DARKTRACE

Comment sécuriser l'IA en entreprise

Guide de sécurisation des modèles, des données et des agents



© 2026 Darktrace Holdings Limited. Tous droits réservés.

Sommaire

02	**Préface**	**07**	Analyse de l'IaC
02	**À propos de Darktrace**	**07**	Analyse du code généré par l'IA
03	**Sécuriser l'IA : comprendre les risques**	**08**	**Sécuriser la chaîne d'approvisionnement de l'IA**
04	**5 catégories de risques liés à l'IA et comment les traiter**	**08**	Utilisation de l'IA fantôme
05	**Se défendre contre les utilisations abusives et les comportements émergents**	**08**	Fournisseurs d'IA proposant directement des agents
05	Analyse des prompts malveillants	**08**	IA dans les services des fournisseurs
05	Cas d'utilisation non intentionnels	**08**	Analyse du code et des dépendances des agents IA
05	Accès à l'historique des prompts	**09**	Approvisionnement en modèles d'IA
05	Effets des résultats malveillants et des hallucinations	**09**	Données d'entraînement de l'IA
06	**Surveiller et contrôler l'IA en fonctionnement**	**10**	**Renforcer la préparation et la supervision**
06	Actions et connexions des agents IA	**10**	Reporting
06	Gateway IA	**10**	Pentesting de l'IA, red teaming et purple teaming
06	Capture de l'état des agents IA	**10**	Formation et exercices des équipes de sécurité
06	Transmission des données de l'IA	**11**	**Repenser la sécurité de l'IA pour les dirigeants**
07	**Protéger le développement et l'infrastructure de l'IA**	**12**	**Conclusion**
07	Configuration de la sécurité de l'IA		
07	Architectures d'agents IA		

Préface

Nos échanges avec nos clients et nos propres observations montrent que les responsables de la sécurité ne disposent pas d'une définition claire et concise de ce que recouvre la « sécurisation de l'IA ». S'il existe des ressources utiles sur le sujet, nous n'en avons encore trouvé aucune qui établisse un lien cohérent entre les risques et les fonctions de sécurité. Des cadres plus généraux, tels que la norme ISO/IEC 42001, fournissent d'importants principes de gouvernance, mais n'offrent pas les conseils pratiques nécessaires pour mettre en œuvre la sécurité de l'IA au sein d'une entreprise.

Ce livre blanc vise à combler cette lacune. Il propose un guide indépendant de tout fournisseur pour mieux comprendre le périmètre de la cybersécurité appliquée à l'IA. Il présente les principales catégories, les principaux risques et les fonctions de sécurité que les organisations doivent prendre en compte à mesure qu'elles adoptent et intègrent les technologies d'IA dans leurs activités.

À propos de Darktrace

Depuis plus de dix ans, Darktrace est à l'avant-garde de l'application de l'intelligence artificielle à la cybersécurité, avec des innovations pionnières qui ont profondément transformé le secteur. En 2013, nous avons lancé la détection et la réponse réseau (NDR), entièrement basées sur des méthodes de détection fondées sur l'IA, établissant ainsi une nouvelle norme en matière d'identification autonome des menaces.

En 2019, nous avons lancé un agent d'investigation autonome basé sur l'IA, capable de reproduire les processus humains d'investigation et d'analyse à la vitesse et à l'échelle d'une machine. Notre adoption précoce de modèles de langage avancés, notamment d'approches basées sur les transformeurs en 2021 et de modèles de langage spécifiques à un domaine (DSLM) en 2023, a encore renforcé notre capacité à interpréter les cybermenaces complexes.

Aujourd'hui, Darktrace AI fonctionne en continu au sein d'environ 10 000 organisations à travers le monde, offrant une résilience éprouvée à grande échelle. Mi-2025, nous sommes devenus l'une des premières entreprises de cybersécurité à obtenir la certification ISO 42001, témoignant ainsi de notre engagement en faveur d'une gouvernance responsable de l'IA. Cela reflète non seulement notre rôle de premier plan dans l'innovation en matière d'IA, mais aussi notre engagement constant à sécuriser l'IA elle-même, afin que la confiance, la transparence et la sécurité restent au cœur de chaque avancée.



Plus de 10 ans d'application de l'IA à des cas d'usage en cybersécurité



L'une des premières entreprises de cybersécurité certifiées ISO 42001



Lauréat des AI Excellence Awards dans la catégorie Cybersecurity AI (2025)



Environ 10 000 clients, et ce nombre ne cesse d'augmenter

Sécuriser l'IA

Comprendre les risques

L'intelligence artificielle fait son entrée dans les entreprises à un rythme que peu de systèmes de gouvernance et de gestion peuvent égaler.

Les cadres traditionnels peinent à suivre le rythme face aux nouvelles complexités introduites par l'IA et à l'élargissement de la surface d'attaque. Chaque modèle, jeu de données et agent présente de nouveaux risques, tout comme le réseau complexe d'API, d'intégrations tierces et d'interactions humaines qui les alimentent. Autant de nouvelles voies susceptibles d'être exploitées à des fins malveillantes.

Pour certains, l'IA promet des gains de productivité, de rapidité et d'efficacité. Mais elle apporte également son lot d'incertitudes et nécessite une supervision croissante, en particulier dans le domaine de l'IA générative et des agents basés sur l'IA générative. L'IA générative peut produire des résultats peu fiables ou des hallucinations, manquer de transparence dans ses prises de décision et fonctionner de manière autonome, selon des modalités que les outils de sécurité traditionnels n'ont jamais été conçus pour surveiller. Elle étend ainsi l'écosystème numérique, où les frontières entre propriété, contrôle et responsabilité deviennent souvent floues.

