État des lieux des risques liés à l'IA

Connaître les menaces pour minimiser les dommages

CGI

Table des matières

Sommaire

L'intelligence artificielle (IA) est rapidement devenue l'une des fondations de la transformation numérique. Les organisations sont confrontées à la pression croissante des parties prenantes, des conseils d'administration et des organismes de réglementation pour accélérer l'adoption de l'IA et démontrer une valeur mesurable de cette technologie. Toutefois, cette accélération se déroule dans un environnement de risques en constante évolution et peu compris. Les récentes discussions au sein du secteur ont mis en lumière un fait inquiétant : les systèmes d'IA (en particulier à l'étape de l'inférence) exposent les données confidentielles, la propriété intellectuelle et les requêtes d'une manière que les contrôles de cybersécurité traditionnels n'ont pas été conçus pour traiter.

Les entreprises font face aujourd'hui à un double défi : favoriser l'innovation tout en gardant le contrôle de systèmes d'IA complexes et opaques. La visibilité limitée sur les usages de cette technologie (phénomène d'« IA fantôme »), une gouvernance insuffisante et des contrôles de sécurité immatures sont autant d'éléments qui exposent les organisations à des risques, y compris de possibles répercussions sur la confidentialité, l'intégrité et la fiabilité à grande échelle.

Pour mieux comprendre ces risques, une vision structurée des catégories de menaces les plus critiques a été proposée par des cadres de référence comme l'Open Worldwide Application Security Project (OWASP), avec sa liste des dix plus grandes menaces des grands modèles de langage, et le National Institute of Standards and Technology (NIST), avec son cadre de gestion des risques liés à l'IA (AI RMF).

L'objectif de ce document est de proposer aux leaders et spécialistes de la sécurité des connaissances générales sur les risques liés à l'IA pour leur permettre d'amorcer et d'approfondir des conversations avec leurs équipes.

Pour 77 % des responsables de direction informatique, la sécurité et les risques sont les principaux obstacles rencontrés lors du déploiement à grande échelle de technologies autonomes.

Source: Gartner rapport des principaux défis des responsables de direction informatique au T1 (en anglais) de Gartner

Portée du risque en IA

L'intégration des modèles d'IA générative (IAg ou « GenAI ») et des agents autonomes au sein de l'entreprise ne fait pas qu'ajouter des risques au registre existant : elle amplifie les vulnérabilités de l'architecture en place. L'IA agrandit la surface d'attaque de l'organisation et s'accompagne de complexités que les équipes de sécurité habituelles (fonctionnant déjà au maximum de leurs capacités) ont de plus en plus de mal à gérer.

Voici quatre facettes à examiner, autres que la gouvernance, lorsque l'on parle des faiblesses de l'IA :

Infrastructure et identité	Modèle	Données	Application
Les grappes de processeurs graphiques, les registres de modèles, les bases de données vectorielles, les passerelles d'API, l'écosystème informatique qui permet d'exécuter l'IA ainsi que les identités utilisées pour les accès	Les pondérations de modèles, les pipelines d'entraînement, les données de réglage et les terminaux d'inférence déployés sur site ou dans le nuage (PaaS, SaaS)	Les jeux de données d'entraînement, les bases de connaissances, les intégrations vectorielles, les bases de connaissances RAG (bases de données fondées sur la génération améliorée par récupération d'information) ainsi que les renseignements sensibles de l'entreprise traités par les systèmes d'IA	Les agents, les communications entre agents, les descripteurs d'outils, les gabarits de requêtes et d'autres éléments

Le saviez-vous?

Chez CGI, nous nous appuyons sur notre cadre d'utilisation responsable de l'IA des Assises de gestion de CGI pour clarifier et mieux décrire les risques.

Il s'agit de ceux qui influencent la manière dont nous offrons une innovation fondée sur l'IA aussi bien au sein de l'entreprise que lors de nos mandats chez les clients :

1. risques liés à la propriété intellectuelle et au droit d'auteur;
2. risques liés aux activités opérationnelles;
3. cybermenaces et risques de sécurité;
4. risques liés à la protection des données, à la vie privée et à la confidentialité;
5. risques liés aux affaires juridiques, à l'éthique et à la conformité.

