CRPC | Chaire de recherche en prévention de la cybercriminalité

Notes *de synthèse*

Vol. 4, Num. 3
2024

La sécurité de l'IA

Layla Jasic, étudiante au doctorat en criminologie
Benoît Dupont, titulaire de la Chaire

Introduction

L'intelligence artificielle (IA) incarne une révolution technologique remodelant l'ensemble des secteurs d'activité humaine. Cependant, cette transformation fulgurante n'est pas sans risques, particulièrement en termes de sécurité des systèmes et de protection des données. Cette note de synthèse vise à **cadrer les enjeux de sécurité liés à l'IA, en évaluant les risques et les solutions envisageables pour atténuer ces risques de manière proactive**.

Cette note débute par un **exposé des principaux risques de sécurité associés à l'utilisation de l'IA**. L'accent est ici mis sur la sécurité, par distinction avec les défis éthiques et de sûreté – notamment la discrimination algorithmique – qui ont focalisé l'attention sur les risques de l'IA. L'émergence de l'IA s'accompagne de la découverte de vulnérabilités inédites, où des brèches de sécurité pourraient engendrer des violations massives de données privées, accroître la fréquence et la gravité des cyberattaques et faciliter l'apparition de nouveaux types de cybercriminalité. Cette note de synthèse discerne **deux principales catégories de risques**: ceux liés à **l'utilisation de l'IA comme un outil pour faciliter des attaques**, et ceux où **l'IA constitue la cible des attaques**.

Cette première section du texte examine les attaques perpétrées avec l'IA, soulignant l'aggravation des attaques existantes grâce à la faculté de l'IA d'**incarner de manière crédible des traits humains**. La capacité de l'IA à **générer de l'information erronée, mais hautement crédible** crée un risque sans précédent de manipulation des individus, qui pourraient être exposés à des escroqueries sophistiquées. Cette section aborde également **les attaques dirigées contre les systèmes d'IA**, notamment par l'empoisonnement des modèles et les tentatives d'extraction de données qui menacent l'intégrité et la fiabilité des systèmes d'IA. La deuxième section du texte s'intéresse aux **principaux cadres de sécurité élaborés par des organisations en pointe dans les secteurs technologiques et gouvernementaux**, tels que Google, Microsoft, Nvidia, Malwarebytes, Gartner, l'Agence européenne de sécurité et de régulation du secteur numérique (*European Union Agency for Cybersecurity* [ENISA]) , ou le *National Cyber Security Centre* (NCSC) du Royaume-Uni. Celles-ci proposent des stratégies opérationnelles complètes pour encadrer la sécurité des systèmes d'IA, depuis le développement jusqu'au déploiement des systèmes de production, tout en soulignant l'importance de l'intégration de ces cadres dans le cycle de vie complet des systèmes d'IA.

Pour finir, une série de recommandations et de meilleures pratiques de cybersécurité spécifiques à l'IA sont présentées. Ce panorama s'efforce de synthétiser les mesures de cybersécurité pouvant

être appliquées à l'environnement spécifique de l'IA, ainsi que les pratiques de gestion et de sensibilisation des employés à mettre en œuvre. En somme, cette note de synthèse offre aux organisations qui souhaitent anticiper les risques découlant de l'intégration toujours plus grande de l'IA dans les systèmes numériques un aperçu des cadres de référence disponibles dans le domaine.

Principaux risques de sécurité liés à l'utilisation de l'IA

L'intelligence artificielle (IA) est une technologie prometteuse pour l'avancement d'un grand nombre de secteurs d'activités, mais il est indéniable qu'elle vient avec son lot de risques, surtout au stade de faible maturité qui est le sien actuellement. Les **principaux enjeux et risques** entourant l'utilisation de l'IA concernent principalement **la sûreté** et **l'éthique** (par exemple, enjeux de discrimination algorithmique, sûreté physique des individus), ainsi que **la sécurité**, c'est-à-dire la préservation de la confidentialité, de l'intégrité et de l'accessibilité des données d'un système. Les enjeux éthiques associés à l'IA et l'utilisation des données de masse (*Big Data* en anglais) en général [1, 2, 3, 4] font déjà l'objet de nombreuses études et sont bien documentés. C'est pourquoi **cette note de synthèse se focalisera plutôt sur les enjeux de sécurité**, qui nécessitent d'être davantage approfondis.

