Agentic AI : Guide Pratique de Gouvernance pour CISOs

La gouvernance de l'IA agentique est le nouveau défi opérationnel des CISOs en 2026 : les agents IA autonomes déployés dans les entreprises ne se gèrent pas avec les frameworks de gouvernance classiques. Ils nécessitent un nouveau modèle basé sur l'inventaire des agents, la cartographie des permissions, la surveillance comportementale et une chaîne de responsabilité claire. Ce guide pratique donne aux CISOs les outils concrets pour structurer leur programme de gouvernance agentique — du premier inventaire au programme mature en 90 jours.

L'**agentic AI gouvernance pour les CISOs** est devenue une priorité urgente en 2026. Les agents IA autonomes prolifèrent dans les entreprises — certains déployés par l'IT, beaucoup plus déployés en dehors de tout contrôle. Un agent LangChain intégré à Microsoft 365, un assistant CrewAI connecté au CRM, un bot AutoGPT avec accès aux emails internes : chacun représente un acteur autonome qui prend des décisions et exécute des actions au nom de l'entreprise. Le NIST AI Risk Management Framework (AI RMF) et l'EU AI Act posent des exigences croissantes sur la supervision de ces systèmes. Selon Gartner, d'ici fin 2026, plus de 80% des grandes entreprises auront des agents IA en production sans cadre de gouvernance formel. Ce guide pratique donne les étapes concrètes pour que les CISOs construisent ce cadre avant que les incidents ne les y forcent.

À RETENIR

À retenir

L'inventaire des agents est la première priorité : vous ne pouvez pas gouverner ce que vous ne savez pas. Cartographier tous les agents déployés — IT et Shadow — avec leurs outils et permissions est le point de départ obligatoire.

Les permissions doivent être granulaires par agent : chaque agent reçoit uniquement les droits nécessaires à sa mission — lecture seule sur les données non critiques, validation humaine obligatoire pour les actions à fort impact.

Le NIST AI RMF fournit le cadre structurel : ses quatre fonctions (Govern, Map, Measure, Manage) s'appliquent directement aux déploiements d'agents IA et fournissent un vocabulaire commun pour les équipes sécurité et IA.

L'EU AI Act impose des obligations documentées : les agents à fort impact (RH, crédit, infrastructure critique) sont classés haut risque et nécessitent évaluation des risques, traçabilité et supervision humaine documentées.

Un programme de 90 jours est réalisable : inventaire et classification (J1-30), politiques et contrôles techniques (J31-60), monitoring et incident response (J61-90) — un CISO peut déployer un programme minimal viable en trois mois.

Pourquoi les frameworks de gouvernance classiques ne suffisent pas ?

Les frameworks de gouvernance IT et cybersécurité existants — ISO 27001, SOC 2, NIST CSF — ont été conçus pour des systèmes déterministes. Vous définissez une politique, vous la traduisez en contrôles techniques, vous vérifiez la conformité. Ce modèle suppose que les systèmes ont un comportement prévisible et que les risques peuvent être cartographiés à l'avance.

Un **agent IA autonome** brise ce modèle. Son comportement n'est pas entièrement prévisible — il dépend du contexte de chaque session, des données qu'il traite, des décisions qu'il prend en temps réel. Il peut être détourné par des inputs malveillants. Il peut prendre des décisions inattendues face à des situations non prévues. Les contrôles classiques — règles de firewall, politiques IAM, règles DLP basées sur des patterns — ne capturent pas ces risques.

La gouvernance agentique nécessite une couche supplémentaire : des politiques comportementales (que peut faire l'agent, dans quelles conditions, avec quelles limites), des mécanismes d'observabilité sémantique (pas seulement "l'agent a appelé cette API" mais "l'agent a pris cette décision pour cette raison"), et une chaîne de responsabilité claire (qui est responsable si un agent prend une décision qui cause un préjudice). Pour comprendre les risques techniques sous-jacents, consultez notre analyse de la menace agentic AI pour l'entreprise en 2026.