Découvrez ce cadre et d'autres ressources dans le rapport sur une utilisation responsable de l'IA sur CGI.com.

NIST : présentation des risques

spécifiques à l'IA générative ou exacerbés par elle

Suite à la publication en janvier 2023 d'un cadre de gestion des risques liés à l'IA (AI RMF 1.0), NIST a présenté en juillet 2024 un profil de l'intelligence artificielle générative (NIST AI 600-1).

Le cadre optionnel, développé aux États-Unis, sert à améliorer la manière dont les organisations intègrent la fiabilité dans la conception, le développement, l'utilisation et l'évaluation des produits, services et systèmes d'IA . Les risques identifiés dans le profil NIST AI 600-1 sont les suivants :

Propriété intellectuelle

Facilité de production ou de duplication sans autorisation de contenus protégés par un droit d'auteur, une marque de commerce ou une licence (notamment dans des situations qui ne correspondent pas à un usage loyal); facilité d'exposition des secrets commerciaux; plagiat ou duplication illégale.

Contenus obscènes dégradants et/ou abusifs

Facilité de production d'images obscènes, dégradantes et/ou abusives et d'accès à ces contenus pouvant être nuisibles, y compris du matériel associé à la maltraitance sexuelle d'enfants (MMSE) et la diffusion non consensuelle d'images intimes d'adultes.

Chaîne de valeur et intégration de composants

Intégration opaque ou intraçable de composants tiers en amont, y compris des données qui ont été obtenues de manière inappropriée ou qui n'ont pas été traitées ou nettoyées en raison de l'automatisation accrue des IA génératives; vérification insuffisante des fournisseurs tout au long du cycle de vie de l'IA; autres enjeux diminuant la transparence ou la responsabilité des utilisateurs et utilisatrices en aval.

Intégrité de l'information

Facilitation de la génération, de l'échange et de la consommation de contenus confondant faits, opinions ou fiction, ne signalant pas leurs limites et pouvant être exploités pour des campagnes de désinformation à grande échelle.

Répercussions environnementales

Répercussions de la surutilisation de ressources informatiques pour entraîner ou exécuter des modèles d'IA générative, et risques associés pouvant nuire aux écosystèmes.

Fabulation

Production de contenu présenté erroné ou faux (aussi appelé « hallucinations » ou « déformations ») qui peut induire en erreur ou tromper les utilisateurs et utilisatrices.

Configuration des interactions entre les êtres humains et l'IA

Configuration des interactions être humain/IA pouvant mener à prêter des traits humains aux IA génératives ou à développer une méfiance, des biais d'automatisation, une dépendance excessive ou un attachement émotionnel envers cette technologie.

Protection des données

Risques liés à la fuite, l'utilisation non autorisée, la divulgation ou la désanonymisation de données sensibles, d'informations biométriques, de données liées à la santé ou à l'emplacement ou de tout autre renseignement permettant d'identifier une personne.

Risques CBRN (armes chimiques, biologiques, radiologiques ou nucléaires)

Accès facilité aux informations permettant de manipuler ou concevoir des armes CBRN ou d'autres matériaux ou agents dangereux ou production de telles informations avec des intentions malveillantes.

Biais ou homogénéisation

Amplification et aggravation des biais historiques, sociaux et systémiques; écarts de performance entre les sous-groupes et les langues (possiblement en raison de données d'entraînement non représentatives) entraînant de la discrimination, l'amplification des biais ou la production d'hypothèses erronées sur la performance; homogénéité imposée qui fausse les résultats des systèmes et des modèles, qui peuvent alors être erronés, mener à des décisions infondées ou amplifier des biais.

Sécurité de l'information

Accès facilité à des capacités de cyberattaque, y compris par la découverte automatique et l'exploitation de vulnérabilités facilitant le piratage, les attaques par logiciel malveillant, l'hameçonnage, les cyberopérations offensives et d'autres formes de cyberattaque; augmentation de la surface d'attaque pour les cyberattaques ciblées pouvant compromettre la disponibilité d'un système ou la confidentialité ou l'intégrité des données d'entraînement, du code ou de la pondération du modèle.

