LES ESSENTIELS

Modèle de Politique de Sécurité des Systèmes d'Information (PSSI) pour l'Intelligence artificielle

Février 2025

CLUSIF

Table des matières

1 Définitions

1.1 Système d'Intelligence Artificielle (SIA)

Un système d'Intelligence Artificielle est un système qui fonctionne grâce à une machine et est capable d'influencer son environnement en produisant des résultats, tels que des prédictions, des recommandations ou des décisions, pour répondre à un ensemble donné d'objectifs. Il utilise les données et les entrants générés par la machine et/ou apportés par l'homme afin de :

- percevoir des environnements réels et/ou virtuels ;
- produire une représentation abstraite de ces perceptions sous forme de modèles issus d'une analyse automatisée (ex. l'apprentissage automatisé) ou manuelle ;
- utiliser les déductions du modèle pour formuler différentes options de résultats. Les systèmes d'IA sont conçus pour fonctionner de façon plus ou moins autonome.

1.2 Système d'IA à usage général

Il s'agit d'un système d'IA qui est fondé sur un modèle d'IA à usage général et qui a la capacité de répondre à diverses finalités, tant pour une utilisation directe que pour une intégration dans d'autres systèmes d'IA.

1.3 Modèle d'IA à usage général

Un modèle d'IA, y compris lorsque ce modèle d'IA est entraîné à l'aide d'un grand nombre de données utilisant l'auto-supervision à grande échelle, qui présente une généralité significative et est capable d'exécuter de manière compétente un large éventail de tâches distinctes, indépendamment de la manière dont le modèle est mis sur le marché, et qui peut être intégré dans une variété de systèmes ou d'applications en aval, à l'exception des modèles d'IA utilisés pour des activités de recherche, de développement ou de prototypage avant leur mise sur le marché.

1.4 Cycle de vie d'un système d'IA

Les phases du cycle de vie d'un système d'IA sont au nombre de quatre :

- la conception du système et du modèle ;
- le développement ;
- le déploiement ;
- l'exploitation et la maintenance.

Ces phases se déroulent souvent de manière itérative et ne sont pas nécessairement séquentielles. La décision de retirer un système IA de l'exploitation peut intervenir à tout moment pendant la phase d'exploitation et de surveillance.

1.5 IA générative

L'IA générative est un sous-ensemble de l'intelligence artificielle, axé sur la création de modèles qui sont entraînés à générer du contenu (texte, images, vidéos, etc.) à partir d'un corpus spécifique de données d'entraînement.

1.6 *Large Language Model* (LLM)

Catégorie de modèles d'IA générative qui peuvent générer du texte proche du langage naturel d'un être humain, et qui sont généralement entraînés sur un large ensemble de données.

1.7 Modèle d'IA

Un modèle d'IA désigne, dans le contexte de cette politique, un réseau de neurones et ses paramètres (poids, biais).

1.8 Requête

Une requête (ou *prompt*) désigne l'instruction sous forme de texte envoyée par l'utilisateur au SIA.

1.9 Attaque adverse

Une attaque adverse (*adversarial attack*), parfois aussi appelée attaque antagoniste ou attaque par exemples contradictoires vise à envoyer à un système d'IA une ou plusieurs requêtes malveillantes dans le but de tromper ou d'altérer son bon fonctionnement.

1.10 Poids

Dans un réseau de neurones, un poids est un coefficient de puissance de la connexion entre deux neurones, qui s'ajuste pendant toute la phase d'entraînement. Un biais est une constante liée à un neurone permettant une "compensation" dans le calcul du résultat. La connaissance des poids du modèle peut permettre aux attaquants d'améliorer la capacité de certaines attaques.

1.11 Agent conversationnel

Un agent conversationnel est défini comme une application permettant un échange écrit entre l'utilisateur et le SIA et non un échange oral.

1.12 Qualité des données

La qualité des données désigne généralement un critère métier. Des critères de qualité des données d'un point de vue métier peuvent être par exemple l'origine, la quantité, l'exhaustivité, la pertinence, l'exactitude, la représentativité (au sens statistique), ou encore le respect d'une structure donnée.

1.13 Performance d'un modèle d'IA

La performance d'un modèle d'IA est un concept métier très dépendant des objectifs fixés lors de la conception du modèle. Elle peut inclure plusieurs facteurs comme la précision, la pertinence ou encore la rapidité des réponses générés pour les utilisateurs par exemple.

1.14 Explicabilité

L'explicabilité est la capacité de mettre en relation et de rendre compréhensible les éléments pris en compte par le SIA pour la production d'un résultat.

1.15 Hallucination

Phénomène dans lequel un modèle génère du contenu erroné qui n'est pas basé sur des

données réelles.

1.16 Fine-tuning

Technique consistant à spécialiser un modèle d'IA pré-entraîné à l'accomplissement d'une tâche spécifique. Cela consiste généralement à entraîner le modèle dans son ensemble, ou seulement certaines couches d'un réseau de neurones, pour un faible nombre d'itérations sur un ensemble de données spécifiques correspondant à la tâche visée. Cette pratique est parfois traduite par affinage, réglage fin, peaufinage ou encore spécialisation.

2 Périmètre et objet

Cette politique s'applique à tous les systèmes d'intelligence artificielle développés ou utilisés par l'entité.

Conformément au règlement européen du 13 juin 2024 établissant des règles harmonisées concernant l'intelligence artificielle (*AI Act*)[1], les pratiques en matière d'intelligence artificielle décrites dans l'article 5 sont interdites dans l'entité.

Conformément à l'article 4 de l'*AI Act* l'objet de cette politique est de décrire les mesures pour garantir, dans toute la mesure du possible, un niveau suffisant de maîtrise de l'IA pour les collaborateurs de l'entité, en prenant en considération leurs connaissances techniques, leur expérience, leur éducation et leur formation, ainsi que le contexte dans lequel les systèmes d'IA sont destinés à être utilisés, et en tenant compte des personnes ou des groupes de personnes à l'égard desquels les systèmes d'IA sont destinés à être utilisés.

Conformément à l'article 15 de l'*AI Act*, la conception et le développement des systèmes d'IA de l'entité sont tels qu'ils leur permettent d'atteindre un niveau approprié d'exactitude, de robustesse et de cybersécurité, et de fonctionner de façon constante à cet égard tout au long de leur cycle de vie.

