Livre blanc

Sécurisation complète des agents d'IA, de leur développement à leur utilisation dans l'entreprise

Sommaire

Résumé

Les agents d'IA ne se contentent pas de transformer les modes de travail, ils redéfinissent fondamentalement l'identité.

Conçus pour être autonomes, ils sont indépendants, guidés par des objectifs précis et agissent de plus en plus sans supervision humaine. Grands consommateurs de données, ils analysent constamment des informations, écrivent du code, envoient des e-mails et prennent des décisions dans différents systèmes. Programmés pour atteindre rapidement leurs objectifs, les agents repoussent sans cesse les limites pour trouver de nouveaux moyens de consommer encore plus de données. En l'absence de garde-fous, ils peuvent malencontreusement s'égarer, laissant dans leur sillage dégâts et chaos. Pourtant, la plupart des entreprises sont incapables de répondre aux questions les plus simples : Où sont-ils déployés dans mon écosystème ? À quelles données et à quels systèmes peuvent-ils accéder ? Qui est responsable lorsqu'ils décident de faire cavalier seul ? D'après le rapport AI at Work 2025 d'Okta, **91 % des entreprises utilisent des agents d'IA**, mais **44 % n'ont pas mis en place de gouvernance**. C'est la raison pour laquelle on voit apparaître une nouvelle frontière de la sécurité : une explosion des identités non humaines autonomes qui ont besoin d'un framework cohérent à des fins d'authentification, d'autorisation ou de visibilité.

L'essor de l'IA agentique perturbe les fondements mêmes de la gestion des identités et des accès. Les contrôles traditionnels sont conçus pour les humains : ils ne peuvent pas suivre le rythme d'agents capables d'initier des workflows complexes et des chaînes d'API à grande échelle, sans supervision humaine. La prochaine génération de solutions de sécurité des identités **doit évoluer à la vitesse de l'IA**, en s'adaptant à sa vélocité, à son envergure et à son intelligence pour **garder la confiance des clients**.

Ce livre blanc s'intéresse à la façon dont il est possible de **sécuriser tous les agents**, en intégrant la sécurité de la première ligne de code jusqu'au point de contrôle de l'entreprise qui les gère. En effet, à l'ère de l'IA autonome, **l'identité ne représente pas simplement ce que nous sommes, mais aussi la façon dont nous gardons le contrôle.**

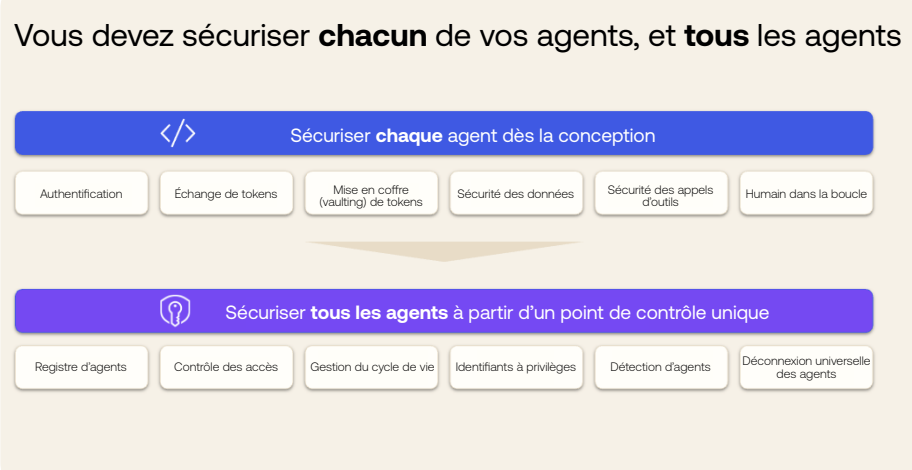
Sujets abordés

Nous proposons un framework complet pour relever le double défi de la sécurité des agents d'IA :

1. Pour les concepteurs — Sécuriser chaque agent dès la conception : découvrez les modèles de sécurité essentiels que les développeurs doivent intégrer lors de la création, notamment une **authentification robuste des utilisateurs**, une **mise en coffre (vaulting) de tokens** sécurisée pour l'accès aux API, une **autorisation d'accès aux données** granulaire pour les systèmes RAG et des contrôles « **humain dans la boucle** » pour les actions critiques.

2. Pour les équipes IT et sécurité — Sécuriser tous les agents à partir d'un point de contrôle unique : maîtrisez les fonctionnalités de haut niveau nécessaires à la gestion des agents à grande échelle. Elles incluent la **détection des agents** (pour identifier l'IA sans supervision), un **registre d'agents** (pour établir l'identité et la propriété), un **contrôle des accès complet avec confiance interdomaine** (permettant aux agents d'accéder en toute sécurité aux ressources au-delà du périmètre organisationnel, tout en préservant le contexte utilisateur), une **gestion complète du cycle de vie** et enfin, la **détection des menaces** avec des fonctions de déconnexion universelle.

3.



Points à retenir

- **Le risque majeur réside dans une gouvernance lacunaire :** le problème fondamental n'est pas l'IA en soi, mais le fait que le déploiement des agents va beaucoup plus vite que la mise en place de la gouvernance nécessaire pour les contrôler et les superviser. La gouvernance englobe à la fois les contrôles préventifs (politiques d'accès, principe du moindre privilège, règles d'autorisation) et la supervision (certifications, évaluations des accès, suivi comportemental). Les entreprises ont besoin des deux aspects pour gérer efficacement les agents d'IA.
- **La sécurité constitue un double défi :** une stratégie complète doit prendre en compte à la fois la sécurité lors de la phase de développement (conception correcte des agents) et la gouvernance au niveau de l'entreprise (gestion de tous les agents à grande échelle).
- **Une plateforme unifiée est essentielle :** seule une plateforme d'identité unifiée qui traite les agents comme des identités de premier ordre, en les gérant depuis la découverte et l'enregistrement tout au long de leur cycle de vie permettra de pallier les lacunes de gouvernance, d'atténuer les risques posés à la confidentialité des données et de déployer l'IA à l'échelle de l'entreprise en toute confiance.

Sécuriser chaque agent dès la conception

Lorsque vous créez des agents d'IA, vous êtes confronté à des exigences de sécurité totalement différentes de celles associées au développement d'applications classiques. La sécurité ne peut pas être ajoutée a posteriori : elle doit être embarquée dans l'architecture de l'agent dès la première ligne de code.

Cette section s'adresse à trois profils de concepteurs distincts :

- **Concepteurs de solutions SaaS B2C** chargés de créer des agents d'IA orientés consommateurs (chatbots, assistants personnels, moteurs de recommandations)
- **Concepteurs de solutions SaaS B2B** chargés de développer des agents d'IA pour les clients professionnels (automatisation des workflows, analytique, outils d'entreprise)
- **Développeurs internes** chargés de concevoir des agents d'IA internes pour les workflows et les processus spécifiques à leur entreprise

Si les détails de l'implémentation peuvent varier légèrement d'un scénario à l'autre, les principaux modèles de sécurité — authentification, gestion des tokens, autorisation et supervision humaine — restent d'application. Cette solution cherche à offrir à tous les développeurs les fonctionnalités essentielles dont ils ont besoin pour créer des agents sécurisés sans sacrifier la vitesse ou l'innovation.

1. Authentification : établissement de l'identité utilisateur

Les agents d'IA doivent identifier les utilisateurs de façon sécurisée pour offrir des expériences personnalisées tout en respectant les limites de sécurité. Il est essentiel de comprendre ce concept : nous n'authentifions pas l'agent lui-même mais l'utilisateur, et l'agent agit au nom de l'utilisateur authentifié. Qu'il s'agisse de chatbots interactifs ou de programmes travaillant en arrière-plan, les agents ont besoin d'une authentification fiable qui s'intègre harmonieusement avec les fournisseurs d'identité actuels.

Fonctionnalités indispensables aux entreprises :

- **Authentification universelle** compatible avec un large éventail de fournisseurs d'identité et prenant en charge les identifiants traditionnels et les options d'authentification sociale
- **Authentification basée sur des standards** qui utilise OpenID Connect et OAuth 2.0 pour assurer l'interopérabilité et la sécurité
- **Identité de l'utilisateur transmise par des tokens sécurisés**, permettant aux agents de déterminer pour qui ils agissent
- **Gestion stricte des sessions** avec des délais d'expiration et des contrôles de sécurité appropriés
- **Prise en charge de l'authentification multifacteur** (MFA) pour les scénarios nécessitant un niveau d'assurance de sécurité élevé

L'expérience développeur doit permettre d'intégrer l'authentification en quelques lignes de code seulement, de façon compatible avec les frameworks les plus répandus, et de gérer automatiquement la complexité des URL de rappel, la gestion des sessions et la validation des tokens.

