AI Technology Research

IGF Internet Governance Forum RIYADH 2024

Lignes directrices pour le développement et le déploiement sécurisés de systèmes d'IA

kaspersky bring on the future

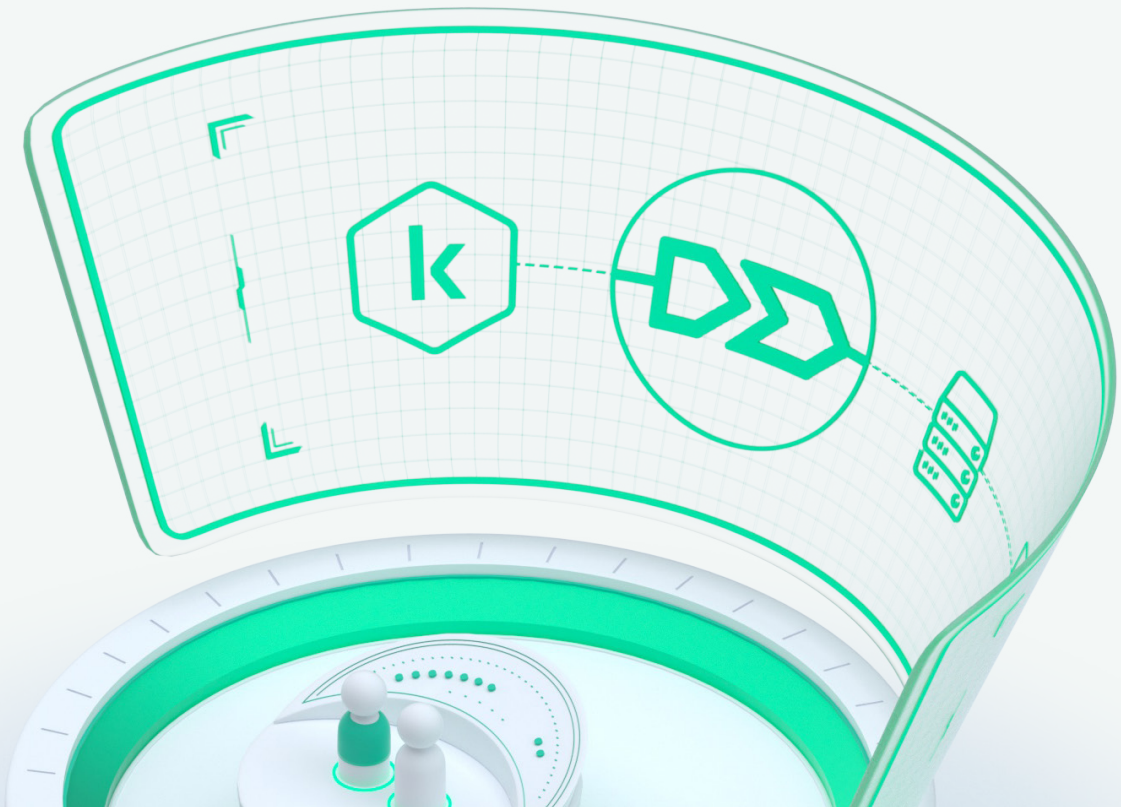
Remerciements

Ce document a été rédigé par Kaspersky et présenté dans le cadre de l'**atelier n° 31 intitulé « Cybersecurity in AI: balancing innovation and risks » (« Cybersécurité et IA : trouver un équilibre, entre innovation et risques »)** lors de la 19e réunion annuelle du **Forum sur la Gouvernance de l'Internet** qui s'est tenu du 15 au 19 décembre 2024.

Les personnes suivantes ont apporté leur contribution à la rédaction de ce document :

- Yury Shelikhov, Head of Cybersecurity and Data Protection chez Kaspersky
- Vladislav Tushkanov, Research Development Group Manager, Machine Learning Technology Research chez Kaspersky
- Alexey Antonov, Lead Data Scientist, Detection Methods Analysis chez Kaspersky
- Allison Wylde, membre de l'équipe de l'UN IGF PNAI (interopérabilité)
- Dr Melodena Stephens, professeure en innovation et en gouvernance des technologies à la Mohammed Bin Rashid School of Government (Émirats arabes unis)
- Sergio Mayo Macías, directeur des programmes d'innovation à l'Institut technologique d'Aragon (ITA) (Espagne)
- Igor Kumagin, expert en cybersécurité
- Dmitry Fonarev, responsable des affaires publiques senior chez Kaspersky

Sommaire



Introduction

Objectif

Les nouvelles technologies numériques s'accompagnent de **nouveaux risques et vecteurs d'attaque en matière de cybersécurité**. Par conséquent, les entreprises doivent veiller à ce que leur intégration de l'IA soit protégée face à ces menaces. Le concept de sécurité dans le développement de systèmes d'IA est aujourd'hui au premier plan de diverses législations, telles que l'AI Act de l'UE ou le cadre de gouvernance de l'IA générative de Singapour, afin de minimiser les risques cyber qui y sont associés. L'UE met en place une réglementation stricte en matière d'IA avec l'AI Act, qui vise à garantir la transparence, la sécurité et le respect de normes éthiques. Les États-Unis s'attachent à développer des normes industrielles et à encourager l'innovation plutôt que des lois strictes. La Chine élabore activement des normes et des réglementations qui soutiennent le développement des technologies d'IA, tout en limitant leur utilisation dans certains domaines.