La présente politique s'applique aux systèmes d'IA à haut risque et aux systèmes d'IA à usage général, développés et mis en œuvre dans l'entité, tels qu'ils sont décrits dans l'*AI Act*.

[1] https://artificialintelligenceact.eu/ai-act-explorer/

3 Cybersécurité du développement de systèmes d'IA

Ce chapitre définit les règles de cybersécurité pour les fournisseurs de systèmes d'intelligence artificielle (SIA), que ces systèmes aient été créés de toutes pièces ou qu'ils aient été construits à partir d'outils et de services fournis par d'autres. La mise en œuvre de ces règles permet l'élaboration de SIA sûrs, disponibles et qui fonctionnent sans révéler de données sensibles à des parties non autorisées.

Les SIA ont le potentiel d'apporter de nombreux avantages à l'entité. Toutefois, pour que les possibilités offertes par l'IA soient pleinement exploitées, celle-ci doit être développée, déployée et exploitée de façon sécurisée. La cybersécurité est une condition préalable à la sécurité, à la résilience, au respect de la vie privée, à l'équité, à l'efficacité et à la fiabilité des SIA.

Les SIA sont sujets à de nouvelles vulnérabilités en matière de sécurité qui doivent être prises en compte parallèlement aux menaces classiques de cybersécurité. Lorsque le rythme de développement est élevé - comme c'est le cas avec l'IA - la sécurité peut souvent être une considération secondaire. La sécurité doit être une exigence fondamentale, non seulement pendant la phase de développement, mais tout au long du cycle de vie du système.

C'est pourquoi ces règles sont divisées en quatre domaines clés du cycle de vie du développement d'un SIA :

- conception sécurisée ;
- développement sécurisé ;
- déploiement sécurisé ;
- exploitation et maintenance sécurisées.

3.1 Pourquoi la sécurité de l'IA est-elle différente ?

La spécificité des SIA est que ce sont des systèmes d'information qui :

- impliquent des composants logiciels (modèles) qui permettent aux ordinateurs de reconnaître et de mettre en contexte des modèles dans les données sans que les règles ne doivent être explicitement programmées par un être humain ;
- générer des prédictions, des recommandations ou des décisions basées sur un raisonnement statistique.

Outre les menaces existantes en matière de cybersécurité, les SIA sont exposés à de nouveaux types de vulnérabilités. Le terme "*adversarial machine learning*" (AML) est utilisé pour décrire l'exploitation de vulnérabilités dans les composants d'un SIA, y compris le matériel, les logiciels, les flux de travail et les chaînes d'approvisionnement. L'AML permet aux attaquants de provoquer des comportements involontaires dans les systèmes d'intelligence artificielle, notamment :

- en affectant les performances du modèle en matière de classification ou de régression ;
- en permettant aux utilisateurs d'effectuer des actions non autorisées ;
- en permettant l'extraction d'informations sensibles sur les modèles.

Il existe de nombreuses façons d'obtenir ces effets, comme les attaques par injection dans le domaine des grands modèles de langage (LLM), ou la corruption délibérée des données d'apprentissage ou du retour d'information de l'utilisateur (connue sous le nom d'empoisonnement des données).

3.2 Rôle du responsable du développement d'un SIA sécurisé

Les chaînes d'approvisionnement modernes de l'IA comptent de nombreux acteurs. Une approche simple suppose deux entités :

- le fournisseur qui est responsable de la conservation des données, du développement algorithmique, de la conception, du déploiement et de la maintenance ;
- l'utilisateur, qui fournit les données d'entrée et reçoit les données de sortie.

Le fournisseur doit mettre en œuvre des contrôles de sécurité et des mesures d'atténuation au sein de ses modèles, pipelines et/ou systèmes et, lorsque des paramètres sont utilisés, mettre en œuvre l'option la plus sûre par défaut. Lorsque les risques ne peuvent être atténués, le fournisseur est responsable de :

- informer les utilisateurs en aval de la chaîne d'approvisionnement des risques qu'ils acceptent ;
- conseiller les utilisateurs sur la manière d'utiliser le composant en toute sécurité.

Lorsque la compromission d'un système peut entraîner des dommages physiques ou de réputation tangibles ou étendus, une perte importante des opérations commerciales, la fuite d'informations sensibles ou confidentielles et/ou des implications juridiques, les risques liés à la cybersécurité de l'IA doivent être traités comme des risques critiques.

3.3 Règles de sécurité applicables pour la conception

Cette section décrit les règles qui s'appliquent à la phase de conception du cycle de vie du développement d'un SIA. Elle traite de la compréhension des risques et de la modélisation des menaces, ainsi que des sujets spécifiques et des compromis à prendre en compte lors de la conception du système et du modèle.

3.3.1 Sensibiliser les collaborateurs aux menaces et aux risques

Réf.	Règle	Informations de mise en œuvre
SIA-01	Les collaborateurs du fournisseur de SIA doivent être formés aux menaces et risques qui pèsent sur un SIA	Cette formation a pour objectif de comprendre les menaces qui pèsent sur la sécurité de l'IA et les moyens de les atténuer. Cette formation s'appuiera sur le document du NIST AI 100-2e2023 et sur le référentiel Atlas du MITRE.

Les collaborateurs du fournisseur de SIA doivent effectuer une veille technique afin de se tenir informés des menaces de sécurité et des modes de défaillance pertinents. Ils aident les gestionnaires de risques à prendre des décisions éclairées.

Réf.	Règle	Informations de mise en œuvre
SIA-02	Les utilisateurs du SIA doivent être sensibilisés aux risques de sécurité spécifiques aux SIA.	Cette sensibilisation concerne tous les utilisateurs d'un SIA. Elle doit être renouvelée tous les 2 ans.

Les développeurs doivent être formés aux techniques de programmation sécurisées et aux pratiques d'IA sûres et responsables.

3.3.2 Modéliser les menaces qui pèsent sur le SIA

Réf.	Règle	Informations de mise en œuvre
SIA-03	Une analyse de risque doit être conduite pour tout développement de SIA.	Cette analyse de risque inclut la compréhension des impacts potentiels sur le système, les utilisateurs et les entités si un composant d'IA est compromis ou se comporte de manière inattendue. Ce processus implique d'évaluer l'impact des menaces spécifiques à l'IA.

