Établir un cadre de gouvernance pour les applications d'IA

Avec les progrès rapides de l'intelligence artificielle, les organisations de tous les secteurs déploient des applications d'IA à un rythme sans précédent. Dans le rapport « Sécurité cloud-native : état des lieux 2024 » de Prisma Cloud[1], 100 % des sondés disent ainsi s'approprier les outils de développement applicatif assisté par IA. Un chiffre surprenant, quand on sait que beaucoup considéraient cette technologie comme de la science-fiction il y encore quelques années. Mais bien que les systèmes d'IA offrent de réels avantages, ils présentent également de nouveaux risques de sécurité et défis de gouvernance, au-delà des capacités des solutions traditionnelles de cybersécurité.

En tant que responsable sécurité, il est impératif de bien comprendre le paysage actuel de l'IA, ses implications potentielles et les risques uniques qu'elle fait peser sur votre organisation. Dans ce livre blanc, nous dressons un état des lieux de l'IA et de ses principaux enjeux de sécurité, et proposons un cadre de gouvernance pour vous aider à mieux vous orienter dans cet univers complexe. Vous aurez ainsi une idée plus claire des principales technologies d'IA disponibles, des problèmes de sécurité qu'elles peuvent engendrer et des éventuelles protections à mettre en place. Enfin, nous ferons le point sur les facteurs à prendre en compte pour élaborer votre stratégie de sécurité IA.

1 https://www.paloaltonetworks.fr/state-of-cloud-native-security

IA : état des lieux et enjeux de sécurité

Le paysage de l'IA a considérablement évolué ces dernières années. Des avancées significatives ont été réalisées, tant en matière de systèmes généralistes que spécialisés, à mesure que les organisations évaluent et implémentent une foule de nouvelles technologies.

Mais pour bien gérer les risques et les opportunités liés à l'IA, les responsables sécurité doivent rester en prise avec cet écosystème en constante évolution. Comme le montre notre étude récente, le risque peut se présenter sous de multiples aspects. Ainsi, 47 % des organisations interrogées se disent préoccupées par les risques de sécurité associés au code généré par IA[2].

Dans ce document, nous nous pencherons plus particulièrement sur les risques qui pèsent sur l'infrastructure cloud hébergeant des applications d'IA.

Pour commencer, dressons ensemble un tour d'horizon du paysage de l'IA.

IA et ML « traditionnels »

Depuis de nombreuses années, les entreprises exploitent les outils d'inférence IA et ML spécifiques à différents domaines de connaissances. L'objectif : renforcer l'efficacité et l'automatisation dans l'industrie, optimiser les supply chains, détecter les fraudes, booster l'efficacité du marketing en ligne, etc. Ces systèmes d'IA étroite (Narrow AI) sont entraînés sur des jeux de données (datasets) particuliers afin d'exécuter des tâches bien définies. Bien qu'ils fassent désormais partie intégrante de nombreux processus métiers, ils sont généralement limités dans leur champ d'application.

Ces outils ne datent pas d'hier et n'ont pas connu d'évolution majeure ces dernières années, mais l'essor de l'IA leur est favorable car il crée un effet d'entraînement : face à l'enthousiasme général, on observe un engouement renouvelé et une augmentation des budgets consacrés au déploiement de ces technologies et des nouveaux outils basés sur les LLM.

Enjeux de sécurité

- Garantir l'intégrité et la sécurité des données utilisées pour entraîner et opérer les modèles d'IA/ML
- Surveiller les systèmes d'IA/ML pour détecter les comportements anormaux, les résultats inattendus ou les dégradations de performances symptomatiques de problèmes de sécurité (« empoisonnement » délibéré des données d'entraînement, attaques par exfiltration de modèle, compromission du système, etc.).
- Auditer régulièrement les modèles pour s'assurer qu'ils n'introduisent pas de biais involontaires, de problèmes de partialité ou de résultats discriminatoires susceptibles de causer un préjudice juridique ou réputationnel
- Mettre en place des contrôles d'accès et un cadre de gouvernance appropriés quant aux personnes autorisées à développer, modifier ou utiliser les systèmes IA/ML

IA et ML traditionnels

Cas d'usage

- Optimisation des processus industriels, supply chain, détection des fraudes, marketing

Risques de sécurité

- Intégrité des données, suivi des performances du modèle, détection des biais, contrôle des accès

2 https://www.paloaltonetworks.fr/state-of-cloud-native-security

Grands modèles de langage (LLM)

Les LLM se sont imposés comme des systèmes d'IA généralistes d'une incroyable polyvalence, capables de répondre aux questions ouvertes des utilisateurs, de générer des textes cohérents et même de créer du code. Les chatbots grand public comme ChatGPT font énormément parler d'eux, mais ils ne représentent que la partie émergée de l'iceberg. Le principal impact des LLM réside dans leur capacité à alimenter tout un éventail d'applications d'entreprise – aussi bien internes qu'en prise directe avec les clients – d'une façon souvent imperceptible pour les utilisateurs.

