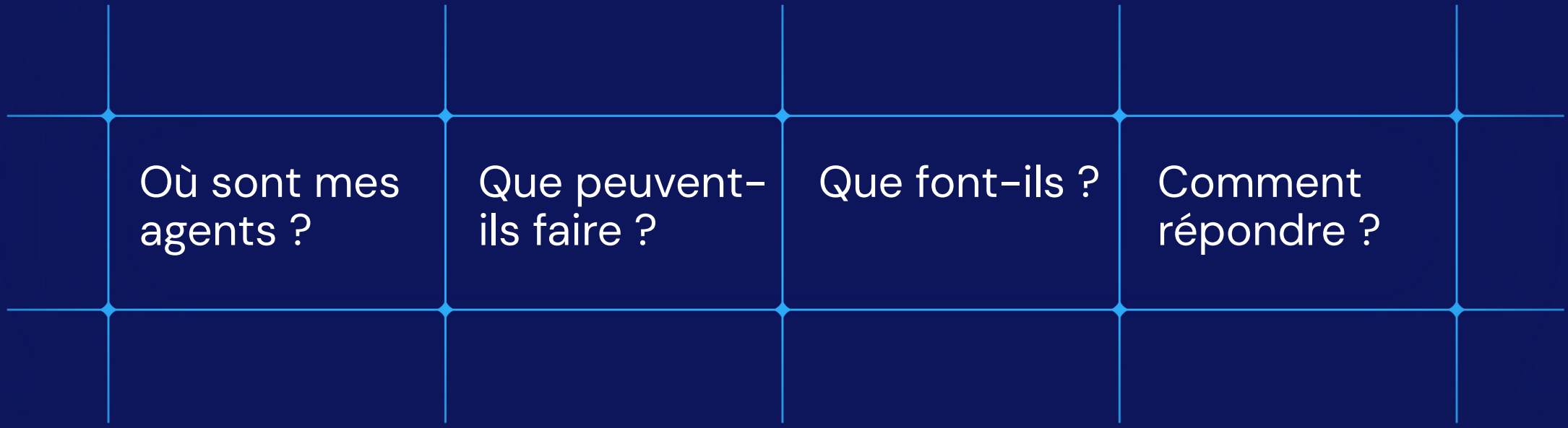
Livre blanc

Gouvernance de l'exécution agentique

Modèle d'architecture pour l'entreprise agentique sécurisée



Rédigé par les membres fondateurs de la Blueprint Alliance
Architecture de référence intersectorielle pour les RSSI, les DSI,
les directeurs techniques et les architectes en sécurité d'entreprise

Blueprint Alliance

Sommaire

Résumé

Le périmètre de sécurité traditionnel des entreprises est remis en question avec l'adoption des agents d'IA. En deux ans à peine, les chatbots LLM expérimentaux ont laissé place à des workflows agentiques autonomes en plusieurs étapes. Par le passé, un logiciel se contentait d'exécuter les commandes reçues ; à présent, il décide, agit et se connecte de sa propre initiative. Si cette autonomie apporte une véritable valeur ajoutée à l'entreprise, elle pose par contre un défi de gouvernance que la plupart des équipes sécurité ne sont pas encore en mesure de relever.

Selon les prévisions de Gartner, d'ici 2028, une entreprise du Fortune 500 comptera en moyenne plus de 150 000 agents actifs, pour moins de 15 en 2025. Cette évolution entraînera une prolifération agentique importante, une complexité accrue de l'IT et de nouveaux défis de gestion. Ces agents ne se contentent pas d'afficher des données. Ils raisonnent, orientent, appellent des API en aval, créent des sous-agents et exécutent des opérations en plusieurs étapes sur de multiples systèmes d'entreprise.

En l'absence d'un standard architectural unifié, les entreprises sont confrontées à un **déficit croissant de gouvernance, de sécurité et d'identité** : les agents non approuvés se multiplient, les identifiants sont transmis au-delà des zones de confiance et l'exécution échappe à tout contrôle. L'architecture de référence Blueprint Alliance est un framework multifournisseur ouvert, publié à l'intention de la communauté dans son ensemble, qui déplace la sécurité du périmètre réseau vers un point de contrôle agentique Zero Trust unifié et offre ainsi aux responsables technologiques un moyen concret d'intensifier les déploiements agentiques sans jamais perdre le contrôle.

Introduction

Pendant des dizaines d'années, la sécurité des entreprises s'est appuyée sur une séparation claire des identités humaines (gérées via l'IAM, le MFA et le SSO) et des identités machines (gérées via des comptes de service statiques, des clés API ou des services cloud de calcul). Les agents d'IA autonomes abolissent cette dichotomie. Lorsqu'un agent d'IA agit pour le compte d'un collaborateur, hérite de privilèges d'accès, interroge des bases de données vectorielles et déclenche des transactions financières, il devient un acteur opérationnel et un principal de sécurité actif qui exige des contrôles explicites d'identité, d'accès et de gouvernance.

La sécurité ne peut pas être ajoutée a posteriori uniquement au niveau du modèle ou de la couche réseau. Elle doit reposer sur un point de contrôle agentique qui assure une gouvernance continue de la visibilité, de la connectivité et de l'exécution.

Logiciel traditionnel (déterministe)	**IA agentique (non déterministe)**
Exécute des chemins de code explicites, étape par étape.	Reçoit une intention humaine de haut niveau et formule des plans d'exécution en plusieurs étapes.
S'appuie sur des autorisations et des comptes de service statiques et prédéfinis.	Sélectionne les API et les applications de façon dynamique, tirant parti des coffres-forts de tokens, des courtiers d'identifiants et des sessions déléguées.
Strictement limité aux chemins d'intégration codés de manière irréversible.	Agit de manière autonome, utilise des outils via le protocole MCP (Model Context Protocol) et crée des sous-agents.
Fonctionne exclusivement au sein de sessions utilisateurs authentifiées.	Fonctionne en mode asynchrone, souvent bien longtemps après la déconnexion de l'utilisateur délégant.

Le déficit de gouvernance :

cinq questions que chaque conseil d'administration se pose

Les agents d'IA se multiplient à un rythme trop rapide pour la plupart des équipes sécurité. Selon les prévisions de Gartner, d'ici 2028, une entreprise du Fortune 500 comptera en moyenne plus de 150 000 agents actifs, pour moins de 15 en 2025. À la clé, une prolifération agentique importante, une complexité accrue de l'IT et de nouveaux défis de gestion. Seuls 13 % des entreprises estiment disposer d'une gouvernance adéquate des agents d'IA. Cet écart explique pourquoi les comités de cybersécurité et de conformité reviennent sans cesse aux cinq mêmes questions.

1. Où se trouvent nos agents d'IA, et qui en est responsable ?
2. Comment limiter les opérations que les agents peuvent exécuter et veiller à ce qu'ils respectent leur mandat ?
3. Pouvons-nous auditer et expliquer la prise de décisions autonome a posteriori ?
4. Quel est notre niveau d'exposition des données et le risque lié aux tiers dans l'ensemble de la chaîne logistique des agents ?
5. Disposons-nous d'un arrêt d'urgence vérifié et d'un plan de restauration si un agent prend des initiatives non approuvées ?

Ces cinq questions correspondent aux quatre piliers de l'architecture de référence Blueprint Alliance présentée à la section 3. La découverte et l'enregistrement des identités répondent à la première question ; la gestion des accès à la deuxième ; la surveillance et l'autorisation à l'exécution aux troisième et quatrième questions ; et le confinement actif à la cinquième. Répondre à ces cinq questions avec des éléments tangibles, et non de simples promesses, permet à un RSSI d'accélérer en toute sécurité l'adoption des agents tout en alignant la profondeur des contrôles à l'exposition réglementaire et opérationnelle. Sans cela, l'entreprise s'expose à des incidents de sécurité importants et de plus en plus fréquents.

Plus qu'une checklist systématique, ces questions servent surtout d'outil de diagnostic pour évaluer les risques. Assez logiquement, les exigences de sécurité varient selon le périmètre opérationnel, car un assistant de productivité en lecture seule n'a pas besoin des mêmes contrôles de confinement qu'un agent chargé d'exécuter des transactions financières. Identifier les risques réellement applicables permet aux équipes sécurité d'adapter les contrôles à l'exposition réelle. Cela évite aux entreprises de complexifier les déploiements à faible risque et de s'exposer à des retards d'approbation de plusieurs mois, tout en veillant à ce que les agents à haut risque soient associés à des limites opérationnelles appropriées.

Principes du framework Blueprint Alliance

Le framework Blueprint Alliance étend le Zero Trust aux agents : ils doivent être considérés comme des identités à part entière, et non comme du code, des identifiants ou des éléments accessoires ajoutés à la pile d'identités humaines. Ces six principes sous-tendent tout ce qui suit.

Chaque agent hébergé constitue une identité de sécurité à part entière, et non un élément accessoire. Les agents sont provisionnés, authentifiés et déprovisionnés avec la même rigueur que les collaborateurs et les comptes de service. Tous les agents hébergés par l'entreprise, y compris les pilotes et les outils internes, doivent fonctionner dans un framework de gouvernance approprié, avec une profondeur des contrôles adaptée à leur périmètre opérationnel et à leur périmètre de risque. Les environnements d'exécution locaux sont régis par un processus de confinement spécialisé des endpoints, plutôt que par une inscription formelle dans un annuaire d'identités.

L'accès est limité et spécifique à chaque tâche, jamais permanent. Si l'octroi d'autorisations dynamiques par tâche demeure l'objectif à atteindre, dans la pratique, les implémentations associent des tokens d'accès à courte durée de vie pour les tâches à des autorisations de base résilientes afin d'équilibrer la charge opérationnelle.