Malgré ces progrès sur le plan législatif, des différences importantes subsistent entre les cadres généraux et leur mise en œuvre pratique à un niveau plus technique. Dans ce document, nous examinons les **exigences de base en matière de cybersécurité**, qui devraient être prises en compte lors de la mise en œuvre de systèmes d'IA. Ces exigences devraient s'appliquer à un **plus grand nombre d'entreprises qui s'appuient sur des modules d'IA tiers** pour développer leurs propres solutions.

Pour mettre en œuvre l'IA en toute sécurité, les entreprises ont besoin de conseils techniques sur le développement et le déploiement de l'IA au sein de leur infrastructure. La mise en œuvre de l'IA sans orientations appropriées peut présenter des risques importants. Ce document vise à fournir des lignes directrices aux développeurs et aux administrateurs de systèmes d'IA, de MLOps et de DevOps d'IA, en s'appuyant sur les modèles fondamentaux existants pour créer des solutions d'IA généralisées, avec une attention particulière pour les systèmes d'IA basés sur le cloud. Ce document aborde les **aspects clés du développement, du déploiement et de l'exploitation des systèmes d'IA**, notamment la conception, les bonnes pratiques en matière de sécurité et l'intégration, sans se concentrer sur le développement de modèles fondamentaux.

L'intelligence artificielle (IA) est aujourd'hui une technologie essentielle à l'économie mondiale, qui fait partie intégrante de notre quotidien. L'IA permet aux entreprises d'automatiser leurs tâches routinières, d'améliorer leur service clientèle et de fournir à leurs employés un accès plus rapide et plus simple aux informations.

> 50 %

Une récente étude menée par Kaspersky a révélé que plus de 50 % des entreprises ont déjà mis en œuvre des solutions basées sur l'IA au sein de leurs infrastructures*.

33 % des entreprises

prévoient d'adopter cette technologie dans les deux années à venir.

* Plus de la moitié des entreprises utilisent l'IA et l'IoT dans le cadre de leur activité, https://www.kaspersky.fr/about/press-releases/plus-de-la-moitie-des-entreprises-utilisent-lia-et-liot-dans-le-cadre-de-leur-activite

Les menaces qui pèsent sur les systèmes d'IA se multiplient à mesure que cette technologie est de plus en plus déployée au sein des entreprises. Les cyberattaques touchent toutes les étapes du développement de l'IA, depuis les ensembles de données jusqu'aux algorithmes et aux résultats des modèles.

Vue d'ensemble du paysage des menaces liées à l'IA

D'après les recherches menées par Kaspersky*, les systèmes d'IA sont confrontés à des problèmes de sécurité uniques et en constante évolution, qui mettent en péril leur fonctionnement. Des acteurs malveillants exploitent les vulnérabilités des données de formation, manipulent les modèles pour influencer leur comportement, et compromettent l'intégrité des systèmes. Il en ressort un besoin urgent de sécurité globale dans les applications d'IA.

Problèmes liés au développement des modèles

Contrairement à la programmation classique, pour laquelle le comportement du code peut être explicitement compris et testé, les modèles de machine learning, et en particulier les modèles de deep learning comprenant des millions ou des milliards de paramètres, sont intrinsèquement complexes et fonctionnent souvent comme une « boîte noire ». Cette complexité rend difficiles la prévision et l'interprétation du comportement des modèles. Par conséquent, le risque d'erreurs non détectées susceptibles d'avoir des conséquences graves (sur la stabilité financière d'une banque ou même sur la vie d'un patient, par exemple) augmente considérablement.

Un autre problème réside dans le fait que les modèles d'IA peuvent parfois fonder leurs décisions sur des **propriétés de données d'entrée peu pertinentes ou négligeables**, plutôt que sur des éléments pertinents. Par exemple, les modèles de reconnaissance d'images peuvent apprendre à classifier des animaux tels que le guépard, le léopard et le jaguar en se fondant uniquement sur les motifs de leurs taches, au lieu de s'appuyer sur leur anatomie dans son ensemble**. À titre d'exemple, un modèle a identifié à tort un canapé à motif tacheté comme étant un léopard, car il a associé le motif tacheté du canapé à cet animal. De telles erreurs de classification peuvent conduire à des résultats erronés dans des applications essentielles des secteurs de la santé, de l'éducation, de la protection sociale, des transports, de l'administration publique, etc.

Des incohérences entre les données utilisées à des fins de formation et celles rencontrées lors du déploiement peuvent conduire à de mauvaises performances des modèles. Par exemple, si un modèle est formé à partir de données collectées via un certain type de dispositif, mais qu'il est appliqué à des données provenant d'un autre appareil, il peut mémoriser des caractéristiques propres à l'appareil, plutôt que les éléments sous-jacents qu'il est censé reconnaître. Cette inadéquation peut entraîner des prédictions ou des classifications inexactes.

La formation de modèles à partir de rien comme le perfectionnement de modèles fondamentaux peuvent être confrontés à de tels défis. Bien que les ensembles de données de perfectionnement soient plus petits et plus faciles à gérer, ils peuvent également introduire des corrélations parasites, qui rendent le modèle qui en résulte non conforme à l'objectif visé.