La sensibilité et les types de données utilisées dans le SIA peuvent influencer sa valeur en tant que cible pour un attaquant. L'analyse de risques doit tenir compte du fait que certaines menaces peuvent se développer à mesure que les SIA sont de plus en plus considérés comme des cibles de grande valeur et que l'IA elle-même permet de créer de nouveaux vecteurs d'attaque automatisés.

3.3.3 Conception sécurisée du SIA

Il est nécessaire d'évaluer la pertinence des choix de conception spécifiques à l'IA. Pour cela, il faut tenir compte du modèle de menace et des mesures d'atténuation de la sécurité associées, ainsi que de la fonctionnalité, de l'expérience utilisateur, de l'environnement de déploiement, de la performance, de l'assurance, de la surveillance, des exigences éthiques et juridiques, entre autres considérations.

Réf.	Règle	Informations de mise en œuvre
SIA-04	La conception du SIA doit prendre en compte la sécurité de la chaîne d'approvisionnement lorsque le choix est fait de développer en interne ou d'utiliser des composants externes	La sécurité de la chaîne d'approvisionnement doit faire l'objet d'une analyse détaillée.

Cette analyse comprend les éléments de décision sur :

- le choix de construire un nouveau modèle, d'utiliser un modèle existant (avec ou sans *fine-tunning*) ou d'accéder à un modèle par le biais d'une API externe ;
- le choix de travailler avec un fournisseur de modèles externe ce qui implique une évaluation de la posture de sécurité de ce fournisseur ;
- dans le cas de l'utilisation d'une bibliothèque externe, il est nécessaire d'en effectuer une évaluation pour s'assurer, notamment, que la bibliothèque dispose de contrôles qui empêchent le système de charger des modèles non fiables sans s'exposer immédiatement à l'exécution d'un code arbitraire ;
- la nécessité de mettre en œuvre une analyse et une isolation, lors de l'importation de modèles tiers, qui doivent être traités comme du code tiers non fiable ;
- dans le cas de l'utilisation d'une API externe, la nécessité d'appliquer des contrôles appropriés aux données qui peuvent être envoyées à des services échappant au contrôle de l'entité, par exemple en demandant aux utilisateurs de se connecter et de confirmer avant d'envoyer des informations potentiellement sensibles ;
- la nécessité de procéder à des vérifications appropriées et à l'assainissement des données et des entrées, y compris lorsque sont incorporées dans le modèle des données relatives au retour d'information de l'utilisateur ou à l'apprentissage continu, en reconnaissant que les données de formation définissent le comportement du système.

Réf.	Règle	Informations de mise en œuvre
SIA-05	La conception du SIA doit prendre en compte l'intégration du développement du système logiciel d'IA dans les meilleures pratiques existantes en matière de développement et d'exploitation sécurisés	Tous les éléments du SIA sont écrits dans des environnements appropriés en utilisant des pratiques de codage et des langages qui réduisent ou éliminent les classes connues de vulnérabilités.

Si les composants d'IA doivent déclencher des actions, par exemple modifier des fichiers ou diriger les résultats vers des systèmes externes, il est nécessaire d'appliquer des restrictions appropriées aux actions possibles (cela inclut les dispositifs de sécurité externes, IA et non IA, si nécessaire).

Les décisions relatives à l'interaction avec l'utilisateur sont éclairées par les risques spécifiques à l'IA, comme par exemple :

- le système fournit aux utilisateurs des résultats utilisables sans révéler des niveaux de détail inutiles à un attaquant potentiel ;
- si nécessaire, le système fournit des garde-fous efficaces autour des résultats du modèle ;
- si l'entité propose une API à des clients ou collaborateurs, il est nécessaire d'appliquer des contrôles appropriés afin d'atténuer les attaques contre le SIA via l'API ;
- l'intégration par défaut des paramètres les plus sûrs dans le système ;
- la mise en œuvre des principes du moindre privilège pour limiter l'accès aux fonctionnalités du SIA ;
- la nécessité d'expliquer aux utilisateurs les capacités les plus risquées du SIA et valider leur acceptation de les utiliser ;
- la communication des cas d'utilisation interdits et si possible des solutions alternatives.

3.3.4 Sélection d'un modèle d'IA

Le choix d'un modèle d'IA implique de trouver un équilibre entre plusieurs exigences. Il s'agit notamment du choix de l'architecture du modèle, de la configuration, des données d'entraînement, de l'algorithme d'entraînement et des hyperparamètres. Les décisions s'appuient sur un modèle de menace et sont régulièrement réévaluées à mesure que la recherche sur la sécurité de l'IA progresse et que la compréhension de la menace évolue.

Lors du choix d'un modèle d'IA, il est nécessaire de prendre en compte les éléments suivants, sans toutefois s'y limiter :

- la complexité du modèle utilisé, c'est-à-dire l'architecture et le nombre de paramètres choisis. L'architecture et le nombre de paramètres choisis pour le modèle auront, entre autres, une incidence sur la quantité de données d'entraînement nécessaires et sur la robustesse du modèle aux modifications des données d'entrée lorsqu'il est utilisé ;
- l'adéquation du modèle au cas d'usage et/ou la possibilité de l'adapter aux besoins spécifiques (par exemple par le fine-tuning) ;
- la capacité d'aligner, d'interpréter et d'expliquer les résultats de du modèle (par exemple pour le débogage, l'audit ou la conformité réglementaire). Il peut être avantageux d'utiliser des modèles plus simples et plus transparents plutôt que des modèles complexes et de grande taille qui sont plus difficiles à interpréter ;
- les caractéristiques du ou des ensembles de données de formation, notamment la taille, l'intégrité, la qualité, la sensibilité, l'âge, la pertinence et la diversité ;

- l'intérêt d'utiliser des techniques de renforcement des modèles (comme l'entraînement contradictoire), de régularisation et/ou d'amélioration de la protection de la vie privée ;
- la provenance et les chaînes d'approvisionnement des composants, y compris le modèle ou le modèle de base, les données de formation et les outils associés.