L'utilisation de l'IA présente à la base un risque accru pour la confidentialité et la protection des données sensibles, simplement en raison de la quantité massive d'informations que son utilisation requiert. Une brèche dans un système IA pourrait avoir des **conséquences particulièrement catastrophiques pour la vie privée des utilisateurs ou des clients**, **ou la propriété intellectuelle des organisations**, si ces données se retrouvent entre de mauvaises mains [5]. De plus, l'IA pourrait contribuer à **augmenter la fréquence et la gravité des cyberattaques**. Elle pourrait également donner la possibilité à des acteurs malveillants de **commettre de nouveaux types de crimes** jusqu'ici impossibles, mais il est raisonnable de croire que cette technologie sera surtout utilisée pour **amplifier certains types de cybercrimes déjà courants**, tels que l'hameçonnage, la cyberfraude ou le vol de données [6]. Bien qu'il existe dans la littérature différentes typologies des risques de sécurité liés à l'utilisation de l'IA, deux principales catégories en ressortent, soit **les attaques où l'IA est utilisée comme un outil** et **les attaques où l'IA constitue une cible**.

L'IA comme outil

Le principal risque découlant d'une utilisation criminelle de l'IA comme outil est **l'amplification des attaques malveillantes et autres types de piratage déjà existants**, notamment **grâce à la possibilité de « personnaliser » les attaques**. La plupart des tentatives d'ingénierie sociale ou d'hameçonnage utilisent habituellement des messages génériques qui sont envoyés aux cibles potentielles. L'utilisation de l'IA pourrait permettre aux cybercriminels de personnaliser les messages, de manière que la victime potentielle ait davantage l'impression qu'il s'agit d'un courriel légitime, par exemple [6, 7]. Par ailleurs, **ces attaques personnalisées réussissent nettement mieux que les attaques génériques**, plus précisément, **elles ont jusqu'à quatre fois plus de chances d'atteindre leurs objectifs qu'une attaque non ciblée** [8].

Un autre risque lié à l'utilisation de l'IA vient de **la possibilité pour des acteurs malveillants de créer de l'information qui n'existe pas ou de supprimer celle qui existe**. Il y aurait là un danger que des données personnelles très sensibles (comme des données médicales) soient compromises, parce qu'un malfaiteur aurait ajouté des éléments créés de toutes pièces au dossier d'un patient ou qu'il aurait supprimé des informations vitales d'un dossier [8]. De plus, **l'utilisation accrue de l'IA générative permettrait une sorte d'optimisation de ce type d'attaques**, car elle permettrait aussi de générer des images et du son synthétiques (les hypertrucages [*deepfakes* en anglais], p. ex. [6, 7].

Ces images et sons, fabriqués de toutes pièces, peuvent faciliter les vols d'identité, certes, mais ils peuvent aussi permettre aux cybercriminels de mettre en œuvre des stratagèmes frauduleux qui seraient pratiquement impossibles sans l'IA. Il peut s'agir notamment de se faire passer pour quelqu'un, comme une figure d'autorité (le cadre dirigeant d'une entreprise par exemple) ou un membre de la famille, afin de persuader les membres de leur entourage ou de leur organisation d'effectuer des paiements ou des transferts d'argent frauduleux en usurpant l'apparence ou la voix de la personne concernée [9].

L'IA comme cible

Il s'agit d'attaques où **le système IA lui-même constitue la cible** et où, le plus souvent, **les acteurs malveillants cherchent à tromper le système afin de compromettre son fonctionnement et/ou son intégrité** [7, 8, 10].

L'un des plus importants risques identifiés est **l'empoisonnement des données** (*data poisoning* en anglais), c'est-à-dire que des données inexactes ou biaisées sont délibérément injectées dans l'algorithme du système afin de nuire à son bon fonctionnement. Cela peut entraîner comme conséquence un bris ou un dysfonctionnement dudit système, des prédictions erronées ou de manière plus générale des résultats problématiques. À l'opposé, plutôt que d'injecter des données, il se peut aussi que **des acteurs malveillants cherchent à les extraire**, **dans le but de tenter d'identifier si des informations spécifiques ont été utilisées pour l'entraînement de l'algorithme**, – et par conséquent se les approprier. En principe, un système d'IA ne devrait pas permettre à ses utilisateurs d'identifier quelles données ont été utilisées pour l'entraîner, mais certaines manipulations sophistiquées peuvent exploiter la fragilité de ces systèmes et contourner leurs mesures de sécurité afin de les amener à divulguer des informations sensibles [7].

