IMPACT AI

LES BRIEFS DE L'IA RESPONSABLE

2

IA générative & sûreté, (cyber)sécurité et vie privée

OCTOBRE 2024

Plus une technologie est puissante, plus les bénéfices ou les risques qu'elle porte sont importants. Les systèmes et modèles d'IA générative ne font pas exception à cette règle.

Ce deuxième « Brief de l'IA responsable » d'Impact AI résume ses travaux et propose des bonnes pratiques et initiatives visant à s'assurer que le développement et l'utilisation de l'IA générative est aussi sûre et sécurisée que possible.

Ce brief s'appuie sur l'expérience des membres d'Impact AI, ainsi que sur un certain nombre d'études et d'initiatives récentes, afin de proposer une première vision des stratégies à mettre en œuvre pour limiter les risques de l'IA générative.

1/ Le contexte

Pour que les entreprises puissent prendre des décisions en toute connaissance de cause et soient en mesure de tirer toute la valeur possible de la mise en œuvre de modèles d'IA générative, elles doivent pouvoir identifier, comprendre et traiter les risques en matière de sûreté et de (cyber)sécurité. Ces risques peuvent se ranger en deux catégories, les risques déterministes et les risques non déterministes.

Les risques déterministes

Aussi révolutionnaires soient-ils, les systèmes d'IA générative demeurent **des briques de logiciels** et doivent être traités comme tels. Ils sont constitués du système lui-même, de l'infrastructure qui l'accueille et le porte et des autres systèmes avec lesquels il interagit. Il existe donc des situations évidentes dans lesquelles les vulnérabilités doivent être identifiées et traitées. Par exemple : « Le modèle tourne-t-il sur une machine virtuelle sécurisée ? » ou « L'accès au modèle se fait-il par un processus sécurisé d'autorisations et quelles informations d'authentification sont utilisées ? ».
Naturellement, ces questions sont traitées par des processus comme le NIST Secure Software Development Practices (SSDF) et le Microsoft Security Development Lifecycle (SDL), pour ne citer qu'eux. Mais l'essentiel est de se conformer à TOUS les principes fondamentaux de sécurité pour une bonne hygiène en matière de cybersécurité. Car les risques cyber que recèlent les applications IT traditionnelles concernent tout aussi bien les systèmes d'IA générative.
Ces IA génératives sont également des systèmes d'IA qui héritent des spécificités de sécurité, comme la sécurité des données et les réponses aux attaques adverses qui peuvent se produire à différents moments du cycle de vie des systèmes (collecte de données, entraînement et modélisation, inférence).

Les risques non déterministes

Les systèmes d'IA générative peuvent être attaqués au travers des contenus qu'ils génèrent, puisqu'il existe des possibilités de les tromper, de les troubler ou de les contraindre ou bien qu'un utilisateur puisse mal interpréter un résultat. Ces attaques sont chaotiques, difficiles à identifier et ne peuvent pas être stoppées de façon déterministe. Pour les experts de la sécurité des logiciels, ce type de risque est nouveau et ils ne peuvent pas être « étiquetés » comme des vulnérabilités traditionnelles : on peut seulement les rendre plus ou moins probables. Ces risques concernent notamment :

→ Les **attaques directes par prompts**, ou jailbreak attacks ou encore user prompt injection attack (UPIA) qui surviennent lorsque quelqu'un exploite de façon intentionnelle les vulnérabilités d'un système ou d'un modèle d'IA générative. Ces techniques peuvent détruire les garde-fous et se combiner avec d'autres types d'attaques par prompts.
L'objectif est clairement de franchir les barrières de sécurité d'un système pour des motifs non autorisés. Ces attaques directes peuvent utiliser des techniques très « humaines » – action psychologique, intimidation, flatterie – qui peuvent attirer le modèle dans une forme de « sweet talking » pouvant lui faire baisser la garde. Mais les attaques peuvent aussi être menées par des techniques « artificielles », qui injectent des chaînes de caractères sans signification humaine évidente, mais tout aussi capables de tromper un système.