Exemple : un chatbot de support client authentifie les utilisateurs via le SSO Google. Lorsque Sarah se connecte, l'agent reçoit ses informations d'identité, ce qui lui permet de formuler des réponses personnalisées tout en préservant les limites de sécurité.

2. Échange de tokens : interconnexion des domaines de confiance

Comme les agents d'IA couvrent plusieurs systèmes et domaines de sécurité, ils ont souvent besoin d'accéder à des ressources situées dans des zones de confiance différentes. L'échange de tokens permet aux agents d'obtenir des tokens d'accès correctement délimités pour des ressources hébergées en dehors de leur domaine immédiat, tout en préservant le contexte utilisateur et les chaînes d'autorisation.

Fonctionnalités indispensables aux entreprises :

- **Échange de tokens standard** pour les scénarios au sein d'un seul domaine de confiance, permettant aux agents de demander différents types de tokens ou champs d'application auprès du même serveur d'autorisation
- **Confiance interdomaine** pour les scénarios nécessitant un accès à des zones de confiance distinctes
- **Mécanismes permettant de préserver l'identité utilisateur** et le contexte d'authentification dans différentes zones de confiance
- **Validation des relations de confiance** entre différents fournisseurs d'identité
- **Translation des champs d'application** assurant une mise en correspondance correcte des autorisations entre les domaines
- **Conversion sécurisée des identifiants** qui n'expose jamais les tokens sensibles en transit

Pour les agents appartenant à un environnement de serveur d'autorisation unique, l'échange standard de tokens OAuth 2.0 permet de gérer efficacement les identifiants. Lorsque les agents doivent bénéficier d'un accès au-delà du périmètre organisationnel, la confiance interdomaine permet d'étendre cette fonctionnalité à d'autres domaines de confiance.

Le choix entre les flux de consentement OAuth standard ou la confiance interdomaine sera déterminé par votre modèle de déploiement :

Scénarios B2C : consentement OAuth standard

- Applications orientées consommateurs avec des utilisateurs finaux propriétaires de leurs propres données
- Utilisateurs autorisant explicitement une application à accéder à une autre (par exemple en autorisant TravelBot à accéder à Google Agenda)
- Option de l'écran de consentement appropriée, car les utilisateurs prennent des décisions personnelles concernant leurs propres données
- **Exemple :** une application de planification des repas demande l'accès aux données du bracelet d'activité d'un utilisateur.

Scénarios B2B et collaborateurs : confiance interdomaine

- Environnements d'entreprise où les administrateurs IT gèrent les politiques d'accès de manière centralisée
- Scénarios B2B2E (Business-to-Business-to-Employee) où les collaborateurs doivent respecter les politiques de l'entreprise
- Option des écrans de consentement inappropriée, car l'accès est régi par des politiques d'entreprise et non par les décisions des utilisateurs
- Fournisseur d'identité (IdP) de l'entreprise assumant le rôle d'intermédiaire de confiance
- **Exemple :** un commercial de l'entreprise accède à la fois au CRM Salesforce et à une base de données de tarifs interne — les collaborateurs n'ont pas de consentement à donner, cet accès est régi par la politique IT de l'entreprise.

Pourquoi c'est important : dans les contextes liés au B2B et aux collaborateurs, la confiance interdomaine permet d'éliminer la lassitude vis-à-vis du consentement et s'aligne sur la gouvernance IT centralisée. Les entreprises établissent au préalable des relations de confiance entre les applications, et le fournisseur d'identité applique les politiques de l'entreprise plutôt que de demander aux utilisateurs individuels de prendre des décisions d'autorisation pour chaque interaction entre les applications.

3. Mise en coffre (vaulting) des tokens : gestion sécurisée des accès aux API

Les agents d'IA doivent souvent accéder à des API tierces (Salesforce, Slack, Google Workspace) pour le compte d'un utilisateur. La mise en coffre des tokens permet de stocker et de gérer en toute sécurité ces tokens d'accès OAuth, la méthode d'authentification la plus courante des API modernes, afin d'éliminer le risque d'exposition des tokens dans le code, les journaux ou les fichiers de configuration. Bien que le coffre-fort (vault) puisse également protéger d'autres types d'identifiants (comme les tokens d'accès personnels ou les clés API requises pour les systèmes hérités), les tokens OAuth devraient être votre modèle par défaut, car ils prennent en charge l'actualisation automatique, la définition granulaire des champs d'action et la révocation sécurisée.

Fonctionnalités indispensables aux entreprises :

- **Stockage dans un coffre-fort sécurisé** des tokens OAuth avec un chiffrement fort des données au repos
- **Gestion automatique du cycle de vie des tokens**, y compris l'actualisation proactive avant l'expiration afin d'éviter toute interruption de service
- **Récupération de tokens à la demande** afin de ne jamais exposer des identifiants au code applicatif
- **Prise en charge de divers types de tokens** dans différents modèles d'authentification
- **Accès correctement délimité** garantissant que les tokens n'incluent que les autorisations nécessaires
- **Modèles d'intégration** compatibles avec les derniers frameworks de développement de l'IA

Principe de sécurité : les tokens ne doivent jamais apparaître dans le code des agents, les journaux ou les fichiers de configuration. Le coffre-fort gère le cycle de vie complet des tokens de manière transparente, ce qui réduit considérablement le risque de vol ou d'exploitation des identifiants grâce à un système centralisé et auditable.

Protection des identifiants dans les flux de sortie des agents : en plus d'empêcher les identifiants d'apparaître dans le code et les journaux, les entreprises doivent s'assurer que les tokens et les secrets ne figurent jamais dans les réponses ou les résultats des agents. Lorsque les agents utilisent des identifiants mis en coffre pour accéder aux API, les données de réponse peuvent contenir des informations sensibles, mais les identifiants eux-mêmes ne doivent jamais être inclus dans les réponses et conversations des agents, les documents générés ou tout autre résultat qui puisse être consulté par les utilisateurs finaux. Implémentez des fonctions de filtrage et de validation des résultats afin de détecter toute exposition accidentelle d'identifiants avant que les réponses ne parviennent aux utilisateurs. C'est tout particulièrement important pour les agents qui génèrent des extraits de code, des exemples de configuration ou des guides de dépannage dans lesquels des identifiants pourraient être accidentellement inclus.

Exemple : un agent d'IA commercial doit accéder au compte Salesforce d'un client. Au lieu de stocker les identifiants Salesforce, l'agent demande un token stocké dans le coffre-fort. Ce dernier fournit un nouveau token, dont le champ d'application est correctement délimité, et l'actualise automatiquement si nécessaire. L'agent accomplit sa tâche sans aucun contact entre les identifiants et son code.

4. Sécurité des données : autorisation granulaire pour la RAG

Lorsque des agents d'IA utilisent la RAG (Retrieval Augmented Generation) pour répondre à des questions, ils doivent uniquement accéder aux données que l'utilisateur est autorisé à consulter. Sans autorisation appropriée, les systèmes RAG peuvent exposer accidentellement des informations sensibles, créant ainsi une vulnérabilité de sécurité critique.

Fonctionnalités indispensables aux entreprises :

- **Contrôle des accès basé sur les relations** définissant les autorisations entre utilisateurs et documents
- **Application de l'autorisation** au moment de la récupération des documents, avant que les données ne soient dans le contexte de l'agent
- **Modèles d'intégration pour les bases de données vectorielles** utilisées dans les architectures RAG
- **Autorisations granulaires** applicables au niveau d'un document ou d'une section
- **Évaluation des autorisations en temps réel** pendant le traitement des requêtes
- **Prise en charge de modèles d'autorisation complexes**, y compris des politiques hiérarchiques et basées sur des attributs

Fonctionnement du modèle :
Les documents sont stockés avec des plongements dans une base de données vectorielle. Un système d'autorisation gère les relations entre les utilisateurs et les documents. Lorsqu'un agent récupère un contexte, les filtres d'autorisation valident les autorisations avant que les documents ne parviennent au contexte de l'agent. Le grand modèle de langage (LLM) génère uniquement des réponses au moyen de données que l'utilisateur est autorisé à consulter.

Exemple : un agent d'IA financier aide les collaborateurs à analyser des rapports. Lorsqu'Alice se renseigne sur les résultats du troisième trimestre, la base de données vectorielle récupère les documents financiers pertinents. Avant de les transmettre au LLM, des filtres d'autorisation vérifient l'accès d'Alice. Alice ne voit que les rapports de sa division, et non les données financières de l'ensemble de l'entreprise, ce qui évite toute exposition de données non autorisées.