* L'IA attaquée, https://content.kaspersky-labs.com/se/media/en/business-security/enterprise/machine-learning-cybersecurity-whitepaper.pdf

** Soudain, un canapé imprimé léopard apparaît, https://web.archive.org/web/20200208171948/http://rocknrollnerd.github.io/ml/2015/05/27/leopard-sofa.html

Attaques contre les modèles d'IA

Des acteurs malveillants peuvent cibler les modèles d'IA en recourant à diverses méthodes. Voici des exemples de la manière dont les pirates informatiques exploitent les vulnérabilités dans la conception, la formation et les mécanismes d'interaction de l'IA :

Moyens d'attaque	Description
 Empoisonnement des données : compromission de l'intégrité du modèle	L'empoisonnement des données implique qu'un pirate informatique injecte des données malveillantes dans l'ensemble de données de formation, afin d'influencer le comportement du modèle. En concevant et en introduisant avec soin des éléments empoisonnés*, les pirates informatiques peuvent amener le modèle à prendre de mauvaises décisions ou à effectuer des classifications incorrectes sur certaines données d'entrée. Ce type d'attaque peut compromettre l'intégrité du modèle et nuire à la fiabilité de ce dernier. Il en va de même pour le perfectionnement de modèles de base.
 Attaques adverses : manipulation invisible de l'IA	Les attaques adverses impliquent des modifications subtiles des données d'entrée, qui amènent le modèle d'IA à les classifier de manière erronée, alors que ces changements passent inaperçus aux yeux d'humains**. Les pirates informatiques ajoutent aux données d'entrée des parasites spécialement conçus à cet effet, ce qui amène le modèle à produire des données de sortie incorrectes, alors que les données d'entrée semblent inchangées aux yeux d'observateurs humains.
 Mémorisation des données de l'IA : risque de divulgation involontaire	Les modèles d'IA modernes peuvent mémoriser par inadvertance certaines informations de leurs données de formation, en particulier si ces données contiennent des éléments uniques ou exceptionnels. Les pirates informatiques peuvent exploiter cette situation en ayant recours à des techniques pour extraire des informations confidentielles que le modèle a stockées par inadvertance. Cela peut conduire à la divulgation de données personnelles d'utilisateurs ou d'informations commerciales confidentielles.
 Injection de prompts : une menace pour les grands modèles de langage	L'injection de prompts est une menace spécifique aux LLM, comme ChatGPT. Les développeurs programment les grands modèles de langage pour qu'ils effectuent des tâches en leur transmettant des prompts initiaux en langage naturel. Étant donné que les utilisateurs interagissent également avec le modèle en utilisant le langage naturel, le modèle ne peut pas intrinsèquement faire la différence entre les instructions des développeurs et les données d'entrée des utilisateurs. Les pirates informatiques peuvent créer des données d'entrée qui annulent ou manipulent le comportement du modèle, l'amenant à effectuer des actions non souhaitées ou à divulguer des informations confidentielles. Ces prompts malveillants peuvent être saisis directement par l'utilisateur ou intégrés aux données traitées par le modèle, comme des documents ou des pages Internet.

Il ne s'agit là que des attaques les plus courantes ; une description complète de toutes les attaques possibles contre les systèmes d'IA dépasse en effet le cadre du présent document.

* Comprendre les attaques par empoisonnement de données, https://bdtechtalks.com/2020/10/07/machine-learning-data-poisoning/
Attaques contre le machine learning : vue d'ensemble, https://elie.net/blog/ai/attacks-against-machine-learning-an-overview/

** Expliquer et exploiter les exemples adverses, https://arxiv.org/abs/1412.6572
Attaques et défenses adverses dans le deep learning, https://arxiv.org/abs/2201.06192
Comment perturber les réseaux neuronaux anti-malwares : attaques adverses et protection, https://securelist.com/how-to-confuse-antimalware-neural-networks-adversarial-attacks-and-protection/102949/

Vulnérabilités **classiques en matière de sécurité**

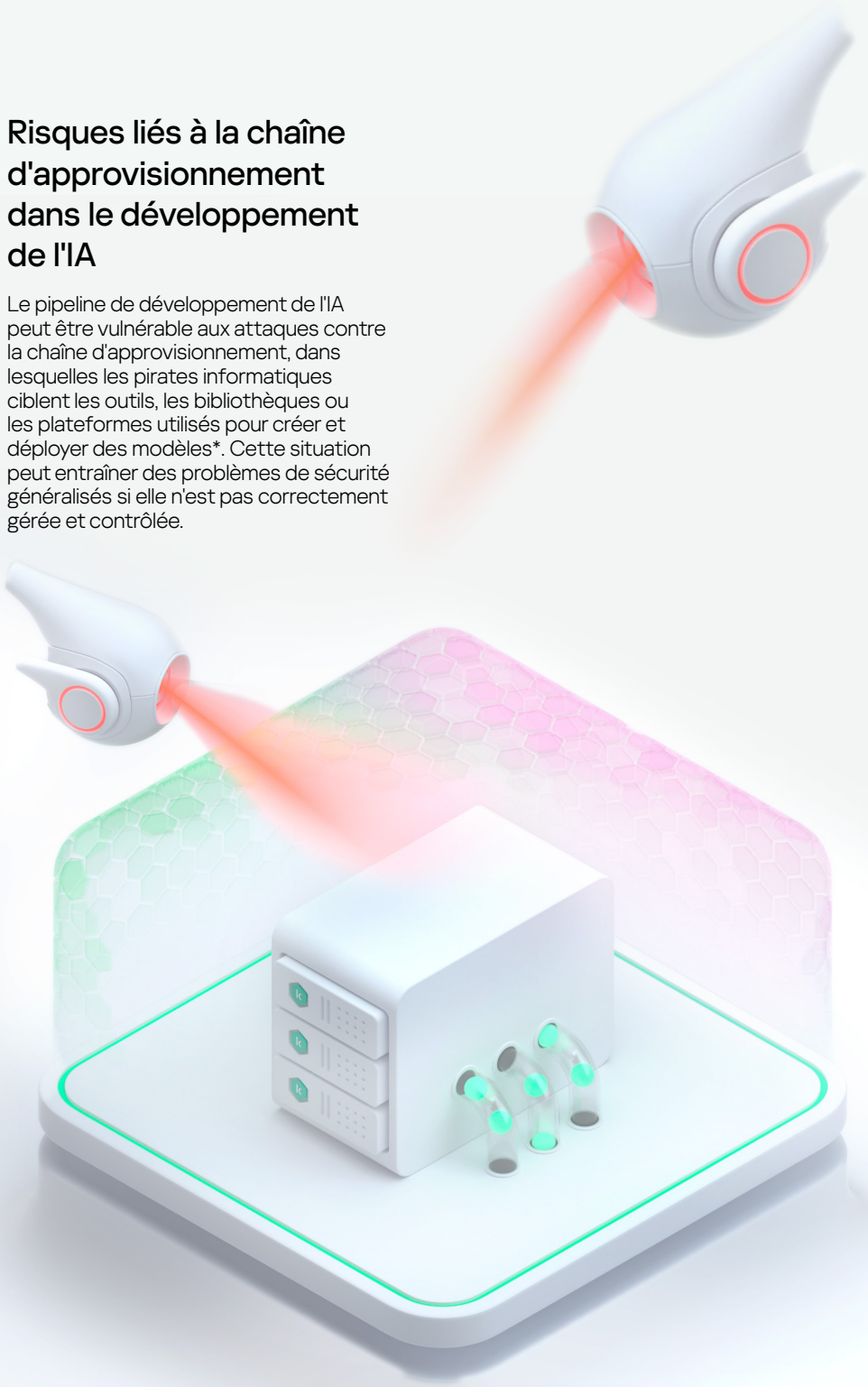
Les modèles d'IA peuvent être sensibles aux faiblesses traditionnelles en matière de sécurité :