→ Les attaques **indirectes par prompts** ou cross-domain prompt Injection attack (XPIA), qui interviennent lorsque l'agresseur embarque un élément malveillant dans des données externes. Il peut s'agir d'instructions cachées, ignorées du rédacteur du prompt, qui peuvent conduire le système à traiter la demande comme si elle émanait du rédacteur lui-même.

→ Les **fuites de prompts**, c'est-à-dire l'exposition non intentionnelle d'informations personnelles, confidentielles ou sensibles au travers de l'écriture d'un prompt. Ces informations peuvent être récupérées par des utilisateurs non autorisés, au travers de technologies comme la génération augmentée de récupération (RAG), un processus qui optimise le résultat d'un LLM en faisant appel à des bases de connaissances externes aux sources de données utilisées pour l'entraîner.

→ Les **fuites de données**, qui peuvent survenir lorsque sont utilisées des informations externes au jeu de données d'entraînement mobilisé pour développer un modèle d'IA générative. Des données « sensibles » peuvent être frauduleusement étiquetées « non sensibles » pour permettre à un utilisateur non autorisé d'y avoir accès.

→ La **génération d'hallucinations** sous la forme de réponses ne reposant sur aucune information sourcée ou de réponses omettant des informations cruciales.

→ Le **désalignement des pratiques** qui peut survenir lorsque des règles spécifiques de confidentialité ou de terminologies souvent écrites en langage naturel, notamment dans certains secteurs comme la santé ou les services financiers, ne sont pas comprises par les utilisateurs, ce qui peut avoir de graves conséquences.

→ La **diffusion d'IA génératives fantômes,** un risque créé par des modèles construits sans règles de gouvernance appropriées ou modalités de contrôle. Lorsque des collaborateurs utilisent ou intègrent ces IA, qui n'ont pas passé de tests rigoureux en matière de sécurité, de protocole et de gouvernance, ils génèrent des risques en matière de confidentialité des données, de sécurité et d'éthique.

2/ Les expériences et pratiques des membres d'Impact AI

Face à ces risques, les membres d'Impact AI s'efforcent d'apporter des réponses concrètes.

Microsoft France

Microsoft France a partagé un guide de modélisation des menaces pour les systèmes d'IA (générative) ainsi qu'un ensemble de bonnes pratiques en matière de red teaming, consistant à tester des modèles d'IA pour les protéger contre des comportements frauduleux.
Ces pratiques, mises au point par le AI Red Team (AIRT) de Microsoft, permettent de tester la sécurité d'un système dans des conditions réelles d'utilisation, de rechercher de façon pro-active les éléments de vulnérabilité, de définir des stratégies de défense et de mettre en place des plans de renforcement de la sécurité.
Un outil open source comme le PyRIT (Python Risk Identification Toolkit for Generative AI), créé par l'AIRT, permet notamment de simuler manuellement ou automatiquement des attaques contre des modèles de GenAI. Naturellement, même testé avec succès au cours d'une soixantaine d'exercices, le PyRIT ne remplace pas la supervision humaine, qui reste absolument nécessaire pour ne pas laisser les systèmes d'AI générative se tester les uns les autres.

Orange

Orange a insisté sur la nécessite de gérer les systèmes d'IA générative, d'abord comme des systèmes d'IA et a rappelé le process de gestion de la sécurité classique, qui suit les étapes d'un projet de développement informatique de la conception à l'exploitation.
Face aux différents risques, les actions principales ont été rappelées en termes de gestion des fournisseurs/librairies tiers, de protection des données (de leur qualité à leur accès), ou de protection des infrastructures. Quelques outils en open source, pour gérer la sécurité des modèles d'IA en cours de test, ont été présentés à l'image de la boîte à outils ART (Adversarial Robustness Toolbox) ainsi que les premières recommandations des régulateurs CNIL et ANSSI en France.

