AI ACTION SUMMIT

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

Annexe

Synthèse des échanges des participants autour des questions directrices

RÉPUBLIQUE FRANÇAISE
Liberté Égalité Fraternité

ANSSI

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

Annexe 2 – Synthèse des échanges des participants par questions directrices

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. Ces questions ont servi de catalyseur pour des discussions approfondies entre les 250 participants, experts en IA et en cybersécurité, visant à explorer les défis spécifiques et les stratégies de réponse face aux menaces 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 clés pour renforcer la sécurité de l'écosystème de l'IA.

Question directrice 1 : Comment améliorer le partage d'informations sur les vulnérabilités entre les fournisseurs de solutions d'IA et leurs clients ?

Les échanges ont convergé vers la nécessité d'établir des mécanismes de partage d'informations sur les vulnérabilités des systèmes d'IA qui soient à la fois plus efficaces, plus clairs et plus adaptés aux spécificités de cette technologie. Il est apparu que les modèles traditionnels de communication des vulnérabilités, bien qu'essentiels, présentent des limites significatives face à la complexité et à l'interdépendance des composants d'IA. La nécessité de mettre en place des canaux de communication explicites, sécurisés et dédiés aux vulnérabilités des systèmes d'IA a été soulignée. Ces canaux devraient être accompagnés de politiques de divulgation des vulnérabilités formalisées et d'un engagement proactif de la part des fournisseurs à partager les informations pertinentes avec leurs clients de manière opportune et compréhensible. L'importance de contextualiser la vulnérabilité au sein de l'architecture spécifique du client et de clarifier les responsabilités de chacun en matière de correction et d'atténuation a également été soulignée. Cette approche s'inscrit dans la continuité des politiques de gestion responsable des vulnérabilités promues en cybersécurité, en intégrant les vulnérabilités propres aux systèmes d'IA dans les dispositifs existants de divulgation coordonnée des vulnérabilités (*coordinated vulnerability disclosure*, ou CVD).

Bien que les bulletins d'information et de vulnérabilité demeurent un outil de communication central, les discussions ont mis en lumière les défis posés par le "*patching*" des modèles d'IA. La complexité des modèles, leur interdépendance et l'impact potentiel des mises à jour sur les personnalisations effectuées par les clients (notamment le « *fine-tuning* ») rendent les approches traditionnelles souvent inapplicables. Des solutions alternatives et complémentaires ont été explorées, telles que la publication de systèmes de rails de sécurité (ou *guardrails*) renforcés pour prévenir l'exploitation de la vulnérabilité sans nécessiter une modification profonde du modèle sous-jacent, le remplacement temporaire du modèle de fondation par une version corrigée (si possible), ou la diffusion d'améliorations ciblées du modèle (ajustement des poids) avec des instructions claires pour leur intégration par les clients.

La difficulté d'identifier avec précision le déclencheur d'une vulnérabilité (mot clé, phrase, entrée spécifique) a également été notée comme un obstacle à la communication et à la remédiation efficace.

De plus, la compréhension de la chaîne de dépendance des modèles d'IA (les librairies, les données d'entraînement, les composants tiers) et la clarification des responsabilités entre le fournisseur et le client sont apparues comme des éléments critiques pour une gestion efficace des vulnérabilités et des risques liés à la chaîne d'approvisionnement des systèmes d'IA. Un défaut de partage d'informations à ce niveau expose à des réactions tardives et à une gestion inadéquate des vulnérabilités des composants tiers. L'analyse de risques doit donc cibler ces points névralgiques et définir des mesures de sécurité adaptées à chaque maillon de la chaîne.

Enfin, les discussions ont insisté sur la nécessité d'améliorer l'accessibilité des bulletins d'alerte aux publics non-cyber, notamment les *data scientists* qui sont des acteurs clés dans le développement, l'utilisation et la maintenance des systèmes d'IA, et d'intégrer activement les communautés de recherche en vulnérabilités IA comme sources d'information. Le cloisonnement actuel entre les communautés de recherche en IA et en cybersécurité a été identifié comme un frein majeur à l'identification et au partage des nouvelles vulnérabilités et menaces ciblant les systèmes d'IA, soulignant la nécessité de favoriser une collaboration plus étroite entre ces domaines d'expertise.

En conclusion, les discussions ont mis en avant la nécessité de mécanismes de partage d'informations adaptés aux spécificités des vulnérabilités des modèles d'IA, avec des canaux sécurisés et des politiques claires, alignées avec les politiques de gestion responsable des vulnérabilités. La complexité du *patching* impose des solutions alternatives et une meilleure compréhension des responsabilités entre fournisseurs et clients. Enfin, il est crucial de favoriser l'accès aux alertes pour les non-experts et de renforcer la collaboration entre les communautés de recherche en IA et cybersécurité pour une gestion plus rapide et efficace des vulnérabilités.

<u>Question directrice 2</u> : 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 ?

Les échanges ont souligné l'impératif d'adopter une approche de détection des exploitations de vulnérabilités sur les systèmes d'IA qui combine une surveillance technique sophistiquée et des processus organisationnels robustes.