À ce stade-ci, il est encore difficile de savoir lesquels de ces crimes liés à l'IA deviendront les plus courants ni à quel point ils seront répandus. Il n'est pas facile de savoir non plus quel degré de risque chacun de ces crimes posera, tant pour les individus que pour les organisations [8]. Néanmoins, dès qu'une organisation choisit d'utiliser l'IA dans le cadre de ses activités, **il importe qu'elle identifie et mette en place différentes pratiques de sécurité pour assurer la protection de ses données et de ses systèmes**. Puisque la sécurité des systèmes d'IA est un domaine en émergence, il est recommandé de se baser dans un premier temps sur des pratiques généralistes bien établies en cybersécurité, pour ensuite appliquer des mesures spécifiques à l'IA [11]. Ainsi, la section qui suit examine quelques cadres de sécurité et lignes directrices élaborées à cet effet par des compagnies bien établies dans le domaine des technologies.

Principaux modèles d'encadrement de sécurité

Afin d'identifier les meilleures pratiques en matière de cybersécurité des systèmes utilisant l'IA, nous avons procédé à des recherches par mots-clés en ligne, afin de recenser les différents cadres et stratégies développés par des organisations connues disposant d'une expertise technique indéniable en matière d'IA et de sécurité. L'objectif principal de cette démarche est d'**identifier des organisations qui ont développé des approches intégrées, applicables**, **et par conséquent généralisables de la cybersécurité de l'IA**, par contraste avec des approches qui restent principalement théoriques ou se contentent de suggérer des listes de conseils dont l'origine et l'articulation avec les pratiques établies en cybersécurité peuvent parfois paraître nébuleuses. **Sept cadres de sécurité de l'IA ont été retenus** dans cette première radiographie : deux proviennent d'entreprises dominantes du secteur technologique (**Google** et **Microsoft**), deux autres proviennent d'entreprises moins connues du grand public, mais jouant un rôle central dans

les écosystèmes de l'IA (**Nvidia**) et de la cybersécurité (**Malwarebytes**). Un cinquième cadre a été élaboré par une entreprise de conseil et de recherche (**Gartner**). Les deux derniers ont été développés par des organismes gouvernementaux, soit l'**Agence européenne de sécurité et de régulation du secteur numérique** (European Union Agency for Cybersecurity [ENISA]) et le **National Cyber Security Centre** (NCSC) du Royaume-Uni. Il est à noter que, dans le processus de sélection, des cadres de sécurité de l'IA d'organisations bien connues telles que le National Institute of Standards and Technology (NIST) ou PwC ont été consultés, mais n'ont pas été retenus puisqu'ils sont centrés principalement sur les enjeux de sûreté et d'équité qui ont déjà été amplement examinés ailleurs. Nous présentons de manière plus détaillée quatre de ces sept cadres dans les paragraphes qui suivent, avant de présenter un tableau synthétique des mesures de contrôle préconisées par ces sept cadres.

Google – Safe AI Framework (SAIF)

Le **Safe AI Framework (SAIF)** développé par **Google** est un cadre employé pour sécuriser les systèmes utilisant l'IA. Il se base sur la prémisse que l'organisation qui le met en place possède déjà des contrôles et pratiques robustes en matière de cybersécurité. Cette approche suggère donc que **le contrôle et l'atténuation de certains risques de sécurité liés à l'IA passent par une extension des mesures de cybersécurité déjà connues**. Néanmoins, le cadre mentionne par la suite l'importance d'identifier les risques spécifiques que pourrait poser l'utilisation de l'IA. Il se base sur **6 éléments clés** :

- **Étendre les fondamentaux solides de cybersécurité générale** à tout l'environnement IA de l'organisation;
- **Renforcer la détection et la réponse aux incidents** en intégrant les spécificités de l'IA aux autres cybermenaces faisant l'objet d'une surveillance constante;
- **Automatiser les mécanismes de défense** pour qu'ils restent à jour tant avec les menaces déjà connues qu'avec les nouveaux risques qui pourraient survenir;
- **Harmoniser les contrôles** au niveau de toutes les plateformes de l'organisation afin d'assurer un niveau de sécurité constant;
- **Adapter les contrôles** pour améliorer les moyens d'atténuation et créer des boucles de rétroaction plus rapides pour le déploiement de l'IA;
- **Contextualiser les risques liés à l'IA** dans les processus d'affaires reliés.