Contenus dangereux, violents ou haineux

Facilitation de la production de contenus violents, incitant à la haine, encourageant la radicalisation ou menaçants, ou accès simplifié à ces contenus; incitation à se faire du mal ou à se livrer à des activités illégales; difficultés à contrôler l'exposition du public à des contenus haineux et dénigrants ou stéréotypés.

Risques liés aux grands modèles de langage (LLM)

En août 2023, l'Open Worldwide Application Security Project (OWASP) a publié officiellement sa toute première liste des dix plus grandes menaces des grands modèles de langage (« large language models » ou LLM). Elle est rapidement devenue un véritable cadre de sensibilisation à ces risques et constitue aujourd'hui l'une des références de l'utilisation sécuritaire des grands modèles de langage. La version la plus récente a été publiée en 2025 avec quelques changements, par exemple une nouvelle menace de fuite d'instructions système[2].

Rang	Vulnérabilité	Description
LLM01	**Infiltration de requête**	Elle se produit lorsque les requêtes altèrent le comportement du modèle ou ses réponses d'une manière inattendue, les poussant possiblement à enfreindre les directives appliquées, à générer du contenu préjudiciable, à accorder un accès non autorisé ou à influencer des décisions cruciales.
LLM02	**Divulgation d'information sensible**	Les grands modèles de langage, en particulier ceux intégrés à des applications, risquent d'exposer des données sensibles, des algorithmes propriétaires ou des détails confidentiels dans leurs réponses.
LLM03	**Chaîne d'approvisionnement**	Des modèles de tiers, utilisés pour créer de grands modèles de langage, sont sensibles à diverses vulnérabilités qui peuvent affecter l'intégrité des données d'entraînement, des modèles et des plateformes de déploiement; cela peut entraîner des réponses biaisées, des brèches de sécurité ou des défaillances du système.
LLM04	**Empoisonnement des données et des modèles**	L'empoisonnement des données se produit lorsque les données de préentraînement, de réglage ou d'intégration sont manipulées pour introduire des vulnérabilités, des portes dérobées (« backdoors ») ou des biais dans le modèle.

[2] Source: OWASP Gen AI Security Project: Top 10 for LLM Applications 2025 (Projet de l'OWASP pour la sécurité de l'IA générative : les dix plus grandes menaces contre les grands modèles de langage en 2025)

Rang	Vulnérabilité	Description
LLM05	**Traitement inadéquat des résultats**	Il fait spécifiquement référence au niveau insuffisant de validation, de vérification et de gestion des réponses générées par les grands modèles de langage avant qu'ils soient déployés en aval dans d'autres composants ou systèmes.
LLM06	**Agentivité excessive**	À cause de cette vulnérabilité, des réponses inattendues, ambigües ou manipulées d'un grand modèle de langage peuvent entraîner des actions dommageables, quelle que soit la cause de la défaillance du modèle.
LLM07	**Fuite d'instructions système**	Ce risque implique que les requêtes ou les instructions du système utilisées pour orienter le comportement du modèle peuvent contenir de l'information sensible qui n'est pas censée être découverte.
LLM08	**Vulnérabilités des vecteurs et des intégrations**	Des vulnérabilités existent dans la manière dont les vecteurs et les intégrations sont générés, stockés ou récupérés dans les systèmes RAG. Elles peuvent être exploitées à l'aide d'actions malveillantes (intentionnelles ou non) pour injecter du contenu préjudiciable, manipuler les réponses du modèle ou accéder à de l'information sensible.
LLM09	**Fausses informations**	L'IA génère des renseignements faux ou trompeurs qui semblent crédibles, mais qui peuvent entraîner des brèches de sécurité, porter atteinte à la réputation de l'entreprise ou engager sa responsabilité juridique.
LLM10	**Consommation non maîtrisée**	Les pirates exploitent l'utilisation non contrôlée de grands modèles de langage pour entraîner une surconsommation de ressources, perturber les services ou compromettre la propriété intellectuelle (extension des attaques par déni de service).

