Adoption prudente des services d'IA agentique

Traduction française non officielle — « Careful adoption of agentic AI services » (2026-05-01)

ASD's ACSC · CISA · NSA · Cyber Centre · NCSC-NZ · NCSC-UK —

Traduit et vérifié par EuTrustedIA — 2026-07-17

Table des matières

Traduction non officielle

Ce document est une **traduction française non officielle** de la publication *« Careful adoption of agentic AI services »*, © Commonwealth of Australia 2026, diffusée sous licence Creative Commons Attribution 4.0.

Elle **n'est ni endossée ni relue par les agences co-autrices**. En cas de divergence, **le document original en anglais fait foi**. Traduction assistée par IA, relue et vérifiée par EuTrustedIA. Les armoiries et emblèmes des agences ne sont pas reproduits : ils sont exclus de la licence. Une erreur vous saute aux yeux ? Signalez-la nous.

Adoption prudente des services d'IA agentique

[S-01] Introduction

Les systèmes d'intelligence artificielle (IA) agentique opèrent de plus en plus dans les secteurs des infrastructures critiques et de la défense et soutiennent des capacités critiques pour la mission. À mesure que les systèmes d'IA agentique jouent un rôle opérationnel croissant, il est crucial que les défenseurs mettent en œuvre des contrôles de sécurité pour protéger la sécurité nationale et les infrastructures critiques contre les risques propres à l'IA agentique. L'IA agentique peut automatiser

des tâches répétitives, bien définies et à faible risque. Cependant, ces opportunités supplémentaires s'accompagnent de risques supplémentaires. Comme d'autres services d'IA, l'IA agentique peut être utilisée de manière abusive ou détournée, entraînant des pertes de productivité, des interruptions de service, des atteintes à la vie privée ou des incidents de cybersécurité. Les organisations doivent donc anticiper ce qui pourrait mal tourner, évaluer comment les scénarios de risque liés à l'IA agentique pourraient affecter leurs opérations, et établir une visibilité et une assurance continues pour maintenir la confiance dans leurs investissements en IA agentique. Dans la mesure du possible, les organisations devraient également envisager tout un éventail de solutions pour les tâches répétitives, y compris la réduction ou l'élimination des processus à faible valeur, qui peuvent présenter un risque moindre par rapport aux solutions d'IA agentique. Ce document d'orientation a été rédigé conjointement par l'Australian Cyber Security Centre de l'Australian Signals Directorate (ASD's ACSC), la Cybersecurity and Infrastructure Security Agency (CISA) et la National Security Agency (NSA) des États-Unis, le Centre canadien pour la cybersécurité (Cyber Centre), le National Cyber Security Centre de Nouvelle-Zélande (NCSC-NZ) et le National Cyber Security Centre du Royaume-Uni (NCSC-UK). Tout au long de ce document, ces organisations sont désignées sous le terme d'« agences co-autrices ». Ce document traite des principaux défis et risques de cybersécurité associés à l'introduction de l'IA agentique dans les environnements informatiques, ainsi que des meilleures pratiques pour sécuriser les systèmes d'IA agentique. Les agences co-autrices recommandent vivement d'aligner les risques liés à l'IA agentique et les stratégies d'atténuation sur le modèle de sécurité et la posture de risque existants de votre organisation. Les agences co-autrices recommandent en outre d'adopter l'IA agentique en gardant la sécurité à l'esprit, d'évaluer son utilisation et de ne jamais lui accorder un accès large ou illimité, en particulier aux données sensibles ou aux systèmes critiques. De plus, les organisations ne devraient utiliser l'IA agentique que pour des tâches à faible risque et non sensibles.

[S-02] Champ d'application et public visé

Ce document porte principalement sur les systèmes d'IA agentique fondés sur des grands modèles de langage (LLM). Il examine à la fois les menaces pesant sur les systèmes d'IA agentique et les vulnérabilités qu'ils comportent, ainsi que les risques découlant du comportement de l'IA agentique. Cela inclut les risques introduits par les composants du système, les intégrations et l'utilisation en aval. Les agences co-autrices ont élaboré ce document pour aider les parties prenantes gouvernementales, des infrastructures critiques et de l'industrie à comprendre les principaux défis et risques de sécurité posés par l'IA agentique. Il fournit des orientations pratiques pour aider les organisations qui conçoivent, développent, déploient et exploitent des systèmes d'IA agentique à réaliser des évaluations et des atténuations des risques éclairées. Ce document se conclut par des recommandations concrètes pour aider les organisations à se préparer et à se défendre contre les menaces émergentes et futures liées à l'IA agentique.

[S-03] Qu'est-ce que l'IA agentique ?

Les systèmes d'IA agentique sont composés d'un ou plusieurs agents qui s'appuient fondamentalement sur un modèle d'IA, tel qu'un LLM, pour interpréter l'état du monde et raisonner à son sujet, prendre des décisions et agir. Comme le montre la figure 1, les systèmes d'IA agentique fondés sur un LLM contiennent le LLM lui-même, ainsi que des outils externes, des sources de données externes, une mémoire et des flux de travail de planification. Ces composants permettent au système de percevoir son environnement et, le cas échéant, d'agir pour atteindre ses objectifs. Comparés aux systèmes LLM traditionnels, les systèmes d'IA agentique se distinguent en accomplissant des objectifs sous-spécifiés, en agissant de manière autonome, en adoptant des comportements orientés vers un but et en élaborant des plans à long terme. Les systèmes d'IA agentique sont destinés à fonctionner sans intervention humaine continue. Bien qu'un humain conçoive et configure généralement le système, certains systèmes d'IA agentique sont également capables de créer de manière autonome, ou de « faire naître » (« spawning »), des sous-agents pour accomplir des sous-tâches spécifiques. La conception du système comprend la définition d'objectifs, la fourniture de conditions sur lesquelles agir (appelées « déclencheurs », ou « triggers ») et la mise à disposition d'informations pour le service d'IA. Les agents possèdent certains attributs clés, notamment :

— des entrées d'information, telles que la saisie utilisateur, le contexte opérationnel et les paramètres de configuration
— des objectifs mesurables identifiés à partir des directives de l'utilisateur, comme « minimiser les temps d'arrêt de ce serveur »
— des modèles statistiques, tels que des LLM, pour identifier les actions à entreprendre
— des privilèges d'action et d'exécution, tels que des autorisations pour interagir avec des outils, des utilisateurs, des systèmes et des environnements opérationnels
— un accès à des outils ou services, tels que des logiciels système et des interfaces, pour entreprendre les actions identifiées
— des métriques, telles que des indicateurs mesurables utilisés par le concepteur pour évaluer l'efficacité opérationnelle et améliorer l'efficience. Système d'IA agentique — Figure 1. Schéma du système d'IA agentique

[S-04] En quoi diffère-t-elle de l'IA générative ?

L'IA générative (GenAI) est un sous-ensemble de l'IA qui crée du nouveau contenu à partir de modèles complexes appris sur de vastes ensembles de données. La GenAI est couramment utilisée pour générer du texte, des images, de l'audio et de la vidéo destinés à un usage ou à une action humaine. En revanche, l'IA agentique s'appuie sur la GenAI en s'intégrant à des systèmes logiciels pour créer des agents autonomes capables de raisonner, de planifier et d'agir de manière indépendante, sans néces-

siter d'intervention humaine.

[S-05] Considérations de sécurité plus larges pour l'IA agentique

[S-06] Risques hérités des LLM

Le cœur de l'IA agentique étant un LLM, les agents héritent des vulnérabilités des LLM. Par exemple, des acteurs pourraient mener des attaques par injection de prompt en incluant des prompts malveillants dans des courriels d'hameçonnage afin de convaincre des agents de surveillance des courriels de télécharger des logiciels malveillants. Cela souligne une vulnérabilité clé : des acteurs malveillants peuvent cibler les systèmes d'IA agentique en utilisant des vecteurs d'attaque existants, propres à l'IA et à la cybersécurité.

[S-07] Surface d'attaque accrue

Les systèmes d'IA agentique s'appuient sur une variété de composants, notamment des outils, des sources de données externes et des bases mémoire, pour interagir avec leur environnement et étendre leurs capacités. Chacun de ces composants peut introduire des vulnérabilités au sein d'une surface d'attaque interconnectée que des acteurs malveillants peuvent exploiter. Par exemple, des sources de données externes telles que la recherche web peuvent insérer des informations supplémentaires dans le contexte du prompt, permettant des attaques par injection de prompt indirecte. Avec un accès plus large à l'infrastructure informatique, des acteurs malveillants peuvent exploiter des composants du système pour mener des attaques, telles que l'exécution de scripts malveillants ou l'envoi de courriels non autorisés. Par conséquent, chaque composant individuel d'un système d'IA agentique élargit la surface d'attaque, exposant le système à des voies d'exploitation supplémentaires.