5. Sécurité des appels d'outils : autorisation « humain dans la boucle »

Souvent, les agents d'IA fonctionnent de manière autonome en arrière-plan, nécessitant plusieurs minutes, heures ou jours pour accomplir leurs tâches. Pour les actions critiques telles que l'approbation des achats, l'envoi de contrats ou l'octroi d'accès, les entreprises ont besoin d'une supervision humaine sans pour autant compromettre l'autonomie des agents pour les tâches de routine.

Fonctionnalités indispensables aux entreprises :

- Modèles d'autorisation asynchrone qui fonctionnent avec des workflows d'agents de longue durée
- Mécanismes de notification riches proposant un contexte transactionnel complet aux approbateurs
- Messages de liaison proposant des détails essentiels tels que les montants, les destinataires et les actions prévues
- Workflows d'approbation accessibles à partir de terminaux mobiles et de l'e-mail sans nécessiter d'accès à un ordinateur de bureau
- Demandes d'autorisation à durée limitée qui expirent automatiquement si elles ne sont pas traitées
- Pistes d'audit complètes documentant toutes les décisions d'approbation et leur contexte

Fonctionnement du modèle :
Les développeurs identifient les actions de l'agent exigeant une approbation humaine. Lorsqu'un agent tente d'effectuer une opération protégée, une demande d'autorisation est envoyée à la personne appropriée avec le contexte complet de l'action prévue de l'agent. L'approbateur examine la demande et accorde ou refuse l'autorisation. En cas d'approbation, l'agent reçoit l'autorisation et poursuit son action ; en cas de refus, l'opération est bloquée et l'agent reçoit un message d'erreur.

Exemple : un agent d'IA d'achat identifie les licences logicielles nécessaires et se prépare à les acheter. Avant d'exécuter la transaction d'un montant de 5 000 euros, il envoie une notification au responsable des achats en indiquant le fournisseur, le montant et la justification. Le responsable examine la demande, l'approuve via un terminal mobile ou par e-mail, et l'agent effectue l'achat, ce qui permet de bénéficier à la fois de l'automatisation et d'un contrôle.

Sécuriser tous les agents à partir d'un point de contrôle unique

S'il est essentiel d'intégrer la sécurité dans les agents individuels au cours de leur développement, les entreprises ont également besoin d'une visibilité, d'un contrôle et d'une gouvernance centralisés pour l'ensemble de leurs agents d'IA. Cette solution permet aux entreprises de gérer des centaines ou des milliers d'agents opérant dans différents services, systèmes et cas d'usage.

1. Registre d'agents : établissement d'identités de premier ordre

Chaque agent d'IA doit être enregistré en tant qu'identité de premier ordre, avec une propriété et une responsabilité clairement définies. En l'absence d'un enregistrement en bonne et due forme, les entreprises manquent de visibilité et sont incapables de répondre aux questions élémentaires suivantes : À qui appartient cet agent ? Qu'est-il autorisé à faire ? Qui est responsable en cas de problème ?

Fonctionnalités indispensables aux entreprises :

- **Profils d'identité IA** pour chaque agent avec des identifiants uniques et persistants
- **Mappage de propriété** qui associe les agents aux équipes ou personnes responsables
- **Systèmes de métadonnées** documentant la finalité de l'agent, le cas d'usage et l'étape du cycle de vie
- **Intégration** aux structures organisationnelles telles que les systèmes RH et les hiérarchies de reporting
- **Monitoring des changements** qui enregistre les modifications apportées aux configurations et aux autorisations des agents
- **Mappage des dépendances** montrant les relations entre les agents et les systèmes auxquels ils accèdent

Pourquoi c'est important : selon le rapport « AI at Work 2025 », seuls 10 % des entreprises disposent d'une stratégie bien définie pour gérer les identités non humaines. L'enregistrement crée la couche d'identités fondamentale qui aide à mettre en place tous les autres contrôles de gouvernance et de sécurité. Sans cela, les agents restent invisibles pour les équipes sécurité, favorisant ainsi une Shadow AI qui échappe à leur contrôle.

Exemple : l'équipe sécurité découvre un agent d'IA accédant à Salesforce avec un compte de service. Elle l'enregistre dans le registre central des agents, en attribuant la propriété à l'équipe des ventes et en documentant sa finalité : « Génération automatisée d'offres pour les comptes d'entreprise ». Cela permet d'établir la responsabilité dès lors qu'en cas de comportement imprévu de l'agent, il existe un propriétaire à contacter pour résoudre le problème.

2. Contrôle des accès : autorisation basée sur des politiques

Les agents d'IA ont besoin d'une autorisation granulaire qui s'adapte au contexte et aux risques en temps réel. Les politiques de contrôle d'accès déterminent ce que chaque agent peut faire, quand et dans quelles conditions. Elles appliquent le principe du moindre privilège tout en permettant à l'agent de fonctionner correctement.

Fonctionnalités indispensables aux entreprises :

- **Moteurs de politiques** qui définissent les autorisations en fonction de l'identité de l'agent, du contexte opérationnel et des signaux de risque
- **Flux d'authentification basés sur des standards** qui prennent en charge les protocoles modernes
- **Gestion des accès aux API** afin de contrôler les modes d'interaction des agents avec les services protégés
- **Confiance interdomaine** dont les fonctionnalités permettent aux agents d'accéder en toute sécurité aux ressources au-delà des frontières organisationnelles et à d'autres domaines de confiance, tout en préservant le contexte de l'utilisateur
- **Évaluation dynamique des politiques** prenant en compte de multiples facteurs tels que le moment, l'emplacement et les modèles de comportement
- **Modèles d'intégration** pour la connexion à l'infrastructure IAM existante
- **Prise en charge d'architectures complexes multifournisseurs** couvrant différents domaines de sécurité

Modèle avancé – Connexions gérées :
Les entreprises doivent pouvoir définir les serveurs d'autorisation auxquels les agents peuvent accéder et les autorisations qu'ils peuvent demander. Cela inclut des cadres de politiques de sécurité qui spécifient les champs d'application accordés automatiquement, ceux nécessitant une approbation supplémentaire et ceux systématiquement interdits, afin que les agents puissent opérer dans des limites clairement définies. Pour les agents qui ont besoin d'accéder à des ressources dans différents domaines de confiance, la confiance interdomaine étend ces connexions gérées pour mettre en place une autorisation sécurisée à travers l'entreprise tout en maintenant un contrôle centralisé sur les politiques.

Exemple : lorsqu'un agent demande un accès, le système d'autorisation valide la demande par rapport aux politiques définies. Les autorisations courantes peuvent être accordées automatiquement, les opérations sensibles nécessitent une justification ou une approbation, et les actions dangereuses sont purement et simplement refusées — le tout étant appliqué par programmation, sans intervention manuelle. Lorsqu'un agent doit accéder à l'API d'une entreprise partenaire ou passer d'un système cloud à un système on-premise, XAA permet cet accès interdomaine tout en veillant à ce que le point de contrôle central continue de bénéficier d'une visibilité complète et d'appliquer les politiques.

3. Gestion du cycle de vie : gouvernance complète des agents

Le cycle de vie des agents d'IA est identique à celui des collaborateurs humains : onboarding, service actif, changement de rôle et, enfin, mise hors service. Une solution de gestion du cycle de vie automatise ces transitions tout en maintenant des contrôles de sécurité tout au long du cycle.

Fonctionnalités indispensables aux entreprises :

- **Workflows de provisioning automatisés** qui créent des identités d'agents avec les autorisations initiales appropriées
- **Modèles basés sur les rôles** qui normalisent les modèles d'accès pour les types d'agents les plus courants
- **Fonctionnalités d'accès en flux tendu (JIT)** pour des autorisations temporaires élevées associées à une expiration automatique
- **Processus d'évaluation planifiés** validant l'adéquation des autorisations au fil du temps
- **Workflows de déprovisioning** qui suppriment systématiquement tous les accès lorsque les agents sont mis hors service
- **Systèmes de gestion des changements** chargés de monitorer les modifications apportées aux autorisations et leur approbation

Cycle de vie complet
Les agents sont provisionnés avec des autorisations initiales qui varient selon leur finalité. Au cours de leur service, ils exécutent des tâches soumises à une validation continue de leurs droits d'accès. Au fur et à mesure que les besoins évoluent, les autorisations sont rectifiées au moyen de workflows d'approbation. Des évaluations régulières permettent de s'assurer que l'accès reste justifié. Lors de la mise hors service des agents, toutes les autorisations sont révoquées et les pistes d'audit sont conservées aux fins de conformité.