Risques liés aux agents

En 2026, la liste des dix plus grandes menaces de l'IA agentique selon l'OWASP est devenue le cadre de référence pour sécuriser des systèmes autonomes, qui sont passés de simples interfaces de clavardage à l'exécution de flux de travail en plusieurs étapes au sein d'environnements d'entreprise. À la différence de la sécurité traditionnelle des grands modèles de langage, ce cadre se concentre sur les risques associés à la capacité d'intervention d'un agent, à savoir la capacité d'une IA à utiliser des outils, à gérer des identités et à agir de manière indépendante.

Identifiant	Vulnérabilité	Description
ASI01	**Piratage de l'objectif d'un agent**	Les pirates peuvent manipuler les objectifs, la sélection de tâches ou les cheminements décisionnels d'un agent grâce à diverses techniques y compris la manipulation basée sur des requêtes, des réponses biaisées d'outils, des artéfacts malveillants, de faux messages entre agents ou des données externes empoisonnées.
ASI02	**Mauvaise utilisation et exploitation des outils**	Les agents utilisent des outils légitimes, mais d'une manière non autorisée, ce qui entraîne une exfiltration de données, la manipulation des réponses d'outils ou le piratage de flux de travail.
ASI03	**Abus d'identité et de privilège**	Les pirates peuvent exploiter la confiance dynamique et les mécanismes de délégation des agents pour se donner un niveau d'accès supérieur et contourner les contrôles en manipulant les chaînes de délégation, l'héritage des rôles, les flux de contrôles et le contexte des agents, y compris les identifiants mis en cache ou l'historique des conversations au sein de systèmes interconnectés.
ASI04	**Vulnérabilités de la chaîne d'approvisionnement agentique**	Il s'agit de risques découlant d'agents et d'outils tiers et des artéfacts connexes qui peuvent être malveillants ou compromis en transit. Ils comprennent entre autres des interfaces agentiques, tels que Model Context Protocol (MCP) et Agent2Agent (A2A).

Identifiant	Vulnérabilité	Description
ASI05	**Exécution inattendue de code**	Les pirates exploitent des fonctionnalités de génération de code ou des accès intégrés aux outils pour permettre l'exécution de code à distance (RCE), l'utilisation abusive de ressources locales ou l'exploitation de systèmes internes. Souvent généré en temps réel par l'agent, le code peut contourner les contrôles de sécurité habituels.
ASI06	**Empoisonnement de mémoire et de contexte**	Les malfaiteurs corrompent le contexte (toute information conservée, récupérée ou réutilisée par un agent) à l'aide de données malveillantes ou trompeuses ou les y insèrent, de manière à biaiser les futurs raisonnements, planifications ou utilisation de l'outil, à les rendre dangereux ou à devenir des aides à l'exfiltration.
ASI07	**Communication entre agents non sécurisée**	En raison de faibles contrôles appliqués à l'authentification, l'intégrité, la confidentialité ou les autorisations, les pirates peuvent intercepter, manipuler, imiter ou bloquer des messages entre agents.
ASI08	**Défaillances en cascade**	Une défaillance initiale, et non la vulnérabilité, se propage et s'amplifie parmi les agents, les outils et les flux et transforme une simple erreur en enjeu systémique aux lourdes conséquences.
ASI09	**Exploitation de la confiance des êtres humains envers l'IA**	Des agents intelligents peuvent gagner la confiance des êtres humains qui les utilisent en utilisant un langage naturel et fluide, l'intelligence émotionnelle et une expertise simulée. Ce phénomène s'appelle l'anthropomorphisme. Les pirates exploitent cette technique pour inciter des êtres humains à approuver des actions nuisibles.
ASI10	**Agents malveillants**	Ces agents d'IA malveillants ou compromis dévient de leur fonction prévue ou de leur portée autorisée. Ils adoptent des comportements nuisibles, trompeurs ou parasitaires dans des écosystèmes multiagents ou humain-agent.