Sur le plan technique, la mise en place d'une surveillance continue et en temps réel est considérée comme primordiale. Cette surveillance doit s'appuyer sur une cartographie exhaustive des composants de l'architecture du système d'IA, incluant les modèles, les données d'entraînement et d'inférence, les API, les *add-ons* et les systèmes d'intégration (comme les systèmes RAG). Les flux de données, l'étendue des droits d'accès aux données par les modèles, et l'historique des résultats des modèles doivent être contrôlés attentivement à la recherche d'anomalies comportementales ou de déviations suspectes. L'intégration de systèmes de détection comportementale avancés, potentiellement basés sur

l'IA elle-même et inspirés des meilleures pratiques de l'industrie en matière de cybersécurité, a été identifiée comme une voie prometteuse pour identifier des tentatives d'exploitation subtiles. De plus, l'intégration des informations provenant du CERT-FR et des CSIRT et des bases de données de vulnérabilités telles que le dictionnaire CVE) dans les systèmes de surveillance est essentielle pour corréler les alertes internes avec les menaces externes connues.

Sur le plan organisationnel, les discussions ont insisté sur la nécessité d'établir des canaux de communication interne clairs et efficaces pour signaler et gérer la découverte d'une vulnérabilité. La rapidité et la clarté de cette communication sont cruciales pour une réponse coordonnée et opportune. Le rôle des fournisseurs de modèles d'IA dans la divulgation responsable des vulnérabilités en se fondant sur les bonnes pratiques de CVD promue au sein de l'OCDE notamment[1] a également été souligné, en tenant compte des chaînes de dépendance potentielles avec d'autres systèmes et composants. L'absence d'une surveillance proactive et d'une gestion des incidents adaptée aux spécificités de l'IA expose l'organisation à des risques significatifs de latéralisation en cas d'exploitation. La surveillance continue doit également s'étendre au suivi des performances des modèles en production, en détectant les dérives de performance ou l'apparition de biais inattendus qui pourraient indiquer une compromission ou une manipulation.

En matière de Cyber Threat Intelligence (CTI), une approche multi-source est recommandée tant pour les fournisseurs que pour les organisations. Les fournisseurs devraient investir dans des tests d'intrusion approfondis (combinant approches manuelles et semi-automatiques itératives), surveiller activement les forums et les échanges en ligne pour anticiper les nouvelles techniques, tactiques et procédures d'attaque, et utiliser des systèmes d'IA pour la découverte automatisée de vulnérabilités sur leurs propres modèles. Du côté des organisations, la remontée d'informations sur les vulnérabilités doit provenir à la fois de la veille technologique des équipes de développement et de science des données, ainsi que des flux de CTI classiques, avec une attention particulière portée aux systèmes et modèles spécifiques utilisés par l'entreprise. La responsabilité des développeurs et des *data scientists* dans la maintenance de cette veille spécifique a été mise en avant comme un élément clé d'une posture de sécurité proactive.

En synthèse, les discussions ont souligné l'importance d'une surveillance technique continue et d'une cartographie précise de l'architecture IA pour détecter toute exploitation de vulnérabilité. L'efficacité repose sur l'intégration de systèmes de détection comportementale avancés et de sources externes de renseignement sur les menaces. Côté organisation, des canaux de communication interne clairs et réactifs sont essentiels pour une gestion rapide des incidents. Enfin, la veille active, la collaboration entre fournisseurs et entreprises utilisatrices, et l'implication des équipes techniques dans la détection et la remontée des vulnérabilités constituent les piliers d'une posture de sécurité robuste pour les systèmes d'IA.

[1] Rapport de l'OCDE sur la divulgation responsable des vulnérabilités: Encouraging vulnerability treatment (EN), https://one.oecd.org/document/DSTI/CDEP/SDE(2021)9/FINAL/en/pdf

<u>Question directrice 3</u> : 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 ?

Le processus de définition des critères de sélection d'un modèle d'IA (qu'il soit *open source* ou propriétaire), de son fournisseur, et de sa modalité de déploiement (sur site, SaaS, etc.) a été identifié comme une étape critique nécessitant une approche multidimensionnelle et rigoureuse. Les discussions ont souligné que ce choix ne peut se limiter à des considérations purement techniques ou fonctionnelles, mais doit intégrer des aspects de sécurité et de gestion des risques. L'établissement de critères d'évaluation clairs, incluant des audits de sécurité techniques et organisationnels du fournisseur, ainsi qu'une analyse approfondie de sa gestion des risques, est apparu comme fondamental. Au-delà de la santé financière et de la réputation du fournisseur, l'évaluation des caractéristiques techniques intrinsèques aux modèles a été jugée essentielle : cela comprend l'examen des références du modèle, sa taille et sa complexité, les caractéristiques des données d'entraînement (provenance, qualité, biais potentiels), la propriété intellectuelle et les implications de licence, ainsi que la robustesse et l'activité de la communauté *open source* le cas échéant. Un point crucial soulevé a été le manque actuel de critères de sécurité standardisés et de métriques d'explicabilité des modèles d'IA, soulignant un besoin de soutien normatif et de développement de bonnes pratiques dans ce domaine. La décision du mode de déploiement, influencée notamment par les exigences de confidentialité des données, a également été discutée, avec une préférence pour l'hébergement interne dans les cas les plus sensibles, impliquant la mise en place de garde-fous techniques et humains stricts et une réflexion approfondie sur le niveau d'autonomie accordé aux systèmes d'IA.