Vulnérabilités de l'IA liées à des ressources tierces

Les systèmes d'IA s'appuient souvent sur des modèles ou des ensembles de données tiers provenant de dépôts ouverts. Ces ressources peuvent contenir des erreurs ou des portes dérobées introduites de manière délibérée par des acteurs malveillants. La présence de modules compromis de ce type peut entraîner des vulnérabilités dans le système d'IA, affectant ainsi sa sécurité.

Risques liés à la chaîne d'approvisionnement dans le développement de l'IA

Le pipeline de développement de l'IA peut être vulnérable aux attaques contre la chaîne d'approvisionnement, dans lesquelles les pirates informatiques ciblent les outils, les bibliothèques ou les plateformes utilisés pour créer et déployer des modèles*. Cette situation peut entraîner des problèmes de sécurité généralisés si elle n'est pas correctement gérée et contrôlée.



Erreurs de code exposant à des vulnérabilités de l'IA

Des erreurs dans le code des interfaces d'accès à l'IA peuvent entraîner des vulnérabilités.

Risques de vol de modules d'IA

Les systèmes d'IA ou leurs modules essentiels, comme les modèles ou les ensembles de données, peuvent être dérobés sans une protection adéquate telle que définie dans le présent document.

* Chaîne de dépendances PyTorch-nightly compromise entre le 25 et le 30 décembre 2022, https://pytorch.org/blog/compromised-nightly-dependency

Lignes directrices

Sensibilisation et formation à la cybersécurité

La mise en œuvre de nouvelles technologies comme l'IA nécessite **le soutien des dirigeants, la mise en place de politiques et d'une gouvernance internes, ainsi qu'une formation spécialisée destinée aux employés sur les risques et les menaces associés à l'IA**. Ce dernier point est essentiel en raison de l'évolution rapide de cette technologie. De nombreuses ressources d'IA mises à la disposition des développeurs ne sont pas suffisamment abouties, ou n'intègrent pas pleinement les principes de Security by Default et de Security by Design. Par conséquent, une charge supplémentaire est imposée aux développeurs de systèmes d'IA pour faire face aux risques qui doivent être expliqués.

À cette fin, une entreprise doit envisager de mettre en œuvre les mesures suivantes, en plus de ses pratiques de sécurité habituelles :

1

Les dirigeants de l'entreprise doivent avoir conscience des risques de sécurité associés à l'utilisation des services d'IA et apprendre à les gérer.

2

Les politiques de sécurité de l'entreprise doivent être mises à jour pour tenir compte des exigences spécifiques des services d'IA, en veillant à ce que tous les employés et tous les sous-traitants en aient connaissance.

3

La politique doit souligner les risques associés au développement et à l'utilisation de services d'IA, ainsi que les restrictions actuelles concernant leur utilisation, conformément à la législation locale.

4

La politique doit définir les rôles et les responsabilités liés à l'utilisation des services d'IA au sein de l'entreprise.

5

Il convient de mettre au point ou d'acheter une formation sur l'utilisation sûre des services d'IA au sein de l'entreprise, en veillant à ce que tous les employés et tous les nouveaux arrivants la suivent. Le programme doit couvrir les politiques de l'entreprise, les menaces existantes et des exemples d'incidents, les mesures de protection contre les menaces, les lois applicables et d'autres sujets pertinents. Il doit aussi inclure des exercices sur table basés sur des scénarios, le cas échéant. Les employés doivent être testés une fois la formation terminée. La formation doit être mise à jour régulièrement.

6

Les formations existantes en matière de sécurité de l'information doivent être mises à jour pour inclure les nouvelles méthodes utilisées par les acteurs malveillants qui exploitent les services d'IA pour attaquer les entreprises, comme la génération de textes. Il est par exemple question de clonage de voix, de manipulation de photos ou de production de fausses vidéos.

