Conseils sur la cybersécurité pour les institutions démocratiques : intelligence artificielle

CENTRE CANADIEN POUR LA CYBERSÉCURITÉ

Juin 2024 | ITSAP.00.135

L'intelligence artificielle (IA) est une technologie en constant développement qui fait appel aux algorithmes informatiques pour effectuer des tâches, prédire les résultats et créer du contenu. L'IA générative est un sous-ensemble particulier de l'IA qui utilise de vastes jeux de données pour créer du nouveau contenu. L'IA a sans contredit compliqué le paysage démocratique. La plus grande disponibilité des modèles de langage de grande taille (LLM pour *Large Language Model*), des agents conversationnels et des technologies de création de médias synthétiques a fait en sorte qu'il est plus facile pour les auteures et auteurs de menace de perturber les campagnes et les élections démocratiques. Les auteures et auteurs de cybermenace peuvent créer et diffuser du contenu fabriqué afin de susciter de la confusion, de discréditer le système électoral et de déformer les faits pour tenter de convaincre l'électorat. Ils peuvent avoir recours à l'IA pour accroître l'ampleur de leurs attaques par déni de service distribué (DDoS pour *Distributed Denial of Service*) ou créer des maliciels que les mécanismes de protection antimaliciels et antivirus réguliers n'arriveront pas à détecter. Ils utilisent également des LLM pour produire et diffuser à grande échelle des communications réalistes et très personnalisées. Les LLM sont peu coûteux et facilement accessibles à quiconque souhaite perturber une campagne électorale. En raison de ces capacités, les personnes qui travaillent pour des institutions démocratiques doivent faire preuve d'une très grande vigilance lorsqu'elles utilisent la messagerie électronique, les médias sociaux et d'autres canaux de communication. L'IA facilite également la collecte de données. Dans les mains des auteures et auteurs de menace, cette technologie peut servir à collecter rapidement l'information publique et à la distribuer. On peut l'utiliser contre des candidates, des candidats et des membres du personnel pour mener des activités comme des attaques par divulgation de données personnelles. La divulgation de données personnelles est un acte qui consiste à diffuser en ligne des renseignements personnels, comme des adresses, à des fins généralement malveillantes. Les technologies d'IA évoluent rapidement et les stratégies d'atténuation proposées dans la présente pourraient ne pas être suffisantes pour assurer une protection dans l'avenir. Des mesures d'atténuation additionnelles et de plus amples conseils seront offerts à mesure que l'IA continue d'évoluer.

Types de contenu faux ou synthétique

- **Hypertrucages** : Images ou enregistrements qui ont été manipulés de manière à montrer une personne disant ou faisant des choses controversées ou sensationnalistes. Conçus dans le but de décevoir, ils sont généralement d'aspect convaincant.
- **Faux comptes de médias sociaux** : Utilisés pour personnifier des politiciennes et politiciens ou des personnes influentes dans la communauté, comme des célébrités.
- **Clonage vocal** : Utilisation d'extraits de discours, d'entrevues et de différents clips audio dans le but de personnifier quelqu'un. On peut avoir recours au clonage vocal pour personnifier des responsables et des membres du personnel afin d'obtenir accès à de l'information sensible, de propager de la désinformation ou de discréditer une candidate ou un candidat.
- **Désinformation** : Utilisée pour fournir de la fausse information sur des sujets liés aux élections, comme l'endroit ou le moment où voter, dans le but de perturber le processus électoral. L'objectif consiste à provoquer de la confusion et à réduire le taux de participation des électrices et électeurs appartenant à un groupe ciblé.

Biais et entraînement de l'intelligence artificielle

Les LLM sont des algorithmes utilisés par des programmes qui arrivent à comprendre les invites et à générer des réponses semblables à l'humain. Ils sont entraînés à partir de vastes jeux de données dans le but d'analyser, de résumer, de traduire et de générer du contenu. La qualité et la précision des résultats ne sauraient toutefois être supérieures à celles des données d'entrée.