Dans ce contexte plus large, on observe différents schémas d'implémentation, chacun associé à des outils et des problématiques de sécurité spécifiques.

LLM

	Cas d'usage	Enjeux de sécurité
Modèles de base préentraînés	• Génération de contenu, chatbots, opinion mining, traduction, assistants de code	• Shadow AI, confidentialité des données, gouvernance des modèles, contrôle des résultats
Fine-tuning et RAG	• Assistants IA spécialisés (service client, RH, IT), applications de questions-réponses (documentation, code, supports de formation)	• Exposition de données sensibles pendant le fine-tuning, gouvernance des données
Entraînement de modèle personnalisé	• Applications de pointe (mise au point de médicaments, science des matériaux, systèmes autonomes)	• Empoisonnement délibéré des données d'entraînement, isolement des ressources informatiques, responsabilité et auditabilité des modèles

Utilisation de LLM préentraînés (propriétaires ou open-source)

Certains fournisseurs cloud, comme OpenAI et Anthropic, proposent un accès API à de puissants LLM, dont ils assurent la gestion et la protection. À l'aide de ces API, les organisations peuvent enrichir leurs applications de capacités LLM sans devoir gérer l'infrastructure sous-jacente.

Autre possibilité : exécuter des LLM open-source, comme le modèle LLaMa de Meta, sur l'infrastructure de l'entreprise. Cette solution présente l'avantage de fournir davantage de contrôle et d'options de personnalisation. Revers de la médaille, elle requiert d'importantes ressources de calcul, tandis que son implémentation et sa maintenance nécessitent une certaine expertise de l'IA.

Il existe différents modèles de déploiement pour les LLM :

- **SaaS basé sur des API –** L'infrastructure est fournie et gérée par le développeur du LLM (par ex. OpenAI) et provisionnée via une API publique.
- **Gestion par les CSP –** Le LLM est déployé sur l'infrastructure d'un fournisseur de services cloud (CSP), et peut être exécuté sur un cloud public ou privé (par ex. Azure, OpenAI, Amazon Bedrock).
- **Gestion par l'entreprise –** Le LLM est déployé sur l'infrastructure on-prem de l'entreprise (uniquement valable pour les modèles open-source ou développés en interne).

Cas d'usage types

Parmi les cas d'usage des LLM, on peut citer la génération de contenu, les chatbots, l'analyse de ressenti (ou opinion mining), la traduction linguistique et les assistants de code. À titre d'exemple, une entreprise d'e-commerce peut utiliser un LLM pour générer des descriptions de produits, et un éditeur de logiciels peut avoir recours à un assistant de code pour booster la productivité de ses développeurs.

Enjeux de sécurité

De par leur disponibilité et leur facilité d'accès, les API cloud et les modèles open-source ont considérablement simplifié l'ajout de capacités IA avancées aux applications. Les développeurs peuvent intégrer facilement des LLM à leurs produits logiciels, même sans expertise approfondie de l'IA et du ML. Côté pile, cela accélère l'innovation. Mais côté face, on observe également un risque accru de Shadow AI, ces projets d'IA échappant à tout contrôle en matière de sécurité et de conformité. De leur côté, les développeurs peuvent expérimenter les LLM sans prendre en compte les questions de confidentialité des données, de gouvernance des modèles et de contrôle des résultats.

> La disponibilité des API cloud et des modèles open-source augmente les risques de Shadow AI.

Fine-tuning et génération augmentée par récupération (RAG)

Afin d'affiner les LLM pour des contextes particuliers, il est possible de les entraîner sur des datasets plus petits mais aussi plus spécifiques au domaine en question. Les organisations peuvent également avoir recours à la génération augmentée par récupération, qui consiste à associer les LLM à des bases de connaissances externes pour améliorer la qualité des réponses et des résumés de contenus.

Cas d'usage types

Parmi les cas d'usage, citons les assistants IA disposant d'un accès aux données internes de l'entreprise (service client, RH, support IT…) ou les applications de questions-réponses (documents, dépôts de code, supports de formation, etc.). À titre d'exemple, une entreprise de télécommunications peut décider d'optimiser son chatbot de service client en entraînant le modèle sur de la documentation produit, des FAQ ou d'anciennes interactions d'assistance.

Enjeux de sécurité

Le fine-tuning et la RAG permettent aux organisations d'adapter les LLM à leur domaine d'activité et à leurs données, avec à la clé des résultats plus ciblés et plus précis. Mais attention : lors de ce processus, il n'est pas rare que le modèle soit exposé à des informations internes sensibles. D'où l'importance capitale de l'encadrer par une solide gouvernance des données pour que seules les données autorisées soient utilisées pour le fine-tuning, et que le modèle qui en résulte soit parfaitement sécurisé.

> Le fine-tuning et la RAG peuvent exposer les modèles à des informations internes sensibles.