Gartner - Artificial Intelligence Trust, Risk and Security Management (AI TRiSM)

L'*Artificial Intelligence Trust, Risk and Security Management* (AI TriSM) est défini par Gartner (2022) comme **un cadre de sécurité qui soutient la gouvernance**, **la fiabilité**, **l'équité** (*fairness*), **la robustesse**, **l'efficacité** et **la confidentialité** (*privacy*) des modèles d'IA. Il s'agit donc d'**un cadre qui combine à la fois la sécurité** (*security*) et **la sûreté** (*safety*) **des systèmes d'IA**. En ce qui concerne la sécurité spécifiquement, l'objectif principal identifié est de mettre en place des pratiques qui assureront la confidentialité des données, afin d'assurer la protection des clients et préserver leur confiance envers l'organisation.

ENISA – Framework for AI Cybersecurity Practices (FAICP)

Le *Framework for AI Cybersecurity Practices* (FAICP) de l'ENISA est **un cadre constitué de trois niveaux ou couches**, où chaque niveau repose sur le socle du précédent. **Le premier niveau comprend les bonnes pratiques de cybersécurité de base**, communes à tous les systèmes de technologies de l'information. **Le second niveau incorpore les pratiques de sécurité propres à l'IA**, et **le troisième niveau concerne les pratiques de sécurité de l'IA** dans ce qu'elles ont de **spécifique à chaque secteur d'activités** (comme les secteurs de

l'énergie, de la santé ou du transport automobile par exemple). Similaire au cadre de Google, l'approche proposée par l'ENISA conçoit l'atténuation des risques liés à l'IA comme une combinaison visant d'abord à étendre les mesures de cybersécurité fondamentales aux systèmes d'IA, puis de mettre en place des mesures qui leur sont spécifiques. Ce cadre est le résultat d'une étude de l'ENISA visant à recueillir les informations sur les différentes exigences nationales sur la cybersécurité de l'IA au sein des pays membres de l'UE et à identifier les lacunes dans les pratiques déjà existantes, afin de permettre à toutes les parties prenantes au déploiement de l'IA d'assurer la sécurité et la fiabilité des systèmes.

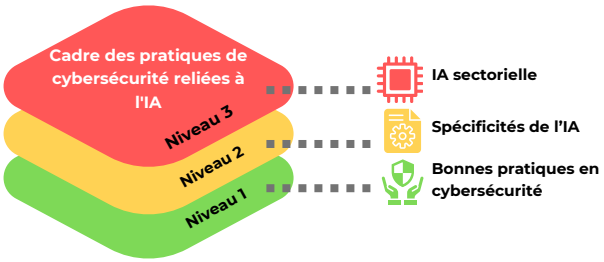

Figure 1. Cadre évolutif des bonnes pratiques de cybersécurité reliées à l'IA selon l'ENISA (2023)

National Cyber Security Centre – Guidelines for secure AI development

Le document *Guidelines for secure AI development* a été développé par le NCSC, en partenariat avec une vingtaine d'organisations du même type de différents pays, tels le Canada, la France, les États-Unis ou l'Australie, pour n'en nommer que quelques-uns. Son objectif est de **fournir des lignes directrices aux quatre phases de développement et de déploiement des systèmes d'IA** afin de s'assurer de leur sécurité, fonctionnalité et performance :

- **Conception sécuritaire** (sensibilisation des parties prenantes impliquées; modélisation des menaces; sécurité, fonctionnalité et performance du système);
- **Développement sécuritaire** (mise en place de mesures de sécurité à travers toute la chaîne d'approvisionnement; identification et protection de tous les actifs; documentation des données et des modèles utilisés dans le cadre du développement);
- **Déploiement sécuritaire** (sécurisation de l'infrastructure IA; protection en continu du modèle IA; développement de procédures de gestion et d'atténuation des incidents; évaluation pour assurer que le système d'IA est responsable; assurer que les utilisateurs peuvent utiliser facilement le système selon les fins prévues);
- **Opération et entretien sécuritaires** (surveiller les intrants et le comportement du système; mises à jour sécuritaires; partage des connaissances et des leçons apprises dans le cadre de la mise en place du système d'IA avec des acteurs de l'industrie, du milieu académique et du gouvernement).