3.4 Règles de sécurité applicables pour le développement

Cette section décrit les règles qui s'appliquent à la phase de développement du cycle de vie du SIA, y compris la sécurité de la chaîne d'approvisionnement, la documentation et la gestion des actifs et de la dette technique.

3.4.1 Sécurité de la chaîne d'approvisionnement

Réf.	Règle	Informations de mise en œuvre
SIA-06	La sécurité des chaînes d'approvisionnement en IA doit être évaluée contrôlée et documentée tout au long du cycle de vie du système.	Les fournisseurs doivent également adhérer aux mêmes normes que celles que l'entité applique à d'autres logiciels (signature du plan d'assurance sécurité).

Lorsqu'ils ne sont pas produits en interne, il est nécessaire d'acquérir et de maintenir des composants matériels et logiciels sécurisés et documentés (par exemple, les modèles, les données, les bibliothèques logicielles, les modules, les middlewares, les Frameworks et les API externes) auprès de développeurs commerciaux, open source et autres tiers vérifiés, afin de garantir une sécurité éprouvée des systèmes de l'entité.

3.4.2 Identifier, suivre et protéger les actifs

Les actifs liés à l'IA ont de la valeur pour l'entité.

En particulier :

- les modèles ;
- les données ;
- les commentaires des utilisateurs ;
- les *prompts* ;
- les logiciels ;
- la documentation ;
- les logs.

Réf.	Règle	Informations de mise en œuvre
SIA-07	Les actifs liés à l'IA doivent être traités comme des données sensibles.	Des mesures de sécurité doivent être mises en œuvre pour assurer leur confidentialité, leur intégrité et leur disponibilité.

La cartographie de ces actifs doit être réalisée pour chaque SIA.

Il existe un processus et des outils pour suivre, authentifier, contrôler les versions et sécuriser ces actifs.

La sauvegarde de ces actifs doit être réalisée et il est possible de restaurer un état connu en cas de compromission.

Il existe un processus et des contrôles pour protéger les données auxquelles un SIA peut accéder et le contenu généré par l'IA en fonction de sa sensibilité et de la sensibilité des données qui ont servi à le générer.

3.4.3 Documentation des données, modèles et prompts

Réf.	Règle	Informations de mise en œuvre
SIA-08	La création, l'exploitation et la gestion du cycle de vie de tous les modèles, ensembles de données et prompts systèmes doivent être documentés.	La documentation comprend des informations relatives à la sécurité, telles que les sources des données d'entraînement (y compris les données de fine-tuning et les feedbacks humains ou techniques), la portée et les limites prévues, les garde-fous, les hachages ou les signatures cryptographiques, la durée de conservation, la fréquence de révision et les modes de défaillance potentiels.

Parmi les structures utiles pour y parvenir, on peut citer les cartes de modèles, les cartes de données et les nomenclatures logicielles (SBOM).

La production d'une documentation complète favorise la transparence et la responsabilité.

3.4.4 Gérer la dette technique

Réf.	Règle	Informations de mise en œuvre
SIA-09	Il est obligatoire d'identifier, de suivre et de gérer la dette technique tout au long du cycle de vie d'un SIA.	La dette technique correspond aux décisions d'ingénierie qui ne respectent pas les meilleures pratiques pour obtenir des résultats à court terme, au détriment d'avantages à plus long terme.

La dette technique n'est pas mauvaise en soi, mais elle doit être gérée dès les premières étapes du développement.

En raison de cycles de développement rapides et d'un manque de protocoles et d'interfaces bien établis, les niveaux de dette technique d'un SIA sont susceptibles d'être élevés. Il est nécessaire de veiller à ce que les plans de cycle de vie (y compris les processus de mise hors service des systèmes d'IA) évaluent, reconnaissent et atténuent les risques pour les futurs systèmes similaires.

3.5 Règles de sécurité applicables pour le déploiement

Cette section décrit les règles qui s'appliquent à la phase de déploiement du cycle de développement des SIA, y compris la protection de l'infrastructure et des modèles contre les compromissions, les menaces ou les pertes, l'élaboration de processus de gestion des incidents et la diffusion responsable.

3.5.1 Sécuriser l'infrastructure

Réf.	Règle	Informations de mise en œuvre
SIA-10	La prise en compte de la sécurité de l'infrastructure utilisée doit être assurée à chaque étape du cycle de vie du système.	Il est nécessaire de mettre en œuvre des contrôles d'accès appropriés aux API, modèles et données, ainsi qu'à leurs pipelines d'entraînement et de traitement, tant au niveau de la recherche et du développement que du déploiement. Cela inclut une séparation appropriée des environnements contenant du code ou des données sensibles.

Ces mesures permettent notamment de réduire le risque d'attaques visant à voler un modèle ou à nuire à ses performances.

3.5.2 Protéger les modèles dans la continuité

Les attaquants peuvent être en mesure de reconstruire la fonctionnalité d'un modèle ou les données sur lesquelles il a été formé, en accédant à un modèle directement (en acquérant les poids du modèle) ou indirectement (en interrogeant le modèle par l'intermédiaire d'une application ou d'un service). Les attaquants peuvent également altérer les modèles, les données ou les prompts pendant ou après l'apprentissage, ce qui rend le résultat non fiable.

Réf.	Règle	Informations de mise en œuvre
SIA-11	Le modèle et les données doivent être protégés de l'accès direct et indirect par la mise en œuvre de contrôles sur l'interface d'interrogation afin de détecter et empêcher les tentatives d'accès, de modification et d'exfiltration d'informations confidentielles.	NA

Réf.	Règle	Informations de mise en œuvre
SIA-12	Pour que les systèmes consommateurs puissent valider les modèles, il est nécessaire de calculer et partager les hachages cryptographiques et/ou les signatures des fichiers de modèles (par exemple, les poids des modèles) et des ensembles de données (y compris les points de contrôle) dès que le modèle est entraîné.	Pour ce faire, seuls les algorithmes de hachage spécifiés dans la PSSIG sont autorisés.