Entraînement des modèles

De grands acteurs de la tech et d'éminents instituts de recherche choisissent d'entraîner leurs propres LLM à partir de zéro. Ce processus mobilise énormément de ressources, une puissance de calcul considérable et d'immenses volumes de données. En contrepartie, ces organisations ont un contrôle total sur l'architecture de leur modèle, sur ses données d'entraînement et sur son optimisation. Autre avantage clé : elles disposent des droits exclusifs de propriété intellectuelle sur les modèles ainsi créés.

Cas d'usage types

Les LLM propriétaires sont utilisés sur des domaines d'application hautement spécialisés : mise au point de nouveaux médicaments, science des matériaux, conception de systèmes autonomes, etc. Un acteur de la santé pourra, par exemple, développer un modèle visant à diagnostiquer les patients à partir de leurs informations médicales et de leurs données d'imagerie.

Enjeux de sécurité

L'entraînement de LLM personnalisés exige à la fois d'immenses datasets et une infrastructure informatique extrêmement puissante, deux éléments susceptibles d'introduire de nouveaux risques. Pour commencer, il est essentiel de s'assurer que les données d'entraînement du modèle ne contiennent aucune information sensible ou personnelle. Autre danger possible : les actes délibérés d'empoisonnement des données d'entraînement. Ces attaques visent à modifier le comportement du modèle en « injectant » des données corrompues.

Le processus d'entraînement étant extrêmement gourmand en puissance de calcul, son environnement doit être isolé et soumis à des contrôles d'accès stricts afin d'éviter toute interférence ou utilisation abusive. Lorsque les modèles incluent des données sensibles difficiles à détecter (on parle alors de « boîtes noires »), la responsabilité et l'auditabilité de leurs comportements suscitent des points d'interrogation. Des questions d'autant plus cruciales quand il s'agit d'un modèle maison entraîné sur des données internes.

> L'entraînement des LLM personnalisés nécessite la vérification approfondie des données d'entraînement pour éviter les informations sensibles et les données personnelles

Comment conceptualiser le risque IA

C'est bien connu : les équipes de cybersécurité ont généralement un temps de retard. Le plus souvent, elles doivent adapter leurs stratégies et leurs systèmes de contrôle en réaction à l'adoption de nouvelles technologies (cloud computing, containérisation, architectures sans serveur, etc.). Mais l'essor de l'IA confronte la cybersécurité à de nouveaux obstacles très différents de ceux qu'elle a pu rencontrer jusqu'alors.

Un changement plus rapide (et souvent, plus extrême)

L'IA a eu l'effet d'un tremblement de terre dans de nombreuses entreprises, tant par son ampleur que par sa vitesse de propagation bien supérieure aux précédents chocs technologiques :

- De nouveaux modèles, techniques et applications font leur apparition à un rythme effréné. Il ne se passe pas un mois, voire pas une semaine, sans qu'une avancée majeure ne soit annoncée.
- Les entreprises se sentent obligées d'adopter rapidement ces technologies pour gagner en compétitivité et booster leur innovation.
- La disponibilité de ces outils via de simples API, ainsi que l'émergence d'un écosystème dédié de solutions et de frameworks ont accéléré leur adoption en éliminant les obstacles liés aux pénuries de compétences IA.

Côté sécurité, cette combinaison de bouleversements technologiques rapides et d'adoption dans l'urgence par les entreprises crée un cocktail explosif. Quand les délais sont serrés, difficile de prendre toute la mesure des risques et de mettre en œuvre les contrôles appropriés en amont du déploiement de l'IA. Dans la précipitation, la sécurité passe au second plan. Quant aux bonnes pratiques et aux processus de mise en conformité, ils ont souvent un temps de retard sur la technologie. Résultat : les équipes de sécurité sont contraintes d'improviser et de prendre des décisions en l'absence de tout consensus sectoriel ou de directives claires.

Heureusement, la situation offre une nouvelle opportunité : celle de sécuriser l'IA dès la conception. En tenant compte des questions de sécurité et de gouvernance dès les premières étapes du développement de leurs modèles d'IA, les organisations peuvent réduire les risques de manière proactive et renforcer la résilience de ces systèmes. Pour ce faire, les équipes juridique, de sécurité et de développement de l'IA doivent collaborer étroitement afin de garantir l'alignement sur les bonnes pratiques et d'intégrer les contrôles nécessaires tout au long du cycle de vie des systèmes IA.

Les ramifications plus larges

L'impact de l'IA est extrêmement vaste, et beaucoup n'en prennent pas toute la mesure. Les modèles d'IA peuvent automatiser les décisions à fort enjeu, générer du contenu ayant des implications légales (par ex. utilisation de contenu protégé par des droits d'auteur), ou encore accéder à d'immenses volumes de données sensibles. Résultats biaisés, violation de la vie privée, exposition de la propriété intellectuelle, utilisation malveillante… les risques liés à l'IA imposent un changement de paradigme radical quant à la gestion des risques.