La délégation est traçable de bout en bout. Lorsqu'une personne délègue une tâche à un agent et que cet agent crée un sous-agent, la chaîne de responsabilité doit être préservée à chaque étape. Chaque action doit pouvoir être rattachée à la personne ou au processus qui l'a autorisée. Bien que la délégation en cascade de bout en bout reste un objectif ambitieux avec les outils actuels, ce modèle d'architecture offre une trajectoire de référence capable d'évoluer parallèlement aux standards.

L'environnement d'exécution de l'agent est isolé et surveillé, pas simplement provisionné. Les décisions d'accès prises lors de la configuration sont nécessaires, mais insuffisantes. Le modèle exige une observation continue de différents workflows d'exécution afin d'établir des bases de référence comportementales, au-delà des autorisations accordées.

Le confinement est instantané et réversible. Chaque agent, quel que soit le fournisseur ou la plateforme, doit être doté d'un arrêt d'urgence ciblé capable de le suspendre, de le mettre en quarantaine ou de mettre fin à son exécution en temps réel, tout en offrant le moyen de revenir en arrière. Associer une surveillance continue du comportement à un périmètre d'exécution structurel permet de réduire l'écart entre ce qui est « autorisé » et « sûr » en limitant immédiatement l'impact dès qu'un agent s'écarte de son comportement de référence.

La gouvernance s'adapte à la vitesse de l'IA. Les politiques de sécurité statiques ne peuvent pas suivre le rythme de l'évolution des capacités agentiques, et la gouvernance doit compenser ce que les politiques seules ne peuvent pas anticiper. Ce modèle fonctionne comme une architecture adaptable qui évolue avec l'apparition de nouveaux comportements d'agents, en associant des plafonds d'autorisation fixes à une détermination dynamique des droits en temps réel. Les entreprises peuvent ainsi maintenir des garde-fous stricts tout en adaptant l'accès au contexte.

Ensemble, ces six principes étendent le Zero Trust à une nouvelle catégorie d'identités : des agents qui opèrent de manière continue, persistante et autonome dans l'entreprise. Une bonne maîtrise de ces principes permet de combler le déficit de gouvernance, de sécurité et d'identité.

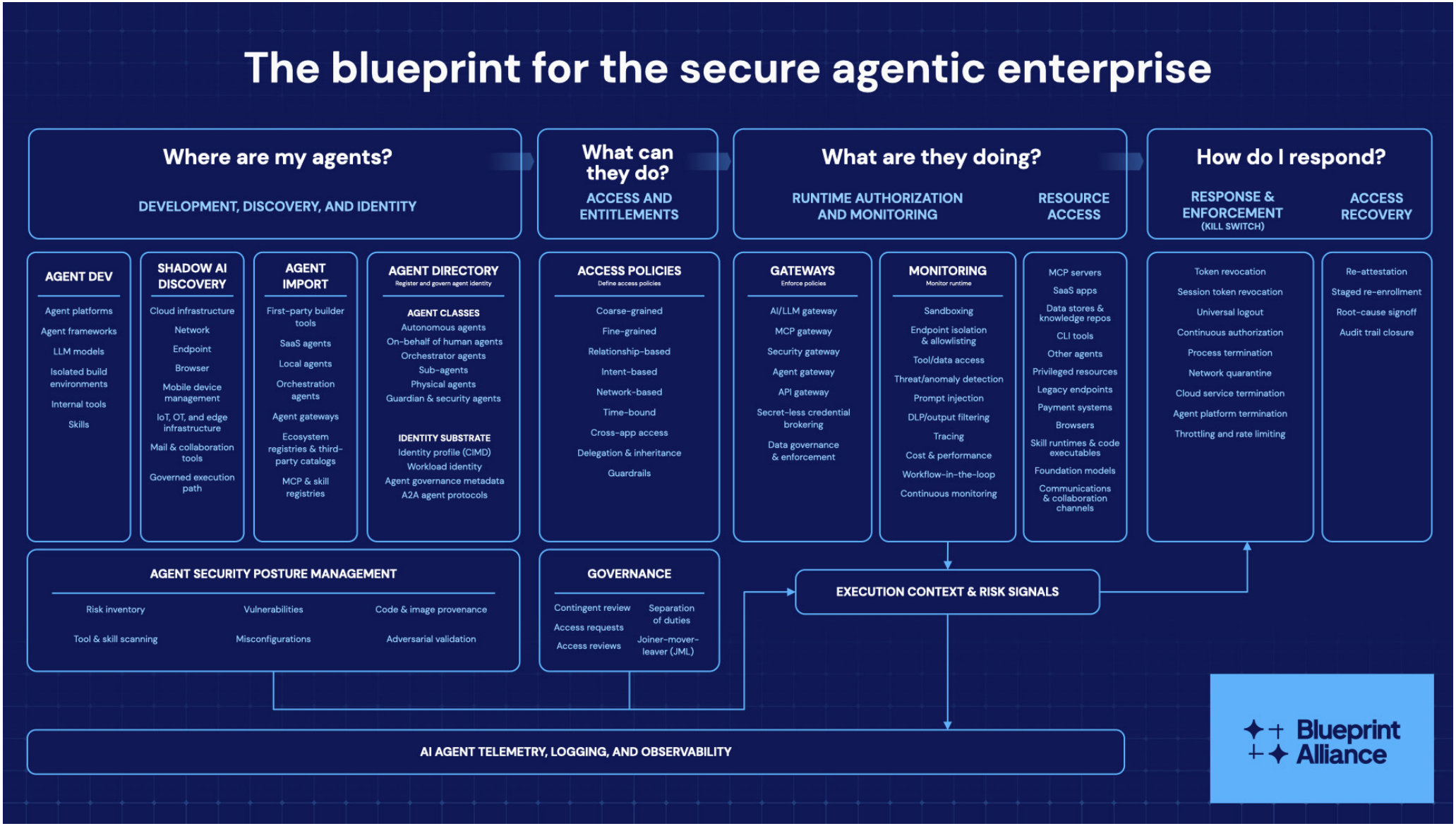
Architecture de référence Blueprint Alliance

L'architecture de référence Blueprint Alliance fonde la sécurité agentique des entreprises sur quatre questions opérationnelles essentielles à la gouvernance de l'exécution autonome, chacune correspondant à un pilier de l'architecture de référence :

1. Où sont mes agents ?
2. Que peuvent-ils faire ?
3. Que font-ils ?
4. Comment répondre ?

Toutes ces questions sous-tendent le besoin d'un socle solide en matière de contexte d'exécution et de signaux de risque, alimenté en continu par la télémétrie, la journalisation et l'observabilité à l'exécution. Ensemble, ces éléments forment le tissu relationnel permettant aux systèmes de détecter les profils de risque et de coordonner la réponse.

Architecture de référence Blueprint Alliance



3.1 Pilier 1 :

Où sont mes agents ? (Développement, découverte et identité)

En matière de sécurisation de l'entreprise agentique, le principal défi est la visibilité. Les entreprises ne peuvent pas assurer la gouvernance ou la protection d'agents dont elles ignorent l'existence. Ce pilier implémente une fonctionnalité de surveillance complète dans l'ensemble des environnements de l'entreprise afin d'identifier, de recenser et d'évaluer chaque agent d'IA opérant au sein de l'infrastructure, qu'il soit approuvé ou non géré. Il se termine par l'enregistrement des agents validés en tant qu'identités gouvernées, chacune rattachée à un propriétaire désigné, afin de dresser l'inventaire de référence auquel le deuxième pilier associe les politiques d'accès.

3.1.1 Découverte des agents d'IA

La découverte des agents d'IA sert de mécanisme de détection continue pour établir l'empreinte des agents au sein de l'entreprise. Elle permet de cartographier l'ensemble du cycle de vie des agents, des pipelines internes d'ingénierie aux intégrations externes, de sorte qu'aucun modèle actif ni processus autonome ne s'exécute en dehors de tout contrôle.

3.1.1.1 Développement d'agents

Ce composant se concentre sur les environnements internes d'ingénierie où des agents propriétaires sont conçus, entraînés et optimisés. Sécuriser le pipeline de développement permet d'intégrer la sécurité dès la conception, avant qu'un agent n'interagisse avec des données en production.

- **Plateformes d'agents :** environnements d'orchestration gérés par l'entreprise, où les agents sont hébergés, entraînés et déployés de manière systématique.
- **Frameworks agentiques :** architectures et bibliothèques logicielles sous-jacentes (telles que LangChain, LlamaIndex ou AutoGen) utilisées par les développeurs pour concevoir la logique agentique.
- **Modèles LLM :** grands modèles de langage fondamentaux et optimisés servant de moteurs cognitifs aux agents.
- **Environnements de développement et d'exécution isolés :** limites CI/CD sécurisées avec isolation des ressources de calcul qui exécutent le code des agents lors des phases de test et d'évaluation, ce qui permet aux équipes sécurité d'identifier les exécutions de tests éphémères au lieu de s'appuyer uniquement sur les déclarations de registre statiques.
- **Outils internes supplémentaires :** référentiels de code personnalisés, assistants de code et registres de modèles internes intégrés au pipeline de développement.
- **Skills :** modules procéduraux définis, extraits de code ou outils développés pour étendre les capacités d'un agent.

3.1.1.2 Importation d'agents

Les entreprises modernes s'appuient rarement sur une seule source d'intelligence. Elles intègrent plutôt différentes architectures d'agents, qui varient par leur origine, leur framework et leur périmètre de calcul, afin de compenser l'évolution rapide des capacités agentiques. L'importation d'agents surveille et gouverne l'introduction de ces différentes classes d'agents dans l'écosystème de l'entreprise, en mettant surtout l'accent sur la façon dont les données circulent dans l'entreprise selon leur modèle de déploiement sous-jacent.