Inventaire des agents IA : la première priorité du CISO

Vous ne pouvez pas gouverner ce que vous ne savez pas. L'inventaire des agents IA déployés dans l'organisation est le point de départ non-négociable — et dans la plupart des entreprises, le résultat de cet inventaire est une surprise. Outre les déploiements IT officiels, les équipes métier ont déployé leurs propres agents : un script Python LangChain développé par un data scientist, un agent Copilot Studio créé par un power user, un bot n8n avec des intégrations API. C'est le **Shadow AI 2.0** — bien plus dangereux que le Shadow IT classique parce que ces agents agissent.

L'inventaire doit couvrir au minimum : l'identité de l'agent (nom, version, framework), ses propriétaires et les équipes qu'il sert, la liste exhaustive de ses outils et des données auxquelles il accède, les utilisateurs ou systèmes qui peuvent l'invoquer, et le contexte d'utilisation (fréquence, volumes, criticité des données traitées). Une classification par niveau de risque doit suivre — agents à faible impact vs agents avec accès à des données critiques ou capacité d'action à fort impact.

> Dans une mission d'audit pour un groupe financier (2025), l'inventaire initial déclaré par l'IT listait 4 agents IA en production. L'inventaire réel, après analyse des logs de trafic API

> et interviews des équipes métier, en a révélé 23 — dont 7 avec accès à des systèmes de données clients sensibles et aucune documentation de sécurité. Trois d'entre eux n'avaient aucune rotation de credentials configurée.
>
> *— Retour d'audit, groupe financier, Q4 2025*

Architecture de permissions minimales pour les agents IA

Le principe du moindre privilège — donner à chaque acteur exactement les permissions nécessaires à sa mission, ni plus — s'applique aux agents IA avec une acuité particulière. Un agent qui a accès à plus d'outils ou de données que nécessaire amplifie le rayon d'impact d'une compromission. Et contrairement à un utilisateur humain, un agent peut utiliser l'intégralité de ses permissions de façon automatique, sans friction.

L'architecture de permissions doit être granulaire à plusieurs niveaux. Au niveau des **outils** : chaque agent déclare explicitement les outils dont il a besoin, et n'a accès qu'à ceux-là. Un agent de synthèse documentaire n'a pas besoin d'un outil d'envoi d'emails. Au niveau des **données** : définir quelles catégories de données l'agent peut lire et écrire. Au niveau des **actions** : distinguer les actions en lecture (faible risque) des actions en écriture (risque moyen) et des actions à fort impact comme l'envoi d'emails externes ou les modifications de configurations système (requérant une validation humaine).

L'OWASP LLM Top 10 2025 (LLM06 — Excessive Agency) recommande spécifiquement de minimiser les permissions des agents et d'éviter les configurations qui permettent à l'agent d'outrepasser les contrôles via ses outils. Cette recommandation est le fondement de toute architecture de gouvernance agentique sérieuse.

Politiques d'autorisation et principe du moindre privilège agentique

Au-delà des permissions techniques, la gouvernance agentique nécessite des politiques claires sur ce que les agents sont autorisés à faire — et ce qu'ils ne le sont pas — dans différents contextes. Ces politiques doivent être documentées, communiquées aux équipes qui déploient des agents, et techniquement enforçables.

Un exemple de politique minimale à mettre en place : *Tout agent ayant accès à des données clients ne peut pas envoyer d'emails vers des destinataires externes non présents dans une liste approuvée. Tout agent pouvant déclencher des actions financières doit requérir une approbation humaine pour tout montant supérieur à 500€. Tout agent a un budget d'appels d'outils par session (ex: max 50 appels) au-delà duquel la session est suspendue et un humain alerté.*

Ces politiques se traduisent en contrôles techniques dans le code de l'agent (guardrails configurés dans le framework) et en monitoring (détection des violations de politique). Elles doivent être revues régulièrement — les agents évoluent, les contextes changent, de nouveaux vecteurs d'attaque émergent. La gouvernance agentique n'est pas un projet ponctuel mais un processus continu, à l'image de la gouvernance IA en entreprise plus large.

Monitoring et détection des comportements d'agents anormaux

Le monitoring des agents IA est fondamentalement différent du monitoring applicatif classique. Les métriques techniques — latence, taux d'erreur, volumes d'appels — ne suffisent pas. Ce qui compte, c'est la **sémantique des actions** : l'agent fait-il ce pour quoi il a été conçu ? Ses décisions sont-elles cohérentes avec son objectif ? Ses actions correspondent-elles aux requêtes qui les ont initiées ?