L'évaluation de la maturité en cybersécurité d'un fournisseur d'IA et l'analyse des risques liés à l'utilisation d'un modèle *open source* au sein des systèmes ont été identifiées comme des composantes essentielles du processus de sélection et de déploiement. Pour les fournisseurs d'IA, les discussions ont insisté sur la nécessité d'évaluer leurs politiques et pratiques de sécurité, leurs certifications, leurs processus de gestion des vulnérabilités, leur transparence quant à la sécurité de leurs modèles et de leurs infrastructures, ainsi que leurs capacités de réponse en cas d'incident. Une attention particulière doit être portée à la sécurité de leur chaîne d'approvisionnement et à leur compréhension du parcours des données utilisées pour entraîner et faire fonctionner les modèles. Concernant l'utilisation de modèles *open source*, les échanges ont souligné l'avantage de la transparence et de la possibilité d'examiner le code source, mais ont également mis en évidence le risque de vulnérabilités potentiellement plus fréquentes et exploitées plus rapidement. La responsabilité incombe alors aux équipes techniques d'acquérir une compréhension approfondie de l'outil *open source* utilisé (y compris ses dépendances), de surveiller activement les alertes de sécurité, d'appliquer les correctifs rapidement et de mener des audits de sécurité spécifiques.

En conclusion, les discussions ont convergé vers la nécessité de conduire une analyse des risques spécifique non seulement à chaque modèle et à chaque mode de déploiement, mais surtout à chaque cas d'usage, la criticité de ce dernier prévalant dans la régulation de l'IA instaurée par le règlement européen. Cette approche guide la mise en place de mesures de

sécurité adaptées et souligne la responsabilité partagée entre le fournisseur (y compris en open source) et l'organisation utilisatrice en matière de sécurité des systèmes d'IA. Il convient de rappeler que le règlement européen sur l'IA interdit certains cas d'usage de systèmes d'IA jugés à risque inacceptable, mais ne vise pas les modèles d'IA en tant que tels.

Question directrice 4 : 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 ?

Les échanges ont insisté sur la nécessité de déployer des systèmes de surveillance et de détection d'anomalies intrinsèquement adaptatifs et contextuels, capables d'évoluer dynamiquement en temps réel en réponse à une attaque confirmée ciblant les modèles d'IA en production. Il est apparu que la simple détection statique d'anomalies est insuffisante face à la sophistication des cybermenaces visant l'IA. Les discussions ont souligné l'importance de disposer de systèmes de sécurité qui peuvent non seulement identifier des comportements déviants par rapport aux modes d'opération normaux, mais également ajuster leurs seuils de sensibilité et leurs règles de corrélation en fonction du type d'attaque détectée et de l'évolution de la menace. Cela implique l'intégration de mécanismes d'apprentissage automatique au sein des systèmes de surveillance eux-mêmes, leur permettant de reconnaître de nouvelles tactiques, techniques et procédures (TTPs) utilisées par les attaquants ciblant spécifiquement les modèles d'IA. De plus, les échanges ont mis en lumière la nécessité d'une orchestration de la réponse incidentelle qui soit spécifique aux systèmes d'IA, allant au-delà des procédures génériques pour inclure des actions telles que l'isolement granulaire des composants du modèle affecté, la mise en quarantaine des flux de données suspects et le déclenchement d'alertes ciblées vers des équipes d'intervention spécialisées en sécurité des systèmes d'IA.

Concernant l'analyse des données passées pour retracer une attaque et la vérification du périmètre des données, les discussions ont souligné la criticité de capacités de criminalistique numérique robustes et adaptées à l'opacité potentielle des systèmes d'IA. Face à la complexité des interactions au sein des modèles et entre les modèles et les données, les méthodes traditionnelles d'investigation peuvent s'avérer inefficaces. Les échanges ont insisté sur le développement et l'application d'approches spécifiques pour retracer les activités malveillantes, telles que l'analyse des logs d'exécution des modèles, l'examen des flux de données d'entrée et de sortie, et potentiellement l'utilisation de techniques d'audit des modèles (si disponible). La vérification du périmètre de données auquel un système d'IA a accès a été identifiée comme une mesure de sécurité fondamentale. Les discussions ont mis en avant la nécessité de mécanismes stricts de contrôle d'accès, appliqués non seulement aux données d'entraînement et d'inférence, mais également aux configurations des modèles, aux API et aux autres composants de l'infrastructure d'IA. Ces contrôles doivent être dynamiques et auditables, avec une surveillance continue des tentatives d'accès et des modifications de permissions. L'intégration de politiques de moindre privilège et de segmentation des données a été recommandée pour limiter la surface d'attaque potentielle en cas de compromission.

En résumé, les discussions ont convergé vers une approche de sécurité dès la conception pour l'IA, où la surveillance adaptative, l'investigation numérique spécialisée et des contrôles d'accès dynamiques et auditables sont intrinsèquement intégrés pour protéger les modèles et les données associées.

<u>Question directrice 5</u> : 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…) ?

Face à une situation de compromission avérée d'un système d'IA, les discussions ont mis en avant la nécessité d'une réaction coordonnée et rapide, articulée autour de plusieurs actions prioritaires et simultanées. La première action, largement partagée et recommandée, concerne l'isolement du système potentiellement compromis, avec une nuance cruciale : couper l'accès à Internet pour circonscrire la menace et permettre une analyse plus sereine de la situation, sans pour autant interrompre l'alimentation électrique de la machine afin de préserver les preuves et faciliter l'investigation. En parallèle, le lancement immédiat d'une investigation approfondie a été soulignée comme essentiel pour comprendre la nature de l'attaque, identifier son origine, évaluer l'étendue de la compromission et déterminer les systèmes affectés. Cette enquête devrait mobiliser des équipes multidisciplinaires incluant des experts en sécurité de l'IA, des analystes de logs, et potentiellement des *data scientists* pour examiner le comportement du modèle et identifier toute activité anormale. Simultanément, les discussions ont insisté sur l'importance d'activer une *checklist* de crise ou un plan de réponse spécifique préétabli pour ce type d'incident impliquant un système d'IA. Ce plan devrait détailler les rôles et responsabilités des différentes équipes, les procédures de communication interne et externe, les étapes d'analyse technique et de remédiation, ainsi que les critères de décision pour les actions ultérieures.