Le type d'agent, la catégorie et le canal sont trois questions distinctes que vous pouvez poser à propos de chaque agent que vous importez. Un même agent aura généralement une réponse pour chacune d'elles, par exemple un agent SaaS (développé par un fournisseur) utilisé pour des workflows RH internes (B2E) et intégré via une passerelle d'agents.

Type d'agent : qui a développé l'agent ?

- **Agents propriétaires :** capacités d'IA développées en interne, allant d'implémentations entièrement personnalisées conçues à partir de code brut (comme Python et LangChain) et exécutées dans des environnements de calcul généralistes (par exemple, des fonctions sans serveur ou Kubernetes) aux agents créés à l'aide de frameworks agentiques spécifiques à un fournisseur, qui gèrent de façon native l'environnement d'exécution, la mémoire et l'orchestration des workflows.
- **Agents SaaS :** produits d'IA tiers, conçus pour un usage spécifique, qui fonctionnent comme des entités entièrement indépendantes au sein de l'écosystème logiciel de l'entreprise.
- **Agent local :** architecture client installée directement sur l'appareil local du collaborateur, chaque étape du raisonnement étant envoyée à un fournisseur de LLM dans le cloud. Il agit comme une couche non déterministe entre l'intention humaine et l'action de l'outil, les données franchissant le périmètre de l'entreprise à chaque appel d'inférence.
- **Agents d'orchestration :** agents conçus sur des plateformes d'orchestration visuelle qui servent à relier différentes API d'entreprise entre elles.

Catégorie d'agent : à qui l'agent est-il destiné ?

- **Agents B2E (Business-to-Employee) :** agents d'IA internes, orientés collaborateurs et déployés pour automatiser les workflows de l'entreprise, les opérations IT ou RH et la productivité quotidienne des équipes, dans le respect des contrôles d'accès de l'entreprise basés sur l'identité et les rôles.
- **Agents B2B :** agents d'IA orientés entreprises, qui opèrent au-delà du périmètre de l'entreprise pour automatiser des workflows entre entreprises, des transactions de chaîne logistique et des intégrations d'API entre tenants.
- **Agents B2C :** agents d'IA orientés consommateurs, déployés pour interagir directement avec des utilisateurs externes dans le cadre de services, de transactions commerciales ou de demandes d'assistance, exigeant des mécanismes de protection robustes contre les injections de prompts publiques, l'exposition de la marque et les fuites de données personnelles des consommateurs.
- **Agents B2B2C :** agents d'IA en marque blanche ou fournis par un prestataire, déployés via une entreprise cliente pour servir les consommateurs de ce partenaire et soumis à une isolation multiniveau et à des limites de confiance transitive.
- **Agents d'assistance personnelle :** assistants d'IA grand public, pilotés par l'utilisateur. Ils agissent pour le compte d'une personne dans des contextes personnels et professionnels, et exigent des contrôles pour séparer l'accès aux données personnelles et l'accès aux données de l'entreprise.

Canal d'ingestion : comment l'agent est-il découvert ?

- **Passerelles d'agents :** points d'ingestion et proxys par lesquels ces types d'agents externes ou importés interagissent avec les API internes de l'entreprise.
- **Registres de l'écosystème et catalogues tiers :** pipelines de connecteurs automatisés et adaptateurs d'API qui surveillent en continu, extraient et associent les définitions d'agents issues des registres externes natifs des fournisseurs, afin que les agents soient automatiquement recensés, sans intervention humaine.
- **Registres MCP et de Skills :** points d'ingestion à travers lesquels les agents extraient des schémas d'outils, des Skills et des définitions de contexte à partir de registres MCP et de Skills publics ou hébergés par des fournisseurs. Les registres constituent eux-mêmes un vecteur d'injection : une entrée malveillante ou déguisée peut être récupérée et utilisée automatiquement, de sorte que chaque outil ou Skill provenant d'un registre est considéré comme une donnée non fiable, soumise à l'analyse de la chaîne d'approvisionnement décrite à la section 3.1.2 avant qu'un agent puisse l'appeler.

3.1.1.3 Découverte de la Shadow AI

Les collaborateurs se tournent spontanément vers des outils d'IA non approuvés pour optimiser leurs workflows, et introduisent ainsi des risques de conformité, d'exfiltration de données et de sécurité opérationnelle. La découverte de la Shadow AI analyse activement les périmètres d'entreprise afin de détecter le déploiement et l'utilisation non autorisés d'agents. Les fonctionnalités de découverte et de surveillance doivent être implémentées conformément aux exigences applicables en matière de confidentialité, d'emploi, du droit du travail, des comités d'entreprise et de protection des données, avec un niveau approprié de transparence, de proportionnalité, de minimisation et de conservation des données, et de contrôle des accès.

- **Infrastructure cloud :** instances de calcul cloud, environnements sans serveur et API endpoints non gérés, identifiés au moyen d'une inspection active du réseau ou des API.
- **Réseau :** analyse du trafic en temps réel et surveillance des sorties pour identifier les appels API non reconnus à des fournisseurs externes de LLM et à des endpoints d'IA.
- **Endpoint :** surveillance au niveau des appareils pour monitorer l'exécution non autorisée de binaires d'IA, de scripts ou de poids de modèles locaux.
- **Navigateur :** suivi des extensions et surveillance des sessions de navigateur pour détecter le moment où des utilisateurs se connectent à des systèmes de chat ou plateformes d'IA sur le Web non approuvés.
- **Gestion des terminaux mobiles (MDM) et gestion des endpoints :** audit continu au niveau des appareils pour détecter les exécutables d'IA locaux non approuvés, les poids de modèles locaux, les outils d'agent en ligne de commande et les extensions IDE, tout en appliquant une posture de sécurité de base et en bloquant les environnements d'exécution d'agent non gérés au niveau du système d'exploitation.
- **IoT, OT et infrastructure en périphérie :** surveillance de la sécurité industrielle et inspection des protocoles réseau pour détecter les petits modèles de langage (SLM) et les agents autonomes sur les équipements de production, les automates programmables industriels (PLC) et le matériel embarqué sur lesquels il n'est pas possible de déployer des logiciels traditionnels pour endpoints.
- **Flux de messagerie et de collaboration :** analyse du trafic de messages, de chats et de calendriers afin de détecter les agents non autorisés qui envoient des messages, répondent automatiquement ou gèrent des calendriers pour le compte d'un utilisateur avant que l'activité ne soit visible dans les journaux de sortie des endpoints ou du réseau.
- **Chemin d'exécution gouverné en tant que surface de détection :** périmètre architectural au sein duquel tous les agents autorisés s'exécutent au travers d'un point de contrôle centralisé ; tout processus agentique exécuté hors périmètre est automatiquement identifié comme un signal de Shadow AI, sans qu'il soit nécessaire d'implémenter des règles de détection spécifiques à chaque environnement.

3.1.2 Gestion de la posture de sécurité des agents d'IA

Une fois identifiés, les agents doivent faire l'objet d'une évaluation continue. La gestion de la posture de sécurité des agents d'IA fournit la couche de diagnostic qui évalue les profils de risque, la qualité des pratiques de configuration et les vulnérabilités logicielles connues, et ce, dans l'ensemble de l'inventaire des agents.

- **Inventaire des risques :** registre centralisé et dynamique destiné à surveiller la posture de sécurité, les vulnérabilités, les dépendances logicielles et les erreurs de configuration des agents découverts, indépendamment de l'étape de leur cycle de vie, en apportant une perspective axée sur les risques de sécurité pour compléter l'annuaire d'agents axé sur l'IAM.
- **Vulnérabilités :** gestion des vulnérabilités logicielles standard étendues aux dépendances agentiques et aux endpoints des modèles ; elle couvre les vulnérabilités CVE au sein de la base de code de l'agent, de sa chaîne logistique et des modèles fondamentaux, ces CVE étant associées à des frameworks standardisés pour garantir une évaluation cohérente des risques.
- **Provenance du code et des images :** vérification cryptographique de l'origine, de l'intégrité de la build et des signatures du code des agents, des images de conteneurs, des dépendances et des poids des modèles (par exemple, attestations du framework SLSA) avant l'exécution en production. L'utilisation d'images de base minimales, préalablement certifiées, sans paquets superflus, et récréées en fonction des données CVE actuelles, réduit la surface d'attaque de la chaîne logistique et simplifie la vérification de la provenance lors de la phase de compilation.
- **Analyse de la chaîne logistique des outils et des Skills :** inspection statique et dynamique automatisée des ensembles d'instructions des agents locaux, des capacités d'exécution et des schémas d'outils MCP (Model Context Protocol) distants, y compris les éléments extraits des registres MCP et de Skills, afin de détecter la présence de code malveillant incorporé, d'opérations masquées et de voies de sortie non autorisées avant l'enregistrement.
- **Erreurs de configuration :** analyse continue pour détecter des paramètres peu sûrs, des prompts système vulnérables, des identifiants à longue durée de vie codés de façon irréversible, des vecteurs d'usurpation d'identité et des dérives de configuration apparues après la validation initiale d'un agent.
- **Validation contradictoire :** test contradictoire continu et automatisé des agents et de leurs garde-fous par rapport aux voies connues d'injection de prompt, de débridage et d'exfiltration de données, au sein d'un périmètre d'exécution éphémère et isolé, afin d'évaluer en toute sécurité les possibilités d'exploitation sans exposer l'environnement de production à des risques.

3.1.3 Annuaire d'agents (Enregistrement et gouvernance des identités des agents)

Si l'importation d'agents (section 3.1.1.2) permet de capturer la provenance et les métadonnées lors de la découverte, l'annuaire d'agents est le référentiel d'identités faisant autorité dans lequel les identifiants cryptographiques, la propriété et les droits issus des politiques sont formellement enregistrés et conservés pour l'ensemble des agents validés, hébergés par l'entreprise. L'enregistrement représente la frontière entre un agent identifié et un agent gouverné, et il constitue la condition préalable à toute décision d'accès du deuxième pilier.