Par ailleurs, le document met également l'accent sur le fait que **la sécurité doit demeurer un élément essentiel tout au long du cycle de vie des systèmes d'IA**, et non seulement durant les phases de mise en œuvre. Ceci est parfaitement logique, puisque les risques et les attaques ciblant ces systèmes continuent d'évoluer sans cesse (voir figure 2).

Autres

D'autres entreprises du secteur technologique, telles que Microsoft, Malwarebytes ou Nvidia, proposent des listes de meilleures pratiques ou de lignes directrices pour une implantation sécuritaire de l'IA, sans qu'il s'agisse explicitement d'un modèle d'encadrement à part entière. Microsoft et Malwarebytes misent également sur **l'existence de bonnes pratiques de sécurité** déjà en place et de mesures spécifiques à l'IA, tandis que Nvidia propose des **lignes directrices pour développer des systèmes utilisant l'IA à partir de zéro**. Il est à noter que Nvidia propose un modèle d'encadrement pour l'implantation de l'IA, mais il est surtout technique et destiné davantage aux développeurs et aux équipes de sécurité offensive (*Red Teams*).

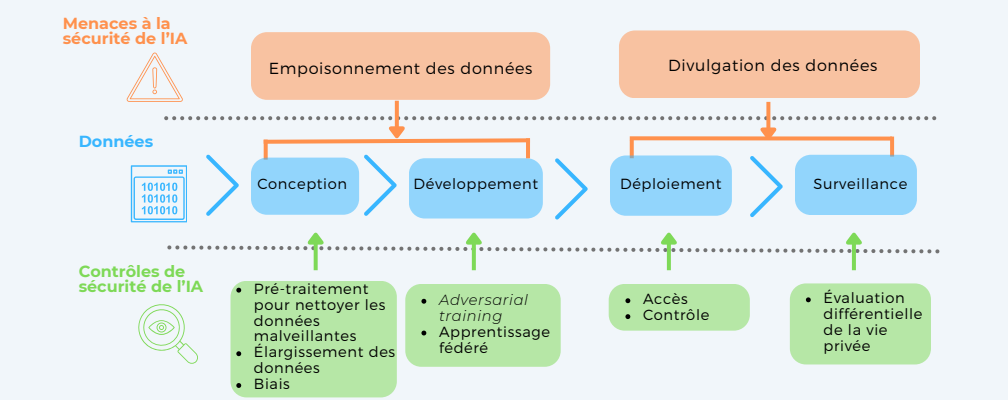

Figure 2. Modèle d'interactions entre les risques de sécurité pesant sur l'IA et les contrôles afférents selon l'ENISA (2023)

Recommandations et meilleures pratiques pour la sécurité des systèmes d'IA

À la lumière des recommandations formulées pour un déploiement sécuritaire de l'IA dans les différents modèles d'encadrement et listes de meilleures pratiques cités à la section précédente, il en ressort que **la plupart se basent sur la prémisse que l'organisation possède déjà de bonnes pratiques ou mesures de cybersécurité générales à partir desquelles elle peut développer celles qui seront spécifiques à l'IA**. Ces pratiques plus générales peuvent se résumer ainsi : **évaluer quels contrôles de sécurité existants peuvent être appliqués à l'IA et analyser quels contrôles spécifiques à l'IA doivent être ajoutés**; **renforcer la détection et l'intervention** en intégrant les spécificités de l'IA aux autres cybermenaces et enfin, **effectuer des mises à jour régulières** de tous les systèmes et applications. De plus, il est important de **gérer l'accès aux données d'entraînement**, par exemple en ne permettant qu'aux utilisateurs autorisés d'y accéder et en limitant la diffusion publique des informations techniques et organisationnelles qui s'y rattachent. Par ailleurs toutes données servant à des fins d'entraînement devraient être « nettoyées » au préalable, ce qui implique de détecter et de retirer celles qui sont suspectes ou compromises [12].

