RÉPUBLIQUE FRANÇAISE
Liberté
Égalité
Fraternité

Cybersécurité des systèmes d'intelligence artificielle : définition de l'ANSSI

Dans le cadre des travaux menés par l'ANSSI sur l'intelligence artificielle (IA), il apparait nécessaire de définir clairement la notion de cybersécurité d'un système d'IA afin de circonscrire les vulnérabilités d'un système IA sur le plan de la cybersécurité et les mesures à mettre en place pour y remédier. L'ANSSI propose ainsi une définition de la cybersécurité des systèmes d'intelligence artificielle et quelques précisions techniques sur cette définition.

La cybersécurité d'un système d'IA (SIA) peut être définie de manière analogue à celle d'un système classique, comme l'étude des vulnérabilités du système et des mesures visant à y remédier afin de garantir, face à un attaquant, les propriétés de sécurité : confidentialité, intégrité, disponibilité. Les vulnérabilités, leurs impacts et leur remédiation peuvent se situer à trois niveaux : les composants individuels du système, son architecture et son intégration au sein d'un SI. La spécificité d'un SIA tient à ce qu'au moins l'un de ses composants implémente un modèle d'IA obtenu par apprentissage statistique, à partir de données d'entraînement. La cybersécurité d'un SIA ne prend pas en compte les risques liés aux aspects métiers, comme la fiabilité du résultat ou la performance du modèle (résultats erronés, hallucinations, etc.). Ces risques relèvent généralement de la sureté de fonctionnement (« *safety* »).

Précisions sur la définition de cybersécurité des systèmes d'intelligence artificielle :

Un système d'IA peut être défini comme **un ensemble de composants matériels et logiciels** agencés selon une **architecture** afin de remplir une fonction donnée. La particularité des systèmes d'IA est qu'au moins un de leurs composants implémente un **modèle** issu d'un processus d'**apprentissage statistique**. L'apprentissage statistique (ou apprentissage automatique) désigne l'application d'un **algorithme d'apprentissage** à des **données d'entrainement** pour produire un modèle. Enfin, un système d'IA peut être **intégré** ou **interconnecté** à un système d'information (SI) plus large.

La cybersécurité d'un système d'IA peut alors être définie de manière analogue à celle d'un système classique, comme l'étude des vulnérabilités du système et des mesures visant à y remédier afin de garantir, face à un attaquant, les propriétés de sécurité : confidentialité, intégrité, disponibilité. Les vulnérabilités, leur impact et leur remédiation peuvent se situer à trois niveaux : les **composants individuels du système**, son **architecture** et **son intégration au sein d'un SI**. L'attaquant peut disposer de moyens plus ou moins étendus : accès au système partiel (à travers une API) ou complet (y compris au niveau matériel), connaissance du modèle partielle (architecture, entraînement initialisé à partir d'un modèle connu) ou complète, accès aux sorties et états intermédiaires du modèle pour certaines entrées, accès en lecture ou écriture à une partie des données d'entraînement.

Du point de vue de la sécurité, la particularité des systèmes d'IA est la présence en leur sein de composants implémentant des modèles d'IA. Les conséquences de cette propriété sont les suivantes :

- Au niveau des composants, en plus des vulnérabilités usuelles des composants matériels et logiciels classiques, il est nécessaire de **prendre en compte des vulnérabilités spécifiques aux modèles d'IA**. L'impact de ces vulnérabilités et les moyens d'y remédier doivent être considérés non seulement à l'échelle des composants individuels, mais également du point de vue de l'architecture du système d'IA et de son intégration au sein d'un SI.
- Les modèles d'IA étant issus d'un processus d'apprentissage statistique, **leur phase d'entraînement doit être intégrée à l'analyse** :
    - Certaines propriétés de sécurité doivent être garanties durant la phase d'entrainement elle-même, notamment pour protéger les données utilisées qui peuvent être sensibles (données métier, données personnelles, etc.) ;
    - Pour un système déployé en production, la phase d'entraînement fait partie de la chaine d'approvisionnement. Ainsi, une compromission des données d'entrainement implique une compromission du modèle entrainé, et donc du système d'IA en production.