Chaque agent hébergé et chaque sous-agent disposant de droits distincts doivent être enregistrés ici avec un profil vérifié, une identité cryptographique, un propriétaire clairement identifié et des relations structurelles explicites avant de pouvoir demander l'accès aux systèmes de l'entreprise. Le processus d'enregistrement doit rester suffisamment simple pour éviter que les équipes ne cherchent à le contourner en utilisant des identifiants humains partagés. Les agents locaux et les sous-agents éphémères qui héritent du périmètre d'autorisation d'un agent parent sont gouvernés par les contrôles de confinement et de délégation décrits ci-dessous, et non par des entrées individuelles de l'annuaire. Une fois l'agent importé, l'annuaire répond à deux autres questions concernant son fonctionnement dans votre entreprise : quel rôle joue-t-il et comment prouve-t-il son identité ?

Catégories d'agent : quel est le rôle de l'agent ?

- **Agents autonomes :** agents entièrement indépendants qui sont exécutés de manière asynchrone afin d'atteindre des objectifs métier à long terme sans supervision humaine immédiate.
- **Agents OBO (On-Behalf-Of) :** agents exécutant des tâches directement liées à la session d'un utilisateur spécifique, agissant comme intermédiaires et limités par les autorisations associées à celui-ci.
- **Agents d'orchestration :** agents orchestrateurs de niveau parent qui décomposent un objectif en sous-tâches et créent, coordonnent et supervisent des sous-agents en aval ; l'annuaire doit surveiller les chaînes de délégation et les relations de création comme attributs de premier niveau.
- **Sous-agents :** agents enfants spécialisés, créés ou appelés de façon dynamique par un agent orchestrateur parent pour exécuter des sous-tâches spécifiques, gouvernés par des périmètres d'autorisation de session hérités plutôt que par un enregistrement individuel dans l'annuaire. Comme les limites de tâches définies au sein d'un LLM ne sont pas infaillibles, il est nécessaire d'encapsuler les sous-agents dans un périmètre d'exécution isolé qui impose structurellement le champ d'application des autorisations héritées, pour éviter que des erreurs de raisonnement n'entraînent une élévation des privilèges.

- **Agents physiques/incarnés :** agents pilotant des robots, des actionneurs ou d'autres systèmes physiques. Cette architecture contribue à une bonne gouvernance des identités et des accès, mais elle complète, plutôt qu'elle ne remplace, les contrôles existants en matière de sécurité des installations, sécurité OT et gestion des changements.
- **Agents de protection et de sécurité :** agents d'IA administratifs spécialisés opérant au sein du point de contrôle de sécurité (par exemple, agents de tri SOC, moteurs de politiques dynamiques, scanners ASPM continus). Comme ces agents disposent de privilèges élevés pour examiner la télémétrie et exécuter des actions de confinement, leur identité exige une attestation cryptographique obligatoire de la charge de travail, une stricte séparation des tâches et une validation continue par les propriétaires.

Substrat d'identité : comment prouver qu'il s'agit bien de cet agent ?

- **Profil de métadonnées d'identité de l'agent :** profils de métadonnées enrichis (CIMD) qui définissent l'intention, l'origine, la traçabilité et le contexte opérationnel de l'agent afin d'améliorer les décisions d'accès basées sur les risques.
- **Identités des charges de travail :** frameworks de production d'identités cryptographiques, indépendants des plateformes et fondés sur des standards ; ils sont utilisés pour délivrer des identifiants sécurisés et vérifiables aux agents dans les environnements cloud dynamiques.
- **Métadonnées de gouvernance des agents :** métadonnées structurées associées à l'enregistrement de l'agent dans l'annuaire, comprenant le propriétaire humain responsable, l'équipe opérationnelle désignée ou la file d'astreinte, l'étape du cycle de vie (par exemple, actif, suspendu, désactivé), le niveau de risque et le champ d'application opérationnel. Ces métadonnées sont utilisées en continu par les moteurs de politiques pour la création des contrôles d'accès et leur évaluation à l'exécution.
- **Protocoles d'agent A2A :** agents d'IA qui s'appuient sur des frameworks d'interopérabilité standardisés pour la découverte dynamique des fonctionnalités, la délégation des tâches et la collaboration entre pairs au-delà du périmètre de l'entreprise. Lorsqu'ils interagissent entre différentes entreprises sans annuaire partagé ni racine de confiance commune, les agents utilisent des mécanismes d'attestation cryptographique et des standards d'échange de tokens (par exemple, WIMSE/AIMS, OAuth Token Exchange, HTTP Message Signatures) afin de préserver les chaînes de délégation et de vérifier l'identité du principal à travers les intermédiaires.

3.2 Pilier 2 :

Que peuvent-ils faire ? (Accès et droits)

Comme les agents autonomes sont conçus avant tout pour agir, ils doivent être en mesure de se connecter aux données, API et systèmes. En s'appuyant sur les identités vérifiées établies dans le premier pilier, celui-ci définit les limites et les autorisations gouvernant les interactions des agents avec les ressources d'entreprise, aussi bien lors de sessions déléguées par des personnes que dans des workflows multiagents complexes, en enrichissant les modèles traditionnels centrés sur l'humain par une gouvernance machine-to-machine de l'IA.

3.2.1 Politiques d'accès (Définition des politiques d'accès)

Les politiques d'accès définissent les limites opérationnelles des actions autorisées des agents (lecture, écriture ou exécution). Compte tenu du caractère imprévisible des résultats générés par l'IA générative, ces politiques doivent être fondamentalement contextuelles et adaptatives.

- **Granularité limitée :** contrôles étendus, basés sur les rôles, précisant à quels environnements de haut niveau ou macro-services un agent peut se connecter.
- **Granularité élevée :** autorisations de niveau objet basées sur les attributs qui limitent l'accès d'un agent à certaines lignes de base de données, à des fichiers ou à des méthodes API spécifiques, en tenant explicitement compte des labels de classification des données de l'entreprise (par exemple, données personnelles ou confidentielles) afin d'appliquer des limites de données conformes. Les politiques peuvent englober la classification des données, l'identité des agents, le périmètre de risque, le contexte environnemental et le type d'action.
- **Accès basé sur les relations :** accès dérivé des relations du graphe d'un agent : qui l'a créé, au nom de qui il agit et sous quel élément il est imbriqué.
- **Accès basé sur l'intention :** moteurs de politiques avancés qui analysent l'intention sémantique du prompt ou du plan de l'agent avant de lui accorder l'accès aux ressources. L'inspection d'un plan est réalisable avec les outils actuels ; considérer un prompt utilisateur littéral comme unique déclencheur reste un objectif théorique, puisque ce ne sont pas toujours les prompts qui déclenchent l'action d'un agent. Des privilèges éphémères et limités dans le temps sont créés uniquement pour la durée d'une tâche précise, puis révoqués immédiatement une fois celle-ci terminée.
- **Accès basé sur le réseau :** segmentation limitant le trafic des agents entre les clouds virtuels privés d'entreprise, associée à des contrôles de microsegmentation L3/L4 qui limitent la communication agent-cible aux plages d'adresses IP ou aux maillages de services autorisés.

- **Accès limité dans le temps :** autorisations limitées à la fois dans le temps et à une seule tâche ou session, et automatiquement révoquées à la fin ou à l'expiration de la session ou tâche. Comme les autorisations à approbation et à expiration automatiques n'offrent qu'un contrôle limité, associez la restriction temporelle à la détection d'anomalies décrite dans la section 3.3.1.2, plutôt que de considérer l'expiration seule comme une garantie suffisante de protection.
- **Accès interapplications :** frameworks de politiques et contrôles du périmètre gouvernant la capacité d'un agent à naviguer entre différents écosystèmes applicatifs. Ils sécurisent la conversion des données, l'échange de tokens et l'exécution de fonctions lorsque l'agent orchestre des workflows sur différentes plateformes d'entreprise.
- **Délégation et héritage :** la limite effective des autorisations pour un sous-agent est calculée comme l'intersection entre ses propres droits et ceux de l'agent ou de l'utilisateur délégant, afin d'éviter toute élévation de privilèges par manipulation d'une entité de confiance (« confused deputy »). La politique doit également limiter l'étendue et la profondeur de la délégation afin d'éviter toute propagation incontrôlée.
- **Garde-fous :** contraintes de sécurité déterministes strictes et limites comportementales qui court-circuitent les boucles de raisonnement autonomes afin d'empêcher globalement toute action non conforme, résultat toxique ou opération à risque élevé, indépendamment de l'intention de l'agent ou de ses privilèges granulaires.

3.2.2 Gouvernance

La gouvernance définit la base de référence de conformité et le cycle de supervision continue, en s'appuyant sur des outils automatisés plutôt que sur des vérifications manuelles superficielles, afin que les privilèges des agents restent alignés sur le principe du moindre privilège, à la vitesse d'exécution des agents.

- **Vérification conditionnelle :** déclenchement de validations de workflows automatisées ou « humain dans la boucle » dès que l'action d'un agent dépasse des seuils financiers ou opérationnels définis (par exemple, un achat ou une réservation dépassant 1 000 €), que cette action soit ou non couverte par les autorisations techniques accordées.
- **Évaluation des accès :** analyse automatisée et continue des droits permettant aux propriétaires des systèmes de valider les autorisations non conformes.
- **Demandes d'accès :** workflows formels et audités pour le provisioning de nouveaux modèles, fonctionnalités ou droits d'accès à un agent existant. Bien que les demandes puissent émaner de propriétaires humains ou être générées de façon dynamique par des agents à l'exécution, les demandes initiées par des agents doivent être strictement évaluées en fonction de plafonds d'autorisations statiques et exigent une vérification obligatoire par le propriétaire pour toute extension du périmètre d'autorisation, afin d'éviter que les agents ne manipulent les workflows automatisés pour obtenir des droits d'accès excessifs.
- **Séparation des tâches :** règles algorithmiques empêchant un seul agent de détenir des privilèges incompatibles, par exemple générer et approuver des transactions financières.
- **Processus liés au cycle de vie des agents :** la même rigueur doit être appliquée à l'évolution du périmètre, de la propriété ou de la version du modèle d'un agent qu'à l'onboarding, à la gestion du changement et au déprovisioning des identités humaines.