Pour les responsables de la sécurité, le premier défi à relever pour protéger une entreprise en période de transformation consiste à déterminer ce qui doit être sécurisé. L'IA s'intègre aux organisations par de multiples canaux. Les employés utilisent désormais des outils SaaS et de productivité intégrant l'IA. Les fournisseurs de solutions cloud et sur site enrichissent leurs infrastructures et leurs services de fonctionnalités d'IA intégrées. Les fournisseurs et les logiciels acquis intègrent souvent des fonctionnalités d'IA qui augmentent l'exposition aux risques sans offrir de supervision claire. En interne, les développeurs peuvent créer ou adapter leurs propres modèles et agents afin d'intégrer directement l'IA aux activités de l'entreprise.

Pour protéger un écosystème d'IA sur l'ensemble de ces points d'entrée, il est essentiel de disposer d'une visibilité sur chaque couche, du développement et de l'entraînement au déploiement et au fonctionnement en temps réel. Chaque étape présente des risques spécifiques, de l'empoisonnement des données et des vulnérabilités des dépendances aux comportements malveillants des agents ou à l'utilisation involontaire des résultats des modèles.

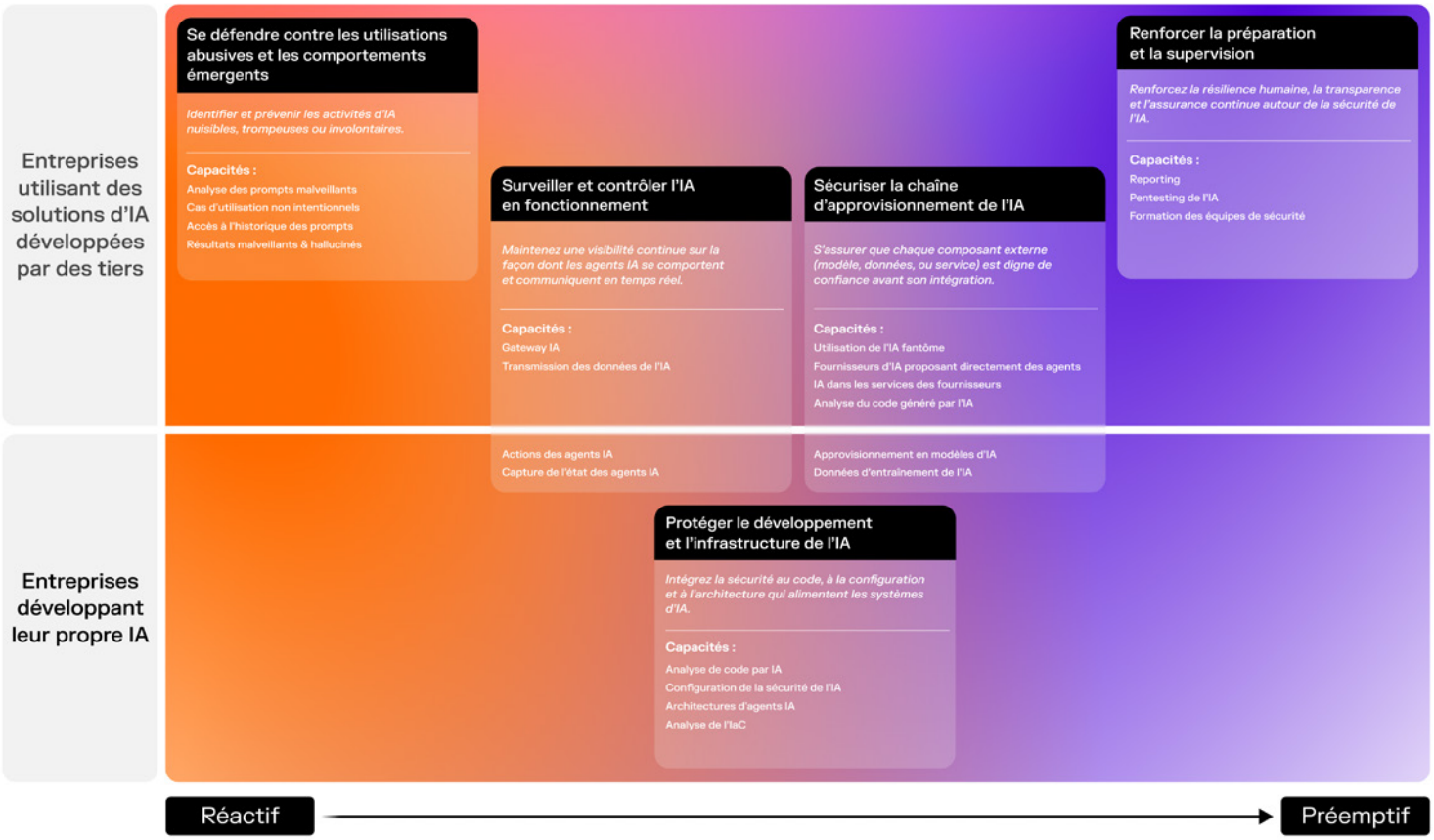


5 catégories de risques liés à l'IA et comment les traiter

Chacune des cinq catégories décrites dans ce cadre représente une couche distincte de la sécurité de l'IA, des fournisseurs externes et des pipelines de développement aux opérations en production et à la gouvernance organisationnelle.