En complément de ces actions initiales, les discussions ont souligné l'importance d'une gestion de la communication structurée et transparente, tant en interne qu'en externe, ainsi que le respect des obligations légales en matière de notification aux autorités compétentes (par exemple, les autorités de protection des données, les agences de cybersécurité). En interne, une communication claire et régulière est importante pour informer les employés concernés de la situation, des mesures prises et des consignes à suivre, afin d'éviter la propagation de fausses informations et de maintenir un climat de confiance. En externe, une stratégie de communication de crise bien définie est indispensable pour informer les clients, les partenaires et le public de manière factuelle et rassurante, minimisant ainsi les dommages potentiels à la réputation de l'organisation. Compte tenu de la sensibilité médiatique particulière des incidents impliquant des systèmes d'IA, une communication proactive et transparente est d'autant plus nécessaire. Par ailleurs, les discussions ont rappelé le rôle du RSSI/CISO dans la compréhension du fonctionnement des systèmes d'IA tout au long de leur cycle de vie[2], l'identification des données auxquelles ces systèmes ont accès et leurs interconnexions avec le reste du système d'information d'une entité, l'adaptation des plans de gestion de crise aux spécificités de l'IA, et l'enregistrement des logs d'activité, qui constituent des sources d'information pour l'enquête post-incident.

2 Les principales étapes du cycle de vie d'un système d'IA incluent l'entraînement (paramétrisation du modèle à partir des données), la validation (tests de performance et de robustesse), la mise en production (génération d'outputs en environnement opérationnel) et l'adaptation continue (apprentissage par renforcement ou mises à jour régulières)

Enfin, l'importance de tirer les leçons de chaque incident et de mettre à jour les plans de réponse en conséquence a été soulignée comme un élément essentiel d'une stratégie de sécurité adaptative et résiliente face aux menaces ciblant les systèmes d'IA.

En somme, les discussions ont insisté sur la nécessité d'isoler rapidement le système compromis, de préserver les preuves et de lancer une enquête interne approfondie en cas d'incident. L'activation immédiate d'un plan de gestion de crise, incluant des procédures de communication interne et externe structurées, est jugée essentielle. Si le respect des obligations légales de notification est obligatoire par définition, il convient surtout que les entités connaissent bien le cadre juridique applicable aux systèmes d'IA, qui reste récent et dont les exigences réglementaires cyber entreront en vigueur d'ici août 2026. Enfin, l'analyse post-incident et l'adaptation continue des plans de réponse sont considérées comme des leviers clés pour renforcer la résilience de l'organisation face aux incidents de sécurité sur les systèmes d'IA.

Question directrice 6 : 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 ?

Les discussions ont porté sur l'établissement de protocoles d'évaluation et de communication d'impact en cas de vulnérabilité affectant les modèles d'IA, avec une attention particulière pour les clients stratégiques dont les opérations dépendent de ces systèmes. Il est apparu que la simple identification d'une vulnérabilité est insuffisante ; une compréhension fine de ses répercussions potentielles est essentielle. Les échanges ont souligné la nécessité de développer un processus d'évaluation structuré allant au-delà des aspects techniques pour intégrer la criticité du cas d'usage du modèle pour chaque client stratégique, la sensibilité des données qu'il manipule, l'impact potentiel sur leurs processus métier et leur réputation, ainsi que les conséquences financières et réglementaires d'une compromission.

Par ailleurs, les échanges ont mis en évidence l'importance d'une communication proactive, transparente en cas de vulnérabilité avérée. Cette communication devrait non seulement alerter rapidement les parties prenantes concernées (qu'il s'agisse des clients stratégiques ou de la chaîne de sous-traitance des fournisseurs de système d'IA), mais également fournir une évaluation claire et détaillée de l'impact potentiel spécifique à leur contexte, c'est-à-dire en considérant l'état de l'usage et du déploiement du système d'IA chez le client, ainsi que l'étendue de l'indisponibilité potentielle de son SI global et les interactions du système d'IA compromis avec le reste de son infrastructure informatique.. De plus, cette communication devrait préciser les mesures d'atténuation immédiates à mettre en œuvre, et un calendrier prévisionnel des actions correctives à long terme, éléments définis par l'ensemble des équipes impliquées dans la gestion de crise, incluant les équipes cyber et IA. La mise en place de canaux de communication dédiés et sécurisés pour ces échanges critiques permet de maintenir la confiance et assurer une collaboration efficace pendant la gestion de crise.

Concernant l'évaluation des risques potentiels associés au déploiement de modèles d'IA spécialisés dans des environnements de production, les échanges ont souligné l'intérêt, pour

les industriels, d'adopter une approche de gestion des risques rigoureuse. La nature spécifique de ces modèles, et leur intégration dans des contextes opérationnels critiques avec des interconnexions au système d'Information existant, créent de nouvelles surfaces d'attaque et des opportunités de latéralisation pour les cyberattaquants. Il est important de souligner que la spécialisation d'un modèle d'IA n'amplifie pas directement son exposition à l'exploitation d'une vulnérabilité ; c'est bien son interconnexion et son rôle critique dans le SI global de l'entité qui constituent un risque d'opportunité privilégié pour les attaquants. À cet égard, la robustesse de méthodes comme EBIOS RM a été soulignée : elle s'applique parfaitement à l'analyse des risques cyber et IA sur ces systèmes d'Information d'IA, et il est fortement recommandé aux acteurs économiques de mener activement leurs propres analyses afin de limiter leur exposition.

