

LES ESSENTIELS

SÉCURISER UN SYSTÈME IA AGENTIQUE

Ce document s'adresse aux organisations qui déploient, configurent ou intègrent un système agentique, y compris des produits sur étagère dont elles maîtrisent la configuration et les connecteurs. Il est à l'usage des architectes, RSSI, exploitants et développeurs ayant une compréhension des architectures de systèmes agentiques.

Sécuriser un système IA agentique repose sur le **cloisonnement** de ce système, la **maîtrise des accès** et des **contrôles déterministes** que les traitements agentiques ne doivent pas pouvoir outrepasser. Néanmoins, **une action autorisée peut être dommageable** : un agent peut se tromper, être manipulé ou combiner dangereusement des opérations permises. La mise en place de contrôles déterministes permet de couvrir les comportements prévus à la conception tout en encadrant les décisions non déterministes des modèles, susceptibles de sortir de ce périmètre.

Sept principes regroupant des mesures organisationnelles et techniques sont à mettre en œuvre pour sécuriser un système agentique.

1/ ANALYSER ET TRAITER LES RISQUES DE BOUT EN BOUT

- **Réaliser une analyse de risque dès la conception** du système agentique et des effets qu'il produit en tenant compte du choix du modèle et de son environnement d'exécution. Actualiser cette analyse après toute évolution.
- **Associer à chaque risque identifié des mesures de sécurité adaptées** relevant de l'autorisation, du confinement, du cloisonnement et de la détection. Désigner un **responsable humain** pour chaque mesure. Bloquer la mise en service tant qu'un risque reste inacceptable.
- **Exiger que le cahier des charges impose des exigences de sécurité issues des mesures de l'analyse de risque** pour tout achat de produit ou service composant le système agentique.

V1.0 (10/26)

2/ LIMITER L'AUTONOMIE DU SYSTÈME AGENTIQUE

- **Définir et documenter la mission** du système agentique à sa conception, puis au minimum tous les ans ou à chaque ajout ou modification significative. Évaluer régulièrement la pertinence et les résultats au regard de la mission et de l'analyse de risque.
- **Inventorier les accès, les ressources et les capacités** du système agentique, et les justifier au regard du strict besoin métier. Faire respecter cet inventaire par des contrôles de configuration externes au modèle.
- **Mettre en œuvre progressivement les accès et les capacités validés**. N'élargir l'autonomie **qu'après avoir vérifié l'efficacité des mesures** de sécurité et des procédures de gestion d'incidents.

3/ CLASSER ET CONTROLER LES ACTIONS DES AGENTS

- **Définir les critères de criticité, documenter l'impact métier** en cas de dysfonctionnement et **classer les actions possibles** selon leur finalité dans l'une des trois catégories : exécution autonome, approbation humaine obligatoire et action interdite.
- **Soumettre à une approbation humaine** les actions critiques (notamment dans les systèmes d'IA à haut risque au sens de l'[article 6 du RIA](#)), les accès à des données sensibles, les modifications d'identité ou de droits, et les déploiements en production. Garder le mécanisme d'approbation **hors de portée du système agentique**.
- **Présenter à l'approbateur humain l'action exacte**, son objectif et ses effets attendus. **N'exécuter que l'action approuvée**.

4/ SÉPARER ET CLOISONNER LES ENVIRONNEMENTS

- **Dédier des identités, secrets et connecteurs** pour chaque environnement : développement, évaluation, production. Si l'analyse de risque le justifie, dédier des ressources physiques à ces environnements.
- **Définir des espaces mémoire du système agentique** par organisation cliente et par environnement. Distinguer les mémoires individuelles des mémoires partagées en identifiant les utilisateurs ou agents y ayant accès.
- **Cloisonner l'environnement d'exécution** des agents, de leurs outils et de leurs connecteurs dans un environnement dédié, isolé, durci et revu régulièrement afin de restreindre l'accès aux données accessibles à l'hôte.
- **Faire transiter les communications externes exclusivement par des interfaces maîtrisées.** Authentifier, chiffrer et filtrer ces flux par destination, protocole et type de données.
- **Considérer que toute donnée transmise à un modèle hébergé peut être accessible à son fournisseur.**

5/ FONDER L'AUTORISATION SUR UN CONTRÔLE DÉTERMINISTE

- **Placer, lorsque cela est réalisable, un mécanisme de filtrage déterministe et indépendant de l'agent,** en amont de chaque outil ou ressource, **sans possibilité de contournement** ou de modification par l'agent.
- **Authentifier le demandeur et vérifier automatiquement ses droits,** l'action et ses paramètres effectifs, sa cible et le respect des règles métier selon une politique inaltérable par le modèle. **Par défaut, refuser l'action.**
- **Vérifier qu'un enchaînement d'actions autorisées individuellement ne produit pas d'effet interdit** par les règles métier ou de sécurité.
- **N'accorder aux agents aucune confiance ni aucun privilège implicites.**

6/ LIMITER L'IMPACT D'UN SYSTÈME AGENTIQUE

- **Considérer comme non fiables les entrées et sorties du système agentique.** Un message ou une mémoire peut détourner l'agent de sa mission. Filtrer les entrées et les sorties, définir leurs conservation et prévoir des mécanismes de suppression, oubli et mise en quarantaine.

- **Limitier le nombre d'appels successifs** entre agents et alerter sur les fréquences anormales.
- **Maintenir les secrets hors du contexte du modèle,** des mémoires et des journaux. Déléguer leur gestion à un coffre fort ou un proxy, et accorder à l'agent des jetons d'accès à durée de validité courte, limités aux droits strictement nécessaires.
- **Prévoir un arrêt d'urgence** avec révocation des accès, déclenchable par un mécanisme de supervision en cas de traitements anormaux. Prédéfinir un mode dégradé, sans affaiblir les contrôles ni élargir les droits.
- **Tester les mécanismes de restauration** pour retrouver un état maîtrisé. Vérifier qu'après une restauration, une action déjà exécutée avec un effet externe **n'est pas rejouée.**

7/ DÉMONTRER ET MAINTENIR LA MAÎTRISE DU SYSTÈME

- **Maintenir l'inventaire du système agentique** pour tous les composants logiciels, prompts système et dépendances. Versionner les composants développés en interne. Vérifier régulièrement à partir de cet inventaire la provenance, l'intégrité et les vulnérabilités connues pour tout composant.
- **Tester le système avant déploiement, puis régulièrement** pour éprouver sa sécurité et détecter les dérives du modèle. S'assurer de l'existence d'un mécanisme de retour en arrière.
- **Superviser en production** les actions de l'agent en les comparant à la mission documentée et aux outils autorisés afin de détecter les changements d'objectif et les enchaînements d'actions inhabituels (UEBA – *User and Entity Behavior Analytics*).
- **Maintenir les journaux du système agentique hors de portée des modifications de l'agent** afin de garantir leur intégrité et d'être capable de reconstituer, en cas d'incident, la séquence complète agentique : *demande → autorisation → action → résultat.*
- **Tenir à jour le référentiel des responsables** des politiques, des alertes et des exceptions du système agentique. Mettre à jour ce référentiel après tout changement ou incident, et *a minima* annuellement selon sa criticité.