3.3 Pilier 3 :

Que font-ils ? (Autorisation et contrôle à l'exécution)

La visibilité et l'identité sont inefficaces sans surveillance et application de contrôles en ligne à l'exécution. Ce pilier se concentre sur l'interception, l'inspection et l'autorisation en temps réel des actions des agents lors de l'exécution de leurs tâches pour éviter qu'ils n'adoptent des comportements malveillants, ne manipulent des données sensibles et ne compromettent des cibles critiques de l'entreprise.

3.3.1 Autorisation et surveillance à l'exécution

L'autorisation et la surveillance à l'exécution représentent la couche d'inspection active pour les sessions d'agent en cours. Directement placée dans le chemin d'exécution, cette couche analyse à la fois les entrées transmises à l'agent et les actions ou productions résultantes.

3.3.1.1 Passerelles (Application des politiques d'accès)

Les passerelles sont les points d'application des politiques dans une chaîne de défense en profondeur. Elles interceptent le trafic entre la couche agentique et le reste de la pile de l'entreprise, analysent les protocoles et bloquent les appels non autorisés avant qu'ils n'atteignent leur destination. Les passerelles API en aval conservent l'autorité d'autorisation finale, en effectuant leur propre validation de chaque appel, indépendamment des contrôles déjà réalisés par les passerelles IA/d'agents en amont. Les passerelles IA/LLM, MCP, de sécurité, d'agents et API sont souvent déployées ensemble, mais chacune est responsable d'un point de contrôle distinct.

Les passerelles IA/LLM gouvernent le trafic entre les modèles et les API ; les passerelles MCP gouvernent le trafic des protocoles de partage de contexte et d'outils ; les passerelles de sécurité offrent une inspection des menaces indépendante des protocoles ; les passerelles d'agents gèrent les problématiques d'identité et de délégation propres aux agents ; les passerelles API appliquent les contrôles classiques au niveau de la couche API pour les outils et services appelés par les agents.

- **Passerelle IA :** intercepte les appels API de bas niveau vers les modèles fondamentaux, et gère la limitation du débit, le comptage des tokens et la modification des prompts système.
- **Passerelle MCP (Model Context Protocol) :** proxy spécialisé, optimisé pour la gestion des protocoles de partage de contexte et des intégrations basées sur des standards ouverts entre agents et applications, avec un accent particulier sur les serveurs MCP.
- **Passerelle de sécurité :** couche d'atténuation des menaces en ligne qui offre une sécurité dédiée des contenus et une vérification des protocoles pour les interactions d'agents. Directement placée dans le flux de trafic, elle procède à une inspection approfondie des charges utiles en temps réel, avec terminaison TLS. La passerelle intercepte activement et filtre les injections de prompts complexes, les débridages, les chaînes de texte contradictoires cachées et les charges utiles de code malveillant avant qu'elles n'atteignent l'agent ou les ressources en aval. Elle se charge également de l'analyse de la classification des données afin d'appliquer l'anonymisation des données sortantes, les vérifications d'intégrité structurelle des résultats des modèles, ainsi que l'inspection cryptographique des charges utiles chiffrées des agents.

- **Passerelles d'agents :** point de contrôle à l'exécution spécifique aux agents : elle associe l'identité de l'agent à chaque demande, maintient le registre MCP/l'outil utilisé par les agents et gouverne les actions des agents (pas seulement l'accès) en fonction de l'identité, du contexte et de l'action spécifique demandée. Elle interrompt les interactions non sûres en temps réel, en bloquant ou en assainissant le trafic à risque des modèles, et en refusant ou en soumettant à approbation les actions d'outils sensibles via des politiques intégrées et personnalisées. Elle injecte également des identifiants en aval, éphémères et sans secret, dans les appels d'API au moment de l'exécution, de sorte que les agents ne puissent jamais manipuler, stocker ou exposer des clés API ou secrets de comptes de service d'entreprise permanents. Elle peut fournir une télémétrie gouvernée des interactions multiagents prises en charge. De même, elle est en mesure de détecter ou d'anonymiser les données sensibles, y compris les données personnelles conformément aux fonctionnalités et politiques configurées, tout en offrant une protection contre les comportements dangereux et les attaques malveillantes par prompt.
- **Passerelle API :** applique les contrôles habituels au niveau de la couche API, la limitation du débit, les quotas, la validation des schémas, mTLS, WAF, la gestion du trafic et la révocation au niveau passerelle pour les outils et services utilisés par les agents et serveurs MCP, en effectuant sa propre vérification d'autorisation finale à chaque appel au lieu de supposer que les passerelles IA/d'agents en amont ont déjà détecté toutes les anomalies.

3.3.1.2 Surveillance (Surveillance de l'environnement d'exécution)

La surveillance combine observabilité approfondie, détection des menaces et application active des politiques, en analysant la télémétrie d'exécution pour identifier les comportements anormaux, les fuites de données et les tentatives d'exploitation externe, et en appliquant des contrôles tels que le sandboxing et l'isolation des endpoints en temps réel.

- **Sandboxing :** isolation de l'exécution des agents (en particulier les agents non approuvés, tiers ou récemment mis à jour) dans des environnements d'exécution sécurisés et restreints afin d'empêcher les déplacements latéraux et de maîtriser les risques liés à l'exécution grâce à des limites structurelles.
- **Isolation des endpoints et gestion des listes d'autorisation :** contrôles locaux à l'exécution pour les agents hébergés sur les endpoints, incluant l'isolation au niveau du système d'exploitation, la gestion des listes d'autorisation des applications de l'endpoint et les vérifications locales de la posture de l'appareil afin de limiter l'exécution des commandes et l'accès aux fichiers de l'agent local.
- **Accès aux outils et aux données :** inspection structurelle et audit en temps réel des arguments, charges utiles, schémas et requêtes transmis en aval aux outils et aux référentiels. Une inspection au sein du périmètre d'exécution ou à proximité de celui-ci permet de s'assurer que les plans d'exécution dynamiques ne s'écartent pas des paramètres autorisés et n'introduisent pas de modifications imprévues en cours de session.
- **Détection des menaces et des anomalies :** moteurs de machine learning qui monitorent les comportements de référence des agents afin de signaler des pics inhabituels dans l'accès aux données, des boucles API rapides ou des modèles d'exécution atypiques.

- **Injection de prompts :** scanners intégrés conçus pour détecter les chaînes de texte malveillantes directes ou indirectes visant à détourner la logique fondamentale de l'agent ou à remplacer les instructions du système.
- **DLP / filtrage des sorties :** outils de prévention des pertes de données chargés d'analyser les résultats générés par les agents pour limiter le risque de divulgation de code source propriétaire, de données personnelles ou de secrets commerciaux à des prestataires externes ou à des utilisateurs non autorisés.
- **Traçabilité :** cartographie complète des transactions qui retrace précisément le chemin d'exécution, les piles d'appels et les appels de sous-agents déclenchés par un prompt utilisateur principal.
- **Coût et performance :** surveillance de l'utilisation des tokens, des coûts opérationnels, de la latence et de la consommation de ressources afin d'éviter les boucles d'agents incontrôlées ou les attaques par épuisement des ressources.
- **Workflow dans la boucle :** déclencheurs automatisés qui suspendent l'exécution de l'agent pour les actions à haut risque, en l'orientant vers des vérifications croisées automatisées des politiques ou une validation humaine explicite avant de poursuivre.
- **Surveillance continue :** surveillance continue et exportation des activités de l'agent, incluant une traçabilité granulaire, la journalisation des pistes d'audit et la non-répudiation des actions de l'agent.

3.3.1.3 Application continue de l'intention et de l'alignement

L'application continue de l'intention et de l'alignement permet de déterminer si un agent reste dans les limites de la tâche qui lui a été déléguée tout au long de l'exécution, ou s'il a dévié vers une extension non autorisée de sa finalité. Elle complète les politiques d'accès initiales : une politique basée sur l'intention détermine ce à quoi un agent peut accéder au départ, tandis que l'application continue de l'alignement évalue si des comportements en évolution et en plusieurs étapes restent appropriés dans la durée.

Parmi les principales fonctionnalités :

- **Respect de la finalité :** application d'une corrélation stricte entre l'utilisateur déléguant l'action, l'identité de l'agent, le workflow approuvé et la tâche demandée.
- **Évaluation du plan par rapport à l'action :** comparaison à l'exécution entre les appels d'outils prévus et les actions observées afin de détecter une dérive des tâches, des déviations causées par des injections et les comportements manipulateurs de type « confused deputy ».
- **Application contextuelle des actions :** exécution d'actions en temps réel, notamment l'autorisation, le refus, l'anonymisation, l'approbation renforcée, la réduction du champ d'application, la suspension ou le confinement.
- **Preuve de décision traçable :** maintien de journaux d'audit infalsifiables expliquant pourquoi chaque action importante a été autorisée, modifiée ou bloquée.

3.3.2 Accès aux ressources

L'accès aux ressources délimite la véritable portée d'un agent. Il catégorise les différentes cibles en aval avec lesquelles un agent peut interagir, pour s'assurer que des wrappers de sécurité appropriés entourent chaque interface.

3.3.2.1 Ressources cibles (Composants auxquels l'agent accède)

Les ressources cibles représentent les destinations des actions d'un agent. Qu'il s'agisse d'extraire du contexte ou de propager des modifications, ces endpoints doivent être constamment recensés et protégés.