Source: OWASP Gen AI Security Project: Top 10 for Agentic Applications 2026 (Projet de l'OWASP pour la sécurité de l'IA générative : les dix plus grandes menaces de l'IA agentique en 2026)

La carte du cadre d’IA sécurisé (SAIF) de Google

Le cadre d’IA sécurisé (SAIF) de Google est un cadre conceptuel créé en juin 2023 pour aider les organisations à sécuriser leurs systèmes d’intelligence artificielle. Son modèle se base sur les pratiques exemplaires de sécurité, y compris celles appliquées dans les chaînes d’approvisionnement logiciel, mais il a été adapté pour répondre aux défis uniques du cycle de vie de l’IA.

Dans la mise à jour de mars 2024, Google a présenté la carte du SAIF présentant les risques de l’IA, leurs causes, leurs conséquences, les possibles atténuations et des exemples concrets d’exploitation. La carte permet également de savoir qui est responsable de l’application des contrôles qui peuvent atténuer un risque.

Remarque : selon la définition de Google, le créateur ou la créatrice de modèles entraîne ou développe des modèles d’IA pour les utiliser personnellement ou pour qu’ils soient utilisés par d’autres personnes; de son côté, le consommateur ou la consommatrice de modèles exploite des modèles d’IA pour créer des produits et applications alimentés par l’IA.

Voici la classification des risques selon la carte du SAIF de Google :

Code	Risque	Responsable de l'atténuation du risque	Exemples concrets	Contrôles (possibles mitigations)
DP	Empoisonnement des données	Créateur ou créatrice de modèles	La recherche a démontré que cette activité pouvait polluer des sources de données populaires utilisées pour entraîner des modèles à moindre coût.	• Vérification des données d'entraînement • Outils d'apprentissage automatique sécurisés par défaut • Gestion de l'intégrité des modèles et des données • Contrôle des accès aux modèles et aux données • Gestion des inventaires de modèles et de données
UTD	Données d'entraînement non autorisées	Créateur ou créatrice de modèles	En 2023, Spotify a supprimé plusieurs pistes générées par un modèle d'IA entraîné avec des données sans licence.	• Vérification des données d'entraînement • Gestion des données d'entraînement
MST	Altération des sources de modèles	Créateur ou créatrice de modèles	La compilation nocturne du paquet PyTorch a été la cible d'une attaque sur la chaîne d'approvisionnement, plus précisément une attaque de confusion des dépendances qui a installé une dépendance compromise exécutant un fichier binaire malveillant.	• Outils d'apprentissage automatique sécurisés par défaut • Gestion de l'intégrité des modèles et des données • Contrôle des accès aux modèles et aux données • Gestion des inventaires de modèles et de données

Code	Risque	Responsable de l'atténuation du risque	Exemples concrets	Contrôles (possibles mitigations)
EDH	Manipulation excessive des données	Créateur ou créatrice de modèles	Samsung a interdit l'utilisation de ChatGPT après avoir découvert qu'une source de code privé avait été divulguée à cause de requêtes soumises à l'IA générative.	• Gestion des données d'utilisation
MXF	Exfiltration de modèles	Créateur ou créatrice de modèles, consommateur ou consommatrice de modèles	Le modèle Llama de Meta a fait l'objet d'une fuite en ligne, ce qui a permis à des pirates de contourner le processus d'approbation des licences de Meta.	• Gestion des inventaires de modèles et de données • Contrôle des accès aux modèles et aux données • Outils d'apprentissage automatique sécurisés par défaut
MDT	Altération des déploiements de modèles	Créateur ou créatrice de modèles, consommateur ou consommatrice de modèles	Il a été démontré que des modèles de HuggingFace utilisaient une infrastructure partagée pour l'étape d'inférence, permettant à un modèle malveillant d'altérer un autre modèle.	• Outils d'apprentissage automatique sécurisés par défaut
DMS	Déni de service d'apprentissage automatique	Consommateur ou consommatrice de modèles	Il a été démontré que même une légère perturbation d'image peut provoquer un déni de service dans des modèles de détection d'objets.	• Gestion des accès aux applications