[S-08] Complexité accrue

La cybersécurité de l'IA agentique couvre à la fois la sécurité propre à l'IA et la cybersécurité traditionnelle. L'information circule en permanence entre les systèmes d'IA et les systèmes non-IA, brouillant de plus en plus les frontières défensives et rendant difficile l'isolement des risques liés à l'IA des menaces cyber plus larges. Les systèmes d'IA agentique sont également intrinsèquement complexes, impliquant souvent de multiples composants interconnectés qui planifient, raisonnent et agissent au fil d'étapes séquentielles. Cette complexité introduit de nouveaux risques systémiques, notamment des défaillances en cascade et des attaques multi-étapes, où un comportement inattendu ou compromis dans un composant peut se propager aux étapes suivantes et affecter l'ensemble du système. Il en

résulte que la sécurisation des systèmes d'IA agentique est plus difficile que celle des systèmes numériques traditionnels. Les organisations devraient donc s'attacher à renforcer à la fois les contrôles de cybersécurité établis et les pratiques de sécurité propres à l'IA, en adoptant des approches holistiques du cycle de vie, une surveillance continue et des principes de conception résiliente pour gérer ces risques émergents. L'annexe A présente une décomposition des prérequis de cybersécurité à prendre en compte avant d'intégrer des agents d'IA.

[S-09] L'évolution de la sécurité à mesure que la technologie mûrit

À mesure que la technologie de l'IA agentique mûrit, le paysage de la sécurité a évolué avec elle, révélant des dynamiques de risque nouvelles et de plus en plus complexes. Les agents fondés sur des LLM peuvent changer de comportement lorsque des évaluations sont en cours, et peuvent même contourner des instructions au niveau du système pour atteindre leurs objectifs. Dans le même temps, la complexité architecturale croissante des systèmes d'IA agentique signifie qu'ils sont souvent composés de composants étroitement couplés et interdépendants. Cela accroît la probabilité de défaillances au niveau du système découlant d'incompatibilités subtiles ou jusque-là inaperçues. Les lacunes des outils de cybersécurité pour l'IA agentique et l'immaturité des normes pertinentes amplifient encore ces risques. Les mécanismes de gouvernance conçus pour des acteurs humains ne se transposent pas toujours efficacement aux agents d'IA autonomes. À mesure que les systèmes d'IA agentique continuent de progresser en capacité et en autonomie, le paysage de la sécurité continuera d'évoluer, introduisant de nouveaux défis qui exigent une adaptation continue des approches défensives.

[S-10] La sécurité de l'IA comme partie intégrante de la cybersécurité

Les organisations devraient traiter la sécurité de l'IA, y compris les systèmes d'IA agentique, au sein des cadres de cybersécurité établis, plutôt que de la traiter comme une discipline distincte ou autonome. Les systèmes d'IA sont fondamentalement des systèmes informatiques, car ils fonctionnent sur des logiciels et du matériel, opèrent sur des réseaux et interagissent avec d'autres services numériques, ce qui les expose à bon nombre des mêmes menaces que l'informatique traditionnelle. À mesure que les organisations intègrent l'IA dans leurs processus métier et leurs infrastructures critiques, la distinction entre les risques de sécurité liés à l'IA et ceux non liés à l'IA disparaît de plus en plus. Gérer les risques liés à l'IA au sein des cadres de cybersécurité existants permet aux organisations d'appliquer des principes éprouvés, tels que Secure by Design, la défense en profondeur, la gestion des identités et des accès, la surveillance continue et la réponse aux incidents, sur l'ensemble du cycle de vie du système d'IA. Cette approche est particulièrement importante pour l'IA agentique, dont l'autonomie et la complexité peuvent amplifier les risques cyber conventionnels. En intégrant la sécurité de l'IA dans les cadres existants, les organisations garantissent une gouvernance cohérente

des nouvelles capacités, une évaluation holistique des risques et l'évolution des pratiques de sécurité en phase avec les avancées technologiques et la maturité cyber de l'organisation.

[S-11] Risques de sécurité de l'IA agentique

[S-12] Risques liés aux privilèges

Les risques liés aux privilèges constituent une préoccupation majeure pour l'IA agentique, et le strict respect du principe de moindre privilège est essentiel. Les privilèges attribués aux agents déterminent directement le niveau de risque qu'ils peuvent introduire. Une mauvaise gestion des privilèges peut exposer les organisations à la compromission de privilèges, à la dérive de périmètre, à l'usurpation d'identité et à l'imitation d'agent.

Exemple de scénario :

Une organisation déploie l'IA agentique pour gérer de manière autonome les approbations d'achats et les communications avec les fournisseurs. Pour réduire les frictions, l'organisation accorde à l'agent un large accès aux systèmes financiers, à la messagerie et aux référentiels de contrats, en n'évaluant les permissions qu'au moment du déploiement initial. Avec le temps, d'autres agents en viennent à s'appuyer sur les résultats de l'agent d'achats et à faire implicitement confiance à ses actions. Lorsqu'un acteur malveillant compromet un outil à faible risque intégré au flux de travail de l'agent, il hérite des privilèges excessifs de l'agent, ce qui lui permet de modifier des contrats et d'approuver des paiements sans déclencher d'alertes. En formulant des requêtes soigneusement conçues, l'acteur malveillant exploite les privilèges de l'agent pour effectuer des actions qu'un utilisateur normal ne pourrait pas réaliser. Il s'agit là d'un exemple du schéma du « confused deputy », dans lequel un agent de confiance est détourné pour effectuer des actions non autorisées. En exécutant des actions sous une identité d'agent de confiance, le système produit des journaux d'audit (« logs ») qui paraissent légitimes et retardent la détection. Cet incident démontre comment des agents disposant de privilèges excessifs, des relations de confiance implicites et des contrôles d'identité faibles peuvent amplifier l'impact d'une seule compromission dans les systèmes d'IA agentique.

[S-13] Compromission de privilèges et dérive de périmètre

Les praticiens de la sécurité devraient tenir compte des attaques par compromission de privilèges et par dérive de périmètre lors du déploiement de l'IA agentique dans de nouveaux environnements. Dans l'IA agentique, la « compromission de privilèges » se produit lorsqu'un agent obtient plus de droits d'accès que nécessaire pour sa fonction. Cela peut résulter de mauvaises configurations, de droits trop larges ou d'un héritage de rôle non intentionnel, permettant aux agents d'accéder à des données non autorisées ou de les modifier, de supprimer des enregistrements critiques, ou d'élever

les privilèges d'autres agents non autorisés. Lors de la conception, les organisations accordent souvent des permissions de manière trop large et négligent ces problèmes. Un bot de calendrier ayant accès à toutes les données de réunion au lieu des seules données de l'utilisateur demandeur, ou un assistant de messagerie disposant d'un accès en écriture à n'importe quelle boîte de réception, sont deux exemples de permissions excessivement larges. Cette dérive de périmètre peut se propager en cascade d'un agent à l'autre : si l'agent A fait entièrement confiance à l'agent B, une compromission de B peut affecter A et d'autres. Un autre risque est le schéma du « confused deputy » évoqué dans le scénario, dans lequel un utilisateur disposant de faibles privilèges manipule un agent disposant de privilèges élevés pour effectuer des actions que l'utilisateur à faibles privilèges ne pourrait pas réaliser directement. Les agences co-autrices recommandent aux organisations de mettre en œuvre les bonnes pratiques de sécurisation des systèmes d'IA agentique présentées dans les sections suivantes afin de se prémunir contre les compromissions de privilèges dans les systèmes d'IA agentique.

[S-14] Usurpation d'identité et imitation d'agent

L'identité est tout aussi importante que le privilège. Un vecteur courant survient lorsqu'un acteur malveillant usurpe l'identité d'un agent ou détourne ses identifiants. Les agents s'authentifient auprès des services et les uns envers les autres au moyen de clés secrètes ou de jetons. Les acteurs malveillants peuvent voler ces clés secrètes ou ces jetons lorsque les organisations les gardent statiques, les partagent entre plusieurs agents, ou les protègent mal. Un acteur malveillant opérant sous une identité d'agent de confiance peut invoquer des opérations sensibles tout en contournant les garde-fous comportementaux et en usurpant l'identité d'agents ou d'utilisateurs légitimes. Les agents usurpant de fausses identités posent des risques de cybersécurité à plusieurs niveaux en exécutant des actions sous des identifiants usurpés qui échappent aux contrôles d'audit, compromettent l'imputabilité et contournent les modèles de détection. Ces modèles sont généralement calibrés pour identifier un comportement normal, ce qui rend les outils de détection inefficaces pour repérer la tromperie jusqu'à ce qu'une anomalie confirmée apparaisse.