Exemple : un agent d'IA marketing est provisionné avec un accès à la plateforme e-mail et à la base de données clients. Au bout de six mois, une évaluation trimestrielle révèle que l'agent n'a plus besoin de l'accès à la base de données (le cas d'usage a changé). L'accès est automatiquement révoqué. À la fin de la campagne, l'agent est déprovisionné, toutes les autorisations sont supprimées, mais les journaux d'audit sont conservés aux fins de conformité.

4. Identifiants à privilèges : gestion sécurisée des secrets

Les agents d'IA ont souvent besoin d'identifiants à privilèges pour accéder aux systèmes : clés API, mots de passe de base de données, identifiants de comptes de service et certificats. Une mauvaise gestion des identifiants, des clés codées de façon irréversible et des secrets jamais renouvelés créent des risques de sécurité considérables que les attaquants n'hésitent pas à exploiter.

Fonctionnalités indispensables aux entreprises :

- **Stockage sécurisé dans un coffre-fort** avec un chiffrement puissant protégeant les identifiants au repos
- **Calendrier de rotation automatisée** pour actualiser les identifiants à intervalles réguliers
- **Modèles de provisioning JIT** qui réduit au minimum les périodes d'exposition des identifiants
- **Prise en charge de diverses méthodes d'authentification**, y compris des modèles basés sur des clés et des certificats
- **Gestion automatisée du cycle de vie des certificats**, y compris le renouvellement et la distribution
- **Isolation stricte** assurant que les secrets n'apparaissent jamais dans le code, les journaux ou les fichiers de configuration
- Modèles d'intégration pour les systèmes de gestion des secrets externes

Impact sur la sécurité
Les identifiants volés sont très prisés des acteurs malveillants. Par conséquent, l'élimination des secrets à longue durée de vie grâce à la rotation automatisée des identifiants peut réduire considérablement la surface d'attaque.

Exemple : un agent de pipeline de données utilise des identifiants de base de données renouvelés tous les 30 jours. Lors de la rotation, l'agent récupère automatiquement de nouveaux identifiants dans le coffre-fort, sans intervention humaine ni interruption de service. En termes de sécurité, le principal avantage réside dans le fait que les identifiants n'apparaissent jamais dans les journaux ou le code (isolation stricte). Toutefois, si les identifiants étaient exposés d'une manière ou d'une autre, la rotation automatisée limite la période d'exposition à un maximum de 30 jours au lieu d'une validité indéfinie, ce qui réduit considérablement la surface d'attaque.

5. Détection d'agents : découverte de la Shadow AI

Les entreprises ne peuvent pas sécuriser ce qu'elles ne voient pas. La détection des agents offre une visibilité sur tous les agents d'IA opérant dans l'environnement, y compris la Shadow AI — c'est-à-dire l'IA qui peut avoir été déployée par certains services sans l'approbation de l'équipe IT ni évaluation de la sécurité par celle-ci.

Fonctionnalités indispensables aux entreprises :

- **Mécanismes de découverte automatisés** permettant d'identifier les comptes non humains au sein des plateformes cloud et SaaS
- **Détection de la Shadow IT** qui permet d'identifier les agents déployés en dehors des processus de gouvernance formels
- **Inventaire complet des identifiants** indiquant quels agents ont accès à quels systèmes
- **Méthodes d'évaluation des risques** basées sur les autorisations, les tendances en matière d'activité et les niveaux d'exposition
- **Analyse des configurations** révélant les erreurs de configuration et les comptes associés à des privilèges excessifs
- **Modèles d'intégration** avec les plateformes cloud, les fournisseurs d'identité et les outils de sécurité

Approches de découverte
Une détection efficace des agents combine plusieurs techniques : l'analyse des modèles d'utilisation d'API pour identifier les comportements non humains ; la corrélation des journaux d'authentification pour détecter les activités des comptes de service ; l'analyse des ressources cloud pour identifier les déploiements d'agents ; la surveillance du trafic réseau pour identifier les communications entre agents et services ; et l'utilisation de l'analytique comportementale pour distinguer les agents des utilisateurs humains.

Pourquoi c'est d'une importance critique : le rapport « AI at Work 2025 » a révélé que 91 % des entreprises utilisent déjà des agents d'IA, mais que seuls 10 % d'entre elles possèdent des stratégies de gouvernance bien définies. Le fossé entre le déploiement et la gouvernance peut créer des agents de Shadow AI fonctionnant sans contrôle de sécurité, ce qui entraîne des risques dont les équipes sécurité ne soupçonnent même pas l'existence.

6. Déconnexion universelle des agents : réponse rapide aux menaces

Lorsqu'une menace (compromission d'identifiants, comportement anormal, violation de politique, etc.) est détectée, les entreprises doivent être en mesure de révoquer immédiatement l'accès de tous les agents dans l'ensemble des systèmes. La fonction de déconnexion universelle des agents permet d'activer cet « arrêt d'urgence » tout en conservant des pistes d'audit détaillées.

Fonctionnalités indispensables aux entreprises :

- **Mécanismes de révocation instantanée** pour tous les tokens et sessions d'agents actifs
- **Propagation intersystèmes** garantissant la déconnexion de toutes les applications intégrées
- **Rotation d'urgence des identifiants** pour remplacer les secrets potentiellement compromis
- **Workflows de confinement des menaces** empêchant toute autre action non autorisée
- **Préservation des preuves numériques** destinée à conserver des journaux d'audit complets à des fins d'investigation après l'incident
- **Intégration** avec les centres SOC et les systèmes de résolution des incidents

Détection des menaces grâce à l'analyse comportementale
L'efficacité de la déconnexion universelle dépend de la robustesse de la solution de détection des menaces. Celle-ci doit permettre d'établir des bases de référence du comportement normal des agents, détecter les anomalies telles qu'une période ou un volume inhabituel de demandes, calculer des scores de risque basés sur des modèles comportementaux, déclencher des réponses automatisées lorsque le risque dépasse les seuils définis et avertir les équipes sécurité en temps réel en cas d'activité suspecte.

Exemple : un agent d'IA du service clientèle accède normalement à 10-15 dossiers clients par jour. Subitement, il accède à 500 dossiers en 10 minutes, ce qui constitue une anomalie évidente. L'analyse comportementale détecte l'écart, déclenche automatiquement une déconnexion universelle révoquant tous les accès de l'agent et avertit l'équipe sécurité. L'investigation révèle le vol d'identifiants d'API. L'attaque est neutralisée en quelques minutes, empêchant l'exfiltration complète de la base de données.

Principes de fonctionnement

Ces deux solutions sont conçues pour sécuriser chaque agent dès la conception, mais aussi sécuriser tous les agents à partir d'un point de contrôle unique. Il ne s'agit pas de solutions distinctes nécessitant une intégration. Okta les a conçues comme des fonctionnalités complémentaires au sein de plateformes d'identité unifiées, conçues spécifiquement pour les agents d'IA.

Au cours du développement, la sécurité est intégrée à chaque agent dès la première ligne de code. L'authentification établit l'identité utilisateur. Ainsi, les agents connaissent l'identité des utilisateurs pour le compte desquels ils agissent et les limites de sécurité appropriées sont maintenues. La mise en coffre des tokens permet de gérer l'accès aux API sans jamais exposer les identifiants au code de l'agent, en assurant automatiquement l'actualisation des tokens et la gestion du cycle de vie. Les contrôles d'autorisation empêchent l'accès non autorisé aux données en implémentant des autorisations granulaires qui respectent les droits d'accès au niveau utilisateur. Les approbations « humain dans la boucle » assurent la supervision des actions critiques, ce qui permet aux agents de fonctionner de manière autonome tout en gardant un contrôle sur les opérations sensibles.

Dans la pratique :

- **Authentification universelle avec Universal Login —** Facilite l'authentification des utilisateurs au travers des fournisseurs de réseaux sociaux, des workflows sans mot de passe et du MFA pour les agents d'IA.
- **Accès sécurisé aux API avec Token Vault —** Stockez et gérez les tokens OAuth pour les API tierces, en actualisant automatiquement les identifiants sans qu'ils ne soient jamais exposés au code des agents.
- **Autorisation granulaire avec Auth0 FGA —** Implémentez un contrôle d'accès basé sur les relations pour les systèmes RAG de sorte que les agents récupèrent uniquement les documents que les utilisateurs ont le droit de consulter.
- **Autorisation « humain dans la boucle » avec l'autorisation asynchrone —** Exigez une approbation via mobile et e-mail des actions critiques des agents en utilisant CIBA (Client-Initiated Backchannel Authentication) avec RAR (Rich Authorization Request.