Un programme de monitoring efficace pour les agents IA doit couvrir plusieurs dimensions. L'**observabilité des traces** : enregistrer chaque étape de raisonnement de l'agent, chaque appel d'outil avec ses paramètres, chaque décision prise. Le **profiling comportemental** : établir un baseline du comportement normal de l'agent (outils utilisés, fréquences, types de données accédées) et alerter sur les déviations. La **détection d'anomalies** : identifier les patterns suspects

comme un pic d'accès à des données sensibles, des envois vers des destinations inhabituelles, ou un nombre d'appels d'outils anormalement élevé.

LangSmith (LangChain) : observabilité native des traces pour les agents LangChain

OpenTelemetry + Semantic Conventions pour LLM : standard émergent pour l'instrumentation des agents

Microsoft Azure AI Studio : monitoring intégré pour les déploiements Copilot/Semantic Kernel

Weights & Biases : suivi des expériences et du comportement des agents ML

Arize AI / WhyLabs : observabilité ML avec drift detection applicables aux agents

Incident response spécifique aux agents IA autonomes

Les plans d'incident response existants ne couvrent pas les scénarios spécifiques aux agents IA. Il faut les compléter avec des procédures adaptées à trois types d'incidents : la compromission par prompt injection (l'agent a été détourné de sa mission), le comportement anormal non-malveillant (l'agent a pris des décisions inattendues sans y être poussé), et l'exploitation de permissions excessives (un utilisateur ou un processus légitime a utilisé l'agent pour des actions non prévues).

Pour chaque type d'incident, la procédure doit définir : comment suspendre l'agent immédiatement (kill switch), comment préserver les logs de la session compromettante pour l'investigation, comment évaluer l'étendue des actions malveillantes réalisées, comment notifier les parties affectées, et comment corriger la cause racine avant de redémarrer l'agent. La rapidité de containment est critique — un agent compromis peut exécuter des dizaines d'actions malveillantes en quelques minutes si on ne l'arrête pas.

L'IR doit également couvrir la communication : qui est notifié en interne (RSSI, DPO, direction métier), quand et comment notifier les autorités (CNIL si des données personnelles sont impliquées, ANSSI pour les OIV/OSE), et comment communiquer avec les utilisateurs affectés. La gouvernance cybersécurité entre le RSSI et le COMEX doit intégrer ce nouveau vecteur de risque dans les procédures existantes.

Le NIST AI RMF appliqué aux agents autonomes

Le NIST AI Risk Management Framework (AI RMF) publié en 2023 et mis à jour en 2024 fournit le cadre structurel le plus complet pour gérer les risques des systèmes IA, y compris les agents autonomes. Ses quatre fonctions — Govern, Map, Measure, Manage — s'appliquent directement aux déploiements agentiques.

Govern : établir la structure de gouvernance, les rôles et responsabilités, les politiques. Pour les agents IA, cela signifie désigner des "agent owners" responsables de chaque déploiement, créer un comité de validation pour les nouvelles demandes de déploiement, et documenter les politiques d'usage acceptables. **Map** : identifier et contextualiser les risques. Pour les agents, cartographier les vecteurs d'attaque (prompt injection, tool abuse), les données en jeu, et les scénarios d'impact. **Measure** : évaluer les risques avec des métriques concrètes — nombre d'agents sans supervision humaine pour les actions critiques, couverture du logging, résultats des tests de robustesse. **Manage** : mettre en place et maintenir les contrôles, gérer les incidents, améliorer en continu.

Ce que l'EU AI Act impose aux entreprises utilisant des agents IA

L'**EU AI Act** (Règlement 2024/1689) est entré en vigueur en août 2024. Pour les agents IA autonomes, les obligations dépendent de leur classification de risque. Les agents "haut risque" — définis par leur domaine d'application (emploi, crédit, infrastructure critique, contrôle d'accès, etc.) — sont soumis aux exigences les plus strictes : évaluation de conformité documentée, système de management des risques, traçabilité des décisions, supervision humaine effective, et enregistrement auprès des autorités compétentes.

Pour les CISOs, les implications pratiques sont importantes. Tout agent IA utilisé dans des décisions affectant des droits ou des accès (attribution de missions, scoring de fournisseurs, contrôle d'accès à des ressources) est potentiellement classé haut risque. La documentation doit inclure l'architecture technique de l'agent, les données d'entraînement et de configuration utilisées, les mesures de supervision humaine en place, et les procédures de test et de validation.