- **Serveurs MCP :** référentiels de données et serveurs utilitaires explicitement configurés pour partager des informations contextuelles (données structurées, documents, embeddings, outils et état) avec les agents via le protocole MCP.
- **Applications SaaS :** applications métier et de productivité qui contiennent d'importants volumes de données sensibles de l'entreprise.
- **Data et référentiels de connaissances :** référentiels structurés et non structurés, y compris les bases de données vectorielles, les data lakes, les data warehouses et les systèmes de stockage transactionnels, utilisés par les agents pour extraire le contexte métier ou valider des résultats opérationnels. Ces référentiels sont gouvernés par des pratiques de gestion de la posture de sécurité des données qui suivent la traçabilité et la classification à mesure que les agents accèdent à chaque référentiel.
- **Outils CLI :** interfaces de ligne de commande permettant aux agents d'exécuter des scripts, des configurations ou des commandes opérationnelles au niveau du système.
- **Autres agents :** sous-agents en aval ou agents spécialisés chargés d'exécuter différentes parties d'une charge de travail plus importante.
- **Ressources à privilèges :** environnements cibles hautement sensibles, notamment les répertoires racine, les référentiels d'identités et les consoles de configuration principales.
- **Endpoints standard et hérités :** API REST/GraphQL d'entreprise, bases de données et applications héritées qui n'offrent pas une prise en charge native des protocoles de contexte agentique tels que MCP.
- **Systèmes de paiement :** les réseaux financiers, les passerelles de transactions et les livres comptables exigent le plus haut niveau de validation cryptographique et de supervision humaine.
- **Navigateurs :** sessions web automatisées utilisées par des agents pour extraire des informations publiques ou interagir avec des consoles web externes.
- **Environnements d'exécution de Skills et exécutables de code :** interpréteurs de code, sandbox d'exécution de scripts et modules procéduraux appelés dynamiquement par des agents pour exécuter du code généré, nécessitant un périmètre de calcul isolé afin de réduire les risques d'évasion à l'exécution.
- **Modèles fondamentaux :** fournisseurs d'intelligence brute qui hébergent les poids et nécessitent des clés API de modèles sécurisées et validées.
- **Canaux de communication et de collaboration :** e-mail, chat et systèmes de gestion des tickets à travers lesquels les agents interagissent et représentant des points d'entrée de prompts non fiables soumis à une inspection et à une analyse approfondies des charges utiles.

3.3.2.2 Sécurité des données axée sur la finalité

La sécurité des données axée sur la finalité évalue si un agent donné doit accéder à des données spécifiques, les traiter, les transformer, les résumer, les divulguer ou les transmettre en fonction de la finalité déléguée, de l'autorité de l'utilisateur, de l'identité de l'agent, de l'état du workflow, de la classification des données, du destinataire, de la destination et de l'action en aval. Elle applique des contrôles aux données au moment de leur utilisation, et pas uniquement lors de l'accès. C'est important car les agents combinent des données issues de différents systèmes, déduisent des informations sensibles à partir d'éléments non sensibles, et divulguent les résultats via des sorties, des API, des messages, des fichiers et des appels à des sous-agents. Les fonctions de classification et de prévention des pertes de données déterminent la nature des données et leur caractère sensible. La sécurité des données axée sur la finalité évalue, quant à elle, si leur utilisation est appropriée dans le contexte. Les principales fonctionnalités incluent le respect de la finalité entre la demande utilisateur, la tâche de l'agent, la source de données et l'utilisation autorisée ; l'évaluation à l'exécution de la pertinence de l'accès et de la divulgation selon l'étape du workflow en cours ; des contrôles sur la synthèse, la transformation, l'exportation, les mouvements interapplications et la divulgation à des agents tiers ; la détection d'une collecte excessive de données, d'une extraction excessive, d'inférences non autorisées et de partages inappropriés ; et l'application de mesures telles que l'anonymisation des données, le masquage, la réponse partielle, le workflow d'approbation, la journalisation, le blocage ou le confinement.

3.4 Pilier 4 :

Comment répondre ? (Confinement actif)

Le dernier pilier du modèle d'architecture transforme la télémétrie en confinement immédiat. Lorsque les systèmes de surveillance détectent un comportement anormal, des identifiants compromis ou une attaque par injection de prompt, l'entreprise doit disposer de mécanismes de correction hautement automatisés et ciblés qui privilégient des réponses adaptées et peu intrusives (limitation du débit, mise en quarantaine) plutôt que des mesures destructrices susceptibles de provoquer une interruption de service auto-déclenchée.

3.4.1 Réponse et application (Limitation du débit et confinement)

La boucle de contrôle opérationnel automatisée est conçue pour répondre aux signaux de risque en temps réel. Elle agit comme une couche de contrôle coordonnée, et non comme un point de blocage centralisé unique, en déclenchant des réponses de sécurité ciblées et adaptées dans l'ensemble de la pile technologique.

L'application des mesures est la plus efficace lorsqu'elle est associée à un protocole de notification standardisé : le système qui détecte un risque publie un événement d'arrêt d'urgence, les systèmes abonnés réagissent, puis le résultat est publié en retour sous forme de signal de confirmation.

3.4.1.1 Actions de correction (Mécanismes d'atténuation)

Les actions de correction représentent les leviers multiniveaux concrets que les équipes sécurité et les orchestrateurs automatisés peuvent actionner pour isoler, confiner ou neutraliser totalement un agent problématique ou compromis. Ces actions constituent une bibliothèque de réponses gérée par le SOC, et non des dispositifs d'arrêt ponctuels isolés : les gestionnaires d'incidents choisissent l'action la moins perturbatrice par rapport au signal de risque agrégé.

Les actions de confinement déclenchées par les agents de protection, ou gardiens (suspension, isolation, arrêt) exigent que l'agent ciblé opère dans un périmètre d'exécution prenant en charge un contrôle granulaire au niveau de la session, que ce soit dans un service cloud géré, une application SaaS ou un environnement d'exécution conteneurisé. Sans prise en charge d'une plateforme ou environnement d'exécution sous-jacent permettant de suspendre un agent sans effet collatéral sur les workflows adjacents, le confinement piloté par API procure un faux sentiment de sécurité.

Lors du confinement d'agents associés à des équipements physiques ou OT, les déclencheurs automatisés doivent être calibrés de telle sorte qu'une isolation soudaine ne crée pas de risques opérationnels ou pour la sécurité physique dans l'usine.

- **Révocation de tokens :** invalidation instantanée des tokens d'accès OAuth ou API actifs, bloquant immédiatement la capacité de l'agent à appeler les applications connectées.
- **Révocation de tokens de session :** invalidation des tokens OAuth actifs et révocation des contextes de session en aval dans différents services d'entreprise connectés et/ou interruption de la connexion d'exécution active d'une chaîne d'interaction précise, sans pour autant supprimer l'identité globale de l'agent.
- **Autorisation continue :** réévaluation dynamique en temps réel des contextes de sécurité, qui limite instantanément les autorisations dès que la posture de risque change.

- **Arrêt de processus :** arrêt forcé des environnements d'exécution en dernier recours lorsque la mise en quarantaine échoue. Toutefois, comme l'arrêt entraîne la perte des données télémétriques d'investigation en mémoire, il est préférable d'opter pour la mise en quarantaine réseau, la suspension de session ou de l'environnement d'exécution lorsque c'est possible.
- **Mise en quarantaine :** utilisation de règles de microsegmentation pour isoler l'environnement d'hébergement de l'agent et ainsi interrompre ses communications à la fois avec le réseau interne et avec Internet.
- **Arrêt du service cloud :** arrêt immédiat des instances cloud sous-jacentes, des fonctions sans serveur ou des clusters de machines virtuelles hébergeant l'agent non autorisé.
- **Arrêt de la plateforme d'agents :** commande administrative de niveau supérieur qui indique à la plateforme centralisée d'orchestration des agents de suspendre le compte opérationnel de l'agent et d'arrêter toutes les tâches planifiées.
- **Limitation du débit :** restriction progressive de la fréquence des requêtes, du budget de tokens ou de la simultanéité d'un agent en tant que première réponse proportionnée, afin de réduire ses capacités sans interrompre totalement la session.

3.4.2 Rétablissement de l'accès

Un arrêt d'urgence ne constitue qu'une partie de la boucle de contrôle : chaque action de confinement doit s'accompagner d'un chemin permettant de revenir au dernier état validé et vérifié. Le rétablissement de l'accès gouverne la manière dont un agent isolé, mis en quarantaine ou supprimé est remis en service, et veille à ce que la réactivation soit volontaire, documentée et traçable, plutôt qu'une simple annulation automatique de l'action de correction. Des workflows de récupération standardisés et entièrement automatisés restent encore au stade de la théorie ; la plupart des entreprises exigent toujours une validation manuelle pour une ou plusieurs des étapes ci-dessous.

- **Réattestation :** nouvelle vérification de l'identité, du code et de la configuration de l'agent par rapport à une base de référence validée avant de rétablir l'accès dans une nouvelle instance d'exécution isolée.
- **Réinscription progressive :** rétablissement de l'accès par étapes (par exemple, la lecture seule avant l'écriture, l'exécution en sandbox avant la mise en production) plutôt qu'un rétablissement immédiat de tous les privilèges.
- **Validation de la cause profonde :** la personne responsable ou l'équipe sécurité doit documenter la cause profonde et la mesure corrective appliquée avant de réactiver l'agent.
- **Clôture de la piste d'audit :** regroupement de l'événement d'arrêt d'urgence initial, des actions de correction et de la décision de réinscription au sein d'un même enregistrement d'incident clos auditable.