L'atténuation du risque de perte de confidentialité dépend du cas d'utilisation et du modèle de menace. Certaines applications, par exemple celles qui impliquent des données très sensibles, peuvent nécessiter des modèles de garanties qu'il peut être difficile ou coûteux d'appliquer. Le cas échéant, des technologies d'amélioration de la confidentialité (telles que la confidentialité différentielle ou le chiffrement homomorphique) peuvent être utilisées pour explorer ou garantir les niveaux de risque associés à l'accès des consommateurs, des utilisateurs et des attaquants aux modèles et aux résultats.

3.5.3 Élaborer des procédures de gestion des incidents

Réf.	Règle	Informations de mise en œuvre
SIA-13	L'inévitabilité des incidents de sécurité affectant les SIA doit être prise en compte dans les plans de réponse aux incidents, d'escalade et de remédiation.	Ces plans tiennent compte de différents scénarios et sont régulièrement réévalués au fur et à mesure de l'évolution du système et de la recherche en général.

Les ressources numériques critiques du SIA sont stockées dans des sauvegardes hors ligne. Les intervenants sont formés à l'évaluation et au traitement des incidents liés à l'IA.

3.5.4 Diffuser l'IA de manière responsable

Réf.	Règle	Informations de mise en œuvre
SIA-14	L'entité ne publie des modèles, des applications ou des systèmes qu'après les avoir soumis à une évaluation de sécurité appropriée et efficace, telle que l'analyse comparative et le *red teaming*.	L'entité indique clairement à ses utilisateurs les limites connues ou les modes de défaillance potentiels.

L'entité applique des contrôles pour empêcher l'utilisation ou le déploiement de son SIA à des fins malveillantes.

L'entité fournit aux utilisateurs des conseils sur l'utilisation appropriée de son modèle ou de son SIA, notamment en soulignant les limites et les modes de défaillance potentiels.

L'entité indique clairement aux utilisateurs les aspects de la sécurité dont ils sont responsables.

L'entité est transparente sur l'endroit et la manière dont leurs données peuvent être utilisées, consultées ou stockées (par exemple, si elles sont utilisées pour le recyclage du modèle ou examinées par des employés ou des partenaires).

3.6 Règles de sécurité applicables pour l'exploitation et la maintenance

Cette section décrit les règles qui s'appliquent à l'étape de l'exploitation et de la maintenance sécurisées du cycle de vie du développement d'un SIA. Elle décrit notamment les règles à mettre en œuvre une fois qu'un système a été déployé, y compris la journalisation et la surveillance, la gestion des mises à jour et le partage d'informations.

3.6.1 Surveillance du comportement du système

Réf.	Règle	Informations de mise en œuvre
SIA-15	Les résultats et les performances du modèle et du système doivent être mesurés de manière à pouvoir observer les changements de comportement soudains ou progressifs qui affectent la sécurité.	Cette surveillance prend en compte et identifie les intrusions et les compromissions potentielles, ainsi que la dérive naturelle des données.

3.6.2 Surveillance des entrées du système

Réf.	Règle	Informations de mise en œuvre
SIA-16	Conformément aux exigences en matière de protection de la vie privée et des données à caractère personnelle, les entrées dans le système (telles que les demandes d'inférence, les requêtes ou les prompts) sont loggées et surveillées.	Ces logs doivent permettre l'audit, l'analyse et la correction en cas de compromission ou d'utilisation abusive.

3.6.3 Sécuriser les mises à jour

Réf.	Règle	Informations de mise en œuvre
SIA-17	Des mises à jour automatisées sont incluses par défaut dans chaque produit et des procédures de mise à jour sécurisées et modulaires sont utilisées pour les distribuer.	Ces processus de mise à jour (y compris les régimes de test et d'évaluation) tiennent compte du fait que les modifications apportées aux données, aux modèles ou aux prompts peuvent entraîner des changements dans le comportement du système.

4 Gestion du risque lié à l'IA

Comme pour les logiciels traditionnels, les risques liés à la technologie basée sur l'intelligence artificielle peuvent avoir des impacts sur l'entité. Les systèmes d'IA comportent également un ensemble de risques qui ne sont pas entièrement pris en compte par les méthodes courantes de traitement du risque.

Par rapport aux systèmes d'informations traditionnels, les risques spécifiques à l'IA sont les suivants :

- les données utilisées pour créer un système d'IA peuvent ne pas être une représentation vraie ou appropriée du contexte ou de l'utilisation prévue du système d'IA, et la représentation du terrain peut soit ne pas exister, soit ne pas être disponible. De plus, des biais préjudiciables et d'autres problèmes de qualité des données peuvent affecter la fiabilité du système d'IA, ce qui peut entraîner des impacts négatifs ;
- dépendance au système d'IA et dépendance aux données pour les tâches d'entraînement, combinées à un volume et une complexité accrus généralement associés à ces données ;
- des changements intentionnels ou non pendant l'entraînement peuvent altérer fondamentalement les performances du système d'IA ;
- les ensembles de données utilisés pour entraîner les systèmes d'IA peuvent se détacher de leur contexte d'origine ou devenir obsolètes par rapport au contexte de déploiement ;
- l'échelle et la complexité des systèmes d'IA (de nombreux systèmes contiennent des millions, voire des milliards de points de décision) hébergés dans des applications logicielles plus traditionnelles ;
- l'utilisation de modèles pré-entraînés capables de faire progresser la recherche et d'améliorer les performances peut également augmenter les niveaux d'incertitude statistique et entraîner des problèmes de gestion des biais, de validité scientifique et de reproductibilité ;
- degré de difficulté plus élevé dans la prévision des modes de défaillance pour les propriétés émergentes des modèles pré-entraînés à grande échelle ;
- risque de perte de confidentialité dû à la capacité améliorée d'agrégation de données pour les systèmes d'IA ;
- les systèmes d'IA peuvent nécessiter une maintenance plus fréquente et des déclencheurs pour effectuer une maintenance corrective en raison d'une dérive des données ou du modèle ;
- opacité accrue et soucis de reproductibilité ;
- normes de tests logiciels insuffisantes et incapacité à documenter les pratiques basées sur l'IA selon les normes attendues des logiciels d'ingénierie traditionnelle, sauf dans les cas les plus simples ;
- difficulté à effectuer régulièrement des tests de logiciels basés sur l'IA ou à déterminer ce qu'il faut tester, car les systèmes d'IA ne sont pas soumis aux mêmes contrôles que le développement de code traditionnel ;
- coûts de calcul pour le développement de systèmes d'IA ;
- incapacité de prédire ou de détecter les effets secondaires des systèmes basés sur l'IA au-delà des mesures statistiques.