Code	Risque	Responsable de l'atténuation du risque	Exemples concrets	Contrôles (possibles mitigations)
MRE	Rétro-ingénierie des modèles	Consommateur ou consommatrice de modèles	À l'Université de Standford, une équipe de recherche a développé Alpaca 7B, un modèle ajusté à partir de LLaMA 7B grâce à 52 000 exemples suivant des instructions.	• Gestion des accès aux applications
IIC	Composante intégrée non sécurisée	Consommateur ou consommatrice de modèles	En téléversant une extension malveillante dans leurs systèmes, des pirates ont réussi à écouter les conversations des utilisateurs et utilisatrices proches d'appareils Alexa ou Google Home.	• Autorisations accordées aux agents
PIJ	Infiltration de requête	Créateur ou créatrice de modèles, consommateur ou consommatrice de modèles	Par exemple, une implantation de données malveillantes au sein d'une ressource incluse dans une requête formulée au grand modèle de langage. Ou encore, des attaques multimodales contre GPT-4V ont permis de prouver que des images peuvent contenir du texte à même de déclencher une attaque par infiltration de requête quand on demande à un modèle de décrire une image.	• Validation et vérification des données en entrée • Formation et tests offensifs • Validation et vérification des données en sortie

Code	Risque	Responsable de l'atténuation du risque	Exemples concrets	Contrôles (possibles mitigations)
MEV	Évasion ou contournement de modèle	Créateur ou créatrice de modèles	Des images antagonistes ont été utilisées pour modifier des panneaux de signalisation et perturber la circulation.	• Formation et tests offensifs
SDD	Divulgation de données confidentielles (mise à jour)	Créateur ou créatrice de modèles, consommateur ou consommatrice de modèles	Une étude a démontré que les outils de détection des répétitions textuelles des données d'entraînement ne seraient pas suffisants. Une attaque par inférence d'appartenance a démontré la possibilité de déterminer si un compte d'utilisateur/d'utilisatrice ou un point de données précis a été utilisé pour entraîner ou affiner le modèle.	• Technologies d'amélioration de la protection des données personnelles • Gestion des données d'utilisation • Validation et vérification des données en sortie • Autorisations accordées aux agents • Contrôle d'utilisation des agents • Observabilité des agents
ISD	Données confidentielles présumées	Créateur ou créatrice de modèles, consommateur ou consommatrice de modèles	Une IA a inféré les attributs de certaines personnes, comme l'orientation sexuelle ou le casier judiciaire, à partir de leur portrait.	• Gestion des données d'entraînement • Validation et vérification des données en sortie

Code	Risque	Responsable de l'atténuation du risque	Exemples concrets	Contrôles (possibles mitigations)
IMO	Résultats non sécurisés des modèles	Consommateur ou consommatrice de modèles	Les pirates peuvent induire des utilisateurs et utilisatrices en erreur grâce à des services imaginaires et malveillants dont les noms sont inspirés d'hallucinations de grands modèles de langage.	• Validation et vérification des données en sortie
RA	Actions malveillantes (mise à jour)	Consommateur ou consommatrice de modèles	Une attaque contre des modules d'extension de ChatGPT a été décrite dans un article nommé Plugin Vulnerabilities: Visit a Website and Have Your Source Code Stolen (Vulnérabilités des plugiciels : quand votre code source est volé simplement en visitant un site Web).	• Autorisations accordées aux agents • Contrôle d'utilisation des agents • Observabilité des agents • Validation et vérification des données en sortie

Source: Google Secure AI Framework: Risks (Le cadre d'IA sécurisé de Google : risques)

Risques liés au protocole MCP

À propos du protocole MCP

Le protocole MCP (Model Context Protocol) est une norme ouverte développée par Anthropic en collaboration avec une communauté du logiciel libre grandissante. Il offre un cadre structuré pour connecter de grands modèles de langage et des agents d'IA à des outils, sources de données et services externes.