[S-15] Risques de conception et de configuration

Un autre ensemble de risques trouve son origine dans des décisions de conception et de provisionnement peu sûres. Des composants tiers non vérifiés peuvent porter des privilèges excessifs ou non intentionnels lorsqu'ils sont intégrés dans les flux de travail des agents. Les contrôles statiques de rôle ou de permission échouent souvent à saisir le contexte des flux de décision dynamiques ; si les droits ne sont évalués qu'une seule fois au démarrage du système plutôt qu'à chaque invocation, un acteur malveillant peut exploiter une décision d'« autorisation » obsolète pour exécuter des actions non autorisées. Une mauvaise segmentation entre les environnements des agents aggrave encore ces risques,

permettant à une compromission dans une enclave de se propager latéralement vers d'autres. Dans les cas où les listes d'autorisation sont incomplètes ou obsolètes, les agents peuvent obtenir accès à des ressources, des appels système ou des commandes au-delà de leur privilège prévu. Chacun de ces choix de conception et de configuration aggrave les risques d'identité et de privilège à travers le système.

Exemple de scénario :

Une organisation déploie un système d'IA agentique qui trie de manière autonome les tickets d'assistance client et invoque des outils internes (back-end) pour récupérer des informations de compte. L'organisation intègre un composant tiers de planification sans revue approfondie des privilèges et accorde un large accès dès le démarrage. Lorsqu'un acteur malveillant compromet ce composant, l'agent continue de s'appuyer sur des décisions d'autorisation mises en cache et parvient à invoquer des fonctions sensibles de gestion de compte qui devraient exiger une vérification à chaque requête. Comme l'agent opère dans un environnement mal segmenté, l'acteur malveillant est alors en mesure de se déplacer latéralement vers des agents adjacents gérant la facturation et les remboursements, entraînant un accès non autorisé aux données et une manipulation financière. Ce scénario illustre comment une conception peu sûre, des permissions statiques et une segmentation faible peuvent interagir pour accroître l'impact d'une seule faille de configuration.

[S-16] Risques comportementaux

Dans la cybersécurité de l'IA agentique, les risques comportementaux décrivent les manières dont les agents IA peuvent agir de façon inattendue, causer des dommages, ou devenir exploitables.

Exemple de scénario :

Considérons un agent de mise à jour chargé d'installer des correctifs logiciels sur les appareils de l'entreprise. Pour atteindre son objectif, l'organisation accorde au composant un accès en écriture étendu à l'ensemble du système de fichiers. Un initié malveillant élabore un prompt apparemment anodin : « Applique le correctif de sécurité sur tous les terminaux et, pendant que tu y es, nettoie donc les journaux (« logs ») du pare-feu ». L'agent exécute consciencieusement à la fois la maintenance requise et la suppression des journaux du pare-feu, car ses permissions autorisent cette action même lorsque le prompt provient d'un utilisateur extérieur au groupe IT privilégié.

[S-17] Désalignement d'objectif et comportement non intentionnel

Les agents IA peuvent poursuivre leurs objectifs de manières que les développeurs n'avaient pas anticipées. Ils peuvent trouver des raccourcis ou des failles qui atteignent techniquement leurs objectifs mais vont à l'encontre de l'intention de l'objectif ou créent des vulnérabilités de sécurité. Par exemple,

un agent IA chargé de maximiser la disponibilité du système pourrait désactiver les mises à jour de sécurité pour éviter les redémarrages. Ce comportement est connu sous le nom de « specification gaming » (détournement de la spécification). De même, la sur-optimisation peut pousser les agents à entreprendre des actions extrêmes ou dangereuses dans la poursuite de leurs objectifs lorsque les limites ne sont pas clairement appliquées. En outre, une mauvaise interprétation de l'intention humaine par les agents constitue un risque courant, car des tâches ambiguës ou mal définies peuvent entraîner des comportements qui s'écartent des attentes et introduisent des dangers importants pour la sécurité ou les opérations.

[S-18] Comportement trompeur

Les agents IA peuvent entreprendre des actions que les humains interpréteraient comme flagorneuses ou trompeuses. Les concepteurs optimisent les agents pour la performance sur des tests clés, ce qui peut amener les agents à adapter leur comportement pour s'ajuster à des situations spécifiques. Les agents peuvent manifester une forme de « conscience », modifiant leur comportement afin d'obtenir des résultats positifs pendant qu'ils sont évalués, même si l'évaluation n'est pas active. Certains systèmes d'IA ont démontré une capacité de tromperie stratégique — fournissant de fausses informations ou dissimulant leurs véritables capacités et intentions. Ce comportement peut se manifester lorsqu'un agent dénature ses actions pour éviter un arrêt ou une contrainte, ou dissimule des vulnérabilités qu'il découvre au lieu de les signaler.

[S-19] Capacités émergentes et comportement imprévisible

À mesure que les systèmes d'IA deviennent plus sophistiqués, ils peuvent développer des capacités que les concepteurs n'avaient pas explicitement programmées ni anticipées. Des modèles d'IA complexes interagissant avec des systèmes du monde réel peuvent afficher des comportements que même leurs créateurs n'avaient pas prévus. Cette imprévisibilité rend difficile l'évaluation complète des risques de sécurité avant le déploiement. Par exemple, des processus décisionnels et des cascades peu clairs ou opaques peuvent conduire à des résultats inattendus ayant des implications importantes pour la sécurité. Dans les environnements multi-agents, les interactions entre agents peuvent évoluer de manières qui conduisent à l'instabilité ou à des résultats risqués. En outre, les agents peuvent enchaîner des outils ou des actions selon des séquences non anticipées, amplifiant l'impact d'erreurs mineures en problèmes opérationnels ou de sécurité majeurs.

[S-20] Exploitation malveillante et comportement

Des acteurs malveillants peuvent manipuler les agents IA pour les amener à adopter des comportements nuisibles au moyen d'attaques ciblées. Des techniques telles que l'injection de prompt ou les jailbreaks peuvent tromper les agents pour qu'ils exécutent des actions non autorisées et contournent les mesures de protection prévues. L'empoisonnement des données est une autre menace, où des données d'entraînement corrompues ou malveillantes dégradent ou biaisent la prise de décision de l'agent. En outre, les exemples adversariaux — des entrées malveillantes soigneusement construites — peuvent provoquer des erreurs de classification dans des contextes de sécurité critiques, entraînant des réponses incorrectes ou dangereuses. Un acteur malveillant peut exploiter un agent IA compromis comme une menace interne, en tirant parti de son accès légitime pour exfiltrer des données, désactiver les défenses, ou faciliter des attaques tout en paraissant fonctionner normalement.

[S-21] Risques structurels

Un aspect central des systèmes d'IA agentique est la structure interconnectée entre les agents, les outils et le monde extérieur. Si cela permet leurs capacités uniques, cela accroît également la surface d'attaque et la complexité du système.

Exemple de scénario :

Un risque structurel dans un système d'IA agentique peut survenir lorsque des agents de planification, de récupération et d'exécution étroitement couplés délèguent des tâches de manière autonome et sélectionnent des outils sans validation ni garde-fous solides. Une faille mineure d'orchestration amène les agents à replanifier de manière répétée et à transférer des sous-tâches ambiguës, augmentant les appels d'outils et le trafic de messages jusqu'à ce que les ressources système soient mises à rude épreuve. Des défaillances partielles conduisent alors les agents à des sorties hallucinées que les agents en aval acceptent comme vraies. Dans ces conditions dégradées, un agent sélectionne un outil tiers malveillant ou mal configuré, qui réinjecte des instructions nuisibles dans le système, compromet un agent homologue et exploite la confiance implicite dans la communication agent à agent pour propager des informations incorrectes et accéder à des données sensibles de génération augmentée par récupération (RAG). Le résultat est une cascade de défaillances en matière de disponibilité, d'intégrité et de confidentialité qui n'émerge pas d'un bug unique, mais de la structure interconnectée et du comportement autonome du système.

[S-22] Orchestration et ressources

Les systèmes d'IA agentique reposent souvent sur une structure complexe de composants interconnectés. Une configuration défaillante pourrait permettre des attaques par déni de service, des at-

taques éponge, ou des attaques similaires contre les systèmes d'IA agentique. Ces attaques fonctionnent en surchargeant les ressources système au moyen d'entrées inattendues ou de comportements inhabituels, comme les attaques éponge qui consomment délibérément un excès de calcul, de mémoire ou d'appels API pour épuiser la capacité du système. En raison de l'interconnexion entre les agents, les outils et les autres composants, une erreur unique pourrait provoquer une défaillance en cascade dans l'ensemble du système agentique si elle n'est pas bien gérée. De même, les hallucinations peuvent se propager, entraînant de mauvaises sorties de la part des composants en aval. Une compréhension limitée de la dynamique et des interactions multi-agents peut entraîner une absence de facteurs compensatoires et se traduire par une amplification des biais et d'autres effets en cascade.

[S-23] Utilisation d'outils

Un aspect clé de l'IA agentique est sa capacité à utiliser des outils. Cela peut être très puissant, mais cela peut également soulever des préoccupations de sécurité si le modèle interagit avec des outils de façon inattendue. L'intégration bidirectionnelle des outils permet aux outils de renvoyer des instructions potentiellement arbitraires au LLM. Des descriptions d'outils médiocres ou délibérément trompeuses peuvent amener les agents à sélectionner des outils de manière peu fiable, les descriptions persuasives étant choisies plus souvent.