Au sein de chaque catégorie, des capacités spécifiques permettent d'identifier les sources potentielles de risques et les fonctions de sécurité à mettre en œuvre pour réduire l'exposition. La section suivante examine ces capacités en détail, en associant chacune d'elles aux risques correspondants et aux fonctions de sécurité conçues pour y répondre. Nous avons identifié cinq catégories critiques d'utilisation de l'IA. Ces catégories reflètent les différentes façons dont l'IA s'intègre et fonctionne au sein d'une organisation : par l'intermédiaire des fournisseurs, des pipelines de développement, des systèmes opérationnels, des interactions avec les utilisateurs et des processus de gouvernance.

En résumé, les cinq catégories sont :



Le tableau ci-dessus offre une vue d'ensemble de ces domaines et montre comment les différents risques sont associés aux principales fonctions de sécurité tout au long du cycle de vie de l'IA. Il offre une vue synthétique des domaines sur lesquels concentrer les efforts de défense, avant que les pages suivantes n'examinent chaque capacité plus en détail.

Se défendre contre les utilisations abusives et les comportements émergents

Capacités

Analyse des prompts malveillants

Détection et analyse des prompts nuisibles ou manipulateurs visant à contraindre les systèmes ou agents IA à effectuer des actions indésirables ou dangereuses.

Description du risque : les agents IA peuvent recevoir des instructions ou être manipulés afin d'effectuer des actions indésirables en exploitant leurs autorisations et leur accès aux données. Ces actions peuvent être difficiles à surveiller ou à bloquer, en particulier lorsqu'il s'agit d'agents IA externes. S'il dispose d'un accès suffisant, un acteur de la menace peut explorer et extraire des données d'un modèle d'IA qui étaient censées y rester dissimulées.

Fonction de sécurité : détecter et, dans la mesure du possible, bloquer les prompts malveillants, qu'ils soient internes ou externes, directs ou indirects. Surveiller les tentatives d'extraction ou de manipulation des données des modèles afin de prévenir les utilisations abusives et de protéger les informations sensibles lors des interactions avec l'IA.

Cas d'utilisation non intentionnels

L'utilisation abusive ou le détournement de systèmes d'IA à des fins non prévues ou non autorisées à l'origine, entraînant souvent des comportements ou des risques imprévus.

Description du risque : l'utilisation abusive des agents et systèmes IA peut avoir de nombreux effets indésirables. Ces risques sont d'autant plus importants qu'il peut être difficile de détecter une utilisation abusive par des utilisateurs autorisés lorsque les garde-fous ne sont pas efficaces.

Fonction de sécurité : identifier et bloquer les usages non autorisés ou non approuvés des systèmes d'IA. Surveiller toute exposition involontaire des données ou utilisation abusive afin de garantir que les modèles d'IA fonctionnent uniquement dans le respect des paramètres approuvés et des exigences de conformité.

Accès à l'historique des prompts

Stockage et gestion des historiques de prompts d'IA, qui peuvent contenir des données sensibles, des informations contextuelles ou des schémas comportementaux.

Description du risque : un accès non approuvé à l'historique des prompts peut révéler des informations confidentielles, même lorsque le service auquel les prompts ont été soumis est sécurisé et ne conserve pas ces informations.

Fonction de sécurité : restreindre les accès et surveiller les signes de compromission, car un acteur de la menace ayant pris le contrôle d'un compte pourrait exploiter ces informations pour mener d'autres attaques.

Effets des résultats malveillants et des hallucinations

La génération de résultats d'IA inexacts, trompeurs ou intentionnellement nocifs, susceptibles d'influencer les actions humaines.

Description du risque : indépendamment des risques liés à l'automatisation reposant sur des résultats hallucinés ou malveillants, (tels que le package-squatting ou les actions inattendues des agents IA), ces résultats peuvent également conduire les utilisateurs à effectuer des actions indésirables. Le champ des actions potentielles est bien plus vaste que celui des actions automatisées. Par exemple, un utilisateur peut être incité à transférer de l'argent ou à appliquer des paramètres malveillants à un système informatique critique auquel les agents IA n'ont délibérément pas accès.

Fonction de sécurité : surveiller les résultats nuisibles ou fabriqués par l'IA susceptibles d'induire les utilisateurs en erreur ou de déclencher des actions involontaires. Mettre en œuvre des mécanismes de détection des hallucinations et de validation des résultats afin d'éviter la diffusion d'informations erronées ou les comportements dangereux influencés par du contenu généré par l'IA.

Surveiller et contrôler l'IA en fonctionnement

Capacités

Actions et connexions des agents IA

Pour les entreprises développant des outils d'IA

La surveillance du comportement en temps réel des agents IA, notamment les connexions sortantes, les échanges de données et les interactions avec d'autres agents ou services.

Description du risque : les agents IA peuvent utiliser leurs autorisations de manière inappropriée ou accéder à des données ou à des actions auxquelles ils ne sont pas censés avoir accès. Ils peuvent également divulguer des données ou provoquer des actions indésirables lorsqu'ils font appel à d'autres agents IA ou à des services (notamment des serveurs MCP). Contrairement aux employés, les agents IA n'ont pas la même capacité de prudence si les garde-fous n'anticipent pas les problèmes potentiels ou s'ils s'avèrent inefficaces.