7

Il convient d'organiser un suivi de la législation relative à l'utilisation sûre des services d'IA. Les politiques internes et les programmes de formation doivent être mis à jour en temps utile.

La modélisation des menaces permet d'identifier, de comprendre et de réduire les risques pour la sécurité, dès les premiers stades du développement des systèmes d'IA.

Modélisation des menaces/ évaluation des risques

Ce processus est particulièrement important pour les systèmes d'IA, car il s'agit d'une technologie émergente présentant des risques qui évoluent et s'adaptent en permanence. Procéder à une évaluation des risques peut permettre de relever ces défis. De plus, la modélisation des menaces peut aider les développeurs qui débutent dans ce domaine à mieux se préparer aux défis associés au développement de systèmes d'IA. Elle peut également leur permettre d'identifier et de réduire de manière proactive les faiblesses des systèmes d'IA avant qu'elles ne soient exploitées.

Pour garantir l'efficacité du processus de modélisation des menaces, une entreprise doit suivre les **recommandations suivantes** :

Choisir une méthode d'évaluation des risques (par exemple, STRIDE, DREAD, LINDDUN, PASTA, TRIKE) et développer des procédures d'évaluation des risques et de modélisation des menaces pour les services d'IA. La méthode d'évaluation doit également comprendre un cadre pour déterminer les niveaux de risque, les stratégies de gestion des risques au sein de l'entreprise (y compris les seuils de risque acceptables), les procédures de surveillance des risques et l'affectation du personnel en charge de superviser le processus.

Dans le cadre de la gestion des risques, envisager de classer les menaces dans les catégories suivantes :

- les menaces liées à la non-utilisation des services d'IA ;
- les menaces liées à la non-conformité ;
- les menaces liées à une mauvaise utilisation des services d'IA par les utilisateurs ;
- les menaces pesant sur les modèles d'IA et les ensembles de données utilisés à des fins de formation ;
- les menaces pesant sur les services à cause des modèles d'IA ;
- les menaces pesant sur les données associées ;
- les menaces pesant sur les domaines environnementaux, sociétaux et de gouvernance (ESG).

Une évaluation des risques doit être menée pour tous les services d'IA existants et nouvellement développés.

Une évaluation des risques doit comprendre une identification des acteurs de la menace ou des pirates informatiques potentiels, ainsi que les menaces et les risques identifiés. Des supports de référence comme le NIST-AI-600-1, le MITRE ATLAS, le top 10 de l'OWASP en matière d'applications de grands modèles de langage, le DASF et le CSA doivent être exploités afin d'identifier les menaces et les risques connus.

Les informations relatives aux risques identifiés dans les services d'IA doivent être communiquées à la direction de l'entreprise.

Sécurité des infrastructures (cloud)

Les services d'IA sont généralement fournis sous forme de services cloud et nécessitent souvent une infrastructure spécialisée, par exemple des serveurs équipés de GPU, de FPGA, d'ASIC ou de TPU. Compte tenu de la sensibilité des systèmes d'IA, il convient de les protéger conformément aux **cadres de cybersécurité les plus avancés**, comme le cadre de cybersécurité du NIST ou d'autres normes similaires. Les services d'IA utilisent généralement des logiciels open source ou gratuits comme TensorFlow, PyTorch ou Keras, ainsi que des bibliothèques comme Pandas, NumPy et SciPy. Pour sécuriser cet environnement, les **exigences suivantes** doivent être prises en compte :

1

Identifier toutes les ressources et maintenir un inventaire des ressources informationnelles comme les ensembles de données pour la formation et le test des modèles, les données pour le perfectionnement de la formation, les bases de données, les cartes de données, les modèles et les risques, les données d'entrée et de sortie vers et depuis le service, les poids des modèles et les hyperparamètres, ainsi que les données de journalisation des systèmes des LLM.

2

Contrôler l'accès à tous les niveaux, y compris le réseau, les systèmes d'exploitation, les bases de données, les logiciels, les données et les modèles. Mettre en œuvre une authentification à deux facteurs (2FA) pour l'accès administratif.

3

Enregistrer tous les événements et veiller à ce que les données de journalisation soient protégées. Surveiller les incidents de sécurité et les violations potentielles.

4

Mettre en œuvre des mesures de protection contre les programmes malveillants et d'autres types d'attaques. Appliquer régulièrement des correctifs de sécurité aux modules de l'infrastructure.

5

Segmenter le réseau pour protéger les zones vulnérables. Utiliser le chiffrement pour les données en transit et hors transit.

6

Garantir l'intégrité des données essentielles et vérifier l'authenticité des bibliothèques et des modèles utilisés.

7

Assurer la redondance des serveurs et des canaux de communication. Effectuer des sauvegardes régulières et veiller à leur bon fonctionnement.

8

Stocker les clés en toute sécurité dans un coffre-fort à clés.

9

Appliquer les principes de moindre privilège et de zero trust dans l'ensemble de l'infrastructure.

En fonction de l'infrastructure prenant en charge les services d'IA, des exigences supplémentaires peuvent être requises :

Utiliser une passerelle API pour gérer l'accès aux modèles et l'authentification via les API.

Mettre en œuvre des mesures de sécurité spécifiques aux environnements Kubernetes.

Suivre les bonnes pratiques en matière de sécurisation des services basés sur le cloud.

Assurer l'intégrité du code source, des données de formation, des modèles et des scripts d'automatisation.

Isoler les données de formation, les modèles et les environnements de formation, afin d'éviter toute fuite ou contamination des données.