Pour les équipes IT et sécurité de l'environnement de production, le point de contrôle offre une supervision continue de l'ensemble des agents. Les fonctionnalités de découverte permettent de localiser tous les agents opérant dans l'environnement, y compris la Shadow IT, ce qui donne aux équipes sécurité une visibilité complète sur le paysage de l'IA de leur entreprise. L'enregistrement crée des identités de premier ordre avec une propriété clairement définie. Vous avez ainsi l'assurance que chaque agent est associé à une personne ou à une équipe responsable de ses actions.

La gouvernance des accès applique de façon dynamique des politiques basées sur le principe du moindre privilège et adapte donc les autorisations en fonction du contexte et des risques en temps réel. La détection des menaces identifie les anomalies et y répond grâce à l'analytique comportementale, ce qui permet de confiner automatiquement les attaques avant qu'elles n'aboutissent.

Dans la pratique :

- **Registre des agents dans Universal Directory —** Enregistrez chaque agent en tant qu'identité de premier ordre avec une visibilité sur la propriété, la responsabilité et les métadonnées.
- **Détection d'agents avec Identity Security Posture Management —** Découvrez les agents d'IA gérés et non gérés (Shadow AI) dans les différents environnements afin d'éliminer les zones sans visibilité.
- **Contrôle des accès et gestion du cycle de vie avec Okta Identity Governance —** Définissez, évaluez et certifiez les autorisations des agents grâce à une gouvernance du cycle de vie basée sur des politiques.
- **Mise en coffre des identifiants à privilèges avec Okta Privileged Access —** Sécurisez et renouvelez les identifiants des agents qui nécessitent des autorisations élevées afin de réduire la surface d'attaque.
- **Universal Logout pour les agents —** Révoquez instantanément les sessions, les tokens et les identifiants des agents dans l'ensemble des systèmes en cas de détection d'un risque ou d'une compromission.

La connexion entre ces solutions crée un framework de sécurité complet. Les agents conçus selon des pratiques de développement sécurisées sont automatiquement détectés par le point de contrôle, ce qui offre une visibilité immédiate sans nécessiter d'intégration supplémentaire. Le point de contrôle applique des politiques relatives à l'authentification et à l'accès des agents aux ressources, étendant les contrôles de sécurité implémentés dans l'environnement de développement à la gouvernance déployée dans l'environnement d'exécution. L'analytique comportementale surveille les actions des agents autorisées par les fonctionnalités de développement sécurisé, en détectant les cas où les agents s'écartent des modèles comportementaux attendus. Les workflows de gouvernance s'appliquent à la fois aux identités des agents et aux ressources auxquelles ils accèdent, ce qui permet de garantir une application cohérente des politiques dans l'ensemble de l'écosystème.

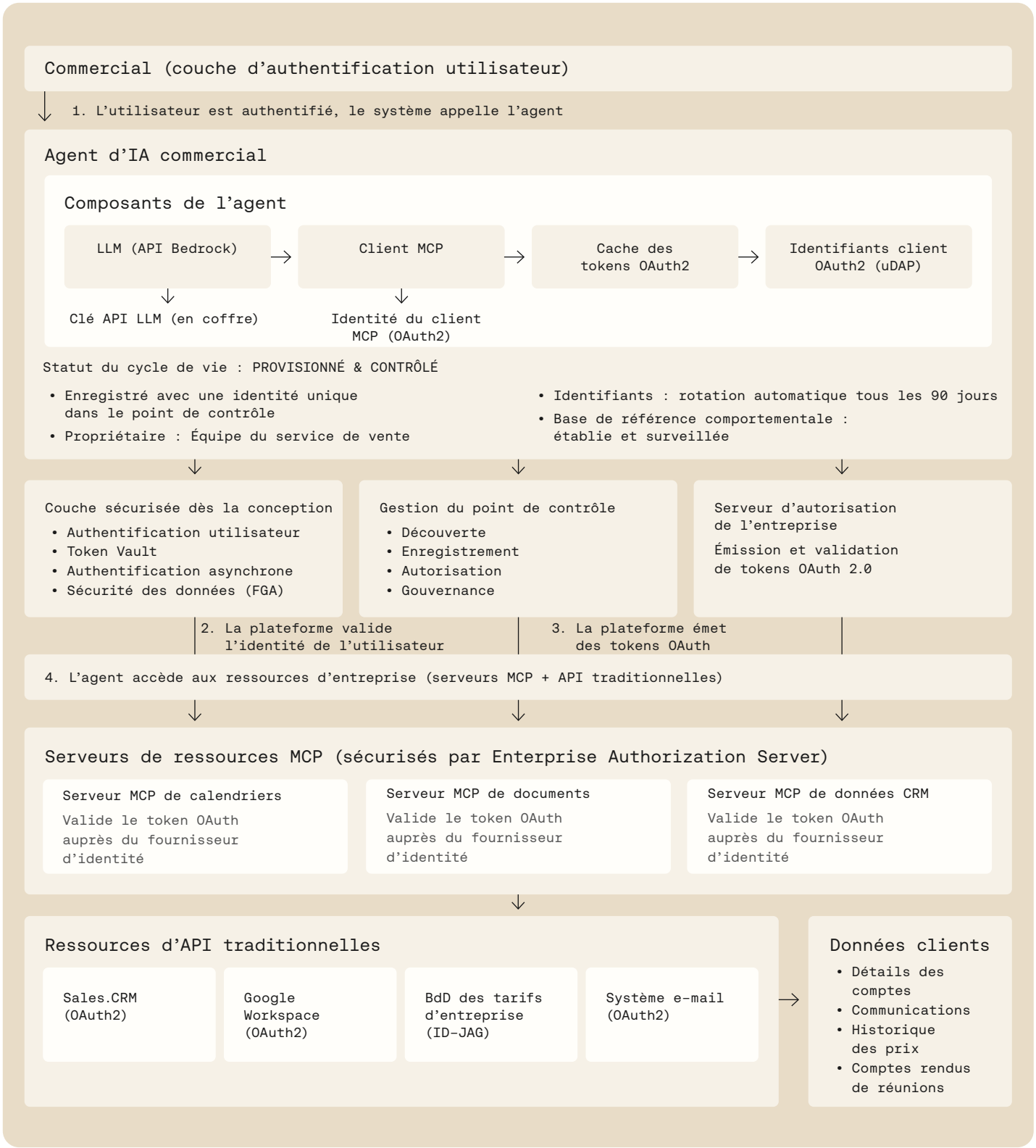
Les entreprises ne doivent pas nécessairement choisir entre un développement sécurisé et la gestion du cycle de vie : elles doivent implémenter les deux dans le cadre d'une stratégie cohérente. L'architecture de référence suivante illustre cette approche unifiée dans un scénario d'entreprise réel.

Architecture de référence : la plateforme unifiée en action

Pour illustrer comment la plateforme unifiée sécurise les agents d'IA tant au niveau du développement que de la gestion de leur cycle de vie, prenons l'exemple d'un agent commercial d'entreprise qui aide les commerciaux en automatisant la recherche de clients, en générant des offres, en accédant aux données CRM et en planifiant des réunions de suivi. Cet agent illustre le fonctionnement conjoint des deux solutions : un développement sécurisé dès la conception combiné à un contrôle centralisé du cycle de vie.

L'agent utilise le protocole MCP (Model Context Protocol) pour accéder en toute sécurité au contexte à partir de plusieurs sources tout en respectant les autorisations des utilisateurs, l'échange de tokens (ID-JAG) pour la confiance interdomaine ainsi qu'une gouvernance globale tout au long de son cycle de vie.

Présentation de l'architecture



La plateforme prend en charge la sécurisation des implémentations MCP (Model Context Protocol). Les serveurs MCP fonctionnent comme des serveurs de ressources OAuth 2.0, et délèguent l'authentification et l'autorisation au serveur d'autorisation de l'entreprise.

Modèle de sécurité MCP
L'agent (client MCP) est enregistré en tant que client OAuth 2.0 auprès du serveur d'autorisation de l'entreprise et reçoit des identifiants client (id_client et secret_client ou identifiants basés sur un certificat). Avant d'accéder à une ressource MCP, l'agent obtient un token d'accès avec les champs d'application appropriés (par exemple, mcp:crm:read, mcp:docs:read, mcp:calendar:read). Lorsque l'agent demande une ressource telle que crm://contacts/acme-corp, le serveur MCP valide le token d'accès auprès du serveur d'autorisation, en vérifiant la validité de la signature, l'expiration, le profil cible et les champs d'application requis avant de fournir la ressource.

