Sécuriser le traitement

05 avril 2022

Analyser les risques et empêcher les failles et attaques.

Identifier les modèles d'attaque et s'en prémunir

De nombreuses recherches récentes prouvent que les systèmes d'IA peuvent être attaqués ou détournés de leur finalité. Ces modèles d'attaque émergents doivent être connus du fournisseur et de l'utilisateur du système d'IA. Étant donné le caractère inédit de ces attaques, le principe de précaution doit être appliqué autant que possible.

☐ Une analyse de risques a-t-elle été conduite ?

☐ Prend-t-elle en compte les modèles d'attaque spécifiques aux algorithmes d'IA ?

☐ Les modèles d'attaque afférents aux méthodes d'IA utilisées sont-ils connus du fournisseur et de l'utilisateur ? Une étude de la littérature scientifique ainsi qu'une veille sont-elles réalisées ?

☐ Des mesures ont-elles été prises afin de se prémunir contre des attaques par empoisonnement (*data poisoning*), par exemples contradictoires (*adversarial attack*), par exfiltration de modèles (*model evasion*) ou par attaque par inférence d'appartenance (*membership inference*) ?

Lesquelles ?

Journaliser pour mieux superviser

En maintenant une journalisation tout au long de la chaîne, l'utilisateur du système d'IA doit être en mesure d'identifier et d'expliquer les comportements anormaux.

☐ Une journalisation des actions est-elle mise en place ? Couvre-t-elle les modifications apportées au système d'IA, logicielles ou matérielles, les requêtes qui lui sont envoyées, les données d'entrées et de sortie du système ?

☐ Une analyse automatique des journaux existe-t-elle ?

☐ Permettrait-elle d'identifier des tentatives d'attaques de type inférence d'appartenance ou empoisonnement de modèle (dans le cas d'un apprentissage continu notamment) ?

D'autres mesures mises en place afin de contrôler, en aval, la qualité des sorties du système existent-elles et permettent-elle de se prémunir contre des attaques ?

Maîtriser les accès

Que le système d'IA soit utilisé dans un système physique ou logiciel, les accès pouvant permettre une modification du système doivent être réduits et encadrés.

☐ Les modifications apportées au système d'IA sont-elles soumises à un protocole particulier ?

Lequel ?

☐ Différents niveaux d'habilitation sont-ils prévus afin de contrôler les modifications apportées au système et de limiter les accès ?

☐ Les modifications du code sont-elles versionnées ?

☐ Permettent-elles un retour rapide à la dernière version fonctionnelle ?

Sécuriser toutes les étapes du traitement

Sans préjudice des mesures de sécurité propres au système d'IA utilisé, les mesures de sécurité habituelles pour un traitement utilisant des données personnelles ou ayant des conséquences pour les personnes doivent être mises en place.

Quelles mesures de sécurité du système ont été prises ?

☐ Une redondance est-elle prévue afin de garantir la disponibilité du système ?

☐ Un serveur secondaire pourrait-il assurer le traitement si le système principal était rendu non fonctionnel ?

☐ Un audit (interne ou externe) a-t-il été mené ?

Si oui, par quel organisme ?

Quelles techniques et méthodologies ont été utilisées pour éprouver le traitement ?

☐ Un système de management des risques a-t-il été mis en place ?

☐ Les recommandations du guide de sécurité de la CNIL ont-elles été appliquées ?

Examiner la nature des modèles

Dans certains cas, les paramètres et modèles issus de l'apprentissage du modèle peuvent être considérés comme des données personnelles. Les mesures de sécurité doivent alors être adaptées.

☐ Si le modèle a été entraîné à l'aide de données personnelles, une étude sur les risques de réidentification/d'inférence d'appartenance a-t-elle été menée sur les agrégats issus de l'apprentissage ?

☐ Quelles méthodes sont utilisées pour limiter ces risques ?

- [ ] Les paramètres du modèle sont-ils alors considérés comme à caractère personnel ?
- [ ] Le niveau de sécurité appliqué aux paramètres du modèle est-il adapté et suffisant au regard de l'obligation de sécurité imposé par le RGPD ?
- [ ] Une fois l'entraînement terminé, les paramètres de l'algorithme contiennent-ils des échantillons des données d'entraînement (cas de certains algorithmes de partitionnement (ou *clustering*) qui identifient, enregistrent puis s'appuient sur certaines données clés de l'ensemble d'entraînement) ?
- [ ] Dans ce cas, les paramètres de l'algorithme font-ils l'objet des mesures de sécurité applicables aux données personnelles ?

Imprimer cette check-list

Aucune information n'est collectée par la CNIL.

< Fiche précédente : utiliser un système d'IA en production

Sommaire

Fiche suivante : permettre le bon exercice de leurs droits par les personnes >

Vous souhaitez contribuer ?

Écrivez à ia[@]cnil.fr

Haut de page