Pour résoudre ces problématiques dans toutes leurs ramifications possibles, les responsables cybersécurité doivent se concerter avec les parties prenantes des fonctions juridique, éthique et métier. En ce sens, il convient d'établir des structures de gouvernance collaboratives afin de définir la tolérance au risque, les bonnes pratiques et les directives, et de mettre en place une surveillance et des audits continus. Enfin, les équipes cybersécurité doivent travailler au contact direct des équipes d'ingénierie et de science des données pour intégrer la gestion de la sécurité et des risques au cycle de vie de l'IA.

Les équipes cybersécurité doivent travailler au contact direct des équipes d'ingénierie et de science des données pour intégrer la gestion de la sécurité et des risques au cycle de vie de l'IA.

De nouveaux mécanismes de supervision de la sécurité et de la conformité s'imposent

La plupart des stratégies de cybersécurité traditionnelles visent à protéger la confidentialité, l'intégrité et la disponibilité des données. Or, avec l'IA, il faut compter avec de nouvelles contraintes comme l'impartialité, la transparence et la responsabilité.

À l'instar de la loi sur l'IA de l'Union européenne, de nouveaux cadres réglementaires imposent aux entreprises une obligation de mise en place de mécanismes de surveillance et de gouvernance des systèmes d'IA. En vertu de ces réglementations, elles doivent évaluer et réduire les risques associés à leurs applications d'IA, en particulier dans des domaines sensibles (recrutement, solvabilité, maintien de l'ordre public, etc.). Notons par ailleurs que ces obligations peuvent différer en fonction du cas d'usage et du niveau de risque. Les systèmes d'IA utilisés dans le cadre du recrutement seront par exemple soumis à des exigences plus strictes en matière d'audit et de transparence afin de garantir l'absence de tout biais discriminatoire dans les modèles.

Les équipes sécurité ne peuvent donc plus se contenter de contrôler les accès et de protéger les données. En collaboration avec les équipes conformité et juridique, elles doivent établir des mécanismes de surveillance et de validation des réponses et décisions fournies par les modèles d'IA. Il peut s'agir d'implémenter des techniques d'IA explicable, de réaliser des audits réguliers pour détecter les éventuels biais, ou de consigner en détail les réponses et résultats du modèle, ainsi que sa logique décisionnelle, pour étayer les rapports et enquêtes de conformité.

Détection et neutralisation des menaces : les nouveaux défis

À la fois inédites et complexes, les problématiques techniques liées à la sécurité des systèmes d'IA concernent aussi bien la détection que la neutralisation des menaces.

Les menaces émergentes imposent de nouvelles méthodes de détection

Comme évoqué plus haut, les équipes de sécurité doivent surveiller non seulement les données sous-jacentes et les artefacts des modèles, mais également les résultats et comportements des systèmes d'IA en production. Ils doivent pour cela analyser une véritable mine de données non structurées et détecter de nouveaux types de menaces (empoisonnement des données d'entraînement, attaques par exfiltration de modèle, biais potentiellement préjudiciables, etc.) pour lesquelles la plupart des outils de sécurité actuels n'ont pas été conçus.

Une résolution particulièrement complexe

Quand des logiciels traditionnels présentent des vulnérabilités, il suffit généralement de quelques lignes de code pour les corriger. Mais quand un problème apparaît sur un modèle d'IA, il faut souvent le réentraîner de A à Z. Explication : les modèles d'IA apprennent de leurs données d'entraînement, qui se retrouvent profondément ancrées dans leurs paramètres. Or, s'il s'avère que les données d'entraînement contiennent des informations sensibles, des biais ou des données corrompues, il est tout simplement impossible de supprimer ou corriger uniquement les éléments qui posent problème. Il faut obligatoirement réentraîner le modèle à partir d'un dataset « nettoyé » – un processus qui peut prendre des semaines, voire des mois entiers, et qui mobilise des ressources humaines et informatiques à hauteur de plusieurs centaines de milliers de dollars.

Par ailleurs, cette solution est loin d'être une panacée : il peut en résulter une dégradation des performances du modèle, voire de nouveaux problèmes. Bien que diverses techniques comme le « machine unlearning » et la suppression des données fassent l'objet de recherches, elles n'en sont qu'à leurs prémices et leur champ d'application reste limité. Les mesures de résolution réactives étant coûteuses et chronophages, il est essentiel de se concentrer sur la prévention et la détection précoce des vulnérabilités de l'IA.

Cadres de gouvernance recommandés pour les applications d'IA

Pour aider les organisations à gérer efficacement les risques et les opportunités de leurs applications d'IA, nous leur conseillons d'adopter de nouveaux cadres de gouvernance axés sur les deux piliers que sont la visibilité et le contrôle.

La visibilité consiste à bénéficier d'un tableau clair de l'utilisation de l'IA dans votre organisation. Pour cela, vous devez dresser l'inventaire de tous vos modèles d'IA déployés, suivre les données utilisées pour entraîner et opérer ces modèles, et documenter les capacités et les autorisations d'accès de chaque modèle. Sans cet élément fondamental, impossible d'évaluer les risques ou de faire appliquer vos politiques.