Des amendes pouvant atteindre 30 millions d'euros ou 6% du chiffre d'affaires mondial sont prévues pour les manquements aux obligations des systèmes haut risque.

Tableau : maturité de gouvernance agentic AI par niveau

Niveau	Caractéristiques	Risque	Actions prioritaires
Niveau 0 — Inexistant	Pas d'inventaire, pas de politiques, agents déployés ad hoc	Critique	Inventaire d'urgence, kill switch sur agents non documentés
Niveau 1 — Initial	Inventaire partiel, politiques informelles, logging minimal	Élevé	Formaliser l'inventaire, créer les politiques de base, activer le logging
Niveau 2 — Géré	Inventaire complet, politiques documentées, permissions granulaires	Moyen	Déployer l'observabilité sémantique, former les équipes
Niveau 3 — Défini	Monitoring comportemental, IR spécifique agents, red teaming régulier	Contrôlé	Automatiser la détection d'anomalies, intégrer NIST AI RMF
Niveau 4 — Optimisé	Gouvernance en temps réel, conformité EU AI Act documentée, amélioration continue	Résiduel	Partage de threat intel, contribution aux standards

Comment construire un programme de gouvernance agentic AI en 90 jours ?

Un programme de gouvernance agentique minimal viable peut être déployé en 90 jours. C'est suffisant pour passer du niveau 0 au niveau 2, et réduire significativement l'exposition au risque pendant que les équipes montent en compétence.

Phase 1 — J1 à J30 : Inventaire et classification. Cartographier tous les agents IA déployés ou en cours de déploiement. Pour chaque agent : framework, outils, données accessibles, propriétaire, criticité. Classer les agents par niveau de risque. Identifier les "quick wins" : agents sans logging, sans rotation de credentials, avec des permissions excessives évidentes. Activer en urgence un minimum de contrôles sur les agents haut risque identifiés.

Phase 2 — J31 à J60 : Politiques et contrôles techniques. Formaliser les politiques d'usage des agents (ce qui est autorisé, ce qui ne l'est pas). Configurer les permissions granulaires pour chaque agent. Activer le logging des actions d'outils. Mettre en place les human-in-the-loop pour les actions critiques. Communiquer les politiques aux équipes qui déploient des agents.

Phase 3 — J61 à J90 : Monitoring et incident response. Déployer un outil d'observabilité sémantique pour les agents prioritaires. Créer les procédures d'incident response spécifiques aux agents. Former les équipes SOC aux scenarios d'incidents agentiques. Réaliser un premier exercice de red teaming sur les agents en production les plus critiques. Pour les agents haut risque EU AI Act, initier la documentation de conformité.

Gouvernance agentique et Zero Trust : une alliance indispensable

Le modèle **Zero Trust** — "ne jamais faire confiance, toujours vérifier" — s'applique naturellement aux agents IA autonomes. Contrairement aux utilisateurs humains, les agents IA n'ont pas d'identité sociale vérifiable, pas d'historique comportemental établi, et peuvent être compromis ou détournés sans signe extérieur évident. Le Zero Trust appliqué aux agents impose trois principes : identité forte pour chaque agent (certificats courts, tokens rotatifs), autorisation par contexte (chaque action réévaluée dynamiquement), et surveillance continue de toutes les interactions.

Concrètement, cela signifie que l'accès d'un agent à un système de production ne doit pas être accordé de manière permanente mais de manière contextuelle : l'agent obtient un accès temporaire pour une tâche définie, cet accès expire automatiquement, et toute action hors du périmètre autorisé déclenche une alerte. Ce modèle est fondamentalement différent de la gestion des droits utilisateurs traditionnelle, qui repose sur des groupes statiques et des rôles permanents. Selon l'NIST AI RMF Playbook, l'application des principes Zero Trust aux systèmes IA autonomes est désormais classée comme contrôle de gouvernance de niveau 2.