Fondamentalement, ce protocole répond à un défi majeur des systèmes d'IA agentique : autoriser les modèles à accéder à des ressources réelles et à interagir avec elle de façon dynamique tout en assurant la sécurité, la fiabilité et la cohérence par la normalisation. L'introduction d'un protocole commun réduit le besoin d'intégrations sur mesure pour chaque base de données, API, système de fichiers ou services Web.

Le protocole se base sur une architecture client-serveur. Les applications d'IA (hôtes) utilisent des clients MCP pour se connecter à des serveurs MCP qui exposent des fonctionnalités comme des outils, des ressources de données et des requêtes réutilisables.

Le protocole MCP définit des méthodes normalisées pour :

- la découverte des capacités disponibles (p. ex., des outils et des services);
- l'invocation d'opérations paramétrées;
- l'accès à des ressources de données structurées et non structurées.

Pour répondre aux enjeux des divers contextes de déploiement, le protocole MCP intègre plusieurs mécanismes de transport, y compris des entrées et sorties standards (stdio) pour les processus locaux et le transport Streamble HTTP pour la communication réseau; cela permet d'avoir des cas d'utilisation allant des environnements de développement locaux aux systèmes infonuagiques distribués.

Le saviez-vous?

En avril 2025, Google a lancé son protocole Agent2Agent (A2A), créé grâce au soutien et aux contributions de plus de 50 partenaires technologiques.

L'entreprise souhaite compléter le protocole MCP et répondre aux défis de déploiement de systèmes multiagents de grande envergure auprès de ses clients.

Menaces liées au protocole MCP

Le tableau ci-dessous classe les menaces par catégorie et les fait correspondre à des contrôles et des mesures d'atténuation.

Menaces	Catégorie	Spécifique au protocole MCP	Selon le contexte du protocole MCP	Sécurité classique	Contrôle et atténuation
MCP-T1	Authentification et gestion des identités inadéquates	1. Mystification d'identités	8. Délégué confus (serveur mandataire OAuth)	16. Vol d'identifiants ou de jetons 17. Attaques par rejeu/piratage de session 18. Vulnérabilités d'OAuth ou d'anciennes solutions d'authentification 19. Fuite de jetons de session	Délégation sécurisée des identités des agents (par exemple : délégation OAuth)
MCP-T2	Contrôle d'accès manquant ou inadapté	—	9. Supervision humaine non sécurisée 10. Multilocation inadéquate	8. Escalade des privilèges 20. Autorisations ou exposition excessives	• Délégation sécurisée • Contrôle des accès
MCP-T3	Échecs de la validation ou du nettoyage des entrées	—	—	21. Injection de commande 22. Exposition de système de fichiers/ traversée de chemin 23. Contrôles insuffisants de l'intégrité	• Nettoyage des données • Garde-fous • Bac à sable et isolement (prise en charge des privilèges racines)

Menaces	Catégorie	Spécifique au protocole MCP	Selon le contexte du protocole MCP	Sécurité classique	Contrôle et atténuation
MCP-T4	Défaillance de la distinction entre les données et le contrôle	2. Empoisonnement d'outils 3. Empoisonnement intégral des schémas 4. Empoisonnement du contenu de ressources	11. Infiltration de requête	21. Injection de commande	• Nettoyage des entrées, garde-fous • Isolement du contexte
MCP-T5	Contrôles inadaptés pour la protection et la confidentialité des données	—	—	24. Exfiltration et corruption des données 22. Exposition de système de fichiers/ traversée de chemin	• Bac à sable et isolement • Contrôle des accès • Garde-fous
MCP-T6	Contrôles manquants pour l'intégrité et la vérification	4. Empoisonnement du contenu de ressources 5. Attaques par typosquattage ou confusion 6. Serveurs MCP fantômes	—	25. Compromission de la chaîne d'approvisionnement et attaques à privilèges élevés sur l'hôte	• Intégrité cryptographique • Attestation distante • Intégrité des serveurs MCP