Les considérations et approches de gestion des risques liés à la confidentialité et à la

cybersécurité sont applicables à la conception, au développement, au déploiement, à l'évaluation et à l'utilisation de systèmes d'IA.

L'IA et les technologies associées sont soumis à une innovation rapide. Les avancées technologiques doivent être surveillées et déployées pour tirer parti de ces évolutions et œuvrer à une évolution de l'IA à la fois sécurisée et responsable.

Réf.	Règle	Informations de mise en œuvre
SIA-18	Chaque système d'IA développé et/ou mis en œuvre au sein de l'entité doit faire l'objet d'une analyse de risques adaptée.	La gestion des risques lié au développement et à l'utilisation des SIA doit s'appuyer sur le Framework du NIST: Artificial Intelligence Risk Management Framework[2]

Ce Framework fournit des résultats et des actions qui permettent le dialogue, la compréhension et les activités pour gérer les risques liés à l'IA et développer de manière responsable des systèmes d'IA sécurisés.

Ce Framework s'appuie sur quatre fonctions :

- gouverner ;
- cartographier ;
- mesurer ;
- gérer.

Chacune de ces fonctions de haut niveau est décomposée en catégories et sous-catégories. Les catégories et sous-catégories sont subdivisées en actions et résultats spécifiques. Les actions ne constituent pas une liste de contrôle, ni nécessairement un ensemble ordonné d'étapes.

La gestion des risques doit être continue, opportune et effectuée tout au long du cycle de vie du système d'IA. Les fonctions principales du Framework doivent être exercées d'une manière qui reflète des perspectives diverses et multidisciplinaires, incluant potentiellement les points de vue des acteurs de l'IA extérieurs à l'organisation.

4.1 La fonction « gouverner »

La fonction "gouverner" :

- développe et met en œuvre une culture de gestion des risques au sein des entités qui conçoivent, développent, déploient, évaluent ou acquièrent des systèmes d'IA ;
- décrit les processus, les documents et les schémas organisationnels qui anticipent, identifient et gèrent les risques qu'un système peut poser, y compris pour les utilisateurs et les autres acteurs de l'entité, ainsi que les procédures permettant d'atteindre ces résultats ;
- intègre des processus pour évaluer les impacts potentiels;
- fournit une structure grâce à laquelle les fonctions de gestion des risques liées à l'IA peuvent s'aligner sur les principes organisationnels, les politiques et les priorités stratégiques ;
- relie les aspects techniques de la conception et du développement de systèmes d'IA aux valeurs et principes organisationnels, et permet des pratiques et des compétences organisationnelles pour les personnes impliquées dans l'acquisition, l'entraînement, le déploiement et la surveillance de ces systèmes ;

[2] https://www.nist.gov/itl/ai-risk-management-framework

- traite du cycle de vie complet du produit et des processus associés, y compris les questions juridiques et autres concernant l'utilisation de systèmes et de données logiciels ou matériels tiers.

"Gouverner" est une fonction transversale qui est imprégnée dans toute la gestion des risques liés à l'IA et qui active les autres fonctions du processus. Les aspects de gouvernance, notamment ceux liés à la conformité ou à l'évaluation, doivent être intégrés dans chacune des autres fonctions. L'attention portée à la gouvernance est une exigence continue et intrinsèque pour une gestion efficace des risques liés à l'IA tout au long de la durée de vie d'un système d'IA et du management de l'entité.

Après avoir mis en place les structures, les systèmes, les processus et les équipes décrits dans la fonction "gouverner", les entités devraient bénéficier d'une culture axée sur les objectifs et axée sur la compréhension et la gestion des risques. Il incombe aux entités de continuer à exécuter la fonction "gouverner" à mesure que les connaissances, les cultures et les besoins ou attentes des acteurs de l'IA évoluent au fil du temps.

4.2 La fonction « cartographier »

La fonction "cartographier" établit le contexte pour encadrer les risques liés à un système d'IA. Le cycle de vie de l'IA comprend de nombreuses activités interdépendantes impliquant un ensemble diversifié d'acteurs. Dans la pratique, les acteurs de l'IA en charge d'une partie du processus n'ont souvent pas une visibilité ou un contrôle total sur les autres parties et leurs contextes associés. Les interdépendances entre ces activités et entre les acteurs concernés de l'IA peuvent rendre difficile une anticipation fiable des impacts des systèmes d'IA. Par exemple, les décisions précoces visant à identifier les buts et les objectifs d'un système d'IA peuvent modifier son comportement et ses capacités, et la dynamique du contexte de déploiement (comme les utilisateurs finaux ou les individus concernés) peut façonner les impacts des décisions du système d'IA.

Cette complexité et ces différents niveaux de visibilité peuvent introduire de l'incertitude dans les pratiques de gestion des risques. Anticiper, évaluer et traiter les sources potentielles de risque peut atténuer cette incertitude et améliorer l'intégrité du processus décisionnel.

Les informations recueillies lors de l'exécution de la fonction "cartographier" permettent de prévenir les risques et éclairent les décisions concernant des processus tels que la gestion des modèles, ainsi qu'une décision initiale sur l'opportunité ou la nécessité d'une solution d'IA. Les résultats de la fonction "cartographier" constituent la base des fonctions "mesurer" et "gérer". Sans connaissance contextuelle et conscience des risques dans les contextes identifiés, la gestion des risques est difficile à réaliser. La fonction "cartographier" vise à améliorer la capacité d'une organisation à identifier les risques et les facteurs contributifs plus larges.

Recueillir des perspectives aussi larges peut aider les entités à prévenir de manière proactive les risques et à développer des systèmes d'IA plus fiables en leur permettant de :

- améliorer leur capacité à comprendre les contextes ;
- vérifier leurs hypothèses sur le contexte d'utilisation ;
- reconnaître quand les systèmes ne sont pas fonctionnels dans ou hors de leur contexte prévu ;
- identifier les utilisations positives et bénéfiques de leurs systèmes d'IA existants ;
- améliorer la compréhension des limites des processus d'IA ;
- identifier les contraintes dans les applications du monde réel qui peuvent conduire à des impacts négatifs ;
- identifier les impacts négatifs connus et prévisibles liés à l'utilisation prévue des systèmes d'IA ;

- anticiper les risques liés à l'utilisation des systèmes d'IA au-delà de leur utilisation prévue.