Just-In-Time (JIT) access pour agents : l'agent reçoit les droits nécessaires uniquement pour la durée de la tâche — automatiquement révoqués après exécution

Micro-segmentation des workloads IA : isoler les agents dans des namespaces réseau dédiés, avec règles egress strictes pour empêcher les communications non autorisées

Attestation d'identité : chaque agent s'authentifie via un mécanisme cryptographique fort (SPIFFE/SPIRE, JWT avec rotation courte) plutôt que par des clés API statiques

L'intégration du Zero Trust dans la gouvernance agentique est détaillée dans notre guide sur l'IA et Zero Trust avec micro-segmentation dynamique. Pour les organisations souhaitant aller plus loin dans la gouvernance des outils IA internes, notre analyse des vulnérabilités des copilotes IA d'entreprise est un complément direct à ce guide.

> Dans un projet de gouvernance agentique pour un groupe bancaire français (T1 2026), nous avons implémenté un modèle Zero Trust complet pour 14 agents IA de production. Le changement le plus impactant : passer de clés API statiques à des tokens SPIFFE avec durée de vie de 15 minutes. Résultat : deux tentatives d'exploitation détectées et bloquées en 6 semaines — elles n'auraient pas été visibles avec les anciennes configurations d'accès permanents.
>
> *— Retour de mission, gouvernance agentique, T1 2026*

Questions fréquentes des CISOs sur la gouvernance agentic AI

Qui est responsable de la sécurité d'un agent IA déployé par une équipe métier ?

La responsabilité doit être partagée mais clairement définie. L'équipe métier qui déploie l'agent est "agent owner" — responsable de son usage conforme aux politiques. L'IT/DSI est responsable des contrôles techniques (infrastructure, logging, permissions). La SSI/RSSI définit les politiques et standards de sécurité et audite la conformité. En cas d'incident, la chaîne de responsabilité doit être documentée à l'avance — une découverte au moment de la crise est toujours une mauvaise idée.

Comment évaluer si un agent IA existant est conforme à l'EU AI Act ?

La première étape est la classification : le domaine d'application de l'agent le fait-il entrer dans la liste des systèmes haut risque de l'Annexe III de l'EU AI Act ? Si oui, vérifier si une évaluation de conformité a été réalisée, si la supervision humaine est documentée, si les logs de traçabilité existent. Un audit de conformité EU AI Act pour les agents haut risque est recommandé — notre service RSSI externalisé accompagne cette démarche.

Les agents IA doivent-ils être déclarés à la CNIL ?

Si l'agent traite des données personnelles (accès à des emails, des dossiers RH, des données clients), il entre dans le périmètre du RGPD et doit figurer dans le registre des activités de traitement. Si l'agent prend des décisions automatisées affectant des personnes (art. 22 RGPD), des obligations supplémentaires s'appliquent — information des personnes concernées, droit d'opposition, garanties spécifiques. L'ANSSI et la CNIL ont publié des recommandations conjointes sur les systèmes IA et le RGPD à consulter.

Quelle est la différence entre Shadow AI et Shadow Agents ?

Le Shadow AI désigne l'usage non autorisé d'outils IA (ChatGPT personnel, Copilot non géré) par des employés. Les **Shadow Agents** vont beaucoup plus loin : ce sont des agents IA autonomes déployés sans validation de la SSI, qui agissent — accèdent à des systèmes, exécutent des tâches, prennent des décisions. Le Shadow AI consulte des informations ; les Shadow Agents agissent sur des systèmes. Le risque est d'un ordre de magnitude supérieur. Notre analyse détaillée des Shadow Agents IA et leur gouvernance couvre les méthodes d'identification et de remédiation.

Comment impliquer le COMEX dans la gouvernance des agents IA ?

Le COMEX doit comprendre que les agents IA sont des acteurs autonomes qui agissent au nom de l'entreprise — avec les mêmes conséquences légales et réputationnelles qu'un employé qui ferait des actions non autorisées. Traduire le risque agentique en langage business : scénarios d'impact concrets (exfiltration de données client, actions financières non autorisées), implications

réglementaires EU AI Act, coût d'un incident. Présenter le programme de gouvernance comme un investissement de protection, pas un frein à l'innovation.

Vous êtes CISO et souhaitez structurer votre gouvernance des agents IA ? Notre équipe accompagne la mise en place de programmes de gouvernance agentique — inventaire, politiques, monitoring, conformité EU AI Act. Contactez-nous pour un premier cadrage.