Fonction de sécurité : surveiller en temps réel les connexions sortantes, les requêtes et les prompts envoyés par les agents IA à des services externes (notamment des serveurs MCP), ou à d'autres agents. Suivre l'utilisation des identifiants et bloquer les activités lorsque cela est nécessaire afin d'empêcher les fuites de données ou les actions non autorisées.

Gateway IA

Point d'accès centralisé ou couche de contrôle permettant de gérer les interactions avec les systèmes d'IA, notamment l'envoi des prompts, le routage des modèles, le contrôle des coûts et l'application des règles de confidentialité des données.

Description du risque : les systèmes d'IA peuvent ne pas être accessibles via des gateways, ou une gateway peut être mal configurée ou tomber en panne. Cela peut entraîner une protection insuffisante des prompts, un routage incorrect des modèles, des coûts non maîtrisés ou encore une surveillance et un contrôle en temps réel insuffisants.

Fonction de sécurité : déployer et gérer de manière centralisée les gateways IA, et imposer leur utilisation dans la mesure du possible. Intégrer les journaux, les alertes et les métriques aux systèmes et aux outils de surveillance de la sécurité.

Lorsque des systèmes d'IA sont délibérément déployés de manière à conserver les données en local (par exemple, sur site), les capacités de sécurité doivent être déployées en conséquence et offrir un niveau de protection équivalent.

Capture de l'état des agents IA

Pour les entreprises développant des outils d'IA

Processus consistant à préserver l'état opérationnel et les informations de composants d'IA potentiellement temporaires ou éphémères, en particulier dans les environnements cloud, afin de permettre leur investigation et leur analyse après un incident.

Description du risque : lorsqu'un composant éphémère d'un agent IA (courant dans les systèmes hébergés dans le cloud) effectue une action qui doit faire l'objet d'une investigation ou être conservée, il peut disparaître avant que les processus de sécurité aient pu le capturer. L'équipe de sécurité peut même ne disposer d'aucun moyen pour capturer et préserver certains composants.

Fonction de sécurité : capturer et préserver automatiquement l'état complet de tout agent IA suspect, y compris les instances cloud temporaires. Cela permet de mener les investigations et d'intervenir en cas d'incident, tout en veillant à conserver les preuves et les éléments de contexte nécessaires à l'analyse.

Les secteurs fortement réglementés doivent accorder une attention particulière à ce risque

Transmission des données de l'IA

Transfert de données entre des composants ou systèmes d'IA, ou vers des points de terminaison externes, qu'il s'agisse d'échanges automatisés ou initiés par des utilisateurs.

Description du risque : les données envoyées aux systèmes d'IA peuvent enfreindre les exigences relatives à la souveraineté des données ou les obligations légales (telles que le RGPD). Les données peuvent être envoyées par des utilisateurs, des systèmes automatisés ou à la suite de décisions prises par des agents IA. Les transmissions de données vers ou depuis des systèmes d'IA peuvent reposer sur des méthodes cryptographiques peu robustes et doivent être prises en compte dans les plans de transition vers la cryptographie post-quantique (PQC).

Fonction de sécurité : sécuriser les données en transit entre les composants d'IA à l'aide de méthodes de chiffrement, notamment la cryptographie post-quantique lorsque cela est pertinent. Veiller à ce que tous les transferts de données respectent les réglementations applicables en matière de circulation et de confidentialité des données, telles que le RGPD.

Protéger le développement et l'infrastructure de l'IA

Capacités

Configuration de la sécurité de l'IA

Pour les entreprises développant des outils d'IA

Configuration et gestion continue des paramètres de sécurité, des garde-fous et des contrôles d'accès qui régissent le fonctionnement des composants d'IA.

Description du risque : Les configurations attribuées aux agents IA ainsi qu'aux composants individuels associés et de support peuvent être défectueuses et autoriser un accès inapproprié ou présenter une faible sécurité contre les attaques. Une mauvaise configuration peut également supprimer les garde-fous prévus.

Fonction de sécurité : garantir une configuration sécurisée de tous les composants d'IA, notamment des garde-fous, des politiques et des autorisations. La surveillance et l'analyse automatisées, ainsi que des examens réguliers des politiques et de leur application, contribuent à maintenir une posture de sécurité de l'IA robuste et conforme.

Architectures d'agents IA

Pour les entreprises développant des outils d'IA

Conception et structure des agents IA et de leurs composants interconnectés, qui déterminent la manière dont ils interagissent, partagent des données et appliquent les limites de sécurité.

Description du risque : l'architecture à plusieurs composants d'un agent IA en cours d'utilisation, ou celle d'un ensemble d'agents IA conçus pour fonctionner conjointement, peut présenter des failles dans son ensemble et permettre des accès inappropriés ou ne pas disposer de garde-fous contre les utilisations involontaires, indépendamment des problèmes propres à chaque composant.

Fonction de sécurité : effectuer une analyse statique des architectures des agents IA afin d'identifier les failles de sécurité. Évaluer l'exposition au réseau au niveau local, les identifiants des services et les autorisations afin d'empêcher les accès non autorisés ou les déplacements latéraux. Une segmentation appropriée et une gestion rigoureuse des identifiants sont essentielles pour réduire les risques liés à l'architecture.

Les équipes de développement peuvent utiliser des services d'aide à la création d'agents IA. Dans la mesure du possible, limiter leur utilisation à un seul service offrant des capacités de sécurité satisfaisantes.

Analyse de l'IaC

Pour les entreprises développant des outils d'IA

Analyse automatisée des modèles d'Infrastructure-as-Code afin de détecter les mauvaises configurations ou les failles de sécurité avant le déploiement.