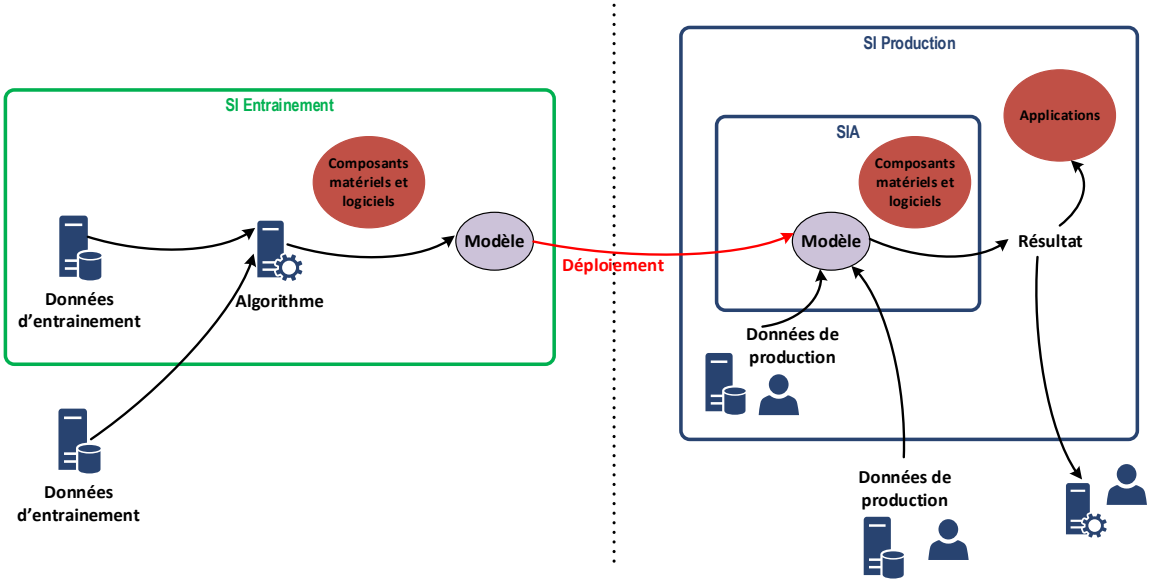


Figure 1 : schémas d'architecture haut-niveau d'un système d'IA et des éléments associés

Au-delà des menaces classiques en matière de SSI (altération du SI, etc.) que l'ANSSI doit couvrir, les vulnérabilités spécifiques aux modèles d'IA peuvent mettre en péril les propriétés de sécurité suivantes :

- Confidentialité et intégrité des données d'entrainement en entrée de l'algorithme ;
- Confidentialité des données d'entrainement face à un attaquant connaissant les paramètres du modèle ou ayant un accès au modèle ;
- Confidentialité, intégrité et disponibilité du modèle en production ;

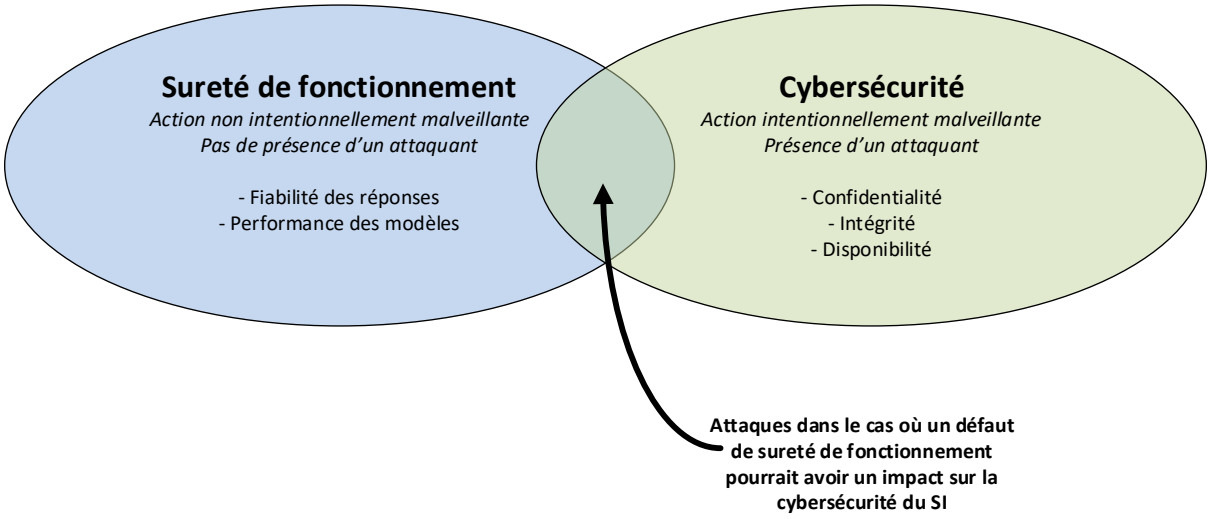
- Confidentialité des données de production et du résultat ;
- Confidentialité, intégrité et disponibilité des applications interconnectées au SIA.

Distinction entre la *safety* et la cybersécurité

Dans la définition présentée ici, la cybersécurité de l'IA n'inclut pas directement les risques liés aux aspects métiers, comme la fiabilité des réponses ou la performance du modèle (résultats erronés, « hallucinations », etc.).

Ces risques relèvent de la sureté de fonctionnement (« *safety* ») et non de la cybersécurité. Ils pourraient néanmoins, dans certains cas spécifiques, avoir un impact indirect sur la cybersécurité. Ces risques seront donc traités, mais uniquement sous cet angle.

En outre, la sûreté de fonctionnement pourrait être un enjeu pour l'ANSSI dans le cas où le métier est précisément la cybersécurité (ex : sonde de détection d'intrusion), mais nous basculons alors dans ce cas précis sur un autre axe de travail : la cybersécurité **par l'IA**.



Exemple 1 : Un SIA permet de catégoriser, d'un point de vue métier, la sensibilité des documents bureautiques d'une entité. En cas de défaillance du modèle, des règles de protection inadéquates pourraient être appliquées sur certains documents dans l'espace documentaire, et il y aurait alors une potentielle atteinte en confidentialité sur le système d'information de l'entité.

Exemple 2 : Un SIA fonctionne comme un assistant au développement de logiciels pour une entité et génère du code source à destination des développeurs. En cas de défaillance du modèle, du code vulnérable pourrait être généré et être intégré dans des applications métiers en production. Ces vulnérabilités pourraient ensuite être exploitées par un attaquant.

Exemple 3 : Un SIA permet de générer automatiquement des règles de pare-feu à destination d'un produit donné, à partir de requêtes réalisées en langage naturel. En cas de défaillance du modèle, des règles inadaptées ou trop laxistes pourraient être intégrées directement dans le système de gestion des pare-feux de l'entité. Un attaquant pourrait alors profiter de ce défaut de configuration pour pénétrer dans le système d'information.