Le contrôle renvoie aux politiques, processus et protections techniques à implémenter pour garantir une utilisation responsable de l'IA, en phase avec les valeurs de votre entreprise. Il englobe les politiques de gouvernance des données (qui spécifient quelles informations peuvent servir à entraîner l'IA), les contrôles d'accès (qui déterminent qui est autorisé à développer et déployer les modèles), ou encore les mesures de surveillance et d'audit en continu (qui valident les comportements et performances des modèles).

L'objectif : établir une approche structurée qui permet aux responsables sécurité de collaborer avec les parties prenantes des différentes fonctions de l'organisation – conformité, ingénierie, etc. – afin de concevoir et d'implémenter les mécanismes de gouvernance appropriés pour l'IA. Les détails de cette implémentation et les politiques associées peuvent varier selon les impératifs et les priorités de chaque organisation, et en fonction des réglementations locales en vigueur.

	Visibilité	Contrôles (politiques)
Modèles	Quels modèles sont utilisés ?	01 Modèles autorisés et non autorisés 02 Chaîne d'approbation pour l'entraînement/le déploiement/ l'utilisation des nouveaux modèles
Données	Quelles données servent à l'entraînement, à l'inférence et au fine-tuning du modèle ?	01 Politiques relatives à l'utilisation acceptable des données sensibles 02 Surveillance du stockage, du traitement et des flux de données
Cas d'usage	À quelles fins l'IA est-elle utilisée ?	01 Cas d'usage autorisés et non autorisés 02 Politiques relatives à l'utilisation des agents IA
Accès	Par qui l'IA est-elle utilisée ?	01 Surveillance renforcée des applications d'IA grand public 02 Politiques de contrôle
Conformité	Quelles sont les réglementations en vigueur ?	01 Prise en compte des risques actuels et futurs 02 Responsabilité assumée en cas d'infractions

Modèles

Visibilité sur l'inventaire des modèles

Commencez par dresser l'inventaire complet des modèles actuellement déployés au sein de votre organisation. Vous disposerez ainsi d'un référentiel centralisé pour tout l'écosystème IA de votre entreprise, base indispensable à l'évaluation des risques et à l'application des politiques. Veillez à inclure dans votre inventaire les métadonnées essentielles comme le fournisseur du modèle, ses finalités métiers et ses cas d'usage. Pensez également à mentionner le type du modèle (par ex. LLM), les modèles de vision par ordinateur, les données tabulaires, les sources des données d'entraînement, les autorisations et restrictions d'accès, et les diagrammes de flux de données.

Les outils de détection automatique peuvent vous aider à identifier les modèles déployés dans vos environnements de production. Cependant, vous devrez peut-être recourir à des processus manuels pour capturer les modèles en cours de développement ou hébergés en externe.

Politiques relatives aux modèles autorisés et non autorisés

Vous devez déterminer quels modèles peuvent être utilisés, évalués et déployés dans des environnements en prise directe avec le client. Assurez-vous d'aligner vos politiques sur la tolérance au risque globale de votre organisation et sur vos impératifs en matière de conformité.

Les modèles autorisés doivent répondre aux critères suivants :

- Provenance fiable (par ex. fournisseur réputé vs open-source)
- Tests et validation suffisants
- En accord avec les valeurs et principes de votre entreprise pour une IA responsable
- Respect des réglementations et des normes sectorielles en vigueur
- Intégration avec les contrôles de sécurité et de gouvernance

Bloquez le déploiement des modèles qui ne respectent pas ces critères, ou soumettez-les à des processus d'approbation bien plus stricts. Pour les modèles autorisés, établissez des directives portant sur les cas d'usage appropriés, les contrôles de sécurité et de conformité requis, et les exigences en matière de surveillance et de maintenance continues. Ces directives vous aideront à garantir la cohérence et la gestion des risques lors du déploiements de vos modèles.

Politiques de vérification et validation des nouveaux modèles d'IA

Les équipes sécurité et conformité doivent pouvoir s'appuyer sur des processus structurés pour valider les modèles d'IA, que ceux si soient nouvellement créés ou pré-déployés. Il s'agit de vérifier que chaque modèle répond aux critères d'autorisation définis, puis d'identifier ses risques spécifiques ou les mesures de contrôle à appliquer.

Une fois les modèles validés, il est essentiel de suivre des processus bien définis pour tout leur cycle de vie : du développement à la production, en passant par la préproduction. Chaque étape doit prévoir des tests et recevoir le feu vert avant de progresser au stade suivant. Vous veillerez également à créer des processus de surveillance continue des modèles déployés, avec examen automatique en cas de changement important (réentraînement, dérive des données, dégradation des performances, etc.)

Données

Détection et classification des données d'entraînement et de déploiement des modèles d'IA

La data est la matière grise des modèles d'IA. Pour cette raison, il est crucial de disposer d'une visibilité totale sur les données utilisées pendant tout le cycle de vie de l'IA : données d'entraînement, données d'inférence ou de prédictions en production, et données utilisées pour le fine-tuning ou le réentraînement des modèles au fil du temps. Assurez-vous de maintenir à jour l'inventaire de vos assets IA. Celui-ci doit recenser tous vos datasets, catégorisés selon leur niveau de sensibilité, leurs exigences réglementaires et leurs cas d'usage autorisés. Cet inventaire doit s'inscrire en complément de celui utilisé pour les modèles, et ce afin d'assurer une traçabilité totale entre chaque modèle et ses données associées.