Description du risque : lorsque l'Infrastructure-as-Code est utilisée, des problèmes de configuration de la sécurité de l'IA ou d'architecture des agents IA peuvent être introduits en raison d'un accès inapproprié au code, de l'introduction de code malveillant en interne ou depuis une source externe, ou encore d'erreurs.

Fonction de sécurité : utiliser des outils d'analyse spécialisés pour analyser l'Infrastructure-as-Code (IaC) afin de détecter les mauvaises configurations ou les risques présents dans le code. La détection des problèmes de sécurité dès les premières étapes dans les scripts de déploiement permet d'empêcher les vulnérabilités de se propager dans les environnements d'IA.

Analyse du code généré par l'IA

Pour les entreprises développant des outils d'IA

L'examen du code produit par des systèmes d'IA (copilotes de codage) pour détecter les failles de sécurité, les inclusions malveillantes ou les dépendances non fiables.

Description du risque : le code généré peut présenter des failles de sécurité, qu'elles soient accidentelles ou qu'elles résultent d'une manipulation malveillante du modèle qui génère le code. Il peut notamment intégrer des dépendances malveillantes. Par exemple, des acteurs de la menace peuvent créer des packages portant des noms susceptibles de faire l'objet d'hallucinations par les modèles. Le code généré peut être nettement plus difficile à comprendre et à déboguer que le code écrit manuellement. Les problèmes sont davantage susceptibles de survenir à mesure que l'ampleur et la complexité du code généré augmentent.

Fonction de sécurité : analyser le code généré par l'IA afin de détecter les failles, les logiques non sécurisées ou l'insertion de code malveillant. Détecter les problèmes tels que les dépendances faisant l'objet d'hallucinations ou le package squatting afin d'empêcher tout code compromis ou non sécurisé d'être déployé dans des environnements de production.

Sécuriser la chaîne d'approvisionnement de l'IA

Capacités

Utilisation de l'IA fantôme

Utilisation par des employés ou des services d'outils d'IA non approuvés, souvent en dehors des contrôles officiels de sécurité et de conformité. Ces services fonctionnent généralement en externe et peuvent collecter des données sensibles de l'entreprise ou les utiliser pour entraîner leurs modèles sans aucune supervision.

Description du risque : des données peuvent être divulguées à des tiers externes non approuvés, notamment des informations protégées, des données financières, des informations sur les clients ou des éléments confidentiels de propriété intellectuelle. Les services non approuvés sont susceptibles d'utiliser toutes les données qui leur sont transmises à des fins d'entraînement.

Fonction de sécurité : détecter et bloquer l'utilisation de services d'IA non approuvés, tels que des outils d'IA générative non approuvés ou des serveurs MCP externes non fiables auxquels accèdent des outils approuvés. Une utilisation non surveillée peut exposer des données sensibles et créer des angles morts dans la supervision de la sécurité.

Fournisseurs d'IA proposant directement des agents

Agents IA externes ou services d'IA générative intégrés aux systèmes de l'entreprise et disposant d'un accès interne ou d'autorisations leur permettant d'effectuer des tâches automatisées pour le compte d'utilisateurs ou d'applications.

Description du risque : les activités de ces agents peuvent ne plus être visibles ni contrôlables une fois lancées, que ce soit au niveau des prompts ou de la connectivité.

Fonction de sécurité : Surveillez les agents IA externes bénéficiant d'un accès interne, car ils opèrent souvent avec des privilèges étendus et une visibilité limitée. Appliquez des contrôles d'accès, une journalisation et une surveillance stricts afin de prévenir toute utilisation abusive ou toute action non intentionnelle au sein des environnements internes.

IA dans les services des fournisseurs

Recours à des fournisseurs externes ou à des services tiers qui utilisent l'IA pour traiter ou analyser les données de l'entreprise. Cela inclut aussi bien les offres d'IA déclarées que les situations dans lesquelles des fournisseurs intègrent l'IA à leurs produits ou à leurs flux de travail sans en informer explicitement leurs clients.

Description du risque : les fournisseurs proposant un service d'IA déclaré peuvent ne pas protéger correctement les données de l'entreprise, que ce soit en raison de garanties contractuelles insuffisantes ou de défaillances opérationnelles. Les fournisseurs peuvent également utiliser l'IA en interne sans en informer l'entreprise qui leur confie des données confidentielles, échappant ainsi à sa supervision.