Dans ce cadre, les discussions ont mis en lumière l'impératif, pour l'utilisateur, d'intégrer une analyse des risques cyber prenant en compte l'ensemble de la chaîne d'approvisionnement des composants d'IA. Cette démarche, qui peut s'appuyer sur des outils comme un SBOM (*Software Bill of Materials*) ou AIBOM (*AI Bill of Materials*), doit inclure une évaluation des risques liés aux modèles pré-entraînés, aux librairies logicielles, aux plateformes d'exécution et aux API externes. Elle vise à identifier les vulnérabilités potentielles chez les fournisseurs, à évaluer leurs pratiques de sécurité et à établir des plans de contingence en cas de compromission d'un élément de cette chaîne, en considérant tant l'aspect technique des composants que le niveau de maturité cyber de l'organisation fournisseuse.

De plus, les échanges ont insisté sur l'importance, pour l'utilisateur, d'intégrer la sécurité lors du déploiement des modèles spécialisés. Cela se traduit par l'application de mesures de sécurité spécifiques et adaptées (telles que des mécanismes robustes de contrôle d'accès aux données d'entraînement et d'inférence, et une surveillance comportementale continue pour détecter les anomalies et les tentatives de manipulation). Il est également essentiel que l'utilisateur exige de ses fournisseurs l'application de la 'sécurité dès la conception' tout au long du cycle de vie du développement des modèles, afin de garantir un SIA sécurisé par défaut. Enfin, la mise en œuvre de tests d'intrusion réguliers et ciblés, simulant des scénarios d'attaque spécifiques aux vulnérabilités de ces modèles spécialisés (comme l'extraction de connaissances propriétaires ou la perturbation de fonctionnalités critiques), a été considérée comme une bonne pratique pour une identification proactive des faiblesses et une gestion efficace des risques associés à leur déploiement en production

Les discussions ont mis en avant l'importance d'établir des protocoles d'évaluation et de communication d'impact en cas de vulnérabilité affectant les modèles d'IA, en particulier pour les clients stratégiques. Il a été souligné que la simple identification d'une vulnérabilité ne suffit pas : il est essentiel d'évaluer précisément ses répercussions en prenant en compte la criticité du cas d'usage, la sensibilité des données, l'impact sur les processus métier, la réputation, ainsi que les conséquences financières et réglementaires.

La nécessité d'une communication proactive, transparente et contextualisée a été rappelée, notamment via des canaux sécurisés dédiés. Cette communication doit fournir une évaluation claire de l'impact, des mesures d'atténuation immédiates, un calendrier des actions correctives et tenir compte de l'état de déploiement du système d'IA chez le client.

Concernant le déploiement de modèles d'IA spécialisés, il est recommandé d'adopter une gestion des risques rigoureuse et globale, en s'appuyant sur des méthodes comme EBIOS RM et des outils de traçabilité tels que le SBOM ou l'AIBOM.

Enfin, il a été rappelé que le risque principal ne provient pas de la spécialisation du modèle, mais de son interconnexion et de son rôle critique dans le système d'information global. Une analyse des risques intégrant l'ensemble de la chaîne d'approvisionnement et des mesures de sécurité adaptées tout au long du cycle de vie du modèle sont essentielles pour limiter l'exposition aux menaces.

Question directrice 7 : Quelles stratégies de gestion de crise mettriez-vous en œuvre ? Sont-elles spécifiques en raison de la nature du système impacté ?

Les discussions ont exploré la nécessité d'une refonte significative, plutôt qu'une simple adaptation, des stratégies de gestion de crise existantes face à l'intégration croissante et à la complexification des systèmes d'IA.

La nature intrinsèque de l'IA—capacité d'apprentissage autonome, opacité ou transparence du processus décisionnel selon les modèles, et potentiel d'impact sur des fonctions critiques—exige une approche fondamentalement nouvelle. Premièrement, il a été jugé que l'adaptation des cadres traditionnels ne constitue qu'un point de départ : il est indispensable de développer des dispositifs spécifiquement conçus pour les risques propres à l'IA.

Ces dispositifs doivent aller au-delà des vulnérabilités cyber classiques pour prendre en compte :

- Les attaques par empoisonnement, pouvant générer des biais algorithmiques ou des dérives comportementales majeures ;
- Les attaques par évasion, qui relèvent de la manipulation malveillante des systèmes d'IA et peuvent compromettre des fonctions critiques ;
- Les attaques par extraction, visant à dérober des connaissances propriétaires ou des données sensibles, avec des conséquences opérationnelles, réglementaires ou éthiques.

Ainsi, la gestion de crise en contexte IA doit intégrer la surveillance continue du comportement des modèles, des protocoles d'investigation adaptés à leur niveau d'explicabilité, et une coordination renforcée entre équipes techniques, métiers et juridiques pour répondre aux enjeux spécifiques de sécurité, d'éthique et de conformité réglementaire

Deuxièmement, la mise en place de systèmes d'alerte précoce a été envisagée sous un angle beaucoup plus sophistiqué. Il ne s'agit plus seulement de détecter des anomalies techniques, mais d'intégrer des capacités d'analyse comportementale avancée pour identifier des signaux faibles précurseurs de crises potentielles, en s'appuyant sur des indicateurs de performance spécifiques à l'IA et des techniques de *monitoring* en temps réel de son fonctionnement interne.