Ainsi, les développeurs de serveurs MCP n'ont pas besoin de développer une logique d'authentification personnalisée. Ils valident les tokens OAuth émis par la plateforme en utilisant la validation de tokens OAuth 2.0 standard. La plateforme gère l'authentification des utilisateurs, le cycle de vie des tokens, la gestion des champs d'application et les contrôles d'autorisation granulaire, garantissant une application cohérente des politiques de sécurité sur tous les serveurs MCP ainsi que des pistes d'audit complètes dans le journal système.

Cette architecture de référence illustre comment la plateforme sécurise la récupération du contexte MCP tout en fournissant une solution de gouvernance du cycle de vie pour les agents d'IA eux-mêmes.

Flux détaillé — Génération d'offres assistée par un agent avec MCP

Phase 1

Découverte et enregistrement (point de contrôle – détection/ provisioning)
Étape 1.1 : Détection de la Shadow AI

- L'équipe de vente a déployé un prototype d'agent sans l'approbation du service IT.
- Okta découvre que l'agent accède à Salesforce avec des identifiants de compte de service.
- L'équipe sécurité reçoit une alerte concernant un agent d'IA non géré.
- Score de risque : ÉLEVÉ (accès à privilèges, absence de propriété, absence de gouvernance).

Étape 1.2 : Enregistrement de l'agent

- L'équipe sécurité enregistre l'agent en tant qu'identité de première classe dans Okta.
- Un profil d'agent est créé avec l'identifiant unique : `sales-agent-prod-001`
- Un propriétaire est désigné : Équipe Sales Operations (John Smith, VP Sales Operations).
- La finalité est documentée : « Génération d'offres et recherche de clients automatisées ».
- Les identifiants sont transférés dans un coffre-fort sécurisé avec une politique de rotation de 90 jours.

Résultat : l'agent passe de la Shadow IT à une identité gérée avec une responsabilité clairement définie.

Phase 2

Authentification de l'utilisateur (couche sécurisée dès la conception)

Étape 2.1 : Authentification du commercial

- Sarah (commerciale) se connecte au portail des ventes à 9h.
- Auth0 Universal Login présente les options d'authentification.
- Sarah s'authentifie à l'aide de Google SSO (authentification sociale).
- Auth0 valide les identifiants auprès du fournisseur d'identité Google.
- Un token d'identification est émis avec le profil de Sarah et les revendications d'authentification.

Étape 2.2 : Liaison du contexte à l'agent

- Un agent reçoit le contexte utilisateur authentifié d'Auth0.
- Le token d'identification contient les revendications OIDC standard :

```
{
  "iss": "https://acmecorp.auth0.com/",
  "sub": "google-oauth2|108204567890123456789",
  "aud": "sales-agent-client-id",
  "exp": 1730480000,
  "iat": 1730477400,
  "name": "Sarah Johnson",
  "email": "sarah.johnson@acmecorp.com",
  "email_verified": true
}
```

Remarque : des revendications personnalisées supplémentaires telles que `role` ou `territory` peuvent être ajoutées à l'aide des fonctions Actions d'Auth0.

- L'agent connaît maintenant l'IDENTITÉ de l'utilisateur et peut agir en son nom.

Phase 3

Récupération du contexte via MCP (sécurité des données + autorisation)

Étape 3.1 : Requête utilisateur

Sarah demande de « créer une offre pour Acme Corp sur la base de ses achats précédents et de ses besoins actuels ».

Étape 3.2 : Découverte du contexte MCP

L'agent utilise le client MCP pour découvrir les sources de contexte disponibles :
Les serveurs MCP exposent des ressources structurées via les URI de ressources :

```
crm://contacts/acme-corp
docs://proposals/templates
calendar://availability/sales-team
pricing://enterprise-tier
```

Il s'agit d'une **récupération contextuelle structurée via MCP**, et non d'une recherche sémantique sur les plongements (RAG). MCP fournit un accès direct à des ressources spécifiques sur la base de schémas définis et d'URI de ressources. L'agent demande des ressources spécifiques par nom/chemin, et les serveurs MCP renvoient des données structurées.

Étape 3.3 : Vérification des autorisations via Auth0 FGA

L'agent interroge le service FGA (Fine-Grained Authorization) pour déterminer les ressources auxquelles Sarah peut accéder :

- FGA évalue les tuples de relations pour chaque ressource MCP :
- ✓ `user:sarah` a un accès `read` (lecture) à `crm://contacts/acme-corp`
- ✓ `user:sarah` a un accès `read` (lecture) à `docs://proposals/templates`
- ✗ `user:sarah` n'a PAS accès à `pricing://executive-discounts`
- Seules les ressources autorisées sont incluses dans la récupération du contexte.
- Il est ainsi possible d'éviter les fuites de données en appliquant l'accès sur le principe du moindre privilège.
- Chaque demande de ressource MCP est validée par rapport aux autorisations de Sarah avant que cette ressources soit récupérée.

Cela correspond au point « Sécurité des données (FGA) » sous « Couche sécurisée dès la conception » dans la présentation de l'architecture.

Étape 3.4 : Récupération des tokens dans Auth0 Token Vault

L'agent demande des tokens OAuth2 à Token Vault pour accéder à des systèmes externes :

- L'agent indique avoir besoin d'accéder à Salesforce CRM pour les données du compte d'Acme Corp.
- Token Vault valide la demande par rapport aux intégrations autorisées de l'agent.
- Token Vault renvoie un token d'accès valide et correctement délimité pour Salesforce.
- Le champ d'application du token est limité à l'accès aux données clients en lecture seule.
- L'agent utilise le token pour récupérer les données CRM via l'API Salesforce.

Token Vault offre les fonctionnalités suivantes :

- Stockage sécurisé des identifiants (aucun token codé de façon irréversible)
- Actualisation automatique des tokens à leur expiration
- Journal d'audit de tous les accès aux tokens
- Tokens à champ d'application limité (sur le principe du moindre privilège)

Étape 3.5 : Assemblage du contexte MCP

L'agent reconstitue un contexte complet à partir des sources autorisées :

CRM (via Token Vault → Salesforce) :

- Contact Acme Corp : Directrice technique Julie Martinez
- Achats précédents : 280 000 euros en services de formation à l'IA (2024)
- Contrat en cours : expiration du contrat de support en mars 2026

Bibliothèque de documents (via MCP) :

- Modèle d'offre d'entreprise (version approuvée)
- Catalogue de prix avec les tranches tarifaires actuelles
- Conditions générales d'utilisation

Calendrier (via MCP) :

- Disponibilités de Sarah pour les appels de suivi
- Capacité de l'équipe de vente pour l'assistance à l'implémentation

NON inclus (autorisation refusée) :

- Tarifs préférentiels (Sarah n'y a pas accès)
- Notes de négociation confidentielles issues d'autres transactions
- Données sur la structure des coûts internes

L'agent dispose maintenant d'un contexte complet et autorisé pour générer l'offre.

Phase 4

Autorisation interdomaine avec ID-JAG (échange de tokens)

Étape 4.1 : Accès à la base de données des tarifs de l'entreprise

- L'agent doit accéder à la base de données de tarifs interne à l'adresse `pricing.acmecorp.internal`.
- Il s'agit d'un domaine d'autorisation distinct de la plateforme d'identité principale.
- Le système de tarification dispose de son propre serveur d'autorisation qui nécessite des tokens ID-JAG.

Étape 4.2 : Échange de tokens via ID-JAG

L'agent envoie le token d'identification au serveur d'autorisation pour l'échange de tokens :
Requête :

```
POST /oauth2/token
Host: acmecorp.okta.com

grant_type=urn:ietf:params:oauth:grant-type:token-exchange
&subject_token=<Token de l'ID de Sarah>
&subject_token_type=urn:ietf:params:oauth:token-type:id_token
&requested_token_type=urn:ietf:params:oauth:token-type:id-jag
&audience=https://pricing.acmecorp.internal
&scope=pricing:read
```

Déroulement du processus :

- L'agent échange le token d'identification de Sarah contre un token ID-JAG.
- L'ID-JAG (Identity Assertion JWT Authorization Grant) est un token signé de façon cryptographique.
- L'ID-JAG est adressé au serveur d'autorisation de la base de données de tarifs.
- Vous pouvez ainsi bénéficier d'une autorisation interdomaine tout en préservant le contexte utilisateur.