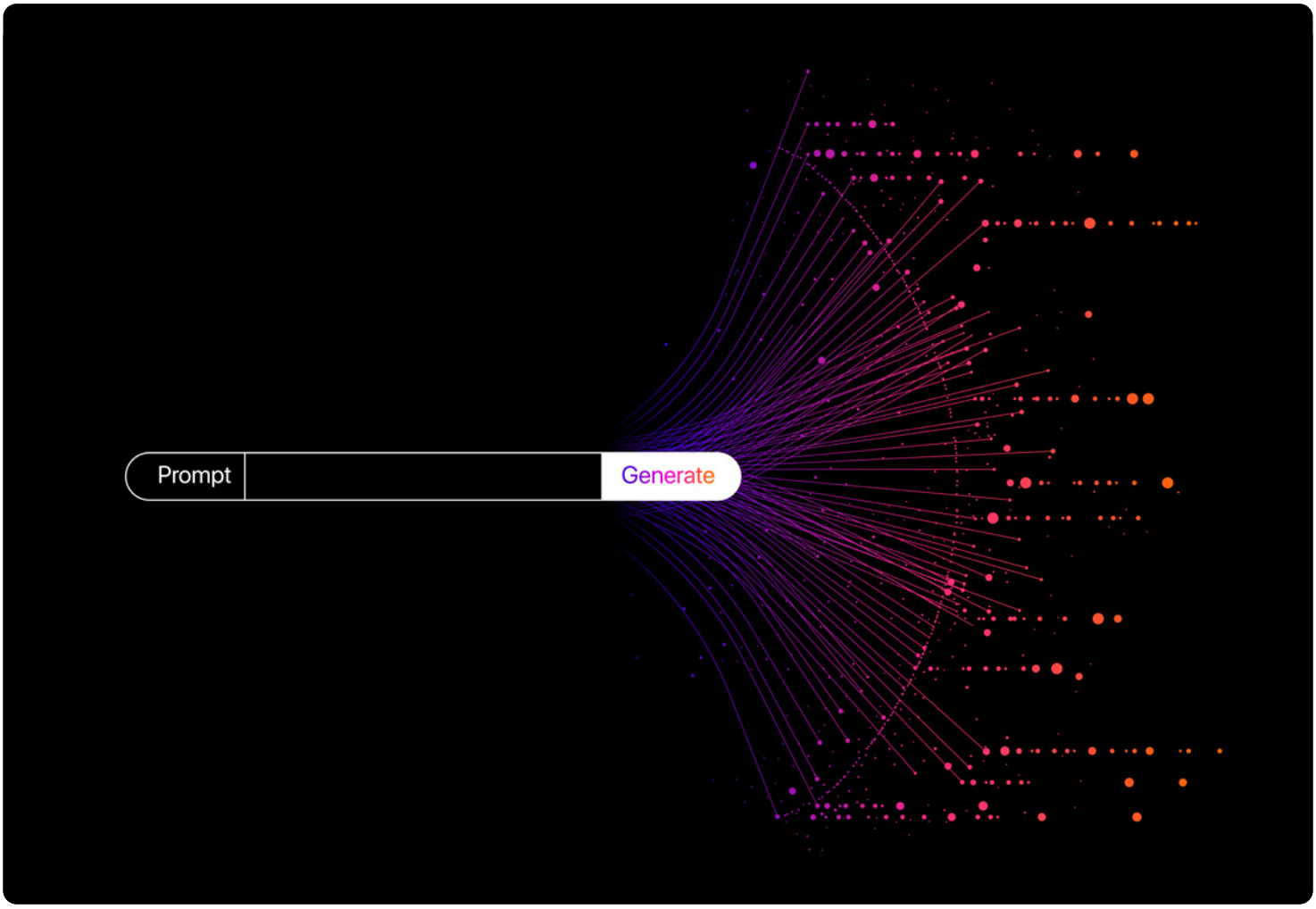
Fonction de sécurité : évaluer les fournisseurs qui utilisent l'IA pour traiter les données de votre entreprise (qu'ils le déclarent ou non). Vérifier comment les systèmes d'IA externes traitent, stockent et protègent ces données afin d'empêcher toute utilisation non autorisée ou exposition involontaire par l'intermédiaire de services tiers.

Analyse du code et des dépendances des agents IA

Processus consistant à examiner le code source des agents IA et les dépendances logicielles associées afin d'identifier les vulnérabilités, les problèmes de licence ou la présence d'éléments malveillants. Cela permet de s'assurer que les composants utilisés pour créer ou étendre les systèmes d'IA sont sécurisés, conformes et fonctionnent comme prévu.

Description du risque : les dépendances logicielles peuvent être soumises à des licences inadaptées ou avoir été compromises à des fins malveillantes, qu'elles soient open source ou commerciales. Le code développé pour les agents IA peut présenter des failles susceptibles de compromettre, voire de rendre inefficaces, les mesures de sécurité prévues.

Fonction de sécurité : Effectuer une analyse de sécurité statique du code des agents IA internes afin d'identifier les faiblesses et les configurations non sécurisées. Examiner toutes les dépendances logicielles afin de détecter les vulnérabilités connues ou les inclusions malveillantes avant le déploiement. Une analyse régulière permet de maintenir l'intégrité du code et de réduire les risques dans le pipeline de développement de l'IA.



Approvisionnement en modèles d'IA

Pour les entreprises développant des outils d'IA

Processus consistant à acquérir ou à intégrer des modèles d'IA externes, qu'ils soient open source, fournis par des tiers ou par des fournisseurs, aux systèmes ou aux pipelines de développement d'une organisation.

Description du risque : Les modèles provenant de sources externes peuvent contenir du contenu et des instructions malveillants ou incorrects, générant des résultats qui incitent les agents IA ou les utilisateurs à effectuer des actions non souhaitées. Ils pourraient inclure des niveaux de biais, de précision, de dérive et d'autres facteurs inacceptables pour l'entreprise ou ses cas d'usage spécifiques. Ils pourraient ne pas disposer de licences adaptées à leur utilisation au sein de l'entreprise.

Fonction de sécurité : sécuriser la chaîne d'approvisionnement des modèles implique d'en valider à la fois l'origine et le contenu. Chaque modèle externe doit être analysé afin de détecter les risques qu'il pourrait contenir, faire l'objet d'une vérification de conformité des licences et être stocké de manière sécurisée avec un contrôle des versions.