L'intelligence artificielle constitue un véritable changement de paradigme en matière de cybersécurité, car elle remet en question la dichotomie traditionnelle entre sécurité et sûreté

de fonctionnement, ainsi qu'entre contenu et contenant. Les systèmes d'IA sont paramétrés par leurs données d'entraînement et évoluent en continu à travers les données qu'ils traitent une fois en production, ce qui complexifie la gestion des risques et la détection des incidents. Les échanges ont ainsi montré qu'il n'est pas nécessaire de repenser en profondeur les stratégies de gestion de crise existantes, mais plutôt d'y intégrer de façon proactive les enjeux propres à l'IA. Il est essentiel que les RSSI/CISO se mettent à jour sur les spécificités des SIA, tant sur le plan technologique que sur les risques cyber associés, afin d'anticiper et de gérer efficacement d'éventuelles crises impliquant ces systèmes. Cela suppose, d'une part, de renforcer la sécurisation ex ante des SIA, notamment en luttant contre les effets de bord liés à leur fonctionnement en « boîte noire », et, d'autre part, d'adapter les dispositifs de gestion de crise pour inclure les problématiques éthiques, juridiques et de gouvernance des données. La constitution d'équipes de gestion de crise pluridisciplinaires est indispensable pour mener l'investigation, déterminer s'il s'agit d'un incident cyber ou de sûreté, et comprendre l'étendue de l'incident. L'intégration de ces dimensions dans les procédures existantes, ainsi que l'utilisation d'outils d'explicabilité et de traçabilité, s'inscrit pleinement dans les recommandations de l'ANSSI et les standards européens en cybersécurité, permettant d'anticiper, de détecter et de traiter plus efficacement les incidents liés à l'IA, sans bouleverser l'architecture globale de la gestion de crise

La diffusion rapide et généralisée des systèmes d'IA dans le tissu économique accroît les enjeux de communication en situation de crise. Dans ce contexte, il est essentiel d'adopter des stratégies de communication adaptées capables de répondre à la complexité technique des incidents et aux attentes de différents publics, tout en préservant la confiance des parties prenantes. Il convient d'éviter toute dramatisation et de rappeler que l'IA ne doit pas être perçue comme un bouleversement fondamental (« *game changer* »), mais comme un outil supplémentaire générant de nouveaux enjeux cyber. La communication de crise doit donc s'attacher à contextualiser les incidents, à expliquer clairement les risques spécifiques liés à l'IA et à rassurer sur la capacité des organisations, notamment des RSSI/CISO, à intégrer ces enjeux dans leur dispositif global de gestion des risques

Enfin, la possibilité de recourir à des stratégies de "rétrogradation contrôlée" ou de "mise en quarantaine" des systèmes d'IA critiques, tout en assurant la continuité des opérations essentielles, a été identifiée comme une mesure de sécurité efficace. L'intégration d'une évaluation continue et dynamique des risques spécifiques à l'IA, ainsi que des formations et des simulations régulières axées sur des scénarios impliquant des systèmes d'IA, sont apparues comme des leviers pour garantir l'efficacité de ces stratégies de gestion de crise adaptées à l'ère de l'IA.

En conclusion, il ne s'agit pas de repenser complètement les stratégies de gestion de crise, mais plutôt de veiller à ce qu'elles intègrent pleinement les spécificités liées à l'intelligence artificielle. Les RSSI et CISO doivent se former aux particularités des systèmes d'IA, afin comprendre leurs risques cyber et de les anticiper. Cela passe, d'une part, par une sécurisation ex ante, en cherchant à limiter au maximum les effets de bord liés au fonctionnement en « boîte noire » des SIA, et d'autre part, par l'inclusion systématique des problématiques IA (éthiques, juridiques, gouvernance des données) dans les stratégies de gestion de crise déjà en place. L'élaboration de procédures dédiés, l'intégration d'expertises multidisciplinaires et la mise en place de d'outils d'explicabilité sont jugées essentielles. Les stratégies doivent inclure des systèmes d'alerte avancés, une communication pédagogique et des mesures telles que la rétrogradation contrôlée des systèmes critiques. Enfin, l'évaluation

dynamique des risques et la formation et la sensibilisation régulières des équipes sont considérées comme des leviers clés pour une gestion de crise efficace à l'ère de l'IA.

Question directrice 8 : Comment les organisations peuvent-elles équilibrer l'exploitation de systèmes 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 ?

L'intégration de l'IA dans les outils de cybersécurité offre des opportunités significatives pour renforcer la défense des organisations. Cependant, elle s'accompagne de défis majeurs, notamment la confidentialité des données, les biais algorithmiques et la dépendance aux systèmes automatisés. Trouver un équilibre entre ces opportunités et ces risques demande une approche stratégique et intégrée, bien au-delà de simples ajustements techniques.

Les discussions ont mis en évidence l'importance de considérer les enjeux de confidentialité, de biais et de dépendance technologique, non seulement pour les systèmes d'IA qui automatisent des fonctions de sécurité, mais pour tout déploiement de SIA au sein d'une organisation. Les RSSI doivent s'assurer que les outils basés sur l'IA sont entraînés sur des jeux de données fiables, que des mesures de protection de la confidentialité sont en place, et que des procédures de contrôle humain et des plans de secours sont définis en cas d'indisponibilité du SIA.

Plusieurs pistes ont été explorées pour naviguer cet équilibre complexe :

L'élaboration de politiques de gestion des données constitue un pilier. Cela implique la mise en place de politiques claires concernant la collecte, le stockage, le traitement et la suppression des données, en accord avec les réglementations sur la protection de la vie privée (comme le RGPD ou le RIA). L'adoption de techniques d'anonymisation et de pseudonymisation avancées aide à minimiser les risques d'identification.