Politiques de prévention des empoisonnements de données

La gouvernance des données d'entraînement est particulièrement cruciale dans le contexte des actes délibérés d'empoisonnement de données. Un attaquant qui parvient à manipuler les données d'entraînement d'un modèle pourra introduire des backdoors ou des biais susceptibles de corrompre le comportement de celui-ci en production. Pour réduire le risque, il est impératif d'adopter de robustes mesures de contrôle et de surveillance continue des flux de données.

Politiques d'utilisation de données sensibles pour l'entraînement, l'inférence et le fine-tuning

L'utilisation des différents types de données pour vos modèles d'IA doit être régie par des politiques claires. Pour définir celles-ci, tenez compte du niveau de sensibilité des données, des contraintes réglementaires et des considérations éthiques.

Vos politiques peuvent stipuler, par exemple, que les données à caractère personnel ou les informations médicales ne peuvent être utilisées pour l'entraînement des modèles que si elles sont anonymisées, et uniquement avec le consentement des personnes concernées. L'utilisation de données tierces ou d'informations de propriété intellectuelle sensibles peut également être soumise à restrictions en l'absence de licences et de droits d'utilisation appropriés. Vos politiques peuvent aussi définir des contrôles de sécurité et de confidentialité pour différents types de données. Parmi ceux-ci : le chiffrement, les contrôles d'accès ou les durées maximales de conservation. Votre équipe conformité doit participer à la création de ces politiques afin de garantir le respect des réglementations en vigueur (RGPD, HIPAA, CCPA).

Cas d'usage

Visibilité sur l'utilisation de l'IA

Les implications en termes de sécurité et de conformité diffèrent grandement selon les cas d'usage de l'IA. Un chatbot IA de support client, par exemple, peut nécessiter des contrôles stricts pour la confidentialité des données et le filtrage des contenus. À l'inverse, un outil IA interne utilisé pour optimiser la chaîne logistique ne soulèvera pas les mêmes préoccupations.

Les éléments suivants doivent impérativement figurer dans la documentation des cas d'usage :

- Finalités métiers et résultats attendus
- Utilisateurs et parties prenantes
- Données d'entrée et de sortie
- Autonomie et latitude pour la prise de décisions
- Surveillance humaine et points d'intervention
- Métriques de performance et critères de succès
- Évaluation des risques et mesures de réduction

Politiques relatives aux cas d'usage autorisés et non autorisés

Les organisations doivent définir des politiques délimitant clairement les cas d'usage autorisés de l'IA, et sous quelles conditions. Principaux critères à prendre en compte pour l'autorisation des cas d'usage :

- Conformité avec les valeurs et principes de votre entreprise pour une IA responsable
- Respect des lois, réglementations et normes sectorielles en vigueur
- Risques de préjudices ou de conséquences involontaires
- Niveau de surveillance humaine et de contrôle
- Transparence et exigences d'explicabilité
- Niveau de risque réputationnel

Les cas d'usage à fort enjeu, qui impliquent des décisions concernant des personnes (crédit, recrutement, diagnostics médicaux, etc.), doivent faire l'objet de contrôles renforcés. Certains peuvent même nécessiter des comités dédiés d'examen ou de supervision. Pour les cas d'usage à faible risque, comme les outils de productivité internes, des processus d'approbation plus simples seront suffisants.

Définir des politiques pour l'utilisation autorisée des agents IA

Les agents IA constituent une catégorie à part de systèmes d'IA. Ils peuvent prendre des décisions de manière autonome et exécuter des tâches dans le prolongement logique de leurs résultats précédents, le tout sans intervention humaine. Les agents IA peuvent notamment écrire, tester et optimiser du code. L'utilisation de ces agents introduit de nouveaux défis de gouvernance, car les risques potentiels et les conséquences involontaires associés sont plus difficiles à prédire et à contrôler. Les organisations doivent donc soumettre l'utilisation de ces systèmes à des politiques clairement définies (quand ? comment ?). Celles-ci doivent notamment :

- Déterminer le périmètre de prise de décision de l'agent
- Limiter le champ d'action et mettre en place des mécanismes de sécurité intégrés
- Implémenter des mesures de surveillance et d'alerte pour les comportements anormaux

En raison de leur forte autonomie et de leur impact potentiel, les agents IA peuvent nécessiter des processus dédiés d'évaluation des risques et de validation.

