UNIVERSITÉ DU QUÉBEC

CESI
CENTRE D'EXPERTISE EN SÉCURITÉ DE L'INFORMATION

TLP : VERT (DIFFUSION PERMISE)

Les dangers de l'intelligence artificielle générative

[Date de publication MM-YYYY]

UNIVERSITÉ DU QUÉBEC

CESI
CENTRE D'EXPERTISE EN SÉCURITÉ DE L'INFORMATION

TLP : VERT (DIFFUSION PERMISE)

TABLE DES MATIÈRES

UNIVERSITÉ DU QUÉBEC

CESI
CENTRE D'EXPERTISE EN SÉCURITÉ DE L'INFORMATION

TLP : VERT (DIFFUSION PERMISE)

INTRODUCTION

L'intelligence artificielle générative (IAG) suscite un intérêt croissant dans divers secteurs. Ses avantages potentiels pour les entreprises en font une technologie attrayante, tant pour les gestionnaires souhaitant réduire les coûts opérationnels que pour les employés cherchant à optimiser et automatiser des processus répétitifs. Toutefois, comme toute technologie émergente, il est essentiel de bien appréhender son fonctionnement ainsi que les risques inhérents à son utilisation.

FONCTIONNEMENT DE L'IAG

L'IAG fonctionne en utilisant des modèles d'intelligence artificielle, souvent basés sur des réseaux de neurones profonds, pour générer du contenu original à partir de données existantes. Ces modèles sont entraînés sur de grandes quantités de données (textes, images, sons, etc.), leur permettant d'apprendre des motifs, des structures et des relations complexes. Grâce à cet apprentissage, l'IAG peut créer de nouveaux contenus en réponse à des invites ou des instructions, que ce soient des textes, des images, de la musique, ou même du code.

DIFFÉRENCES ENTRE L'IA ET L'IAG

L'IA (intelligence artificielle) est un terme générique qui englobe des systèmes capables d'exécuter des tâches spécifiques en imitant l'intelligence humaine, comme la conduite autonome ou le contrôle de la température automatique. L'intelligence artificielle générative (IAG) est une sous-catégorie de l'IA, spécialisée dans la création de nouveaux contenus à partir de données existantes. Tandis que l'IA classique se concentre souvent sur l'analyse et la résolution de problèmes, l'IAG est orientée vers la production créative.

VÉRACITÉ DU CONTENU

Le principal risque lié à l'utilisation d'un outil d'IAG réside dans la véracité de l'information produite. Étant donné que le réseau de neurones apprend en collectant des textes, des images et des vidéos provenant de toutes parts, il est logique qu'il recueille de manière massive des

données contenant de la désinformation, de l'hypertrucage (deepfake) et d'autres contenus générés par l'IAG, en plus du contenu authentique.

La théorie de « l'internet mort » refait surface, indiquant que le réseau internet sera peuplé principalement de robots, et que la majorité du trafic et des contenus seront générés par des algorithmes (Demeure, 2024). L'IAG contribue énormément à ce phénomène, ou ce dernier produit lui-même de la désinformation qui est ensuite publiée, pour être ensuite réabsorbée par le modèle comme source légitime. L'effet « boule de neige » aura comme effet que nous ne pourrons plus faire confiance à ce qu'on retrouve sur internet, puisque les sources d'information utilisées par l'IAG seraient majoritairement du contenu généré par l'IAG.

COMMENT SE PROTÉGER?

Il est primordial de valider l'information fournie en recherchant nous-mêmes des sources sûres qui viennent appuyer et bonifier le contenu généré par l'IAG.

CONFIDENTIALITÉ DES DONNÉES

Un des risques importants pour la sécurité de l'information en lien avec l'utilisation de l'intelligence artificielle générative est lié au concept même de l'IAG : l'apprentissage automatique. Certains modèles d'IAG disponibles apprennent des échanges faits avec leurs utilisateurs. C'est-à-dire que tout ce qui est partagé par l'utilisateur pourrait être ajouté aux connaissances qui composent le modèle, pouvant ainsi être récupérées par d'autres utilisateurs qui utilisent le même modèle.

COMMENT SE PROTÉGER?

Il faut éviter de partager toute donnée nominative ou confidentielle lors des échanges avec un outil d'intelligence artificielle générative.

