Sont-elles spécifiques en raison de la nature du système impacté ?

- Comment les organisations peuvent-elles équilibrer l'exploitation de système d'IA dans les outils de cybersécurité avec les risques potentiels liés à la confidentialité des données, aux biais algorithmiques et à la dépendance aux systèmes automatisés ?
- Comment vos systèmes de surveillance et de détection d'anomalies s'adaptent-ils en cas d'attaque confirmée sur vos modèles en production ? Quelles sont vos procédures de réponse spécifiques pour les anomalies détectées en situation de crise ?
- En cas de cyberattaque ciblant l'une de vos solutions ou d'IA, quelles mesures d'urgence mettez-vous en œuvre pour isoler et sécuriser davantage cet environnement ? Comment gérez-vous les risques associés aux ressources externes dans une telle situation ?
- Comment faites-vous évoluer vos stratégies de sécurité pour la formation et l'isolement des modèles face aux menaces émergentes ? Quelles innovations envisagez-vous pour renforcer la protection des modèles ? Quels types de mécanismes pouvez-vous ajouter pour éviter ce type de situation ?

→ Des ressources documentaires pour aller plus loin

Développer la confiance dans l'IA par une approche par les risques cyber.
https://cyber.gouv.fr/publications/developper-la-confiance-dans-lia-par-une-approche-par-les-risques-cyber

Recommandations de sécurité pour un système d'IA générative
https://cyber.gouv.fr/publications/recommandations-de-securite-pour-un-systeme-dia-generative

L'ANSSI et le BSI publient leurs recommandations de sécurité concernant les assistants de programmation basés sur l'IA
https://cyber.gouv.fr/actualites/lanssi-et-le-bsi-publient-leurs-recommandations-de-securite-concernant-les-assistants-de

Kit de l'exercice du sommet pour l'action sur l'IA
https://cyber.gouv.fr/sommet-de-lia-exercice-de-gestion-de-crise