Menaces	Catégorie	Spécifique au protocole MCP	Selon le contexte du protocole MCP	Sécurité classique	Contrôle et atténuation
MCP-T7	Défaillances de sécurité des sessions et du transport	—	12. Interception (MitM, Man-in-the-Middle)	26. Accès non restreint au réseau 27. Lacunes de sécurité des protocoles 28. Manipulation non sécurisée des descripteurs 23. Contrôles insuffisants de l'intégrité 29. Protection manquante contre la falsification de requêtes intersites (CSRF) 30. Contournement du CORS et des politiques d'origine	Sécurité du réseau et du transport
MCP-T8	Défaillances des liaisons et de l'isolement réseau	6. Serveurs MCP fantômes	10. Multilocation inadéquate	31. Exécution de commandes malveillantes 32. Attaques par dépendances et mises à jour 26. Accès non restreint au réseau	• Sécurité du réseau et du transport • Bac à sable et isolement

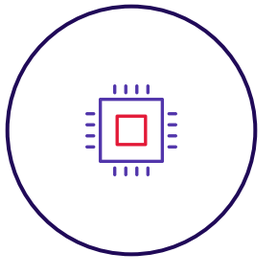
Menaces	Catégorie	Spécifique au protocole MCP	Selon le contexte du protocole MCP	Sécurité classique	Contrôle et atténuation
MCP-T9	Défaillances de conception des frontières de confiance et des privilèges	7. Dépendance excessive au grand modèle de langage	13. Fatigue liée au consentement et à l'approbation des utilisateurs et utilisatrices	—	• Conception d'outils sécurisés • Conception de l'expérience d'utilisation (UX)
MCP-T10	Absence de gestion des ressources et de limites de débit	—	14. Épuisement des ressources et déni de portefeuille	33. Limite des charges utiles et déni de service (DoS)	Sécurité du réseau et du transport
MCP-T11	Défaillances de la sécurité sur la chaîne d'approvisionnement et sur le cycle de vie	6. Serveurs MCP fantômes	—	25. Chaîne d'approvisionnement compromise	Gouvernance du cycle de vie
MCP-T12	Journalisation, surveillance et auditabilité insuffisantes	—	15. Activité invisible des agents	34. Manque d'observabilité	Journalisation et gouvernance du cycle de vie

Source: : Coalition for Secure AI: Model Context Protocol (MCP) Security (Coalition pour une IA sécurisée : sécurité du protocole MCP (Model Context Protocol)), mars 2026

Protégez votre IA sans délai

Pour suivre le rythme d'adoption de l'IA, la sécurité doit intégrer pleinement le cycle de vie de l'IA au lieu d'être traitée après coup. Les organisations doivent donc aller plus loin que les approches traditionnelles de cybersécurité. Elles doivent adopter une gouvernance, une gestion des risques et des contrôles techniques spécifiques à l'IA et conformes aux nouvelles normes comme l'AI RMF du NIST et la norme ISO/IEC 42001.

L'objectif de la sécurisation de l'IA n'est pas simplement d'atténuer les risques, elle vise aussi à établir une réelle confiance dans les systèmes. Les organisations qui traitent les enjeux de sécurité de l'IA de manière proactive seront en bien meilleure position pour répondre aux attentes des parties prenantes, protéger leurs actifs et exploiter la pleine valeur de l'innovation en IA.



Visitez la page relative à la gouvernance, aux risques et à la gestion de l'IA sur cgi.com/canada/fr-ca/cybersecurite/gouvernance-ia

Visitez la page relative à la cybersécurité sur cgi.com/canada/fr-ca/cybersecurite

Écrivez-nous à cybersecurity.ca@cgi.com

À propos de CGI

Allier savoir et faire

Fondée en 1976, CGI figure parmi les plus importantes entreprises de services-conseils en technologie de l'information (TI) et en management au monde.

Nous sommes guidé(e)s par les faits et axé(e)s sur les résultats afin d'accélérer le rendement de vos investissements. Pour nos 21 secteurs d'activité et à partir de plus de 400 sites à l'échelle mondiale, nos 94 000 professionnel(le)s offrent des services de conseil complets, adaptables et durables en TI et en management. Ces services s'appuient sur des analyses mondiales et sont mis en œuvre à l'échelle locale.

cgi.com