En ce qui a trait aux recommandations spécifiques à l'utilisation de l'IA, plusieurs aspects communs ressortent, en plus de quelques recommandations qui sont spécifiques à certains modèles d'encadrement. Les recommandations les plus fréquentes ont été synthétisées dans le **tableau 1**. La première partie présente **des recommandations qui concernent l'ensemble de l'organisation**, qu'il s'agisse de pratiques de gestion, de moyens de protéger les systèmes et, dans une moindre mesure, de bonnes pratiques pour les employés. La seconde partie quant à elle présente **des recommandations spécifiques pour la formation et la sensibilisation des employés**. De manière générale, il est toutefois à noter que certaines recommandations consultées manquent de précision et qu'il serait nécessaire d'approfondir les détails de la mise en œuvre de certaines d'entre elles.

Tableau 1. Synthèse des principales recommandations pour un déploiement sécuritaire de l'IA au sein d'une organisation

Recommandations	Google	Nvidia	Microsoft	Gartner	Malwarebytes	ENISA	NCSC
Recommandations pour l'organisation (systèmes, gestion, employés)							
Identifier toutes les utilisations qui seront faites de l'IA au sein de l'organisation, de même que les complexités et les risques spécifiques qui s'y rattachent. Ensuite, effectuer une évaluation des plus importants risques, menaces et vulnérabilités, de manière à pouvoir développer et mettre en place des mesures de sécurité correspondant au niveau de risque identifié.	✳	✳	✳			✳	✳
Prendre connaissance des différentes exigences en matière de sécurité liée à l'IA; implémenter des standards spécifiques à l'IA et se conformer aux législations en vigueur, lorsque requis.			✳			✳	
Protéger toutes les données passant par le système et porter une attention particulière aux données permettant d'identifier des individus (identifiants personnels). Sécuriser ces données notamment au moyen de la cryptographie et d'un contrôle robuste de gestion des accès.	✳		✳	✳	✳	✳	✳
Rester à jour au sujet des menaces et des nouveaux types d'attaques liées à l'IA. Évaluer régulièrement les vulnérabilités. Mener différents tests de sécurité, tels que des tests conduits par des *Red Team*, mais aussi des tests de pénétration et des techniques d'adversarial training.	✳	✳	✳		✳	✳	
Investir dans des processus de surveillance, de détection et d'intervention en matière de sécurité, en intégrant les spécificités des menaces liées à l'IA et aux autres cybermenaces.	✳	✳	✳	✳	✳		

Recommandations	Google	Nvidia	Microsoft	Gartner	Malwarebytes	ENISA	NCSC
Recommandations pour l'organisation (systèmes, gestion, employés) (suite)							
S'assurer qu'il y a toujours un humain dans le processus (tant pour des raisons de sécurité que d'éthique).	✳						
Programmer le système d'IA pour qu'il soit en mesure de protéger et sécuriser les données même lorsque les humains ne le font pas, c'est-à-dire qu'il doit pouvoir reconnaître lorsque des données sont sensibles et les protéger en conséquence.		✳					✳
Élaborer un plan de réponse aux incidents qui inclut comment endiguer l'incident, l'investiguer et y remédier.				✳			
Développer et mettre en place des pratiques de sécurité propres au secteur d'activité de chaque organisation.					✳		
Recommandations spécifiques à l'égard des employés							
Sensibiliser les employés à la manière dont l'IA est utilisée par l'organisation et les former à la gestion des risques liés à l'IA, notamment aux manières d'identifier du contenu nuisible généré au moyen de l'IA. Sensibiliser également toutes les parties prenantes aux risques dans l'environnement opérationnel, de façon à pouvoir mieux évaluer les risques et intervenir, au besoin.	✳		✳		✳	✳	✳
Présenter une introduction à l'IA à toutes les parties prenantes impliquées dans son déploiement en présentant les concepts essentiels ou de base, afin que même les personnes occupant des rôles non techniques puissent comprendre et évaluer les risques liés à cette technologie, et assurer la protection des systèmes.	✳						✳

Recommandations	Google	Nvidia	Microsoft	Gartner	Malwarebytes	ENISA	NCSC
Recommandations spécifiques à l'égard des employés (suite)							
Créer une équipe ou un groupe de travail dédié au déploiement de l'IA, puis favoriser la rétention des employés spécialisés en sécurité de l'IA et mettre leur formation à jour régulièrement.	*	*		*			
Développer et mettre en place des politiques ou des pratiques organisationnelles en matière de sécurité et de confidentialité des données..	*		*			*	

Références

[1] Bengio, Y., Hinton, G., Yao, A., Song, D., Abbeel, P., Harari, Y. N., ... et Mindermann, S. (2023). Managing AI Risks in an Era of Rapid Progress. *arXiv preprint arXiv:2310.17688*.