Étape 4.3 : Validation du serveur d'autorisation

Le serveur d'autorisation procède à plusieurs contrôles de validation :

- Validation du token d'identification : vérifie la signature du token d'identification et les revendications (relation de confiance préétablie).
- Vérification de la connexion gérée : valide la connexion gérée de l'agent au serveur d'autorisation de la base de données de tarifs.
- Une connexion gérée définit les champs d'application admis :
  - Champs d'application autorisés : `pricing:read`
  - **Champs d'application refusés :** `pricing:write`, `pricing:admin`

Émission du token ID-JAG : le serveur d'autorisation émet un token ID-JAG qui préserve le contexte utilisateur de Sarah.

Revendications du token ID-JAG :

```
{
 "iss": "https://acmecorp.authorization-server.com",
 "sub": "sarah.employee@acmecorp.com",
 "aud": "https://pricing.acmecorp.internal",
 "client_id": "sales-ai-agent",
 "jti": "9e43f81b64a33f20116179",
 "scope": "pricing:read",
 "exp": 1698583800,
 "iat": 1698580200,
 "auth_time": 1698580200,
 "amr": ["pwd", "mfa"]
}
```

Étape 4.4 : Accès aux ressources

L'agent présente le token ID-JAG au serveur d'autorisation de la base de données de tarifs :

- Le serveur d'autorisation de la base de données de tarifs valide la signature ID-JAG à l'aide des clés publiques publiées (JKWS) de la plateforme de gestion des identités.
- Le serveur d'autorisation de la base de données de tarifs vérifie ce qui suit :
  - La revendication `aud` **doit correspondre à sa propre URL d'émetteur.**
  - **La date d'expiration (`exp`) n'est pas dépassée.**
  - **Le champ d'application (`scope`) est dans les limites des autorisations prévues.**
  - L'émetteur (`iss`) est un fournisseur d'identité de confiance.
    - **Accès accordé :** l'agent récupère les données de tarification de l'entreprise avec des autorisations en lecture seule.
    - Cet agent dispose désormais d'un accès autorisé aux informations tarifaires pour la génération d'offres.

Fonctionnalités d'échange de tokens de la plateforme

La plateforme de gestion des identités prend en charge l'échange de tokens RFC 8693 pour les scénarios d'autorisation interdomaine. L'échange de tokens permet aux agents d'IA d'accéder aux ressources de différents serveurs d'autorisation tout en préservant le contexte utilisateur grâce à des tokens ID-JAG signés de façon cryptographique. Cette fonctionnalité est disponible dans toute la plateforme, tant pour les déploiements orientés développeurs que ceux axés sur l'entreprise.

Phase 5

Autorisation asynchrone (« humain dans la boucle »)

Étape 5.1 : Détermination par l'agent qu'une approbation est nécessaire

- L'agent reconnaît que l'envoi d'une offre de 450 000 euros nécessite l'approbation explicite de l'utilisateur.
- Un workflow d'autorisation asynchrone est alors déclenché.
- L'agent initie une demande d'autorisation asynchrone pour suspendre l'exécution dans l'attente de l'approbation.

Pourquoi l'autorisation asynchrone est nécessaire :

- Les offres dont la valeur est élevée dépassent l'autorité dont jouit l'agent autonome.
- La politique de l'entreprise exige une approbation humaine pour les offres d'une valeur supérieure à 100 000 euros.
- Cette mesure assure une prise de responsabilité pour les décisions critiques.
- L'agent ne peut pas exécuter des actions non autorisées.

Étape 5.2 : Demande d'autorisation CIBA

L'agent envoie une demande d'autorisation CIBA (Client-Initiated Backchannel Authentication) :

La demande inclut ce qui suit :

- **Identifiant utilisateur :** ID collaborateur de Sarah
- **Autorisations requises :** `email:send`, `drive:write`, `crm:update`
- **Contexte de l'action :** détails de l'offre à examiner par Sarah
- **Endpoint de rappel :** emplacement de remise du token après l'approbation

```
POST /bc-authorize
Host: acmecorp.authorization-server.com
scope=email:send drive:write crm:update
&login_hint=sarah.employe@acmecorp.com
&binding_message=Approbation d'offre :
Acme Corp - 450,000 €
&client_notification_token=
8d67dc78-7faa-4d41-aabd-67707b374255
```

CIBA offre les possibilités suivantes :

- Workflows d'approbation asynchrone (pas de blocage de l'agent)
- Authentification de l'utilisateur hors bande (notification push sur le terminal mobile)
- Contexte enrichi dans les demandes d'approbation (informations détaillées sur l'offre)
- Mécanisme de rappel sécurisé (remise du token après approbation)

Étape 5.3 : Notification push

Le serveur d'autorisation envoie une notification push à l'application mobile Guardian de Sarah :

Le message de notification enrichi s'affiche :

```
Approbation d'offre requise
Client : Acme Corp
Montant : 450 000 €
Produits : Enterprise AI Suite + Support
Destinataires : dirtech@acmecorp.com, dirfin@acmecorp.com
Action : envoyer l'offre par e-mail et enregistrer dans Drive
[Approuver] [Refuser]
```

Fonctionnalité de notification enrichie

- Contexte détaillé de l'action à approuver
- Nom du client, montant en devise et produits inclus
- Liste des destinataires pour la transparence
- Description claire de l'action
- Simple interface d'autorisation/refus

Remarque : il s'agit de la fonctionnalité de notification enrichie de Guardian, et NON de la spécification OAuth Rich Authorization Requests (RAR - RFC 9396). La notification fournit des informations contextuelles détaillées pour aider Sarah à prendre une décision en toute connaissance de cause.

Étape 5.4 : Approbation de l'utilisateur

Sarah examine et approuve la demande :

- Sarah passe en revue les détails sur son terminal mobile.
- Elle vérifie ce qui suit :
  - Il s'agit du bon client (Acme Corp).
  - Le montant est exact (450 000 €).
  - Les destinataires sont corrects (direction technique et direction financière).
  - Les produits correspondent aux besoins du client.
- Elle approuve la demande en appuyant sur le bouton « Approuver ».

Génération et remise de tokens

- Le serveur d'autorisation génère un token d'accès à champ d'application limité avec les autorisations approuvées.
- Le token n'inclut que les autorisations approuvées par Sarah :
  `email:send` — autorisation d'envoi de l'email contenant l'offre
  `drive:write` — autorisation d'enregistrement de l'offre dans Google Drive
  `crm:update` — autorisation de journalisation de l'activité dans Salesforce
  - Le token est envoyé à l'agent via le terminal de rappel CIBA.
  - L'agent reçoit le token et poursuit l'exécution en utilisant le token récemment émis.

Avantages en matière de sécurité

- Approbation explicite de l'utilisateur requise pour les actions sensibles
- Token à durée limitée (expire une fois l'action terminée)
- Tokens à champ d'application limité (uniquement les autorisations approuvées)
- Piste d'audit complète (approbateur, date/heure et activités approuvées)

Flux du token d'autorisation asynchrone (CIBA) – Détails techniques

Le flux CIBA (Client-Initiated Backchannel Authentication) permet de bénéficier d'une approbation asynchrone des actions de l'agent d'IA par l'utilisateur.

Lorsque Sarah approuve la demande sur son terminal mobile :

- **Validation de l'approbation :** le serveur d'autorisation valide la décision d'approbation provenant du terminal authentifié de Sarah.
- **Génération du token :** le serveur d'autorisation génère un nouveau token d'accès limité à l'action approuvée.
- **Limitation des autorisations :** le token n'inclut que les autorisations approuvées par Sarah (par exemple, `email:send`, `drive:write`).
- **Remise sécurisée :** le token est remis à l'agent par l'intermédiaire d'un terminal de rappel sécurisé spécifié dans la demande CIBA initiale.
- **Exécution par l'agent :** l'agent reçoit le token et exécute l'action approuvée.

Il est ainsi possible de garantir une autorisation « humain dans la boucle » pour les opérations sensibles de l'agent d'IA, l'approbation explicite de l'utilisateur étant requise avant toute action de l'agent.

Avantages du flux CIBA :

- **Non bloquant :** l'agent n'a pas besoin de maintenir la connexion pendant qu'il attend l'approbation.
- **Convivial :** Sarah procède à l'approbation depuis son terminal mobile, et non du chatbot.
- **Sécurisé :** les tokens sont délivrés par rappel sécurisé, et non via le navigateur de l'utilisateur.
- **Auditable :** il existe un enregistrement complet de la demande d'approbation, de la décision de l'utilisateur et de l'émission du token.
- **Flexible :** il prend en charge plusieurs mécanismes d'approbation (notification push, SMS, e-mail).