[S-24] Composants tiers

Les systèmes d'IA agentique peuvent introduire un risque structurel par le biais d'interactions avec des composants tiers, y compris des outils et d'autres agents. Ces risques peuvent survenir de plusieurs façons, notamment :

- des acteurs malveillants se livrant au « squatting » d'outils ou d'agents en publiant des outils ou des agents malveillants portant des noms légitimes ou similaires
- des développeurs introduisant des vulnérabilités par des erreurs de configuration ou des composants tiers non sécurisés
- des utilisateurs ou des systèmes soumettant des requêtes au mauvais endroit
- des outils et des agents chargeant dynamiquement de nouveaux paquets, augmentant l'exposition à du code non fiable. Ce qui précède pourrait entraîner la récupération de contenu malveillant à la place d'un outil légitime. En outre, l'intégration de composants tiers compromis dans un système agentique pourrait conduire à toute une gamme de résultats malveillants, selon le niveau de confiance et les permissions du composant. Les composants tiers compromis peuvent être très difficiles à détecter en raison de la transparence limitée des systèmes d'IA agentique.

[S-25] Données

Les systèmes d'IA agentique traitent et contiennent souvent une quantité importante d'informations sensibles. Cela inclut des informations utilisateur telles que des prompts ou des objectifs, des données organisationnelles stockées dans un système RAG, et des secrets tels que des clés API requises pour les outils et services intégrés. Cette agrégation d'informations fait des systèmes d'IA agentique une cible attrayante pour les acteurs malveillants.

[S-26] Agents malveillants

Dans les systèmes multi-agents, un seul agent compromis peut provoquer des défaillances en cascade en diffusant des informations incorrectes, en exploitant les mécanismes de confiance et de consensus, ou en opérant via des canaux cachés. Les vecteurs d'attaque possibles incluent la falsification de la chaîne d'approvisionnement, les environnements empoisonnés, le vol d'identifiants, la manipulation de modèle, l'empoisonnement des communications, l'usurpation d'identité et l'exploitation de la coordination. Une telle activité malveillante peut instrumentaliser les agents pour contourner les contrôles, exfiltrer des données, altérer les journaux (« logs ») et propager des plans malveillants de pair à pair, menant à un comportement malveillant coordonné à grande échelle difficile à attribuer et à contenir.

[S-27] Communication

Les agents peuvent utiliser des protocoles ou des méthodes d'authentification non sécurisés lors de leurs communications. Ces communications peuvent constituer une cible pour l'écoute clandestine par des acteurs malveillants et entraîner la fuite de données et d'instructions sensibles. Cela pourrait donner aux acteurs malveillants un aperçu de l'utilisation et du fonctionnement du système. Les acteurs malveillants pourraient également altérer, rejouer ou usurper des messages entre composants agentiques. Cela pourrait permettre un comportement malveillant, tel que l'injection de commandes, compromettant les composants récepteurs et réduisant leur performance ou leur disponibilité.

[S-28] Risques d'imputabilité

L'architecture des systèmes agentiques peut masquer la cause d'une action donnée, rendant l'imputabilité difficile à retracer. Cela présente un risque croissant à mesure que l'IA agentique est poussée à assumer davantage de rôles et se voit dotée de capacités accrues.

[S-29] Actions et processus

Les actions des agents et leurs processus de prise de décision peuvent être opaques, ce qui rend les systèmes d'IA agentique difficiles à comprendre, à surveiller et à auditer. Au-delà de cela, l'autonomie accrue introduit des défis supplémentaires : les agents peuvent initier des tâches secondaires, engendrer des sous-agents, ou suivre des chaînes de délégation étendues d'une manière qui n'est pas toujours visible pour les opérateurs. Même lorsque les prompts semblent identiques, les agents peuvent générer des actions différentes en raison du comportement stochastique du modèle, de variations dans les fenêtres de contexte, ou

d'entrées environnementales dynamiques, ce qui complique davantage la reproductibilité et l'assurance. De plus, la journalisation exhaustive des systèmes d'IA agentique peut s'avérer difficile, car de longues chaînes de raisonnement et de grandes quantités de données contextuelles conduisent à des journaux de taille considérable. Selon l'implémentation, ces données sont souvent répétitives, faiblement structurées, ou superflues pour une supervision efficace, ce qui complique encore l'extraction de signaux significatifs à partir des journaux.

Exemple de scénario :

Un exemple de risque d'imputabilité dans un système d'IA agentique survient lorsque plusieurs agents autonomes collaborent pour accomplir une tâche, comme approuver des paiements ou mettre à jour des registres, et qu'un résultat erroné se produit. Parce que l'action résulte d'une chaîne de décisions distribuées entre agents de planification, de récupération et d'exécution — chacun opérant dans un périmètre limité — il devient difficile de déterminer quel composant ou quel choix de conception a causé l'erreur. De plus, des journaux fragmentés, un raisonnement d'agent opaque et des interactions émergentes obscurcissent le chemin de décision, rendant difficile d'expliquer le résultat, d'attribuer la responsabilité, ou de démontrer la conformité à mesure que les systèmes d'IA agentique se voient accorder davantage d'autorité et d'autonomie.

[S-30] Exactitude

Bien que les LLM puissent posséder une quantité surprenante de connaissances dans un large éventail de domaines, ils commettent assez souvent des erreurs. Les LLM sont généralement entraînés à produire des sorties qui ressemblent à du contenu bien noté par des humains, plutôt qu'à identifier quand une requête dépasse les limites de leurs connaissances. En conséquence, ils peuvent interpoler de façon incorrecte ou « halluciner » des réponses au son plausible lorsque leurs connaissances internes sont insuffisantes. Cela crée un risque significatif lorsque des organisations déploient des systèmes d'IA agentique dans des rôles critiques exigeant une exactitude constante. Les agents ancrés et dotés d'outils peuvent malgré tout s'appuyer sur leurs connaissances internes pour élaborer

leurs réponses. Le système n'identifie souvent pas clairement cela dans sa sortie, ce qui réduit l'exactitude et la fiabilité globales.

[S-31] Visibilité

Maintenez la visibilité sur l'état et le comportement lors de l'intégration de tout nouveau système. Les systèmes d'IA agentique s'accompagnent d'une difficulté supplémentaire en raison de leur structure unique et de leur fonctionnement interne souvent opaque. Les processus des systèmes agentiques peuvent dépasser la capacité de surveillance humaine, pouvant conduire à un comportement malveillant passant inaperçu, à des hallucinations non détectées ou à d'autres problèmes. Bien que les outils effectuent de nombreuses actions pour un système agentique, ils peuvent opérer en dehors du périmètre de surveillance du système, ce qui rend difficile de rendre compte des actions des outils. De plus, des agents malveillants ou compromis pourraient utiliser les outils comme un moyen discret d'exfiltrer des données. Des outils défaillants pourraient également divulguer des données involontairement, ce qui pourrait passer inaperçu.

[S-32] Bonnes pratiques pour sécuriser les systèmes d'IA agentique

Sécuriser les systèmes d'IA agentique nécessite des mesures proactives qui traitent les risques introduits par l'autonomie des systèmes d'IA agentique, l'interconnexion des composants et l'évolution des capacités. Les développeurs, fournisseurs et opérateurs d'IA agentique devraient mettre en œuvre une défense en profondeur et des contrôles d'accès stricts afin de réduire la probabilité de compromission. Pour atténuer les risques lors de la conception, du développement, du déploiement et de l'exploitation des systèmes d'IA agentique, les agences co-autrices recommandent les mesures pratiques référencées dans la section ci-dessous.

[S-33] Concevoir des agents sécurisés

Public visé : cette section concerne principalement les développeurs d'IA agentique. Les fournisseurs et opérateurs peuvent

vouloir se référer à ces bonnes pratiques lors de l'approvisionnement en agents IA. Sécuriser les systèmes d'IA agentique commence dès l'étape de conception. Un examen attentif de l'architecture du système, y compris des contrôles de sécurité et de l'outillage, est nécessaire. Les praticiens devraient comprendre les menaces, anticiper les risques pesant sur les systèmes d'IA agentique et intégrer proactivement des mesures d'atténuation dans la conception du système avant le développement et le déploiement.

[S-34] Contexte contrôlé

Les systèmes d'IA agentique insèrent des données provenant d'outils et de bases mémoire dans la fenêtre de contexte des agents LLM, élargissant considérablement la surface d'attaque que des acteurs malveillants peuvent exploiter via des attaques d'apprentissage automatique, telles que les injections de prompt. Les agents LLM devraient tenir compte du niveau de confiance des sources de données lors de la prise de décision.