Après avoir terminé la fonction "cartographier", les entités doivent avoir des connaissances contextuelles suffisantes sur les impacts du système d'IA pour éclairer la décision de l'opportunité de concevoir, de développer, de déployer ou d'arrêter un système d'IA.

Si une décision est prise de poursuivre, les entités doivent utiliser les fonctions "mesurer" et "gérer" ainsi que les politiques et procédures mises en place dans la fonction "gouverner" pour contribuer aux efforts de gestion des risques liés à l'IA.

Il incombe aux entités de continuer à appliquer la fonction "cartographier" aux systèmes d'IA à mesure que le contexte, les capacités, les risques, les avantages et les impacts potentiels évoluent au fil du temps.

4.3 La fonction « mesurer »

La fonction "mesurer" utilise des outils, des techniques et des méthodologies quantitatives, qualitatives ou mixtes pour analyser, évaluer, comparer et surveiller les risques liés à l'IA et les impacts associés. Il utilise les connaissances pertinentes pour les risques IA identifiés dans la fonction "cartographier" et informe la fonction "gérer". Les systèmes d'IA doivent être testés avant leur déploiement et régulièrement pendant leur fonctionnement. Les mesures des risques de l'IA incluent la documentation des aspects de la fonctionnalité et de la fiabilité des systèmes.

Réf.	Règle	Informations de mise en œuvre
SIA-19	La liste des menaces utilisée dans le cadre de la fonction "mesurer" est la base Atlas du MITRE.	ATLAS[3] (Adversarial Threat Landscape for Artificial-Intelligence Systems) est une base de connaissances sur les tactiques et techniques adverses contre les systèmes d'IA. Elle est basée sur des observations d'attaques réelles et des démonstrations réalistes des red teams.

La mesure des risques liés à l'IA comprend le suivi des mesures relatives aux caractéristiques fiables, à l'impact social et aux configurations humain-IA. Les processus développés ou adoptés dans la fonction "mesurer" doivent inclure des méthodologies rigoureuses de tests de logiciels et d'évaluation des performances avec des mesures d'incertitude associées, des comparaisons avec des références de performances, ainsi que des rapports et une documentation formalisée des résultats. Les processus d'examen indépendant peuvent améliorer l'efficacité des tests et atténuer les préjugés internes et les conflits d'intérêts potentiels.

Lorsque des compromis entre les caractéristiques fiables surviennent, la mesure fournit une base traçable pour éclairer les décisions de gestion. Les options peuvent inclure le recalibrage, l'atténuation des impacts ou le retrait du système de la conception, du développement, de la production ou de l'utilisation, ainsi qu'une gamme de contrôles de compensation, de détection, de dissuasion, de directive et de récupération.

Après avoir terminé la fonction "mesurer", les processus de test, d'évaluation, de vérification et de validation objectifs, reproductibles ou évolutifs, y compris les métriques, les méthodes et les méthodologies, sont en place, suivis et documentés. Les mesures et les méthodologies de mesure doivent respecter les normes scientifiques, juridiques et éthiques et être appliquées dans le cadre d'un processus ouvert et transparent.

[3] https://atlas.mitre.org/matrices/ATLAS

4.4 La fonction « gérer »

La fonction “gérer” consiste à allouer des ressources de risque aux risques cartographiés et mesurés de manière régulière et telle que définie par la fonction “gouverner”. Le traitement des risques comprend des plans pour réagir, remédier et communiquer sur les incidents de sécurité.

Les informations contextuelles tirées de la consultation d’experts et des contributions des acteurs concernés de l’IA – établies dans “gouverner” et réalisées dans le “cartographier” – sont utilisées dans cette fonction pour réduire la probabilité de pannes du système et d’impacts négatifs. Les pratiques de documentation systématique établies dans “gouverner” et utilisées dans “cartographier” et “mesurer” renforcent les efforts de gestion des risques liés à l’IA et augmentent la transparence et la responsabilité. Des processus d’évaluation des risques émergents sont en place, ainsi que des mécanismes d’amélioration continue.
Après avoir terminé la fonction “gérer”, des plans de priorisation des risques ainsi qu’un suivi et une amélioration réguliers sont en place. Les entités disposent alors d’une capacité accrue à gérer les risques liés aux systèmes d’IA déployés et à allouer des ressources de gestion des risques en fonction des risques évalués et hiérarchisés.

5 Cas particulier des IA génératives

La mise en œuvre d'un système d'IA générative peut se décomposer en trois phases cycliques :

- une première phase d'entraînement du modèle d'IA à partir de données spécifiquement choisies ;
- une phase d'intégration et de déploiement ;
- une phase de production opérationnelle dans laquelle les utilisateurs peuvent accéder au modèle d'IA entraîné, par l'intermédiaire du système d'IA.

Ces trois phases doivent chacune faire l'objet de mesures de sécurisation spécifiques, qui dépendent en partie du choix de sous-traitance retenu pour chaque composante (hébergement, entraînement du modèle, tests de performance, etc.) ainsi que de la sensibilité des données utilisées à chaque étape et de la criticité du système d'IA dans sa finalité.

En complément des menaces classiques inhérentes à tout système d'information, un système d'IA générative peut-être soumis à des attaques spécifiques visant par exemple à perturber le bon fonctionnement de celui-ci (attaques adverses) ou bien à exfiltrer des données traitées par celui-ci.

La question de la protection des données, notamment des données d'entraînement, est donc un enjeu essentiel d'un système d'IA générative, avec comme corollaire la problématique du besoin d'en connaître des utilisateurs lorsqu'ils interrogent le modèle. En effet, ce dernier est conçu pour générer une réponse à partir de l'ensemble des données auxquelles il a eu accès lors de l'entraînement, ainsi que des données additionnelles qui peuvent être issues de sources internes sensibles.