La provenance et la qualité des données utilisées pour entraîner les modèles d'IA comptent pour prévenir l'introduction et la propagation de biais. Il s'agit notamment de limiter les risques d'empoisonnement du modèle ou d'évasion par la manipulation du jeu de données d'entraînement. De telles manipulations pourraient, par exemple, introduire des portes dérobées exploitables par des attaquants pour pénétrer le système d'information d'une organisation.

En ce qui concerne les biais algorithmiques, les organisations doivent développer des techniques d'atténuation. Cela inclut non seulement la diversification des ensembles de données d'entraînement, mais plus largement un travail d'ingénierie des données ("*data engineering*") visant à retravailler les jeux de données pour limiter les biais, anonymiser ce qui doit l'être, et formater les données pour qu'elles soient interprétables par le modèle d'IA. Le développement d'outils statistiques de correction de biais et la mise en place de processus d'audit réguliers pour évaluer et corriger les biais résiduels tout au long du cycle de vie du modèle s'avèrent aussi utiles.

Il est pertinent de maintenir un contrôle humain significatif, particulièrement lorsque des fonctions critiques, et notamment des fonctions cyber, sont automatisées. Le principe du

"*human-in-the-loop*" doit être appliqué pour éviter que le SIA ne prenne des décisions qui pourraient avoir des conséquences néfastes sur l'organisation. Cela implique de définir clairement les rôles et les responsabilités des opérateurs humains vis-à-vis des systèmes d'IA utilisés pour des fonctions de sécurité, en veillant à ce qu'ils conservent la capacité d'intervenir, de remettre en question les décisions du système d'IA et de comprendre le contexte opérationnel.

Développer une culture d'utilisation responsable de l'IA est aussi un point clé. Cela peut passer par l'établissement de chartes d'usage des SIA au sein d'une organisation. La formation et la sensibilisation du personnel à l'utilisation responsable de l'IA doivent mettre en évidence les risques liés à la confidentialité des données, aux biais algorithmiques et à la dépendance technologique, tout en promouvant une compréhension critique des capacités et des limites de l'IA (notamment les risques d'hallucinations).

L'intégration des principes de "sécurité dès la conception" dans toutes les phases (conception, développement et déploiement) des systèmes d'IA constitue une pratique importante. La mise en place d'audits réguliers et transparents des algorithmes et de leurs performances est également pertinente. Il est important de souligner que la cybersécurité d'un SIA ne garantit pas son éthique. L'évaluation de la cybersécurité des SIA relève des autorités cyber, tandis que l'aspect éthique est une considération distincte.

En résumé, l'équilibre entre l'exploitation de l'IA pour la sécurité et la gestion des risques liés à la confidentialité, aux biais et à la dépendance technologique demande une approche stratégique intégrée. Des politiques de gestion des données solides, des techniques d'atténuation des biais et des audits réguliers sont des éléments à prendre en compte. Il s'agit aussi de maintenir un contrôle humain, de former les équipes à l'utilisation responsable de l'IA et d'intégrer la sécurité dès la conception.

Question directrice 9 : 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 ?

Les systèmes de surveillance et de détection d'anomalies devraient pouvoir s'adapter de façon dynamique et en temps réel lorsqu'une attaque confirmée cible les modèles en production. Si les approches statiques de surveillance restent pertinentes pour détecter certaines anomalies, elles montrent leurs limites face à l'évolution rapide et à la sophistication croissante des attaques. Les discussions ont mis en avant la nécessité de recourir à une surveillance adaptative, capable d'ajuster ses seuils de sensibilité, ses règles de corrélation et ses méthodes d'analyse en fonction de l'attaque détectée. Cette adaptabilité peut s'appuyer, sans s'y limiter, sur des méthodes d'apprentissage statistique ou d'IA, qui permettent d'identifier de nouveaux schémas d'attaque et d'optimiser la détection au fil du temps. Toutefois, toutes les méthodes statistiques n'intègrent pas un apprentissage en continu : il s'agit donc de prévoir des mises à jour régulières des systèmes IA utilisés pour la sécurité, afin de maintenir leur pertinence face à l'évolution des menaces.

Par ailleurs, comme pour toute crise cyber, la gestion d'incidents impliquant un SIA ne doit pas se limiter à la notification d'alertes de sécurité. Il est essentiel de mettre en place des

processus d'investigation structurés, avec des rôles et responsabilités clairement définis pour chaque membre de l'équipe de réponse, ce qui rejoint les principes fondateurs de la gestion de crise en cybersécurité. Pour les SIA, il s'agit en particulier de distinguer ce qui relève d'un incident de sécurité (attaque) et ce qui relève d'un dysfonctionnement ou d'une défaillance du système, la frontière entre sûreté et sécurité étant souvent plus complexe à établir avec ces technologies. Les procédures doivent inclure la validation rapide des alertes, la classification des anomalies selon leur impact, l'analyse des causes et la mise en œuvre de mesures de remédiation adaptées. L'automatisation de certaines étapes, comme l'isolement des systèmes affectés ou le lancement de scripts correctifs, contribue à renforcer la rapidité et l'efficacité de la réponse

Enfin, il est important de tester régulièrement ces dispositifs de surveillance adaptative et les procédures de réponse, notamment via des exercices de simulation d'attaques réalistes, afin d'identifier les points faibles, d'évaluer la pertinence des mises à jour et d'améliorer continuellement la résilience des modèles de production face aux menaces

En synthèse, les discussions ont mis en avant que, face à la sophistication et à l'évolution rapide des attaques, il faudrait compléter les approches statiques de surveillance par des dispositifs adaptatifs, capables d'ajuster leurs paramètres en temps réel selon la nature des menaces. Cette adaptabilité peut s'appuyer sur des méthodes d'apprentissage statistique et sur des mises à jour régulières des systèmes de détection. Comme pour toute gestion de crise cyber, il est indispensable de définir des processus d'investigation structurés, avec des rôles clairement identifiés, afin de distinguer entre incident de sécurité et dysfonctionnement du système, et de mettre en œuvre des mesures de remédiation adaptées. L'automatisation de certaines réponses et la réalisation régulière de simulations d'attaques renforcent la rapidité, l'efficacité et la résilience des dispositifs de sécurité pour les modèles d'IA en production.