INFRACTION AUX DROITS D'AUTEURS

L'utilisation de l'intelligence artificielle générative (IAG) suscite de plus en plus d'inquiétudes concernant les droits d'auteur. Ces modèles sont entraînés à partir de vastes ensembles de données qui incluent souvent des œuvres protégées. Lorsqu'ils génèrent du contenu, il peut

être difficile de déterminer si les créations sont réellement originales ou si elles sont en partie basées sur les œuvres protégées ayant servi à l'apprentissage du modèle. Cela peut entraîner des violations involontaires des droits d'auteur si les productions de l'IAG se rapprochent trop des créations originales. De plus, la complexité juridique entourant la propriété des créations générées par l'IAG pose d'importants défis. Est-ce que le contenu appartient à l'utilisateur, au créateur du modèle ou aux créateurs des données utilisées pour l'apprentissage du modèle? Cette incertitude juridique peut exposer les utilisateurs à des litiges ou des réclamations en cas de violation de droits d'auteurs.

COMMENT SE PROTÉGER?

Il est difficile de se protéger contre l'infraction aux droits d'auteurs en tant qu'utilisateur d'IAG. Cependant, il est possible de demander des garanties aux fournisseurs d'inclure des clauses d'indemnisation dans les contrats de service, transférant la responsabilité de toute violation potentielle.

SOPHISTICATION DE L'HYPERTRUCAGE

La popularité de l'intelligence artificielle générative présente non seulement des avantages, mais aussi des risques potentiellement dangereux, notamment lorsque cette technologie est exploitée par des individus mal intentionnés. Les techniques d'hameçonnage, déjà très efficaces, se voient aujourd'hui renforcées par la simplicité d'utilisation des outils d'IAG. Des enregistrements audios, des photos ou encore des imitations vocales d'un proche, d'une personne influente au sein d'une entreprise ou d'un ami peuvent être utilisés pour créer des situations urgentes et convaincantes, telles que des demandes de transferts bancaires ou la divulgation d'informations sensibles. La confiance suscitée par ces hypertrucages amplifie considérablement le taux de succès de ces tentatives d'hameçonnage.

COMMENT SE PROTÉGER?

Pour se protéger contre les hypertrucages, il est essentiel de vérifier systématiquement l'identité des demandeurs en cas de requête urgente, notamment par un autre canal comme un appel direct. Il faut également rester vigilant face aux comportements inhabituels, provenant même de sources de confiance. Enfin, limiter la diffusion d'informations personnelles en ligne réduit les risques que ces données soient exploitées pour des tentatives de manipulation.

GLOSSAIRE

Hypertrucage (Deepfake) : Procédé de manipulation audiovisuelle permettant l'imitation ultraréaliste de l'apparence et de la voix grâce à l'apprentissage profond de l'intelligence artificielle.

IAG : Technologie qui crée du contenu original, comme du texte, des images ou de la musique, en s'appuyant sur des modèles d'IA entraînés à partir de grandes quantités de données.

Intelligence artificielle (IA) : Domaine de l'informatique qui développe des systèmes capables de simuler l'intelligence humaine en effectuant des tâches comme la reconnaissance de motifs, la prise de décisions ou l'apprentissage à partir de données.

IPI : Les informations personnellement identifiables sont des informations qui peuvent être utilisées pour vérifier l'identité d'une personne.

Modèle d'IAG : Un programme qui a été entraîné sur un ensemble de données pour reconnaître certains schémas ou prendre certaines décisions sans intervention humaine supplémentaire.

RÉFÉRENCES

Risques de l'IA pour les entreprises - KPMG Canada

The rise of AI : Fraud in the digital age - Canada.ca

Dead Internet : et si le Web était sur le point de mourir? (sciencepost.fr)

L'intelligence artificielle générative - ITSAP.00.041 - Centre canadien pour la cybersécurité

Europol Innovation Lab Facing Reality Law Enforcement And The Challenge Of Deepfakes.pdf (europa.eu)

Comment atténuer les risques liés au droit d'auteur à l'ère de l'IA générative? - Skim AI

Europol Innovation Lab Facing Reality Law Enforcement And The Challenge Of Deepfakes.pdf (europa.eu)

RÉVISIONS

Date de révision : 2024-09-17