Données d'entraînement de l'IA

Pour les entreprises développant des outils d'IA

Collecte et préparation des jeux de données utilisés pour entraîner ou affiner les modèles d'IA, ou dans le cadre des activités de test, d'évaluation, de validation et de vérification (TEVV). Ces jeux de données peuvent provenir de sources internes, de référentiels publics ou de fournisseurs tiers, et influencent directement le comportement, la précision et la conformité des modèles.

Description du risque : les données d'entraînement peuvent contenir des données malveillantes susceptibles d'entraîner des comportements indésirables du modèle une fois entraîné. Elles peuvent également contenir des données protégées utilisées de manière illicite ou des données dont la licence n'est pas appropriée. Le jeu de données dans son ensemble peut également ne pas disposer d'une licence adaptée aux usages qui en sont faits au sein de l'entreprise.

Fonction de sécurité : la sécurisation des données d'entraînement commence par une connaissance précise de leur origine et de leur contenu. Vérifier l'origine, les licences et l'intégrité des données avant leur utilisation. Se protéger contre l'empoisonnement des données, qu'il soit d'origine interne ou externe. Identifier et gérer les risques liés aux licences des données, en particulier dans les modèles fondamentaux, ainsi que le risque que les données d'entraînement soient reproduites directement dans les résultats générés par les modèles.

Renforcer la préparation et la supervision

Capacités

Reporting

Processus consistant à recueillir et à communiquer les risques, les métriques et les tendances liés à l'IA afin de favoriser une prise de décision éclairée et une gouvernance efficace.

Description du risque : les objectifs et les métriques sont essentiels à une gestion efficace et à l'amélioration continue. Sans reporting efficace sur la gestion de l'IA, des problèmes critiques ou des tendances importantes peuvent passer inaperçus. Plus largement, l'entreprise peut prendre des décisions inappropriées concernant l'IA si elle ne dispose pas d'un reporting adapté à ses besoins.

Fonction de sécurité : générer des synthèses claires des événements de sécurité liés à l'IA, des métriques de risque et des performances des systèmes. Suivre l'adoption des cas d'usage prévus de l'IA et signaler toute utilisation involontaire ou non autorisée afin de maintenir une supervision et une gouvernance efficaces.

Pentesting de l'IA, red teaming et purple teaming

Test et validation des systèmes d'IA en production au moyen d'attaques simulées et d'exercices adversariaux visant à identifier les vulnérabilités et les lacunes en matière de détection.

Description du risque : les systèmes en production présentent souvent des problèmes latents qui ne peuvent pas être détectés dans la configuration ni dans le code. Sans tests d'intrusion, les acteurs de la menace peuvent découvrir des techniques d'exploitation spécifiques dont les équipes de sécurité n'ont pas connaissance. Sans tests de red teaming ou de purple teaming, des faiblesses plus générales en matière d'exposition et de capacités de détection peuvent être présentes sans avoir été identifiées.

Fonction de sécurité : effectuer des tests d'intrusion et des exercices de red teaming ou de purple teaming afin de simuler des attaques réelles contre les systèmes d'IA en production. Rechercher les faiblesses exploitables, en particulier dans les environnements exposés à l'extérieur, afin d'anticiper les techniques des adversaires et d'en limiter les effets avant qu'elles ne soient mises en œuvre.

Formation et exercices des équipes de sécurité

Formation continue et exercices de simulation conçus pour développer des compétences de détection, d'investigation et de réponse aux incidents spécifiques à l'IA au sein des équipes de sécurité.

Description du risque : une formation individuelle insuffisante ou un manque d'entraînement des équipes à la réponse aux incidents peut entraîner des défauts de détection, des investigations inadéquates ou une incapacité à neutraliser ou à éliminer un problème de cybersécurité ou un acteur de la menace.

Fonction de sécurité : développer les compétences spécifiques à l'IA et la capacité de réaction des équipes de sécurité. Des formations et des simulations régulières renforcent leur aptitude à détecter les incidents liés à l'IA, à y répondre et à rétablir la situation.



Repenser la sécurité de l'IA pour les dirigeants

Cette même intelligence qui accélère la prise de décision en entreprise peut, si elle n'est pas sécurisée, amplifier les risques à une vitesse et à une échelle sans précédent.

Pour les responsables de la sécurité, la mission ne consiste pas seulement à protéger les systèmes d'IA, mais aussi à traduire leur complexité en orientations stratégiques claires à destination de la direction.

Au niveau du conseil d'administration, il peut s'agir dans un premier temps de présenter la sécurité de l'IA sous l'angle de la confiance, de la responsabilité et de la résilience :

Confiance

Les conseils d'administration veulent avoir l'assurance que les décisions fondées sur l'IA sont fiables, interprétables, respectueuses de la confidentialité et protégées contre toute manipulation. Les RSSI doivent mettre en avant les contrôles permettant de vérifier l'intégrité des données, le comportement des modèles et la provenance de chaque composant de la chaîne d'approvisionnement de l'IA.

Responsabilité

Il convient de définir clairement les responsabilités associées à chaque couche de risque lié à l'IA : data scientists, fournisseurs de services cloud, fournisseurs et équipes de sécurité ont chacun des responsabilités bien définies. Les conseils d'administration sont sensibles à une répartition claire et structurée des responsabilités, et non à l'ambiguïté.