[2] Boyd, D. et Crawford, K. (2012). Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon. *Information, Communication & Society*, *15*(5), 662-679.

[3] Blanchard, A. et Taddeo, M. (2023). The ethics of artificial intelligence for intelligence analysis: a review of the key challenges with recommendations. *Digital Society*, *2*(1), 12.

[4] Joh, E. E. (2016). The new surveillance discretion: Automated suspicion, big data, and policing. *Harvard Law & Policy Review*, *10*, 1-15.

[5] Malwarebytes, (s.d.). *AI in Cyber Security: Risks of AI*. https://www.malwarebytes.com/cybersecurity/basics/risks-of-ai-in-cyber-security

[6] Choraś, M. et Woźniak, M. (2022). The double-edged sword of AI: Ethical Adversarial Attacks to counter artificial intelligence for crime. *AI Ethics*, *2*, 631–634

[7] Blauth, T. F., Gstrein, O. J., & Zwitter, A. (2022). Artificial intelligence crime: An overview of malicious use and abuse of AI. *IEEE Access*, *10*, 77110-77122.

[8] Hayward, K. et Maas, M. (2021). Artificial intelligence and crime: a primer for criminologists. *Crime Media Culture An International Journal*, *17*(2), 209-233.

[9] Bateman, J. (2020). *Deepfakes and synthetic media in the financial system: Assessing threat scenarios*. Carnegie Endowment for International Peace, Washington DC.

[10] Dupont, B. et Crosset, V. (2022). The security implications of artificial intelligence (AI) for the justice system. Dans J. A. Pérez Juan et F. J. Sanjuan Andrés (dir.). *Cuadernos Digitales* (p.35-51). Thomson Reuters Aranzadi, Cizur Menor.

[11] Brundage, M., Avin, S., Clark, J., Toner, H., Eckersley, P., Garfinkel, B., ... et Amodei, D. (2018). The malicious use of artificial intelligence: Forecasting, prevention, and mitigation. *arXiv preprint arXiv:1802.07228*.

[12] MITRE, (s.d.). *MITRE Atlas. Mitigations.* : https://atlas.mitre.org/mitigations

Références additionnelles du tableau

Baroni, K., (2023). Six Security Considerations for Machine Learning Solutions. *Microsoft*. https://techcommunity.microsoft.com/t5/fasttrack-for-azure/six-security-considerations-for-machine-learning-solutions/ba-p/3718592

ENISA. (2023). *Multilayer framework for good cybersecurity practices for AI*. https://www.enisa.europa.eu/publications/multilayer-framework-for-good-cybersecurity-practices-for-ai

Gartner, (2022, 19 octobre). *What It Takes to Make AI Safe and Effective*. https://www.gartner.com/en/articles/what-it-takes-to-make-ai-safe-and-effective

Gartner, (2022). *Top strategic technology trends 2023*. https://www.gartner.com/en/articles/gartner-top-10-strategic-technology-trends-for-2023

Google, (s.d.). *Secure AI Framework Approach: A quick guide to implementing the Secure AI Framework (SAIF)*. https://safety.google/cybersecurity-advancements/saif/

National Cyber Security Centre (2023). *Guidelines for secure AI system development*. https://www.ncsc.gov.uk/collection/guidelines-secure-ai-system-development

Pearce, W. et Lucas J. (2023, 14 juin). Nvidia AI Red Team: An Introduction. https://developer.nvidia.com/blog/nvidia-ai-red-team-an-introduction/

Sabin, S. (2023, 8 juin). Exclusive: Google lays out its vision for securing AI. *Axios*. https://www.axios.com/2023/06/08/google-securing-ai-framework

Cette note de synthèse a été réalisée avec le soutien de

ACT AUTONOMY THROUGH CYBERJUSTICE TECHNOLOGIES