Chaîne d'approvisionnement et sécurité des données

Les attaques contre la chaîne d'approvisionnement constituent une menace importante pour l'infrastructure de n'importe quelle entreprise, et l'architecture d'IA ne fait pas exception à la règle. Il existe des cas avérés de bibliothèques spécialisées dans la formation de réseaux neuronaux ayant fait l'objet de telles attaques. Cependant, avec l'IA, une préoccupation particulière se pose quant à la sécurité du fournisseur de services et à la protection des modèles de machine learning.

L'accès aux modèles d'IA avancés, en particulier aux LLM, repose souvent sur des solutions basées sur le cloud. Cependant, l'indisponibilité de certains modèles dans certaines régions ainsi que d'autres contraintes peuvent inciter les employés et les développeurs des entreprises à opter pour des services tiers (proxy) qui revendent l'accès aux modèles d'IA via des API pour leurs tâches quotidiennes. Cette pratique présente des risques importants, depuis l'ouverture d'un vecteur supplémentaire de fuites de données en cas d'incident de sécurité sur le service proxy, jusqu'à la pratique non éthique qui consiste à utiliser abusivement les données obtenues pour les revendre ou pour former ses propres versions de LLM. **Il est important de comprendre ces risques**, d'examiner attentivement les politiques de confidentialité du fournisseur principal et du proxy, et de sensibiliser au moyen d'une formation l'ensemble de l'entreprise aux dangers de l'utilisation de services tiers pour les tâches professionnelles.

Pour réduire ces risques, une entreprise peut choisir de déployer un service de LLM local. Cette approche garantit que les données confidentielles traitées par le LLM resteront au sein de l'entreprise si certaines conditions spécifiques (par exemple, isolement du réseau, désactivation de la télémétrie, etc.) sont remplies. Cependant, outre les risques associés aux vulnérabilités des cadres de machine learning, cette méthode comporte des menaces liées aux portes dérobées dans les modèles. Dans ce contexte, cela signifie que les formats de données utilisés pour distribuer les modèles de machine learning peuvent avoir différents niveaux de sécurité, et certains formats permettent potentiellement d'intégrer un code arbitraire qui peut être exécuté lorsque les modèles sont lancés. Des études ont montré qu'il existe des modèles disponibles dans des dépôts publics, bien que limités en nombre, qui peuvent exécuter un code spécifique lors du chargement. L'utilisation de modèles provenant de dépôts tiers, au lieu de modèles originaux, peut être due à l'indisponibilité de modèles téléchargeables dans certaines régions, ou à des restrictions de licence. L'utilisation de formats sécurisés comme safetensors permet de résoudre ce problème, mais il s'avère nécessaire de sensibiliser les développeurs et les analystes de données à l'importance de sélectionner des sources fiables pour les modèles et d'utiliser des formats sécurisés pour échanger des poids de modèles.

- Veiller à ce que les modèles d'IA proviennent de sources fiables et légitimes. Éviter d'utiliser des stockages tiers.
- Utiliser des formats sécurisés comme safetensors pour échanger des poids de modèles, afin d'éviter le risque d'exécution de code arbitraire.
- Mettre en œuvre des mesures pour détecter et répondre aux attaques contre la chaîne d'approvisionnement sur les modules liés à l'IA.
- Évaluer et examiner les politiques de confidentialité des services tiers et des proxys utilisés pour accéder aux modèles d'IA, afin de veiller à ce qu'ils respectent les normes de sécurité.
- Déployer localement des modèles d'IA dans des conditions garantissant la confidentialité des données, par exemple en isolant le réseau et en désactivant les fonctionnalités de télémétrie.
- Établir des protocoles pour le déploiement sécurisé des modèles locaux afin de réduire les risques associés aux portes dérobées potentielles dans les modèles de machine learning.
- Mettre à jour et corriger régulièrement les cadres de machine learning, afin de remédier aux vulnérabilités connues.
- Mettre en œuvre des mesures pour veiller à ce que les données confidentielles traitées par les modèles d'IA ne quittent pas l'infrastructure de l'entreprise.
- Lorsque des API tierces sont utilisées, effectuer des audits de sécurité conformément aux principales normes internationales, comme le top 10 de l'OWASP en matière de sécurité des API.

Test et validation

Une fois l'évaluation effectuée et les risques identifiés, il est essentiel de comprendre comment se prémunir contre les erreurs accidentelles ou délibérées dans la formation et l'application du modèle. Pour atteindre cet objectif, une entreprise peut envisager de mettre en œuvre les mesures suivantes :

1

Évaluer les dommages potentiels qui pourraient être causés par des erreurs accidentelles ou délibérées dans le système. Évaluer la valeur des données utilisées pour former le modèle d'IA et des données qu'il traite.

2

Déterminer si des cadres, des modèles ou des ensembles de données open source sont utilisés pour développer le système d'IA.

3

Identifier la base d'utilisateurs potentiels : employés de l'entreprise, clients ou accès public.

4

Vérifier le respect des bonnes pratiques en matière de machine learning dans le développement des modèles. Veiller à ce que les ensembles de données soient correctement divisés en ensembles de formation, de test et de validation, en fonction du fonctionnement du modèle.

5

Lors de la validation des modèles d'IA et de leurs mesures (faux positifs et faux négatifs), vérifier que les critères de division des ensembles de données sont adaptés à la nature des données (par exemple, division chronologique pour les données temporelles et prévention des fuites de données).

6

Évaluer les critères utilisés par le modèle pour prendre des décisions et déterminer s'ils sont conformes à l'intuition des experts du domaine. Utiliser des méthodes d'interprétation du modèle, comme les vecteurs SHAP, pour comprendre le processus de prise de décision du modèle.