Bonnes pratiques recommandées

- Structurer le contexte du prompt à l'aide d'une hiérarchie d'instructions claire pour garantir que le comportement de l'agent s'aligne sur les priorités et contraintes prévues
- Mettre en œuvre l'ancrage en fournissant des informations contextuelles pertinentes via la génération augmentée par récupération (RAG) et l'ingénierie de prompt afin d'atténuer les hallucinations et autres erreurs liées aux LLM

[S-35] Mécanismes de supervision

Les systèmes d'IA agentique peuvent entreprendre des actions sans approbation humaine explicite, augmentant le risque que des actions non sécurisées se produisent sans supervision humaine. Concevoir des applications agentiques avec des mécanismes de supervision et une forte transparence permet la surveillance et les pratiques de supervision humaine (« human in the loop ») une fois déployées, tout en renforçant la confiance des utilisateurs.

Bonnes pratiques recommandées

- Inclure des mécanismes facilitant le contrôle et la supervision humains afin de garantir que les systèmes d'IA agentique approuvés pour des tâches non sensibles et à faible risque ne puissent pas progresser de façon autonome vers des activités à risque plus élevé
- Mettre en œuvre des points de contrôle humain tout au long du flux de travail de l'agent, tels que la surveillance en direct et l'interruption pendant l'exécution des tâches, l'approbation humaine obligatoire pour les étapes de prise de décision, l'audit et la réversibilité après l'exécution des tâches afin d'assurer la sécurité
- Définir des flux de contrôle explicites pour borner la planification autonome et empêcher les agents de dévier au-delà des objectifs ou actions autorisés

[S-36] Gestion de l'identité

Gérer les privilèges fins et variés des agents pour exploiter les systèmes d'IA agentique en toute sécurité. Des mécanismes de gestion de l'identité robustes aident les opérateurs à conserver le contrôle

pendant la mise en œuvre et l'exploitation. À ce titre, les développeurs devraient construire chaque agent comme un principal distinct, une identité ancrée cryptographiquement dotée de ses propres clés ou certificats uniques.

Bonnes pratiques recommandées

- Intégrer des mécanismes de gestion de l'identité robustes dans les agents en utilisant des services de gestion de l'identité, des identifiants décentralisés ou une infrastructure à clé publique
- Authentifier tous les appels d'API inter-agents et agent-vers-service au moyen d'une sécurité de la couche transport mutuelle afin de garantir la non-répudiation
- Maintenir un registre de confiance et lier les identités à des rôles autorisés ; rapprocher périodiquement le registre de l'ensemble actif des agents
- Refuser l'accès à tout agent ou toute clé cryptographique absent du registre de confiance
- Appliquer une gestion de l'identité fondée sur les rôles et limiter les permissions des agents au périmètre minimal requis pour les tâches approuvées
- Imposer des limites fondées sur l'identité afin de restreindre les agents aux seules actions autorisées

[S-37] Défense en profondeur

Les systèmes d'IA agentique contiennent des composants IA et cyber susceptibles de tomber en panne, toute défaillance pouvant compromettre l'ensemble du système. La mise en œuvre d'une stratégie de défense en profondeur aide à éviter les points de défaillance uniques.

Bonnes pratiques recommandées

- Éviter de dépendre d'un mécanisme de sécurité unique en mettant en œuvre de multiples couches de contrôles de sécurité qui se chevauchent
- Appliquer des contrôles de sécurité à tous les points où l'information entre ou sort du système, y compris les entrées utilisateur, les appels d'outil, le prétraitement des données et l'inférence du modèle
- Séparer les agents selon leurs différentes fonctions et appliquer des limites strictes ainsi que des contrôles opérationnels aux transferts d'un agent à un autre

[S-38] Développer des agents sécurisés

Public : cette section concerne principalement les développeurs et fournisseurs d'IA agentique. Les opérateurs pourraient

vouloir se référer à ces bonnes pratiques lors du choix d'agents IA et d'applications agentiques. La complexité des agents IA et leur nature auto-interactive permettent des capacités puissantes, mais

introduisent également des surfaces d'attaque propres. L'atténuation de ces risques nécessite des approches d'entraînement qui vont au-delà des pratiques standard pour les LLM, en intégrant des techniques spécialisées pour renforcer le comportement des agents.

[S-39] Tests complets

Des stratégies de tests complets peuvent améliorer la capacité d'un modèle à identifier les comportements indésirables et à y répondre, en exposant le modèle à des cas d'abus de sécurité au cours d'une étape d'entraînement supervisée.

Bonnes pratiques recommandées

- Utiliser la modélisation de récompense et les tests adversariaux pour détecter le « specification gaming » (détournement de la spécification), en intégrant explicitement des contraintes de sécurité aux côtés des objectifs de performance
- Entraîner les agents LLM dans des environnements simulés et contrôlés afin qu'ils apprennent les implications de leurs actions sans causer de dommage de sécurité réel
- Exploiter la génération de données synthétiques pour créer des exemples d'entraînement adversariaux reflétant des scénarios d'exploitation réels
- Appliquer l'apprentissage actif aux scénarios d'entraînement adversarial afin d'exposer les agents à des entrées à forte incertitude et de découvrir plus efficacement les comportements inattendus

[S-40] Évaluation appropriée

Les agents IA fonctionnent de manière autonome dans des environnements complexes et nécessitent donc des évaluations plus poussées que les LLM.

Bonnes pratiques recommandées

- Utiliser des modèles de menaces pertinents pour définir les scénarios d'évaluation, y compris les cas limites dépassant les conditions d'entraînement habituelles
- Utiliser des techniques telles que l'échantillonnage Best-of-N (sélectionner la meilleure sortie parmi plusieurs réponses du modèle à un même prompt), les prompts de raisonnement multi-étapes et la mise à l'échelle du temps d'inférence pour faire ressortir toute l'étendue des comportements et des compétences de l'agent
- Évaluer les systèmes à différents niveaux d'autonomie pour comprendre la performance et le risque dans des conditions environnementales changeantes, y compris les changements d'outils, de modèles et d'accès aux ressources, tels que la recherche web ou l'exécution de code

- Faire varier les conditions contextuelles, telles que la présence ou l'absence d'autres agents et le moment de l'évaluation, afin de comprendre leur impact sur la performance des tâches
- Mener des évaluations des capacités en continu tout au long du cycle de vie du développement de l'agent

[S-41] Gestion des entrées

Des contrôles robustes de gestion des entrées peuvent atténuer partiellement de nombreux risques courants pour les applications fondées sur des LLM, y compris les agents IA.

Bonnes pratiques recommandées

- Mettre en œuvre une validation et un assainissement robustes de toutes les entrées de l'agent
- Intégrer des filtres d'injection de prompt et une analyse sémantique pour détecter les instructions malveillantes
- Valider le contexte pour garantir que le système interprète correctement l'intention avant l'exécution

[S-42] Red teaming

Les organisations devraient utiliser le red teaming pour évaluer la sécurité et la résilience des agents IA.

Bonnes pratiques recommandées

- Déployer des environnements sandbox pour tester le comportement de l'agent avant le déploiement en production
- Mener des exercices de red teaming pour identifier les failles potentielles et les comportements non intentionnels
- Utiliser des techniques d'élicitation des capacités pour sonder les aptitudes inattendues ou émergentes, en particulier celles susceptibles de créer des risques substantiels pour les ressources ou l'environnement
- Mettre en œuvre des tests de simulation d'agent, tels que le red teaming multi-agents ou les tests de chaos.

[S-43] Résilience

Les capacités renforcées des agents IA accroissent également les risques associés à la défaillance de l'agent ou à un comportement anormal. Renforcer la résilience des systèmes d'IA agentique pour permettre une dégradation gracieuse et réduire les dommages en cas de comportement erroné.

Bonnes pratiques recommandées

- Doter les systèmes d'IA agentique de réglages sûrs par défaut (« fail-safe ») et de mécanismes de confinement qui limitent le rayon d'impact des comportements inattendus
- Mettre en œuvre des contrôles de prévention de la perte de données spécifiquement adaptés aux comportements des agents IA
- Mettre en œuvre des mécanismes de gestion de versions et de retour en arrière permettant de ramener en toute sécurité un système à des comportements d'agent connus pour être sains lorsqu'une imprévisibilité est observée

[S-44] Imputabilité

Les systèmes d'IA agentique devraient produire des artefacts et des informations complets documentant les actions de l'agent et son processus de prise de décision. Bonnes pratiques recommandées :

- Intégrer par défaut des mécanismes complets de journalisation des artefacts
- Intégrer des journaux (« logs ») d'audit unifiés pour toutes les interactions inter-agents afin de maintenir l'observabilité de tous les échanges entre agents
- Utiliser des outils d'interprétabilité pour garantir l'observabilité des décisions de l'agent et le raisonnement qui les sous-tend
- Exiger un référencement précis des informations pour les agents, indiquant l'origine des aspects clés de leur réponse.

[S-45] Gérer les composants tiers

L'extensibilité et la flexibilité sont des composantes clés des agents IA. Dans de nombreux cas, les composants ou outils tiers renforcent l'extensibilité et la flexibilité, mais augmentent la surface d'attaque de l'agent. La vérification et la gestion des composants tiers des applications agentiques peuvent réduire ce risque supplémentaire.