3.5

Composants fondamentaux transversaux

Ces couches architecturales ne sont pas cloisonnées ; elles sous-tendent et traversent les quatre piliers, reliant la visibilité, l'identité, la surveillance et la réponse au sein d'un écosystème de sécurité unifié.

3.5.1 Contexte d'exécution et signaux de risque

Le contexte d'exécution et les signaux de risque constituent le tissu analytique relationnel du modèle. Alimenté en continu par la télémétrie d'exécution en temps réel, les évaluations de posture de sécurité et les signaux d'identité des utilisateurs délégants (tels que des indicateurs de compromission de compte ou un risque accru lié au principal), ce moteur d'orchestration évalue de façon dynamique les menaces en corrélant des signaux à faible risque individuellement en combinaisons toxiques. Il publie également un contexte d'exécution positif, une confirmation structurelle des éléments déjà reconnus comme valides, et pas seulement des alertes sur les problèmes rencontrés. Lorsqu'un seuil est franchi (par exemple, une injection de prompt soudaine associée à une tentative d'exfiltration de données financières), les signaux de risque déclenchent directement les mécanismes de réponse/application correspondants de manière automatisée. Pour favoriser l'interopérabilité en temps réel et la coordination des réponses dans des environnements fournisseurs hétérogènes, ce système peut s'appuyer sur des standards d'ingénierie de la sécurité ouverts tels que le framework **SSF (Shared Signals Framework)** et le profil **CAEP (Continuous Access Evaluation Profile)** afin de publier et d'utiliser en mode asynchrone des modifications d'état Zero Trust.

3.5.2 Télémétrie, journalisation et observabilité des agents d'IA

Il s'agit de la couche fondamentale sous-tendant tous les piliers de l'architecture de référence. Il sert de plan de données immuable, capturant les métadonnées d'exécution dans toute la pile de l'entreprise, avec une anonymisation obligatoire et immédiate des charges utiles afin d'éviter que des données personnelles, des secrets et des identifiants ne soient enregistrés dans le log lake. Sans ce log lake fiable et structuré, la détection automatisée des menaces est impossible, les évaluations de posture manquent de données et les équipes d'intervention en cas d'incident n'ont aucun élément pour mener leurs investigations. Pour favoriser la cohérence et limiter le cloisonnement lié aux fournisseurs dans des interactions multiagents complexes, cette couche peut s'appuyer sur des standards ouverts tels que l'**Open Cybersecurity Schema Framework (OCSF)** afin de normaliser des flux de journaux hétérogènes au sein d'une seule taxonomie de sécurité unifiée.

- **Origine de la télémétrie résistante à la falsification :** collecter les données télémétriques au sein du périmètre d'exécution au lieu de s'appuyer sur les journaux auto-déclarés des agents permet de garantir la non-répudiation, et empêche ainsi qu'un agent compromis ne puisse supprimer, modifier ou falsifier son propre historique d'exécution.

4

Opérationnalisation avec des intégrateurs système mondiaux (GSI)

Un modèle d'architecture exige un modèle de déploiement opérationnel pour apporter une valeur ajoutée durable à l'entreprise. Les sections ci-dessous expliquent comment les équipes sécurité et les partenaires conseil opérationnalisent ces quatre piliers dans les workflows existants de l'entreprise. Les entreprises peuvent opérationnaliser cette architecture par l'intermédiaire d'équipes internes d'entreprise, d'intégrateurs système mondiaux (GSI), de fournisseurs de services gérés ou d'autres partenaires qualifiés, selon leurs capacités et exigences. Une architecture de référence publiée sans modèle opérationnel est reléguée au placard en moins d'un an ; les sections ci-dessous expliquent comment cette architecture est réellement mise en place, dotée des ressources nécessaires et pérennisée au sein d'une entreprise.

4.1 Conseil en déploiement et opérationnalisation de l'architecture

Les GSI convertissent les quatre piliers en un déploiement progressif, adapté à chaque entreprise, plutôt qu'en un déploiement simultané et global. Par exemple :

- **Évaluation de l'état actuel :** cartographie de l'IAM, du réseau et de la pile de surveillance existants d'une entreprise par rapport à l'architecture de référence Blueprint Alliance afin d'identifier les véritables lacunes par opposition aux fonctionnalités déjà présentes sous une autre appellation.
- **Conception du modèle opérationnel cible :** définition d'une attribution claire des responsabilités opérationnelles entre les différents piliers afin d'éviter toute lacune en matière de responsabilité entre les équipes. Ce modèle attribue généralement les passerelles et la surveillance à l'équipe d'ingénierie de la sécurité ; l'annuaire d'agents, les politiques d'accès et la gouvernance à l'équipe IAM ; la réponse et l'application à l'équipe SOC ; et les limites d'exécution (sandboxing, environnements d'exécution isolés, gestion du cycle de vie) à l'équipe d'ingénierie plateforme et infrastructure.
- **Planification du déploiement par étapes :** déploiement progressif, adapté aux investissements existants en outils du client et à son calendrier réglementaire.
- **Intégration de l'interopérabilité fournisseur :** configuration des passerelles IA/LLM, MCP, de sécurité, d'agents et API spécifiques déjà sous licence dans l'entreprise en vue de leur interopérabilité via la couche partagée de télémétrie et de signaux de risque, sans imposer de remplacement complet de l'infrastructure existante.

4.2 Matrices standardisées de notation des risques

Tous les agents n'exigent pas le même niveau de contrôle, et les GSI proposent des frameworks d'évaluation des risques adaptés qui évitent aux entreprises de complexifier inutilement les déploiements à faible risque ou de sous-protéger ceux présentant un risque élevé. Par exemple :

- **Modèles de hiérarchisation des risques** qui attribuent un score aux agents en fonction du périmètre opérationnel (lecture seule ou transactionnel), des classifications des données, du niveau d'autonomie (OBO ou entièrement autonome) et de l'impact des ressources ciblées.
- **Seuils adaptés au secteur :** un agent des services financiers qui exécute des workflows de paiement et un agent du secteur industriel qui lit les données de capteurs OT ont besoin de bases de référence des risques sensiblement différentes ; les GSI s'appuient sur la bibliothèque de modèles intersectorielle pour calibrer ces seuils, au lieu de partir de zéro.
- **Évaluation de l'impact métier pour la gouvernance sémantique :** définition des seuils financiers, réglementaires ou de réputation déclenchant une validation humaine obligatoire et adaptés à la tolérance réelle au risque du client plutôt qu'à une valeur par défaut générique.
- **Profondeur d'isolation adaptée au niveau de risque :** la hiérarchisation des risques doit permettre d'ajuster la robustesse du périmètre d'exécution structurel (par exemple, sandboxing au niveau du noyau, cycles de vie d'exécution éphémères) ainsi que les points de contrôle, afin qu'un agent à haut risque ne partage pas le même périmètre d'impact physique qu'un outil à faible risque en cas de compromission.

4.3 Gestion continue du cycle de vie

La gouvernance des agents ne se limite pas à une certification ponctuelle ; les GSI gèrent les processus continus du cycle de vie qui garantissent la fiabilité et l'exactitude de l'annuaire d'agents et des politiques d'accès à mesure que l'environnement évolue :

- **Cycles automatisés de certification des accès :** extractions gérées de l'état de l'annuaire d'agents et des politiques d'accès, signalant les agents dont les propriétaires ont quitté l'entreprise ou dont les droits se sont éloignés du principe du moindre privilège.
- **Parité des processus du cycle de vie pour les agents :** application de la même rigueur de déprovisioning que celle associée à l'offboarding des utilisateurs aux changements de propriété des agents, aux mises à niveau de la version des modèles et à la désactivation des sous-agents, afin d'éviter l'accumulation d'identifiants orphelins, principale source des failles dans l'infrastructure d'identités.
- **Routage des exceptions de séparation des tâches :** exécution de workflows qui transmettent les exceptions liées à des privilèges conflictuels (par exemple, un agent autorisé à générer et à approuver une transaction financière) au propriétaire du système concerné en vue de leur résolution, à la même fréquence que la gouvernance des identités humaines.
- **Gestion de la dérive des environnements d'exécution :** l'application planifiée des correctifs, la mise à jour des dépendances, la réattestation des environnements d'exécution des agents ainsi que les configurations de la plateforme sous-jacente permettent de limiter la dérive des vulnérabilités dans les déploiements hébergés, cloud et conteneurisés.

4.4 Alignement réglementaire et de conformité

Les GSI font le lien entre l'architecture et les réglementations applicables à une entreprise, en convertissant les contrôles techniques en éléments de preuve conçus pour répondre aux exigences réglementaires, de conformité et d'audit :

- **Association des contrôles aux frameworks réglementaires :** alignement de la télémétrie, des pistes d'audit et des résultats des évaluations d'accès sur les exigences propres à chaque secteur (par exemple, gestion des risques liés aux modèles dans les services financiers, gestion des données de santé, protection des infrastructures critiques), à mesure qu'elles évoluent en parallèle des recommandations spécifiques à l'IA agentique.
- **Packages de preuves auditables :** la télémétrie et les données de sortie de la couche de journalisation doivent être structurées de telle sorte que les évaluations des accès, les événements de confinement et de limitation du débit, ainsi que les enregistrements de rétablissement de l'accès (réattestation, validation de la cause première, clôture de la piste d'audit) répondent aux standards de documentation des examinateurs et des auditeurs, et pas seulement aux besoins internes du SOC.
- **Conseils sur la gestion transfrontalière et la résidence des données :** conseils sur l'alignement des canaux d'importation d'agents (agents SaaS, agents propriétaires franchissant les frontières organisationnelles) avec les exigences de résidence des données et de transfert transfrontalier propres à chaque juridiction dans laquelle l'entreprise exerce ses activités.
- **Traçabilité d'audit infalsifiable :** la collecte de données télémétriques est structurée de telle sorte que les autorités de réglementation et les auditeurs reçoivent des journaux d'audit non répudiables, générés directement au sein du périmètre d'exécution, ce qui limite la dépendance vis-à-vis d'une reconstitution secondaire des journaux à partir de l'infrastructure réseau environnante.

