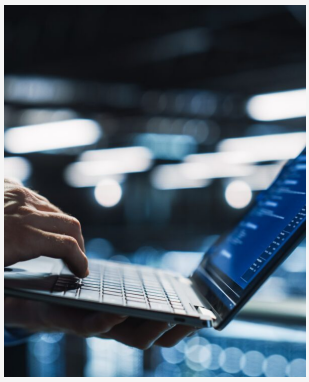
OSLER

Risques et enjeux émergents en matière de sécurité de l'IA : principales conclusions à tirer du rapport du NIST sur l'apprentissage automatique contradictoire



10 MAI 2024 8 MIN DE LECTURE

Expertises Connexes

- Cybersécurité et intervention en cas d'incident lié à la sécurité
- Gestion de risques et réponse aux crises
- Intelligence artificielle
- Litiges
- Opérations commerciales technologiques
- Respect de la vie privée et gestion de l'information
- Sociétés émergentes et à forte croissance
- Technologie

Auteurs(trice): Sam Ip, Naomi Chernos, Joseph Ierullo

À mesure que les systèmes d'IA deviennent de plus en plus répandus et que leur intégration dans les organisations s'affirme, les entreprises sont confrontées à des risques nouveaux et évolutifs en matière de sécurité et de protection de la vie privée. Ces risques découlent en grande partie des systèmes alimentés par des modèles d'apprentissage automatique qui sont exposés à une catégorie de risques de sécurité connus sous le nom d'apprentissage automatique contradictoire (*adversarial machine learning*), qui ciblent les systèmes d'IA, y compris les modèles d'apprentissage automatique et les données utilisées pour former et servir ces modèles. Ces risques nécessiteront des stratégies novatrices à mesure que les organisations apprendront à s'adapter et à répondre aux menaces qu'ils posent.

Pour aider les organisations à se familiariser avec les types d'attaques et de risques auxquels elles peuvent s'attendre, ainsi qu'avec les stratégies permettant de les atténuer, le National Institute of Standards and Technology (NIST) des États-Unis a publié en début d'année un rapport complet, intitulé « *Adversarial Machine Learning : A Taxonomy and Terminology of Attacks and Mitigations* » (Apprentissage automatique contradictoire : classification et terminologie des attaques et des stratégies d'atténuation) [PDF; en anglais seulement], qui se veut un guide de cybersécurité pour les organisations qui élaborent et supervisent la gouvernance des systèmes d'IA. Le rapport, qui s'inscrit dans le cadre des efforts déployés par le NIST pour soutenir le développement d'une IA digne de confiance, contribue à donner une dimension pratique au cadre de gestion des risques liés à l'IA du NIST. Compte tenu de l'influence mondiale des normes et des directives du NIST, les organisations canadiennes peuvent se faire une idée de l'orientation et des directives qui pourraient découler du cadre de référence du NIST.

Plus précisément, le rapport présente une classification des différentes catégories d'attaques pour deux grandes catégories de systèmes d'IA : l'IA prédictive et l'IA générative, sur la base des objectifs et des capacités des attaquants, tout en proposant des stratégies d'atténuation correspondantes. Le rapport aborde également les défis qui subsistent dans ce domaine et les principaux changements à apporter pour assurer la fiabilité de l'écosystème de l'IA.

Types d'attaques

Le rapport du NIST classe les attaques potentielles en fonction de leur capacité et de leur objectif, et indique l'étape du processus d'apprentissage du modèle à laquelle ces attaques peuvent être lancées. Une telle classification vise à créer une terminologie standard à l'appui de l'élaboration de stratégies d'atténuation cohérentes. Les attaques ont été classifiées comme suit :

- Attaques d'évasion : ces attaques, qui se produisent après le déploiement d'un système d'IA, visent à modifier une entrée afin de contourner le comportement prévu du modèle et de changer la façon dont le système y répond, en cherchant essentiellement à générer des exemples contradictoires, qui peuvent être mal classés pendant le déploiement. Un attaquant peut lancer de telles attaques qu'il ait ou non une connaissance préalable de l'architecture du modèle ou des données d'apprentissage. Il peut s'agir, par exemple, d'ajouter des marquages aux panneaux d'arrêt pour qu'un véhicule autonome les interprète mal.
- Attaques par empoisonnement : ces attaques, qui se produisent au cours de la phase d'apprentissage, visent à introduire des données corrompues dans le modèle. Il peut s'agir, par exemple, de glisser des exemples fréquents de langage inapproprié dans les registres de conversations, de sorte qu'un agent conversationnel les interprète comme faisant partie du langage courant et les utilise dans ses propres interactions avec les clients.
- Attaques contre la vie privée :ces attaques, qui se produisent pendant le déploiement, visent à extraire des renseignements sensibles sur l'IA ou des données d'apprentissage. Ce type d'attaque peut par exemple se produire lorsqu'un adversaire pose à un agent conversationnel de nombreuses questions légitimes et utilise les réponses afin de le désosser et, ainsi, de trouver ses points faibles ou de deviner ses sources.
- Attaques malveillantes (*abuse attacks*) ou par injection rapide : ces attaques, qui visent les modèles d'IA générative qui font leur apprentissage en moissonnant une vaste étendue de données, souvent non vérifiées, consistent à insérer des renseignements incorrects dans une source, telle qu'une page Web ou un document en ligne, que l'IA absorbe ensuite comme étant des renseignements légitimes.