<u>Question directrice 10</u> : En cas de cyberattaque ciblant l'une de vos solutions ou systèmes 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 ?

Les échanges ont permis de dégager que, face à une cyberattaque ciblant un système basé sur l'IA, les mesures d'isolement d'urgence doivent s'appuyer sur les bonnes pratiques cyber classiques, tout en adaptant le périmètre d'intervention aux spécificités des SIA. Il a été souligné que ce qui distingue les SIA, c'est la diversité et l'interconnexion de leurs composantes : il ne s'agit pas seulement d'isoler une machine ou un service, mais de segmenter précisément l'ensemble des éléments concernés, incluant les modèles, les données d'entraînement, de test et de production, ainsi que les interfaces et flux associés. Cette granularité est indispensable pour limiter la propagation de l'attaque tout en protégeant l'intégrité du reste du système d'information.

Par ailleurs, les échanges ont insisté sur la nécessité de renforcer immédiatement les contrôles d'accès sur tous les composants du système IA, d'ajuster les droits, et de mettre en place des mécanismes d'authentification adaptés pour empêcher toute escalade de privilèges ou mouvement latéral de l'attaquant et de réévaluer les droits sur tous les composants du SIA, en tenant compte de la multiplicité des points d'entrée et des interdépendances internes.

Concernant la chaîne d'approvisionnement, il a été convenu que la gestion des risques doit intégrer l'ensemble des services et sous-traitants impliqués dans le cycle de vie du SIA, tels que les fournisseurs de modèles pré entraînés, les plateformes cloud ou les prestataires techniques. La cartographie précise de ces dépendances, la mise en place de clauses de sécurité contractuelles, les audits réguliers et la préparation de plans de continuité permettant de contourner ou de remplacer temporairement un fournisseur compromis sont des éléments clés pour assurer la résilience globale.

En synthèse, les échanges ont confirmé que l'efficacité des mesures d'urgence repose sur la combinaison de pratiques cyber éprouvées (isolement, contrôle d'accès, documentation) et d'une adaptation du périmètre d'action aux spécificités et interconnexions des systèmes IA, avec une attention particulière portée à la résilience de la chaîne d'approvisionnement numérique.

<u>Question directrice 11</u> : 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 ?

Les discussions ont porté sur l'adaptation continue des stratégies de sécurité pour protéger les modèles d'IA, en soulignant que ces ajustements doivent couvrir l'ensemble du cycle de vie des modèles : de l'entraînement aux phases de test, de validation, de mise en production et d'exploitation.

Les approches statiques de sécurité restent pertinentes et constituent un socle indispensable, mais elles doivent être complétées par des mesures adaptatives pour répondre aux menaces émergentes. Les discussions ont mis l'accent sur plusieurs axes d'amélioration et de bonnes pratiques, plus que sur des innovations de rupture.

En matière de protection des données utilisées lors de l'entraînement, ont été évoquées des techniques telles que l'apprentissage fédéré (pour limiter la centralisation des données), la confidentialité différentielle (pour protéger les informations sensibles lors du traitement), ou encore le chiffrement homomorphe (pour permettre des calculs sur des données chiffrées). L'intégrité des données d'entraînement, de test et de validation a également été identifiée comme un enjeu majeur, avec l'intérêt de recourir à des signatures numériques ou à la blockchain pour garantir leur authenticité et détecter toute modification non autorisée.

L'isolation des environnements – qu'il s'agisse d'entraînement, de test ou de production – a été présentée comme une barrière supplémentaire, avec la possibilité de mettre en œuvre des architectures de type « *air gap* » logique ou physique, des conteneurs sécurisés et des politiques de transfert de modèles contrôlées et tracées entre les différents environnements.

Pour prévenir l'empoisonnement des modèles, les échanges ont souligné l'importance de mécanismes de détection d'anomalies sur les jeux de données, d'algorithmes robustes aux perturbations et de processus de validation renforcés pour toute contribution externe. En production, la surveillance comportementale en temps réel, l'utilisation de techniques d'obfuscation ou de « *watermarking* », ainsi que le recours à des environnements d'exécution

sécurisés (*Trusted Execution Environments*), ont été évoqués pour limiter les risques de vol, de manipulation ou de fuite des modèles.

Enfin, il a été unanimement reconnu que la sensibilisation et la formation continue des équipes, la réalisation régulière d'audits de sécurité et de tests d'intrusion, ainsi que l'intégration de la sécurité dès la conception des modèles (*security by design*), restent des leviers essentiels pour anticiper et contrer les menaces.

En résumé, les discussions ont mis en avant l'importance d'une approche globale et évolutive de la sécurité des modèles d'IA, couvrant toutes les phases du cycle de vie : entraînement, test, validation, mise en production et exploitation. Les pratiques recommandées incluent la protection et l'intégrité des données, l'isolation des environnements, la surveillance des comportements, ainsi que la formation continue des équipes et l'intégration de la sécurité dès la conception. Les approches statiques et adaptatives sont complémentaires pour garantir la résilience des modèles