Autorisations et accès

Visibilité sur les utilisateurs de l'IA dans l'organisation

Une bonne gouvernance de l'IA passe par une parfaite visibilité sur les personnes qui développent et utilisent ces systèmes dans votre entreprise. Vous devez notamment connaître :

- Leurs rôles et responsabilités (data scientist, ingénieur ML, responsable produit, etc.)
- Les autorisations d'accès aux outils de développement et de déploiement de l'IA
- Les identités machines (comptes de services, clés d'API…) utilisées pour accéder aux systèmes d'IA

Surveillance renforcée des applications d'IA grand public

Les applications d'IA destinées aux clients ou au grand public nécessitent une surveillance et une gouvernance accrues par rapport aux applications d'entreprise à usage interne. Parmi ces contrôles renforcés, on peut citer :

- Des tests rigoureux visant à détecter les éventuels biais pour différents groupes démographiques
- Des simulations d'attaque pour évaluer les risques de sécurité et les abus potentiels
- Des audits de conformité plus fréquents et plus exhaustifs

Les organisations doivent également envisager des protections techniques comme la limitation du débit, le filtrage des contenus et l'arrêt automatique en cas de dépassement des seuils de risque prédéfinis. Par ailleurs, il est important de programmer des audits et des études d'impact réguliers pour identifier et réduire les risques émergents en amont.

Conformité

Surveillance des régimes réglementaires applicables

La gouvernance de l'IA ne fonctionne pas en vase clos. Dans l'organisation, elle doit être intégrée à la structure globale de gestion de la conformité, de manière à aligner vos politiques et contrôles sur les lois, les réglementations et les normes sectorielles en vigueur.

Voici les principaux cadres réglementaires :

- Règlements sur la protection des données (RGPD, CCPA, HIPAA)
- Réglementations sectorielles (par ex. la FINRA pour le secteur financier ou la FDA pour le secteur de la santé)
- Réglementations émergentes ou récentes relatives à l'IA (par ex. la législation new-yorkaise contre les pratiques de recrutement discriminatoires par IA)
- Normes et certifications facultatives (IEEE, ISO, NIST…)

Votre équipe conformité doit travailler de concert avec vos équipes de développement et de gouvernance de l'IA pour dresser la liste de vos obligations, puis établir les politiques et contrôles appropriés. Il peut également s'avérer nécessaire d'effectuer des analyses d'écarts, des évaluations d'impact sur la protection des données et des examens de conformité à divers points clés du cycle de vie de l'IA.

Vos partenaires et fournisseurs d'IA externes ne sont pas exempts : soumettez-les à ces mêmes contrôles de conformité pour vous assurer que leurs pratiques sont bien en phase avec vos obligations. Pour ce faire, intégrez des critères de vérification et des exigences contractuelles spécifiques à l'IA dans vos processus de gestion du risque fournisseur.

Prise en compte continue des risques de conformité actuels et futurs

Le cadre réglementaire de l'IA connaît de profonds bouleversements. La promulgation de nouvelles lois et normes a souvent d'importantes répercussions sur les modes de développement et de déploiement des systèmes d'IA dans les organisations. D'où l'importance de soumettre ces systèmes à des examens et des audits réguliers pour s'assurer qu'ils sont et restent conformes aux réglementations.

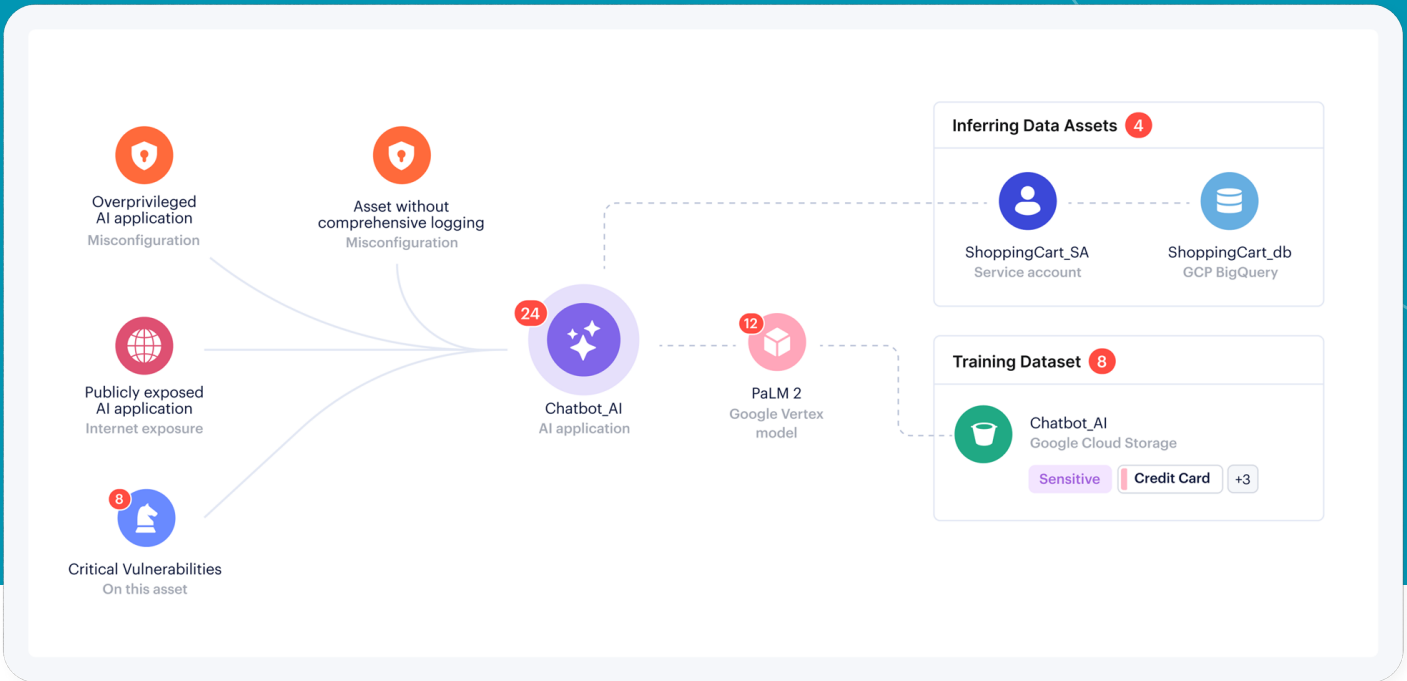
Prisma Cloud AI Security Posture Management : visibilité, contrôle et gouvernance sur les applications d'IA