3/ Les bonnes pratiques identifiées

Il nous parait essentiel – s'il est besoin de le rappeler – de commencer par :
→ Appliquer les processus fondamentaux de sécurité du développement logiciel classique aux systèmes d'IA générative pour une bonne hygiène de base.
→ Initier un état de l'art des menaces des IA génératives et des atténuations connues. La base de connaissance MITRE Adversarial Threat Landscape for Artificial-Intelligence Systems (ATLAS™) peut y contribuer grandement.

Face aux avancées de l'IA générative, les pratiques de red teaming et de stress test sont de plus en plus mises en œuvre, puisqu'elles permettent de challenger un système dans des conditions réelles d'utilisation et de mobiliser des outils, tactiques et procédures, pour identifier les risques, mettre à jour les angles morts, valider des convictions et améliorer la sécurité globale des systèmes.

Appliquée à l'IA générative, la pratique du red teaming a d'ailleurs évolué : **elle ne protège pas seulement contre des vulnérabilités de sécurité mais inclut des dispositifs pour d'autres types de défaillances, comme la génération de contenus potentiellement nocifs.** L'IA générative comporte des risques nouveaux et le red teaming est central pour les comprendre.

Quant à l'être humain, il est aussi sujet à de l'engineering social et à des manipulations, puisque les systèmes d'IA générative sont entraînés sur des contenus produits par des humains, qui peuvent être culpabilisés, menacés, trompés, victimes d'usurpation d'identité ou au contraire faussement mis en confiance au travers de prompts multiples. Ainsi, le red teaming ne consiste pas seulement à vérifier si un modèle refuse un prompt isolé, mais s'il peut résister à la manipulation via des prompts multiples.

4/ Les challenges et questions encore ouvertes

En matière de sécurité, de sûreté et de confiance, il n'y a jamais rien de définitif. C'est un voyage, pas une destination. L'IA générative est un paysage qui évolue rapidement, et il est plus important que jamais de **dévoiler collectivement et de façon responsable ses vulnérabilités**, de partager l'état de l'art en matière d'attaques hostiles, de nourrir les bonnes pratiques, de partager les savoirs et de contribuer à enrichir la connaissance du public.

5/ Pour en savoir plus

→ Recommandations de sécurité pour un système d'IA générative, un document établi par l'ANSSI pour sensibiliser les autorités publiques et les entreprises aux risques de l'IA générative et pour promouvoir les bonnes pratiques

→ Sécurité : Intelligence artificielle - Conception et apprentissage, une fiche pratique de la CNIL détaillant les risques et mesures à prendre recommandées pour garantir la sécurité du développement d'un système d'IA

→ Secure Software Development Practices for Generative AI and Dual-Use Foundation Models : An SSDF Community Profile, destiné aux producteurs et aux utilisateurs d'IA génératives.

→ MITRE Adversarial Threat Landscape for Artificial-Intelligence Systems (ATLAS™), une base de connaissance accessible et mise à jour sur les tactiques et techniques hostiles. Elle repose sur l'observation d'attaques et des démonstrations dans des conditions réelles de la part d'*AI Red Teams* et de groupes d'experts, avec des études de cas spécifiques.

→ Must Learn AI Security, une série pédagogique sur les risques de sécurité des AI génératives.

→ OWASP (Open Web Application Security Project)'s Top 10 for LLM applications, une ressource pour aider les organisations à développer et à déployer de façon sûre des AI génératives.

→ Planning red teaming for large language models (LLMs) and their applications, un guide pour aider à assembler une *AI red team*, définir ses objectifs et délivrer les résultats.

→ The AI Risk Repository (mit.edu) : une base de données regroupant les risques associés aux Système d'IA dont les IA génératives

→ Recommandations de la CNIL

→ IBM Adversarial Toolbox

www.impact-ai.fr
contact@impact-ai.fr