Renforcer la responsabilité en alignant les principes de sécurité de l'IA sur des cadres de gouvernance reconnus, tels que le NIST AI Risk Management Framework (RMF), la norme ISO/IEC 42001 et les recommandations de la CISA et du NCSC. Cet alignement témoigne d'un niveau de maturité élevé, garantit la cohérence entre les différentes obligations de conformité et fournit un référentiel clair pour la supervision.

Résilience

Souligner que l'IA fait désormais partie des infrastructures critiques de l'organisation. La sécuriser implique de garantir la continuité en cas d'attaque, l'auditabilité en cas de contrôle et l'agilité face aux évolutions réglementaires.

À mesure que l'IA s'intègre aux activités de l'entreprise, il sera essentiel d'associer sa sécurité à des enjeux tels que la protection de la propriété intellectuelle, la fiabilité opérationnelle et la confiance des parties prenantes afin de renforcer la résilience de l'organisation. Les organisations qui disposent de systèmes de gestion permettant activement d'étendre les cas d'usage de l'IA tout en garantissant leur sécurité seront les mieux placées pour innover en toute sécurité, s'adapter rapidement et préserver la confiance de leurs clients, des autorités de réglementation et des investisseurs.

Conclusion

Les mêmes caractéristiques qui font de l'IA une technologie transformatrice la rendent également difficile à sécuriser lorsqu'elle est intégrée aux outils SaaS, aux infrastructures cloud, aux systèmes des fournisseurs et aux applications développées en interne. Dans cet environnement, les frontières entre visibilité, propriété et responsabilité peuvent rapidement devenir floues.

Sécuriser l'IA, c'est rétablir cette clarté. Cela nécessite de comprendre non seulement comment l'IA est développée, mais aussi où elle fonctionne, qui interagit avec elle et comment ses décisions influencent les activités de l'entreprise. En envisageant les risques liés à la sécurité de l'IA dans une vision plus globale, englobant la chaîne d'approvisionnement, le développement, le fonctionnement en temps réel et la gouvernance, les organisations peuvent commencer à transformer l'incertitude en confiance.

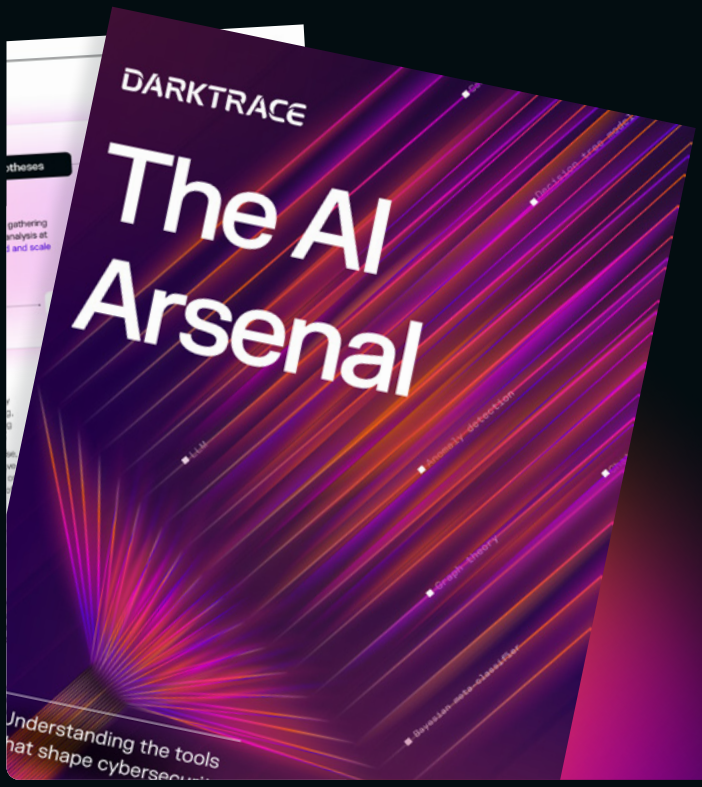
La sécurisation de l'IA est un processus continu, mais elle commence par une prise de conscience : savoir où l'IA est présente, comprendre son comportement et déterminer comment la gouverner de manière responsable.

Développer l'IA de manière sécurisée, c'est la développer de manière responsable.

Découvrez comment concilier innovation et gouvernance dans notre livre blanc « Towards Responsible AI in Cybersecurity »

En savoir plus ↗





Découvrez les différents types d'IA en cybersécurité

Découvrez les technologies qui sous-tendent la résilience moderne, de l'apprentissage automatique supervisé au traitement automatique du langage naturel (NLP), et comment elles fonctionnent ensemble pour contrer les menaces émergentes.

En savoir plus ↗

À propos de Darktrace

Darktrace est un leader mondial de la cybersécurité basée sur l'IA, qui aide chaque jour les organisations à garder une longueur d'avance face à l'évolution des menaces. Fondée en 2013 à Cambridge, au Royaume-Uni, Darktrace fournit une plateforme de cybersécurité essentielle qui protège les organisations contre les menaces inconnues grâce à une IA qui apprend en temps réel à partir de l'environnement propre à chaque entreprise. La plateforme et les services Darktrace s'appuient sur plus de 2 700 employés qui protègent près de 10 000 clients dans le monde. Pour en savoir plus, rendez-vous sur www.darktrace.com.

Amérique du Nord : +1 (415) 229 9100 | Europe : +44 (0) 1223 394 100 | Asie-Pacifique : +65 6804 5010 | Amériqu e latine : +55 11 4949 7696

darktrace.com | info@darktrace.com

© 2026 Darktrace Holdings Limited. Tous droits réservés.