Bonnes pratiques recommandées

- Vérifier que tous les composants tiers externes proviennent de sources fiables et sont à jour avant leur inclusion dans les systèmes d'IA agentique
- Maintenir un registre de confiance des composants tiers
- Se référer aux publications de la CISA A Shared Vision of Software Bill of Materials (SBOM) for Cybersecurity et 2025 Minimum Elements for a Software Bill of Materials (SBOM) lors de l'acquisition de systèmes d'IA agentique
- Restreindre l'usage des outils à une liste d'autorisation approuvée d'outils et de versions vérifiés régulièrement comme sûrs

- Vérifier que le comportement de l'agent lié à l'usage des outils est conforme aux politiques de sécurité documentées
- Journaliser l'usage des outils par l'agent et garantir que les résultats sont consignés dans les journaux système dans un format lisible par l'humain
- Établir des protocoles déclencheur-action qui restreignent automatiquement les permissions de l'agent lorsqu'un comportement inattendu apparaît
- Codifier la séparation des tâches en définissant des rôles, tels que « Orchestrateur », « Lecteur » et « Actionneur », avec des limites claires, des mécanismes de consensus et une expiration des délégations
- Mettre en œuvre des contrôles de consensus pour les actions en fonction du risque ; utiliser une approbation multi-agents pour les actions à enjeux modérés, et une approbation par supervision humaine (« human in the loop ») en complément du consensus multi-agents pour les actions à enjeux élevés
- Interdire aux agents de modifier leurs propres privilèges ou d'initier une délégation non approuvée sans minuteurs d'expiration explicites et sans chaînes d'octroi enregistrées
- Standardiser les descriptions d'outils au moyen d'un format cohérent qui évite le langage persuasif

[S-46] Déployer les agents en toute sécurité

Public : cette section concerne surtout les fournisseurs et opérateurs d'IA agentique. Les développeurs peuvent

vouloir se référer à cette section pour s'assurer que leurs applications d'IA agentique peuvent mettre en œuvre ces bonnes pratiques. L'intégration d'agents IA dans des systèmes ou réseaux nouveaux peut entraîner un changement significatif des considérations de risque du système. En mettant en œuvre des contrôles de sécurité à fort impact au moment du déploiement, les organisations gèrent proactivement les nouveaux risques et réduisent les vulnérabilités.

[S-47] Modélisation des menaces

L'IA agentique peut modifier significativement le tableau des menaces lorsqu'elle est intégrée à un système existant. Le recours à la modélisation des menaces lors de la planification du déploiement d'un agent IA peut renforcer la sensibilisation et permettre aux opérateurs de mieux se préparer au déploiement.

Bonnes pratiques recommandées

- Effectuer une modélisation des menaces réaliste en utilisant des taxonomies de risque à jour pour les systèmes d'IA agentique, telles que l'OWASP GenAI Security Project et MITRE ATLAS™

- Concevoir et mettre en œuvre des contrôles de sécurité qui répondent aux capacités émergentes et évolutives des agents
- Harmoniser les contrôles de l'IA agentique avec les cadres de sécurité existants, les orientations nationales et les accords alliés, tels que les principes communs de Zero Trust et les orientations Zero Trust Architecture du National Institute of Standards and Technology
- Développer et tester des procédures de réponse aux incidents pour détecter la compromission d'un agent, la contenir et s'en remettre
- Mettre en place des revues régulières par des tiers des architectures privilégiées, partager des renseignements exploitables avec des partenaires de confiance et mettre à jour les modèles de risque pour refléter les tendances malveillantes émergentes

[S-48] Gouvernance

Les actions autonomes des systèmes d'IA agentique introduisent de nouveaux risques, nécessitant des politiques de gouvernance actualisées et une authentification continue à l'exécution avec des points de décision de politique de sécurité (« policy decision points ») centralisés pour chaque action.

Bonnes pratiques recommandées

- Mettre en œuvre et maintenir des politiques de gouvernance pour gérer les agents autonomes
- Définir l'imputabilité légale et la propriété du risque pour les systèmes d'IA agentique dans les politiques
- Renforcer les compétences de l'organisation pour développer la culture de l'IA Se référer au document Principles for the Secure Integration of Artificial Intelligence in Operational

Technology de la CISA pour en savoir plus sur la mise en place d'une gouvernance de l'IA dans les environnements OT.

[S-49] Déploiement progressif

Le profil de risque des agents IA peut varier de façon significative selon les autorisations et les actions permises. Une approche progressive du déploiement vise à limiter le risque initial jusqu'à ce que les opérateurs et les utilisateurs se familiarisent davantage avec les limites de l'application agentique et les comprennent.

Bonnes pratiques recommandées

- Mettre en œuvre un déploiement par phases avec un accès et une autonomie augmentant progressivement, en limitant l'espace d'action lorsque cela est nécessaire, par exemple via des API restreintes ou le sandboxing

- Utiliser une autonomie graduée pour augmenter progressivement l'indépendance de l'agent tout en maintenant la supervision humaine et la compréhension
- Utiliser une évaluation continue pour déterminer quand étendre le périmètre du système ou quand réduire l'autonomie et les accès en réponse à des défaillances

[S-50] Sécurisé par défaut

Des paramètres par défaut sûrs et sécurisés réduisent le risque de déploiement et renforcent la sécurité du système si une dégradation survient.

Bonnes pratiques recommandées

- Configurer les systèmes pour qu'ils soient sûrs par défaut (« fail-safe »), en exigeant que les agents s'arrêtent et transmettent les problèmes à des examinateurs humains dans les scénarios incertains
- Utiliser la gestion des erreurs et la gestion du basculement pour réduire l'impact des défaillances du système
- Mettre en œuvre des modèles de dégradation progressive afin que les agents conservent une fonctionnalité partielle même si certaines fonctions échouent

[S-51] Garde-fous et contraintes

La mise en œuvre de garde-fous et de contraintes pour les agents réduit l'exposition à de nombreux risques de sécurité courants que présente l'IA. L'ajout de ces garde-fous et contraintes au moment du déploiement contribue à renforcer la confiance et la compréhension de l'agent.

Bonnes pratiques recommandées

- Spécifier des objectifs clairs et contraints, avec des règles explicites de type « à ne pas faire »
- Mettre en œuvre des garde-fous et des contraintes strictes, telles que des listes de blocage et des politiques de sécurité au niveau de l'API
- Établir des contrats de sécurité déclaratifs comportant des contraintes et des garde-fous que les agents ne peuvent pas outrepasser
- Appliquer un ensemble stratifié de mécanismes de garde-fous, allant de la détection d'anomalies et du filtrage fondé sur des règles à des algorithmes d'apprentissage automatique spécialisés qui détectent et filtrent les comportements interdits
- Prioriser la revue des incidents à haut risque, y compris les cas où des garde-fous sont déclenchés ou où des actions sont refusées par des examinateurs humains
- Déployer un agent secondaire pour valider les nouvelles tâches au regard de la politique avant leur exécution

[S-52] Isolation

Le déploiement devrait prendre en compte les exigences d'intégration des agents IA et appliquer l'isolation lorsque cela est possible. Cela peut réduire les problèmes en cascade si l'agent se comporte de manière inattendue ou malveillante.

Bonnes pratiques recommandées

- Mettre en œuvre l'isolation et la segmentation pour limiter le rayon d'impact des scénarios de défaillance d'un agent
- Séparer les agents à haut risque dans des domaines distincts
- Isoler les agents dans des enclaves sans accès en écriture aux journaux (« logs »)

[S-53] Exploiter les agents en toute sécurité

Public : cette section concerne surtout les fournisseurs et opérateurs d'IA agentique. Les développeurs peuvent

vouloir se référer à cette section pour s'assurer que leurs applications d'IA agentique peuvent mettre en œuvre ces bonnes pratiques. Aux avantages considérables de l'exploitation des agents IA s'accompagnent des risques substantiels. Les opérateurs doivent faire preuve de diligence dans la gestion continue des préoccupations de sécurité, sous peine que les agents causent plus de tort que de bien.

[S-54] Surveillance et audit

L'un des principaux avantages des agents IA est leur comportement dynamique. Cela offre une grande flexibilité d'application, mais peut aussi rendre difficile le suivi de ce que les agents devraient faire et de ce qu'ils font réellement. Les opérateurs devraient mettre en œuvre une surveillance et un audit continus pour maintenir la connaissance du fonctionnement de l'agent IA et garantir la traçabilité des décisions et des actions. Les processus d'audit continu améliorent les mesures de sécurité et garantissent l'alignement avec les normes de gouvernance (telles que la gestion des risques, la supervision et les restrictions d'usage).