7

Évaluer les performances du modèle dans le monde réel pour veiller à ce qu'il produise les résultats escomptés. Surveiller en continu le modèle, car la distribution des données d'entrée peut varier au fil du temps, ce qui risque de dégrader les performances du modèle.

8

Adapter le plan de test de façon à vérifier si le modèle est sensible aux vulnérabilités propres aux modèles de machine learning (par exemple, les attaques adverses et l'empoisonnement des données).

Rapports sur les vulnérabilités

L'IA est un domaine technologique relativement nouveau, qui évolue rapidement. Malgré les avantages considérables qu'ils présentent, de nombreux systèmes d'IA sont susceptibles de souffrir de vulnérabilités propres à ces technologies.

L'une des principales préoccupations est que certains systèmes d'IA peuvent présenter des vulnérabilités susceptibles d'être exploitées pour obtenir un accès non autorisé à leurs données. Un autre exemple de vulnérabilité des systèmes d'IA est le biais, lorsque les modèles sont formés à partir de données qui ne sont pas représentatives ou qui comprennent des biais cachés. Par exemple, les systèmes d'IA peuvent être affectés par des biais liés à des préjugés lorsque des stéréotypes et des présupposés sociétaux erronés s'infiltrent dans l'ensemble des données de l'algorithme, ou par des biais de mesure induits par des données incomplètes. Par conséquent, ces systèmes peuvent prendre des décisions injustes ou discriminatoires, ce qui a un impact négatif sur les utilisateurs et sape la confiance dans les technologies de l'IA.

Pour résoudre ces problèmes, il est nécessaire de mettre en œuvre un mécanisme qui permettra aux utilisateurs de **signaler les vulnérabilités** et les biais identifiés dans les systèmes d'IA. Ce mécanisme d'établissement de rapports permettra aux entreprises de recevoir un retour d'information rapide et de prendre des mesures :

1. Établir une politique accessible publiquement qui définit les vulnérabilités dans les systèmes d'IA et qui indique comment les utilisateurs peuvent les signaler.
2. Fournir aux utilisateurs des méthodes sécurisées pour signaler les vulnérabilités, comme des formulaires Internet chiffrés ou des adresses email dédiées.
3. Définir des procédures pour évaluer rapidement les vulnérabilités signalées, les classer par ordre de priorité et y remédier.
4. Communiquer avec la personne qui a signalé les vulnérabilités au sujet de l'état et de la résolution du problème.
5. Tenir les utilisateurs informés des vulnérabilités connues et des efforts de remédiation, afin d'instaurer une confiance et de démontrer une certaine responsabilité.
6. Collaborer avec des chercheurs en sécurité dans le cadre de programmes de recherche de bugs. Cela vous permettra également de vous tenir au courant des menaces émergentes et des bonnes pratiques en matière de sécurité de l'IA.

Défense contre les attaques spécifiques au machine learning

Compte tenu des progrès actuels dans le développement de l'IA, certains modules de l'IA peuvent être vulnérables à des attaques spécifiques au machine learning. Ces attaques peuvent exploiter les vulnérabilités en introduisant délibérément des données mal formées ou des commandes cachées dans le modèle, par exemple. Les entreprises qui utilisent l'IA en accès libre pour développer leurs systèmes doivent donc avoir conscience de ces risques. La protection contre les attaques spécifiques au machine learning nécessite la **mise en œuvre de diverses mesures de sécurité** :

Introduire des exemples contradictoires* dans l'ensemble de données de formation, afin d'aider le modèle à apprendre à traiter ces données d'entrée de manière plus efficace.

Appliquer des techniques de distillation qui contribuent à rendre le modèle plus résistant aux données contradictoires en simplifiant son processus de prise de décision.

Envisager l'utilisation de modèles monotones**, qui peuvent améliorer la stabilité et réduire la vulnérabilité aux manipulations adverses.

Introduire des systèmes capables de détecter des données d'entrée adverses ou anormales dans les demandes des utilisateurs, ce qui permet au modèle de détecter et de rejeter les tentatives malveillantes avant de traiter les données.

Pour éviter un empoisonnement des données, les développeurs de systèmes d'IA doivent **analyser les échantillons de formation** afin de détecter les objets anormaux, et comparer les performances des nouveaux modèles avec celles des versions précédentes afin d'identifier tout changement important dans les propriétés du modèle.

Pour éviter les attaques par injection de prompts contre les LLM, les développeurs de systèmes d'IA peuvent mettre en œuvre un système qui analyse les demandes entrantes des utilisateurs ou d'autres données tierces introduites dans l'entrée des LLM. Une autre approche consiste à analyser les réponses à ces demandes et à évaluer leur conformité avec la tâche actuelle du système.

* Comment perturber les réseaux neuronaux anti-malwares : attaques adverses et protection, https://securelist.com/how-to-confuse-antimalware-neural-networks-adversarial-attacks-and-protection/102949/

** Modèles monotones pour la détection dynamique en temps réel des malwares, https://arxiv.org/pdf/1804.03643

Mises à jour et maintenance régulières en matière de sécurité

Le domaine de l'IA, et en particulier l'application des LLM, est encore relativement jeune, et la qualité du code ne répond pas toujours à des normes élevées. Par conséquent, de nombreux cadres et outils utilisés pour travailler avec le machine learning peuvent présenter un **nombre important de vulnérabilités**. Heureusement, nous nous trouvons aujourd'hui dans une phase où les cadres populaires sont activement portés à des normes de qualité de production, avec des versions et des mises à jour de sécurité régulières. De plus, des programmes de recherche de bugs permettant de trouver des vulnérabilités dans les infrastructures de machine learning sont en train d'émerger, ce qui permet de résoudre rapidement ces problèmes.