4.5 Playbooks sectoriels

Les exigences de contrôle varient selon les secteurs, ce qui oblige les GSI à disposer de playbooks adaptés à des environnements opérationnels spécifiques :

- **Services financiers :** ce secteur privilégie une stricte isolation de l'exécution, des seuils de transaction à niveau d'assurance élevé et des contrôles des données réglementaires.
- **Industrie et fabrication :** ce secteur met l'accent sur la découverte des agents OT, IoT et en périphérie sur les équipements hérités.
- **Commerce de détail et e-commerce :** ce secteur accorde la priorité à la défense contre l'injection de prompts B2C, à la protection de la marque et à la fuite de données personnelles des clients.

Les membres fondateurs mettent régulièrement à jour ces playbooks au fur et à mesure de la publication des résultats d'interopérabilité conjoints, et vérifient que le modèle opérationnel s'adapte à l'émergence de nouvelles capacités agentiques.

Conclusion : façonner l'avenir de la sécurité de l'IA

L'architecture Blueprint Alliance garantit une bonne gouvernance du comportement des agents à l'exécution en production. Intégrée directement à la pile de l'entreprise plutôt qu'ajoutée a posteriori comme une couche externe, cette architecture offre les garde-fous structurels nécessaires pour déployer l'automatisation agentique à grande échelle, en toute sécurité.

Pour offrir une véritable interopérabilité multifournisseurs, les membres fondateurs valident activement leurs plateformes par rapport aux principaux standards ouverts, notamment le protocole MCP (Model Context Protocol) pour l'interaction avec les outils, le framework OCSF (Open Cybersecurity Schema Framework) pour la journalisation unifiée, le framework SSF et le profil CAEP (événements et signaux partagés) pour l'échange de risques en temps réel, et le standard HTTP Message Signatures (RFC 9421) pour l'authentification des demandes. Cette base de référence télémétrique partagée permet de s'assurer qu'un signal de menace ou une violation de périmètre détectés dans un environnement d'exécution peuvent être intégrés et corrigés de manière native au niveau du point de contrôle. Les fonctionnalités varieront selon le fournisseur et l'implémentation ; les intégrations de référence et les résultats d'interopérabilité validés seront publiés une fois les tests achevés.

En publiant régulièrement des résultats d'interopérabilité conjoints et des intégrations de référence, la Blueprint Alliance propose un écosystème multifournisseur évolutif qui progresse au rythme de l'innovation agentique.

Le modèle en action

Les cinq scénarios ci-dessous illustrent le fonctionnement conjoint des piliers de la section 3 en pratique. Chaque scénario est élaboré à partir de différents modèles observés lors des premiers projets menés par l'Alliance, sans s'inspirer d'un client particulier. Ces scénarios composites illustrent le comportement attendu de l'architecture ; ils ne témoignent pas de déploiements actuels réalisés par des membres de l'Alliance. Considérez-les comme l'expression d'une architecture flexible, et non comme l'étude de cas d'un seul déploiement.

De l'agent clandestin à une identité enregistrée. Un analyste financier connecte un assistant d'IA accessible via navigateur à sa session SSO d'entreprise afin de générer une synthèse de certains contrats fournisseurs. La fonctionnalité de découverte de la Shadow AI signale le trafic sortant non reconnu dès la première connexion. La fonction de gestion de la posture de sécurité des agents d'IA compare l'outil aux vulnérabilités connues et le dirige vers l'annuaire d'agents. Avant même d'accéder aux données sensibles, l'assistant est soit enregistré en tant qu'identité gouvernée avec un profil de métadonnées vérifié et un accès limité, soit bloqué au niveau de la couche réseau. Ce qui restait auparavant invisible pendant des mois est désormais visible, contrôlé et validé avant même d'entrer en contact avec des données sensibles.

Préservation de la délégation dans un workflow multiagent. Un responsable des opérations commerciales demande à un agent orchestrateur de mettre à jour un ensemble d'enregistrements CRM et d'envoyer une notification à une liste de diffusion. L'orchestrateur crée un sous-agent pour gérer la mise à jour et un autre pour rédiger la notification. Chaque saut retransmet une déclaration *actor* imbriquée à la session utilisateur d'origine, afin que l'annuaire d'agents et le moteur des politiques d'accès puissent appliquer les contraintes OBO à chaque étape. Si un sous-agent tente d'accéder à un système auquel le responsable d'origine n'avait pas accès, les limites fondamentales de l'IAM et des politiques, y compris l'accès interapplication, l'en empêchent. De plus, comme chaque sous-agent s'exécute dans son propre contexte d'exécution isolé, une attaque bloquée ne peut pas affecter la session de l'orchestrateur ni les tâches associées, tandis que l'ensemble du workflow reste unifié dans une seule trace d'exécution.

Dimensionnement de l'accès adapté aux tâches que l'agent cherche réellement à accomplir. Un agent du support est provisionné avec un accès en lecture à granularité limitée à une plateforme de données clients. Lorsqu'il tente d'effectuer un remboursement, le moteur de politiques basées sur l'intention détecte la dérive de la consultation à la transaction et exige une approbation « humain dans la boucle » avant d'accorder un privilège finement granulaire et éphémère, limité à cette commande uniquement. L'exécution du remboursement dans un environnement d'exécution éphémère, limité à cette seule tâche approuvée, garantit que les privilèges élevés et l'environnement d'exécution soient supprimés une fois la tâche terminée, ce qui réduit les autorisations permanentes et les surfaces d'attaque persistantes. Le privilège élevé expire dès que la tâche est terminée.

De l'anomalie au confinement automatisé. La surveillance à l'exécution, généralement assurée par un modèle classique de détection d'anomalies basé sur le machine learning plutôt que par un LLM, détecte qu'un agent effectue un volume inhabituellement élevé d'appels à une API interne qu'il sollicite rarement. Parallèlement, un agent de protection (gardien) en ligne signale séparément une signature d'injection de prompt dans sa dernière entrée. Les signaux de risque mettent en corrélation les deux événements via le framework SSF (Shared Signals Framework), franchissent un seuil de risque défini, puis déclenchent automatiquement la révocation du token, la mise en quarantaine réseau et la suspension de l'environnement d'exécution actif de l'agent, tout en préservant son état en mémoire aux fins d'investigation numérique, au lieu de l'arrêter purement et simplement. Tout cela bien sûr sans attendre qu'un utilisateur remarque l'alerte. L'équipe sécurité reçoit la trace complète en tant qu'enregistrement de confirmation en boucle fermée, transformant le confinement en réponse automatisée quasi instantanée, alors qu'il exigeait auparavant une investigation manuelle de plusieurs heures.

Conversion des évaluations des accès en un processus continu. Au cours de chaque cycle continu de certification des accès, le service géré d'un GSI extrait l'intégralité de l'état de l'annuaire d'agents et des politiques d'accès via la couche de télémétrie ouverte du modèle, signale les agents dont les propriétaires ont quitté l'entreprise, et transmet les exceptions liées à la séparation des tâches au propriétaire du système concerné. Une évaluation manuelle, basée sur des feuilles de calcul, devient un workflow continu et auditable.

Bonnes pratiques pour bien démarrer

Vous n'avez pas besoin de toutes les fonctionnalités de ce modèle dès le premier jour. Voici l'ordre dans lequel les déploiements des membres de la Blueprint Alliance ont réellement été effectués.

Commencez par mettre en place une journalisation et une télémétrie unifiées. Vous ne pouvez pas corriger l'application des politiques ou les comportements de référence sans disposer d'un log lake unique et normalisé. Il s'agit d'une condition préalable à chacune des étapes ci-dessous.

Ensuite, passez à la découverte des agents, avant les politiques. Vous ne pouvez pas gouverner ce que vous ne voyez pas. Implémentez une fonction de découverte de la Shadow AI et procédez à un premier inventaire des agents avant de rédiger la moindre politique d'accès. La plupart des équipes sont surprises par le nombre d'agents en cours d'exécution.

Enregistrez les agents avant d'appliquer des restrictions d'accès. Regroupez tous les agents identifiés dans un annuaire d'agents centralisé, avec des identités et des propriétaires vérifiés, avant d'appliquer des politiques granulaires ou basées sur l'intention. Une politique appliquée à un agent non enregistré reste fondamentalement sans effet.

Ajoutez la couche d'autorisation. Des contrôles granulaires (FGA) et fondés sur les relations (ReBAC) doivent être en place avant que l'autorisation basée sur l'intention puisse correctement limiter les autorisations des agents.

Instrumentez l'environnement d'exécution avant d'automatiser la réponse. Déployez des fonctions de surveillance, de traçage et de filtrage DLP/des résultats de manière suffisamment large pour établir une véritable base de référence comportementale avant de connecter l'application automatique des politiques aux signaux de risque. Si vous automatisez le confinement en présence de données télémétriques parasites ou incomplètes, les faux positifs risquent d'éroder la confiance vis-à-vis du système.

Testez votre arrêt d'urgence avant d'en avoir besoin. Validez régulièrement la révocation des tokens, la clôture de session et la mise en quarantaine réseau au moyen d'un agent hors production. Un mécanisme d'arrêt d'urgence que vous n'avez jamais testé reste une hypothèse, pas un dispositif de contrôle.

Considérez la gouvernance comme un processus continu, pas comme un projet ponctuel. Les évaluations des accès, les demandes d'accès et les contrôles de séparation des tâches doivent être effectués avec la même régularité que la gouvernance des identités humaines, et non comme une initiative ponctuelle d'IA agentique. Un intégrateur système mondial (GIS) peut vous aider à transformer ce projet en modèle opérationnel.