Ces programmes dépendent de contenu accessible en ligne. Les informations utilisées comme données d'entrée ou invites peuvent également servir à entraîner les modèles. Vous devez connaître les types de contenu et les mesures de sécurité qui sont transmis aux LLM, ainsi que les types de contenu qui ne sont pas pris en charge ou disponibles. On pourrait retrouver une quantité accablante de données en provenance d'une certaine partie du monde ou d'un groupe démographique en particulier, tandis que d'autres zones et groupes pourraient n'être représentés que par un petit échantillon de données, voire aucun. Cette divergence est ce qu'on appelle un biais. Il est impossible d'éliminer le biais des LLM, mais il convient d'en prendre conscience.

Des groupes sous-représentés, comme des minorités ethniques ou linguistiques, pourraient être représentés de façon inéquitable ou pas du tout dans les résultats des LLM. L'exclusion de certains groupes démographiques fait en sorte que les LLM ne peuvent pas participer à des discussions éclairées à caractère politique.

Conseils sur la cybersécurité pour les institutions démocratiques : intelligence artificielle

CENTRE CANADIEN POUR LA CYBERSÉCURITÉ
Juin 2024 | ITSAP.00.135

Comment atténuer la menace posée par l'intelligence artificielle

- Protégez l'information en mettant en place des contrôles d'accès, des mots de passe rigoureux et l'authentification multifacteur (AMF).
- Déployez des contrôles d'accès dans les courriels du personnel, les plateformes de médias et les autres appareils de l'organisation.
- Renforcez la sécurité des comptes de médias sociaux de votre organisation de la façon suivante :
  - désactivez ou supprimez les profils qui ne sont plus utilisés;
  - supprimez toute information nominative;
  - faites en sorte que les profils personnels soient privés.
- Élaborez des stratégies d'approbation à vérification systématique pour limiter l'accès à certaines informations et applications.
- Prenez le temps et les précautions nécessaires pour vérifier l'information avant d'y répondre.
  - Choisissez judicieusement les publications sur lesquelles vous décidez de formuler des commentaires en ligne.
- Adoptez une stratégie de phrases de passe séquentielles que seul le personnel autorisé connaît pour empêcher les adversaires d'obtenir accès aux systèmes organisationnels au moyen du clonage vocal en temps réel.
  - Une phrase de passe ou un mot de passe séquentiel est souvent généré automatiquement et changé à intervalles réguliers pour prévenir les accès non autorisés.
- Éduquez les membres du personnel, les personnes qui vous suivent et le public par rapport à ce qui suit :
  - les risques potentiels et les vulnérabilités qui touchent l'utilisation et la consommation des médias;
  - les connaissances médiatiques;
  - ce qu'il faut surveiller en ligne;
  - la façon de reconnaître les sources fiables et de vérifier l'information.
- Signalez les incidents d'hypertrucages à la plateforme qui les héberge et au Centre antifraude du Canada (CAFC) au moyen du système de signalement en ligne ou par téléphone au 1-888-495-8501.
- Préparez à l'avance une Foire aux questions ou un ensemble de réponses aux questions concernant le contenu synthétique.
- Assurez-vous que les réseaux qui hébergent les processus électoraux sont conçus et mis en œuvre avec des mécanismes d'authentification rigoureux, comme l'AMF.
- Élaborez un plan de gestion des incidents.

Ressources connexes

- Cybermenaces contre le processus démocratique du Canada : Mise à jour de 2023
- Rapport de l'intelligence artificielle
- Lignes directrices pour le développement de systèmes d'IA sécurisés
- Étapes à suivre pour déployer efficacement l'authentification multifacteur (AMF) (ITSAP.00.105)
- Une stratégie pour l'OTAN en matière d'intelligence artificielle
- La menace posée par les générateurs de texte basés sur des modèles de langage de grande taille
- Risk in focus: Generative A.I. and the 2023 election cycle *(en anglais seulement)*