L'utilisation d'un système d'IA générative doit répondre à des besoins de confidentialité mais également des besoins en intégrité et en disponibilité. Les interactions du système d'IA avec d'autres applications ou composants du SI doivent ainsi être sécurisées, limitées au strict besoin opérationnel, et doivent pouvoir être contrôlées par un humain lorsque celles-ci sont critiques pour entité.

5.1 Scénarios d'attaques sur l'IA générative

Un système d'IA générative est une application métier standard, qui doit disposer du même socle de sécurité que toute autre application métier de l'entité. Toutefois, en complément de ce socle de sécurité, l'entité doit prendre en compte des menaces spécifiques à un système d'IA générative.

Ces menaces peuvent être déclinées en trois grandes catégories d'attaques :

- attaques par manipulation: ces attaques consistent à détourner le comportement du système d'IA en production au moyen de requêtes malveillantes. Elles peuvent provoquer des réponses inattendues, des actions dangereuses ou un déni de service ;
- attaques par infection: ces attaques consistent à contaminer un système d'IA lors de sa phase d'entraînement, en altérant les données d'entraînement ou en insérant une porte dérobée ;
- attaques par exfiltration: ces attaques consistent à dérober des informations sur le système d'IA en production, comme les données ayant servi à entraîner le modèle, les données des utilisateurs ou bien des données internes du modèle (paramètres).

Dans le contexte de l'IA générative, ces attaques peuvent porter atteinte aux besoins de sécurité suivants :

- confidentialité : l'objectif est de protéger un système d'IA contre la fuite d'informations considérées comme sensibles: jeux de données d'entraînement, requêtes des utilisateurs, paramètres des modèles, données additionnelles internes, etc. ;
- intégrité : l'objectif est de protéger un système d'IA contre une modification non prévue de son comportement. L'intégrité peut concerner directement le modèle (paramètres) ou bien viser les jeux de données d'entraînement (empoisonnement) ou encore les composants techniques permettant le bon fonctionnement du système d'IA ;
- disponibilité : l'objectif est de protéger un système d'IA contre des dénis de service ou des actions visant à dégrader ses performances (requêtes malveillantes) ;
- traçabilité : l'objectif est de garantir d'une part l'explicabilité et l'imputabilité des actions réalisées sur un système d'IA. Ces éléments peuvent faciliter le travail d'investigation et de remédiation après un incident de sécurité.

5.2 Génération de code source assistée par l'IA

Les outils d'IA générative peuvent être spécialisés et spécifiquement entraînés pour générer du code source dans plusieurs langages de programmation.

Ces moyens peuvent permettre aux développeurs un gain de temps mais comportent aussi des risques sur la qualité du code (introduction de vulnérabilités) ou d'insertion de porte dérobée dans le cas où un attaquant aurait compromis le modèle.

Il est donc important de faire preuve de vigilance sur le code source généré par IA.

Réf.	Règle	Informations de mise en œuvre
SIA-20	Le code source généré par une IA doit être contrôlé systématiquement et faire l'objet de mesures de sécurité afin de vérifier son innocuité.	* il est interdit d'exécuter automatiquement un code source généré par IA dans l'environnement de développement; * le commit automatique de code source généré par IA dans les dépôts est interdit; * il est obligatoire d'intégrer un outil d'assainissement de code source généré par IA dans l'environnement de développement; * il est obligatoire de vérifier l'innocuité des bibliothèques référencées dans le résultat du code source généré par IA; * il est obligatoire de faire contrôler régulièrement par un humain la qualité du code source généré à partir de requêtes types.

La génération de code source par IA pour des modules critiques d'applications doit être limitée.

Réf.	Règle	Informations de mise en œuvre
SIA-21	Il est interdit de recourir à un outil d'IA générative pour générer des blocs de code source destinés aux modules applicatifs suivants:* les modules de cryptographie (authentification, chiffrement, signature, etc.); * les modules de gestion des droits d'accès des utilisateurs et administrateurs; * les modules de traitement de données sensibles.	N/A

Les développeurs doivent être sensibilisés sur les risques liés au code source généré par IA.

Réf.	Règle	Informations de mise en œuvre
SIA-22	Il est obligatoire d'effectuer des campagnes de sensibilisation sur les risques liés à l'utilisation de code source généré par IA.	En complément, les développeurs doivent également être formés sur les outils d'IA pour l'optimisation de leurs requêtes (*prompt engineering*) afin d'améliorer la qualité et la sécurité du code généré.

5.3 Utilisation de solutions d'IA générative tierces

En raison de leur facilité d'usage et des gains métiers potentiels, il est possible de recourir à des outils d'IA générative disponibles sur Internet pour traiter des données métier, par exemple pour la traduction de textes.

Le fait d'envoyer des informations (texte, images, documents) à un service d'IA générative exposé sur Internet revient à déposer ces mêmes informations sur un espace de stockage appartenant à ce service.

Dans ce cas, le cloisonnement entre les clients ainsi que la protection en confidentialité des données envoyées au système d'IA sur Internet ne sont pas maîtrisés. De plus, dans la majorité des offres, les données envoyées au service sont collectées et utilisées par le prestataire à des fins d'optimisation des modèles.

Proposition de critères de sécurité de confidentialité de données :

Niveau	Intitulé	Description
C0	Public	La donnée est publique.
C1	Interne	La donnée ne doit être accessible qu'aux collaborateurs de l'entité et aux partenaires.
C2	Confidentiel	La donnée ne doit être accessible qu'aux collaborateurs de l'entité concernés.
C3	Secret	La donnée ne doit être accessible qu'aux personnes identifiées et ayant le besoin d'en connaître.

Les règles d'utilisation d'outils d'IA générative, en fonction de la classification des données proposée ci-dessus, est détaillé dans le tableau suivant :

Mode d'accès	Classification	Autorisation
Datacenter Entité	C0, C1, C2	Autorisé
Datacenter Entité	C3	Interdit
SaaS en Europe	C0, C1	Autorisé
SaaS en Europe	C2, C3	Interdit
Saas hors Europe	C0	Autorisé
Saas hors Europe	C1, C2, C3	Interdit

Tour Eria, 5 rue Bellini
92821 Puteaux Cedex
France
✆ +33 1 53 25 08 80
clusif@clusif.fr

clusif.fr