Ce point souligne l'importance d'une surveillance continue de l'état de votre infrastructure, allant des plateformes utilisées pour suivre les expériences aux bibliothèques conçues pour communiquer avec les services cloud. La mise à jour de l'infrastructure peut prendre plus de temps que nécessaire pour les projets relevant de secteurs qui évoluent plus lentement. De plus, l'utilisation des dernières versions des bibliothèques peut entraîner des problèmes de compatibilité, ce qui implique un investissement plus important dans le développement et le maintien de la fonctionnalité du code qui s'appuie sur ces bibliothèques. Ces coûts doivent être pris en compte lors de la planification d'initiatives en matière d'IA.

Un autre risque associé à l'utilisation de modèles d'IA basés sur le cloud, comme les LLM, est le **cycle de vie relativement court de chaque version du modèle**. Le modèle sélectionné pour un projet peut être remplacé par le fournisseur de la plateforme par une nouvelle version, dans un court laps de temps. Si la qualité globale des modèles est censée s'améliorer, leur comportement pour des tâches particulières et leur résistance aux attaques peuvent changer. Il peut par exemple être question d'injections de prompts ou de tentatives d'obtention de résultats non autorisés par le biais de jailbreaks. Une planification avancée est donc nécessaire afin d'assurer une transition en douceur entre les modèles, sans compromettre la qualité des tâches en aval ou le niveau de sécurité :

1. Veiller à ce que l'infrastructure soit maintenue à jour avec les derniers correctifs de sécurité et les dernières mises à jour du cadre.
2. Participer activement aux programmes de recherche de bugs et utiliser des outils d'analyse des vulnérabilités, afin de détecter les faiblesses des cadres de machine learning et de l'infrastructure d'IA.
3. Examiner et appliquer régulièrement les mises à jour de sécurité pour les outils et les bibliothèques de machine learning, afin de réduire l'exposition aux vulnérabilités connues.
4. Prévoir les problèmes de compatibilité potentiels lors de l'utilisation des dernières versions des bibliothèques et des cadres en allouant des ressources de développement et de test.
5. Mettre en œuvre une stratégie de gestion du cycle de vie des modèles d'IA basés sur le cloud, avec des plans de transition vers les nouvelles versions des modèles au fur et à mesure de leur publication par le fournisseur.

Respect des normes internationales

Alors que nous assistons à un développement rapide de la réglementation en matière d'IA, la conformité aux lois en vigueur et le respect des bonnes pratiques deviennent de plus en plus importants. Tout d'abord, les données relatives à la formation de l'IA peuvent être collectées auprès de diverses sources dans différentes juridictions, ce qui complique le traitement et l'utilisation de ces informations. De plus, les modèles proviennent souvent de **dépôts ouverts**, ce qui ajoute un degré d'incertitude quant à leur conformité avec les exigences légales d'une juridiction donnée.

Les développeurs d'IA sont donc confrontés à une tâche difficile, qui consiste à assurer la conformité avec toutes les obligations juridiques dans les pays où le système sera utilisé. Dans cette situation, la meilleure stratégie est de **suivre les normes des principaux acteurs en matière de réglementation de l'IA**, comme la Chine, l'Union européenne ou les États-Unis. De nombreux pays partagent déjà leurs approches et mettent en œuvre des exigences similaires, ce qui permet aux développeurs d'anticiper le déploiement mondial de leur système :

1

Établir des lignes directrices pour l'utilisation et le développement éthiques de l'IA, afin de garantir la transparence et la responsabilité dans les processus associés.

2

Veiller à ce que toutes les données collectées à partir de sources distinctes soient conformes aux lois sur la confidentialité des données dans chaque juridiction, comme le règlement général sur la protection des données (RGPD) en Europe ou la loi californienne sur la protection de la vie privée des consommateurs (CCPA) aux États-Unis.

3

Lors de l'utilisation de modèles d'IA provenant de stockages ouverts, vérifier qu'ils respectent les droits de propriété intellectuelle.

4

Suivre les principaux cadres légaux, comme l'AI Act de l'Union européenne ou la charte des droits de l'IA des États-Unis, car ils sont souvent utilisés comme référence par d'autres pays.

5

Se tenir au courant des nouvelles réglementations en matière d'IA et de leur évolution à travers le monde.

6

Vérifier régulièrement la conformité des modèles et des systèmes d'IA avec les normes internationales, afin d'identifier et de réduire les risques juridiques et éthiques potentiels.

Conclusion

Comme la plupart des innovations technologiques, les technologies de l'IA présentent à la fois de grandes opportunités et des menaces importantes. Les risques de cybersécurité associés à l'IA et l'impact de celle-ci sur la société dépendent du comportement et des intentions du développeur. Pour mettre en œuvre l'IA en toute sécurité, les entreprises doivent suivre des conseils techniques sur la manière de développer et de déployer l'IA dans leur infrastructure, car la mise en œuvre de ce processus sans recommandations adaptées peut présenter des risques importants. Il est essentiel que les entreprises instaurent une culture de la sécurité et de la responsabilité tout au long du cycle de vie de l'IA et qu'elles intègrent des contrôles de sécurité élémentaires, allant de l'évaluation des risques et du test des systèmes à la sécurisation des chaînes d'approvisionnement et à la maintenance en continu. Une mise en œuvre réussie des exigences présentées contribuera à **réduire les risques** liés à l'introduction de systèmes d'IA dans les activités d'une entreprise.

En savoir plus

www.kaspersky.fr

#kaspersky
#bringonthefuture