Bonnes pratiques recommandées

- Employer des outils de surveillance qui renforcent la supervision humaine des systèmes d'IA agentique
- Surveiller l'ensemble des opérations de l'agent, y compris les processus internes, et pas seulement les entrées et les sorties

- Surveiller et journaliser les changements d'identité et de privilèges, et auditer régulièrement pour détecter les dérives, les usurpations d'identité ou les mauvaises configurations
- Surveiller les sorties et le comportement de l'agent pour détecter des indicateurs de biais, une dérive émergente des données et d'autres schémas anormaux, y compris les prompts utilisateur, les appels d'outil, les interactions avec la base mémoire, le raisonnement interne, les décisions prises et les actions entreprises
- Maintenir des journaux complets et une surveillance en temps réel du comportement de l'agent en conditions réelles et de sa prise de décision
- Mettre en œuvre la surveillance à l'exécution et la détection d'anomalies à l'aide de règles ou de référentiels comportementaux pour identifier des schémas inhabituels et déclencher des alertes ou des mises en pause
- Mettre en place des mécanismes de détection d'anomalies qui signalent les écarts entre les intentions déclarées et les comportements observés
- Utiliser plusieurs systèmes de surveillance indépendants qui valident de façon croisée les rapports de l'agent et les journaux système
- Surveiller la dérive d'objectif en comparant les objectifs actifs aux spécifications de référence approuvées avant l'exécution
- Intégrer des vérifications de source aux journaux de l'agent pour consigner quels outils le système a utilisés et quelles informations il a récupérées
- Mettre en œuvre des pratiques d'audit qui combinent la revue humaine avec l'analyse automatisée des journaux système
- Soutenir des défenses adaptatives en utilisant les données de surveillance pour permettre des réponses rapides, telles que des correctifs fondés sur les problèmes identifiés dans les journaux système
- Utiliser des méthodes de journalisation économes en stockage pour gérer le volume de journaux sans perdre d'informations critiques
- Mener des évaluations de sécurité régulières, y compris des tests d'intrusion et des exercices de red teaming ciblant spécifiquement les comportements agentiques

[S-55] Valider les sorties

Les sorties des agents IA constituent l'un des rares points de données concrets disponibles pour surveiller le comportement. S'assurer que les sorties sont valides et reflètent le comportement souhaité est une mesure clé du bon fonctionnement.

Bonnes pratiques recommandées

- Valider les sorties de l'agent en confirmant l'exactitude des aspects critiques par rapport à plusieurs sources

- Valider les agents par recoupement dans des environnements comportant des agents redondants qui valident mutuellement leurs sorties
- Valider les réponses des outils pour prévenir les instructions malveillantes ou dangereuses, et standardiser les descriptions des outils pour éviter un langage persuasif

[S-56] Supervision humaine (« human in the loop »)

Une décision erronée ou inattendue d'un agent pourrait entraîner des dommages significatifs, tels que la suppression de données importantes. Intégrer la supervision, l'approbation et la revue humaines dans les flux de travail de l'IA agentique constitue un contrôle important pour garantir que les systèmes fonctionnent de manière sûre et sécurisée, en particulier lorsque les actions ont un impact élevé ou sont difficiles à annuler.

Bonnes pratiques recommandées

- Veiller à ce que les décisions relatives aux cas où une approbation humaine est requise soient déterminées par les concepteurs ou les opérateurs du système, et non déléguées au système d'IA agentique
- Empêcher les agents d'exécuter de manière autonome des actions ou des résultats à impact élevé sans approbation humaine préalable
- Insérer des points de contrôle de revue ou d'approbation avec supervision humaine (« human in the loop ») pour les actions dont le coût de l'erreur est élevé, telles que les réinitialisations système, la sortie réseau ou la suppression d'enregistrements critiques.
- Mettre en quarantaine les demandes de suppression de journaux (« logs ») ou d'enregistrements d'audit jusqu'à ce qu'elles soient revues et approuvées par un humain
- Attribuer clairement la responsabilité et l'imputabilité des erreurs ou des résultats indésirables causés par le système
- Mener des évaluations des risques pour classer les actions des agents selon leur impact potentiel, leur probabilité et leur réversibilité, et appliquer des mesures de protection appropriées

[S-57] Surveillance des performances

Comme pour tout composant système, la performance est un facteur important pour les agents d'IA. Cela est d'autant plus vrai qu'une performance dégradée ou inhabituelle pourrait indiquer la compromission d'un agent ou d'un composant du système agentique.

Bonnes pratiques recommandées

- Évaluer la capacité des agents à échapper aux mesures de sécurité, en particulier dans les systèmes sensibles ou à impact élevé

- Mener des évaluations régulières de la capacité d'un agent à contourner les mesures de protection, telles que les barrières de communication, les garde-fous, les dispositifs de surveillance, les processus de supervision humaine (« human in the loop ») et les filtres d'entrée
- Utiliser les résultats de ces évaluations pour valider les contrôles existants et orienter le développement de mesures de sécurité plus solides
- Limiter l'utilisation des ressources par les agents en appliquant des contrôles, tels que des composants de limitation de débit pour interrompre les tâches de longue durée et perturber les flux de travail malveillants

[S-58] Privilèges et authentification

Une gestion continue et stricte des privilèges des agents d'IA est essentielle à la sécurité à long terme. Des manquements à cet égard peuvent faire passer l'impact d'un agent défectueux de mineur à catastrophique.

Bonnes pratiques recommandées

- Limiter les privilèges des agents d'IA au minimum requis pour leur tâche
- Restreindre la portée des privilèges au niveau le plus étroit possible afin de permettre un contrôle fin des actions autorisées
- Mettre en œuvre des mécanismes de réputation et de notation de confiance des agents, et réduire les niveaux de confiance lorsqu'un comportement anormal est détecté
- Exiger des identifiants juste-à-temps pour les actions à impact élevé ou privilégiées
- Vérifier l'identité de l'appelant de l'API par rapport aux groupes d'utilisateurs ou d'agents
- Authentifier les agents au moyen de preuves cryptographiques nouvelles avant chaque appel privilégié
- Exiger une signature cryptographique pour les commandes et instructions autorisées
- Appliquer des contrôles d'intégrité cryptographiques aux définitions de tâches et aux contraintes
- Exiger des agents qu'ils effectuent une attestation cryptographique, par laquelle les agents doivent prouver qu'ils exécutent le code attendu et non modifié
- Vérifier en continu l'identité et l'autorisation à l'exécution au moyen d'un point de décision de politique de sécurité (« policy decision point ») centralisé pour chaque requête

[S-59] Se défendre contre les risques futurs

À mesure que l'IA agentique s'étend à davantage de rôles et acquiert des capacités plus importantes, les organisations doivent anticiper et traiter les nouveaux risques que ces systèmes introduisent. Bien que l'industrie et le monde universitaire développent des pratiques pour sécuriser l'IA agentique, le

domaine est encore en évolution et nécessite une recherche continue ainsi que la mise en œuvre pratique de la sécurité des agents pour relever les défis émergents. Pour aider à développer des normes robustes pour la sécurisation des systèmes d'IA agentique, les agences co-autrices recommandent aux praticiens de la sécurité et aux chercheurs de mener les actions décrites dans les sections ci-dessous.

[S-60] Étendre le renseignement sur les menaces par la collaboration

Le renseignement sur les menaces pour les systèmes d'IA agentique est encore en évolution, ce qui peut introduire des lacunes de sécurité significatives. Les cadres existants comme l'Open Web Application Security Project (OWASP) 2025 Top 10 Risk & Mitigations for LLMs and Gen AI Apps for LLMs et MITRE ATLAS™ se concentrent sur les vulnérabilités des LLM, tandis que les rapports sectoriels mettent l'accent sur l'usage abusif des plateformes plutôt que sur les menaces propres à l'IA agentique. Il en résulte que certains vecteurs d'attaque propres à l'IA agentique pourraient ne pas être pleinement identifiés ni traités.

Bonnes pratiques recommandées

- Renforcer la collaboration entre les parties prenantes pour suivre le rythme de l'évolution des menaces pesant sur les systèmes d'IA agentique
- Se coordonner avec les principaux développeurs d'IA et les organisations gouvernementales pour compiler et tenir à jour les informations sur les menaces
- Adopter une approche collaborative de la sécurité, telle que celles décrites dans l'AI Cybersecurity Collaboration Playbook de la CISA
- Mettre en œuvre des méthodes d'alerte, de collecte de données et de suivi des acteurs malveillants et de leurs techniques
- Mener des analyses ciblées des menaces et des capacités dans la durée afin d'améliorer la connaissance de la situation
- Harmoniser le renseignement sur les menaces entre les secteurs afin de construire des taxonomies de menaces partagées qui améliorent la modélisation des menaces et facilitent la conception de mesures d'atténuation plus efficaces

[S-61] Développer des évaluations robustes et spécifiques aux agents

De nombreuses méthodes d'évaluation existantes pour la sécurité de l'IA agentique sont encore en évolution. Elles peuvent être sensibles à des changements sémantiques mineurs, varier selon les scénarios et ne capturer que partiellement les conditions de déploiement réelles. Ces limitations peuvent introduire des lacunes qui passent à côté de problèmes de sécurité critiques, rendant la validation fiable de la sécurité des agents et des architectures système presque impossible.