La gestion des risques et des défis présentés dans ce document nécessite l'adoption de nouveaux outils et de nouvelles approches pour sécuriser un écosystème IA toujours plus foisonnant en entreprise. Interrogées sur leurs priorités pour 2024, 100 % des organisations se disent déterminées à gagner en visibilité sur tout leur pipeline de déploiement IA (cf. notre rapport « État des lieux 2024 de la sécurité du cloud »[3]).

Les modèles d'IA, de machine learning et d'IA générative font émerger de nouveaux risques auxquels les entreprises doivent se préparer. C'est là que Prisma Cloud AI Security Posture Management (AI-SPM) intervient.

Grâce à Prisma Cloud AI-SPM, votre entreprise dispose d'une visibilité sur tout votre écosystème IA – de l'ingestion des données au déploiement, en passant par l'entraînement des modèles. Parce qu'elle analyse le comportement des modèles, les flux de données et les interactions des systèmes, la solution AI-SPM vous aide à repérer les risques de sécurité et de conformité que les outils traditionnels ne savent pas forcément identifier.

3 https://www.paloaltonetworks.fr/state-of-cloud-native-security



Les principales fonctionnalités de notre solution :

- **Détection et inventaire des modèles d'IA –** Cette fonctionnalité vous permet de créer un inventaire de tous les API de modèles, modèles open-source et modèles déployés sur des machines virtuelles dans votre organisation. Vous êtes ainsi mieux armé pour limiter la prolifération des modèles et le Shadow AI, empêcher l'utilisation de modèles non autorisés, et garantir la mise en place de contrôles de gouvernance appropriés.
- **Prévention de l'exposition des données –** AI-SPM détecte et classifie les ensembles de données utilisés pour entraîner et opérer vos modèles d'IA, et vous signale au passage les expositions potentielles de données sensibles. Notre solution surveille en direct les interactions des modèles afin de détecter les abus ou les fuites de données involontaires.
- **Analyse de la posture et des risques –** AI-SPM analyse le pipeline de déploiement IA de bout en bout pour identifier les erreurs de configuration, les contrôles d'accès insuffisants et autres vulnérabilités susceptibles de mettre en danger vos modèles ou vos données. Vous bénéficiez également d'une cartographie visuelle des autorisations d'accès utilisateurs, bien utile pour corriger les privilèges excessifs.

Grâce aux éclairages fournis par AI-SPM, vous avez toutes les cartes en main pour appliquer des politiques de sécurité homogènes, neutraliser de façon proactive les menaces propres à l'IA et maintenir votre conformité aux nouvelles réglementations, comme la loi européenne sur l'intelligence artificielle.

Prisma Cloud AI-SPM s'intègre en toute transparence à la plateforme Code to Cloud™ pour vous fournir des capacités CNAPP complètes (approches CSPM et DSPM inclues). Vous disposez ainsi d'une visibilité et d'un contrôle unifiés sur toute votre stack cloud-native. Grâce à un déploiement sans agent simple et rapide, Prisma Cloud vous aide à sécuriser vos assets IA critiques en l'espace de quelques minutes, favorisant ainsi une innovation responsable à grande échelle.

Pour en savoir plus, rendez-vous sur https://www.paloaltonetworks.fr/prisma/cloud/ai-spm

paloalto NETWORKS

Oval Tower, De Entrée 99 – 197
1101HE Amsterdam, Pays-Bas
Téléphone : +31 20 888 1883
www.paloaltonetworks.fr

prisma_ds_AI Governance_July 2024-fr

Oval Tower, De Entrée 99 – 197
1101HE Amsterdam, Pays-Bas
Téléphone : +31 20 888 1883
www.paloaltonetworks.fr

prisma_ds_AI Governance_July 2024-fr