Stratégies d'atténuation

Bien que le rapport propose diverses stratégies d'atténuation pour chaque catégorie d'attaques, plusieurs d'entre elles ciblent le risque principal, soit le risque que les systèmes d'IA recourent à des données d'apprentissage qui ne sont pas dignes de confiance et que, pour se développer comme il faut, ils doivent s'en remettre à plusieurs sources de données accessibles au public, ce qui les expose au risque d'interférences négatives. Toutefois, comme les ensembles de données d'apprentissage auxquels les systèmes d'IA ont recours sont trop importants pour qu'une personne parvienne à les contrôler et à les filtrer, les stratégies d'atténuation sont difficiles à mettre en œuvre.

Pour la majorité des catégories d'attaques, la stratégie d'atténuation suggérée se concentre généralement sur la purification des données et l'assainissement des modèles dans leur ensemble, ce qui nécessite des audits et des tests fréquents. Toutefois, le rapport précise

que l'on doit associer ces techniques d'assainissement à des techniques cryptographiques permettant de vérifier la source et l'authenticité du système d'IA, plutôt que de s'appuyer sur la détection d'erreurs dans les données elles-mêmes. Pour ce faire, il est habituellement suggéré de faire appel à la méthode de l'équipe rouge (*red teaming*, méthode qui consiste à former à l'interne une équipe chargée de repérer les faiblesses du système), comme élément essentiel du processus d'essai et d'évaluation du système d'IA effectué avant son lancement dans le but de repérer les failles de sécurité potentielles. Cette méthode est conforme au Code de conduite volontaire visant un développement et une gestion responsables des systèmes d'IA générative avancés d'ISDE, qui prescrit le recours à des essais axés sur des positions antagonistes (c.-à-d. la méthode de l'équipe rouge) pour cerner les vulnérabilités.

En outre, bien que le rapport mentionne des stratégies d'atténuation des risques AAC, les menaces doivent être considérées parallèlement aux menaces classiques liées aux modèles de chaîne d'approvisionnement empoisonnés, aux violations de données et aux vulnérabilités de service inhérentes aux systèmes d'apprentissage automatique eux-mêmes. Ces menaces continuent de mettre en péril la confidentialité des données et l'intégrité des réponses et des prédictions créées par le modèle.

Prochaines étapes

Le rapport met en évidence le fait que chaque catégorie d'attaques fait peser des menaces distinctes et insidieuses, et que chaque stratégie d'atténuation peut nécessiter un ensemble distinct de compétences et de techniques. Pour faire face aux risques AAC, les organisations devront donner la priorité à la gouvernance de l'IA et élaborer une stratégie de gouvernance de l'IA comprenant un cadre de gestion des risques approprié. Ces stratégies devraient mettre l'accent sur les risques liés à la sécurité et à la confidentialité des données, étant donné que l'apprentissage automatique privilégie une approche centrée sur les données et, dans certains cas, que le modèle d'apprentissage automatique risque de recourir à des renseignements délicats et personnels dans le cadre de son apprentissage. Pour plus de renseignements sur les questions relatives à la protection des renseignements personnels dans les modèles d'IA, cliquez ici. Pour gérer ces risques avec succès, les organisations devront intégrer à leur culture d'entreprise une politique de sensibilisation.

En outre, les organisations devront veiller à assortir leurs contrats de protections suffisantes, surtout lorsqu'elles se procurent des systèmes d'IA auprès de fournisseurs tiers. De tels contrats devraient préciser les usages appropriés et les mesures de protection des systèmes d'IA et des modèles d'apprentissage automatique, et contenir des garanties et des mesures appropriées pour s'assurer que les modèles et leurs ensembles de données ne sont pas compromis, ainsi que, le cas échéant, exiger des fournisseurs d'IA qu'ils se conforment à des normes en constante évolution, qu'ils fassent la preuve de leurs certifications et qu'ils facilitent les audits de leurs pratiques. Il sera essentiel de répartir adéquatement les risques dans ces contrats afin que la responsabilité qui y est associée soit clairement définie. Les attaques et les risques peuvent également être introduits par des tiers tout au long de la chaîne d'approvisionnement. Il conviendra d'envisager dans ces contrats des stratégies prévoyant des mesures de protection, par exemple en veillant à ce que le contrôle et la surveillance des données soient mis en œuvre à tous les stades du cycle.

En bref, le rapport du NIST souligne l'importance de faire preuve de vigilance et d'innovation en matière de sécurité de l'IA, de se doter d'un vocabulaire limpide et d'élaborer des stratégies d'atténuation des risques. Il fournit des conseils aux organisations qui cherchent à améliorer leurs pratiques en matière d'IA et leur procure un soutien dans le déploiement et l'adoption de systèmes d'IA dignes de confiance.