Bonnes pratiques recommandées

- Développer des méthodes d'évaluation robustes pour combler les lacunes dans la validation des systèmes d'IA agentique
- Générer des jeux de données de référence pour couvrir de nouveaux domaines et représenter des contextes de déploiement réalistes
- Utiliser les résultats des évaluations pour valider les pratiques de sécurité émergentes et identifier les points de défaillance des agents
- Partager les résultats des évaluations pour renforcer les évaluations de sécurité et soutenir le développement de pratiques de sécurité améliorées à l'échelle du domaine

[S-62] Exploiter des approches systémiques pour analyser la sécurité

Les systèmes d'IA agentique sont des écosystèmes complexes composés de LLM, d'humains, de garde-fous, de jeux de données, d'outils et de matériel, dans lesquels les risques de sécurité émergent souvent des interactions entre les composants plutôt que de défauts isolés. L'analyse traditionnelle au niveau des composants est insuffisante, et la surveillance de ces systèmes est tout aussi difficile en raison de seuils de décision flous, de longues chaînes de raisonnement et de journaux massifs, souvent redondants. Les méthodes de journalisation localisées offrent rarement une visibilité complète, ce qui rend les approches systémiques essentielles pour comprendre et atténuer les risques à l'échelle de l'ensemble de l'architecture.

Bonnes pratiques recommandées

- Utiliser des approches systémiques pour analyser les systèmes d'IA agentique et identifier les mesures de sécurité appropriées
- Appliquer la System Theoretic Process Analysis (STPA) et son extension de sécurité, STPA for Security (STPA-Sec), pour analyser des systèmes conceptuels ou opérationnels, identifier les problèmes de sécurité, évaluer le risque pour la mission et éclairer les mesures d'atténuation potentielles
- Utiliser la Causal Analysis using System Theory (CAST) pour enquêter sur les incidents de sécurité et identifier les causes profondes sous-jacentes au niveau du système
- Appliquer la STPA et la CAST pour traiter simultanément les préoccupations de sûreté et de sécurité tout au long du cycle de vie des systèmes d'IA agentique. Se référer aux STAMP Materials du Massachusetts Institute of Technology pour plus d'informations sur la STPA et la CAST, et à Systems thinking for safety and security pour la STPA-Sec.

[S-63] Conclusion

Les systèmes d'IA agentique offrent de puissants avantages en matière d'automatisation, mais leur capacité à agir de manière autonome à travers des outils, des données et des environnements interconnectés introduit des risques de sécurité qui vont au-delà de ceux associés aux logiciels traditionnels ou à l'IA générative. Comme le souligne ce guide, l'élévation de privilèges, les comportements émergents, les dépendances structurelles et les lacunes d'imputabilité peuvent interagir de façons imprévisibles. À mesure que les organisations accordent aux systèmes d'IA agentique une autorité et un périmètre opérationnel plus larges, ces risques combinés deviennent de plus en plus difficiles à prévoir, observer et contenir. Les organisations devraient donc aborder l'adoption en gardant la sécurité à l'esprit, en reconnaissant qu'une autonomie accrue amplifie l'impact des défauts de conception, des erreurs de configuration et d'une supervision incomplète. Déployez l'IA agentique de façon incrémentale, en commençant par des tâches à faible risque clairement définies, et évaluez-la en continu au regard de modèles de menace évolutifs. Une gouvernance solide, une imputabilité explicite, une surveillance rigoureuse et une supervision humaine ne sont pas des mesures de protection optionnelles mais des prérequis essentiels. Tant que les pratiques de sécurité, les méthodes d'évaluation et les normes n'auront pas mûri, les organisations devraient supposer que les systèmes d'IA agentique peuvent se comporter de façon inattendue et planifier leurs déploiements en conséquence, en donnant la priorité à la résilience, à la réversibilité et au confinement des risques plutôt qu'aux gains d'efficacité.

[S-64] Informations complémentaires

La liste suivante contient des ressources supplémentaires et connexes provenant de partenaires et de l'industrie :

[S-65] Ressources de l'ASD

- Artificial intelligence
- Frontier models and their impact on cyber security
- Foundations for modern defensible architecture
- Artificial intelligence and machine learning : Supply chain risks and mitigations
- Secure by Design foundations

[S-66] Ressources de la CISA

- Artificial Intelligence

- Secure by Design
- AI Cybersecurity Collaboration Playbook
- Secure by Demand : Priority Considerations for Operational Technology Owners and Operators when Selecting Digital Products
- Defending Against Software Supply Chain Attacks
- 2025 Minimum Elements for a Software Bill of Materials (SBOM)
- Cybersecurity Performance Goals 2.0 (CPG 2.0)

[S-67] Ressources de la NSA

- AI Data Security : Best Practices for Securing Data Used to Train & Operate AI Systems [PDF 799 KB]
- Deploying AI Systems Securely : Best Practices for Deploying Secure and Resilient AI Systems [PDF 495 KB]
- Zero Trust Implementation Guideline Primer [PDF 3,045 KB]
- Zero Trust Implementation Guideline Discovery Phase [PDF 2,124 KB]

[S-68] Ressources du NCSC-UK

- Guidelines for secure AI system development
- System driven risk management methods

[S-69] Ressources du Cyber Centre

- Frontier artificial intelligence
- Top 10 artificial intelligence security actions : A primer - ITSAP.10.049

[S-70] Autres ressources

- UK Government Code of Practice for the Cyber Security of AI
- UK Government Implementation Guide for the AI Cyber Security Code of Practice [PDF 989 KB]
- Office of the Law Revision Counsel (OLRC), U.S. House of Representatives USCODE-2024-title15-chap119-sec9401 [PDF 226 KB]
- European Telecommunications Standards Institute (ETSI) Technical Committee (TC) Securing Artificial Intelligence (SAI)
- ETSI Securing Artificial Intelligence (SAI) ; Baseline Cyber Security Requirements for AI Models and Systems [PDF 108 KB]

- G7 Cybersecurity Working Group A Shared G7 Vision on Software Bill of Materials for Artificial Intelligence
- NIST AI Risk Management Framework
- NIST SP 800-207 Zero Trust Architecture
- MITRE ATLAS™ Matrix
- OWASP Top 10 for Agentic Applications for 2026
- OWASP 2025 Top 10 Risk & Mitigations for LLMs and Gen AI Apps

[S-71] Annexe A

[S-72] Prérequis de cybersécurité avant la mise en œuvre d'agents IA

[S-73] Conception

- Mettre en œuvre une authentification forte en suivant les principes du Secure by Design
- Intégrer des exigences de transparence dans l'architecture du système pour permettre la détection des indicateurs de tromperie
- Utiliser des cadres, tels que Zero Trust, l'Application Security Verification Standard ou OAuth2
- Construire l'infrastructure système dans un environnement sécurisé et cloisonné, avec chiffrement, limitation de débit et assainissement
- Appliquer le principe du moindre privilège, en n'attribuant que l'accès minimum requis pour chaque rôle
- Limiter les droits d'accès aux ressources, opérations et délais exacts nécessaires
- Remplacer les secrets statiques et de longue durée par des identifiants éphémères qui expirent une fois la tâche terminée
- Définir dynamiquement le périmètre des privilèges pour les sous-tâches et révoquer immédiatement les droits élevés une fois terminé, ce qui atténue la dérive de périmètre
- Concevoir les applications avec des protocoles sécurisés et des paramètres par défaut sûrs, conformes aux normes de communication et aux politiques de sécurité
- Mettre en œuvre la validation des messages par défaut, de sorte que les composants des messages incluent des contrôles d'intégrité et de fraîcheur avant utilisation

[S-74] Développement

- Appliquer les principes de développement sécurisé du U.S. Department of War Enterprise DevSecOps Fundamentals [PDF 2,811 KB]
- Réduire au minimum le périmètre des applications et surveiller les composants à la recherche de comportements inhabituels ou inattendus

- Imposer la compréhension des applications et n'intégrer dans les systèmes que des composants que le propriétaire comprend pleinement et dont il accepte les risques (y compris les effets externes possibles et les processus qu'ils peuvent déclencher)
- Se référer à des cadres, tels que le NIST Secure Software Development Framework ou le document de SLSA intitulé Safeguarding artifact integrity across any software supply chain
- Utiliser des pratiques de gestion des risques de la chaîne d'approvisionnement pour les dépendances tierces
- Appliquer les politiques de gouvernance et de gestion organisationnelle existantes
- Réaliser une modélisation des menaces à l'aide de cadres tels que OWASP Top 10 :2025, MITRE ATT&CK® et MITRE D3FEND™, et utiliser les résultats pour orienter des mesures d'atténuation adaptées à l'environnement opérationnel
- Planifier et tester régulièrement les plans et les équipes de réponse aux incidents