Phase 6

Exécution multi-système (mise en coffre de tokens)

Étape 6.1 : Sauvegarde dans Google Drive

- L'agent récupère un token Google Workspace d'Auth0 Token Vault.
- Le champ d'application du token est limité à l'accès de Sarah à Google Drive.
- L'agent charge l'offre vers : `Sales/Proposals/2025/Acme-Corp-Q1.pdf`
- Autorisations sur le fichier : membres de l'équipe de Sarah + son responsable

Étape 6.2 : Envoi d'un e-mail

- L'agent récupère un token d'API Gmail de Token Vault.
- L'agent rédige un e-mail à partir du compte de Sarah.
  - À : Direction technique et financière d'Acme Corp
  - Corps : Lettre d'accompagnement de l'offre (générée par un LLM)
  - Pièce jointe : PDF de l'offre enregistrée sur Google Drive
- L'e-mail est envoyé avec la signature de Sarah.

Étape 6.3 : Planification d'un suivi

- L'agent récupère un token de Google Agenda de Token Vault.
- L'agent vérifie les disponibilités de Sarah pour les 2 prochaines semaines.
- L'agent propose des plages de réunion aux contacts d'Acme Corp.
- L'agent ajoute un événement au calendrier : « Examen de l'offre Acme Corp – 30 min. ».

Avantages de Token Vault

- L'agent ne voit jamais les tokens OAuth.
- Tous les tokens sont automatiquement actualisés avant leur expiration.
- Les identifiants ne sont jamais stockés dans le code de l'agent ou dans les journaux.
- Isolation complète entre les identifiants Auth0 et l'identité de l'agent Okta.

Phase 7

Gouvernance et surveillance (point de contrôle – gouvernance)

Étape 7.1 : Piste d'audit

Toutes les activités sont enregistrées en détail dans le journal système de la plateforme :

- Événements d'authentification de l'agent
- Opérations d'échange de tokens et attributions de champ d'application
- Tentatives d'accès aux ressources dans l'ensemble des systèmes
- Décisions d'autorisation (approbation/refus)
- Événements de délégation d'utilisateurs
- Appels d'API effectués au nom des utilisateurs
- Horodatages et métadonnées contextuelles destinés à l'investigation numérique

Étape 7.2 : Évaluation trimestrielle des accès

- Déclenchement du workflow de gouvernance : certification d'accès T1 2025
- E-mail à John Smith (propriétaire de l'agent) : « Évaluation de l'accès de sales-agent-prod-001 »
- L'évaluation des accès montre que l'agent possède les accès suivants :
  - Accès au CRM Salesforce
  - Accès à Google Workspace
  - Accès à la base de données des tarifs de l'entreprise
  - Accès au système e-mail
  - Accès à l'agenda
- John confirme que tous les accès sont toujours nécessaires.
- La certification est enregistrée dans un journal d'audit Okta.

Étape 7.3 : Certification des accès

- L'évaluation trimestrielle des accès montre que l'agent dispose des autorisations appropriées pour son rôle.
- Tous les accès sont justifiés et approuvés par le propriétaire de l'agent.
- La certification est enregistrée dans le journal d'audit à des fins de conformité.

Phase 8

Détection des menaces (point de contrôle – surveillance)

Étape 8.1 : Détection d'anomalie

- **Jour 45 :** l'agent accède soudainement à 500 dossiers clients en l'espace de 10 minutes.
- L'analyse comportementale détecte un écart par rapport à la base de référence.
- Le score de risque augmente : NORMAL → ÉLEVÉ
- Type d'anomalie : « Volume inhabituel d'accès aux données »

Étape 8.2 : Réponse automatisée

- La plateforme bloque automatiquement l'accès de l'agent à Salesforce.
- **Révocation globale des tokens :** tous les tokens actifs sont immédiatement invalidés dans tous les systèmes.
- L'équipe sécurité reçoit une alerte en temps réel.
- Sarah (l'utilisatrice) reçoit une notification : « Agent commercial temporairement suspendu ».
- Le verrouillage complet des accès empêche toute activité non autorisée.

Étape 8.3 : Investigation et correction

- L'équipe sécurité examine le journal système pour déterminer l'ampleur de l'incident.
- La cause profonde est identifiée et résolue.

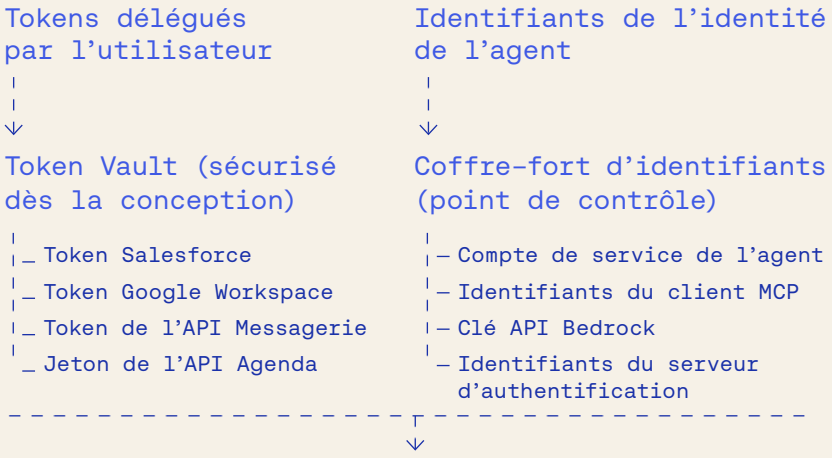
Métrique clé
Attaque détectée et bloquée grâce à la détection et à la réponse aux menaces automatisées.

Points d'intégration : interconnexion des composants de la plateforme unifiée

1. Flux d'authentification

```
Utilisateur ---> Authentification universelle
---> Token d'identification ---> Validation
du token par la plateforme ---> Validation du
token par la plateforme ---> Contexte combiné
(utilisateur + agent) ---> Accès accordé avec
contexte complet
```

2. Cycle de vie des tokens



3. Couches d'autorisation

Couche 1 : Sécurité des données (autorisation granulaire)

Autorisations au niveau document pour le RAG
« L'utilisatrice Sarah peut-elle voir le document offre-acme-2024 ? »

Couche 2 : Mise en coffre (vaulting) des tokens

Autorisations au niveau API pour les outils SaaS de Sarah
« Le token de l'utilisatrice Sarah peut-il accéder à Salesforce ? »

Couche 3 : Contrôle des accès (point de contrôle)

Autorisations au niveau système pour les ressources d'entreprise
« L'agent sales-agent-prod-001 peut-il accéder à la base de données des tarifs ? »

Couche 4 : échange de jetons (ID-JAG)

Confiance interdomaine avec le contexte utilisateur
« Sarah (via un agent) peut-elle accéder à la base de données de tarifs on-premise ? »

Résultat : défense en profondeur avec plusieurs points de contrôle des autorisations

4. Modèle d'autorisation MCP

1. L'agent demande le contexte via le client MCP.
2. Le serveur MCP reçoit la demande.
3. Le serveur MCP vérifie si l'agent dispose d'un token.

Validation des tokens ---> Service d'autorisation Auth0

- Validation de l'identité de l'agent
- Vérification des champs d'application des tokens
- Vérification des autorisations

4. Le serveur MCP vérifie si l'utilisateur est autorisé à accéder aux données.

Vérification des autorisations ---> Autorisation granulaire

- Évaluation des tuples de relations
- Renvoi des documents autorisés uniquement

5. Le serveur MCP renvoie le contexte autorisé à l'agent.

Démonstration des grands principes d'architecture

1. Séparation des responsabilités

- Auth0 gère l'authentification de l'utilisateur et l'accès délégué par l'utilisateur.
- Okta se charge de l'identité de l'agent et de la gestion de son cycle de vie.
- MCP gère la récupération du contexte standardisé.
- Chaque système fait ce qu'il fait le mieux (et ce pourquoi il est conçu).

2. Défense en profondeur

- Les différentes couches d'autorisation permettent d'éviter un point de défaillance unique.
- FGA filtre les documents, Token Vault verrouille les API, Okta contrôle les systèmes.
- Même en cas de contournement d'une couche, les autres continuent d'assurer la protection.

3. Principe du moindre privilège

- Les agents reçoivent les autorisations minimales nécessaires pour chaque tâche.
- Les tokens sont limités à des API et à des actions spécifiques.
- L'accès est limité dans le temps avec expiration automatique.
- Le provisioning JIT (Just in Time) limite les privilèges permanents.

4. Préservation du contexte de l'utilisateur

- L'échange de tokens ID-JAG permet de maintenir l'identité de l'utilisateur entre différentes zones de confiance.
- Les actions de l'agent sont toujours rattachées à un utilisateur spécifique.
- Les décisions d'autorisation prennent en compte le contexte utilisateur, et pas seulement l'identité de l'agent.
- Les pistes d'audit précisent à la fois l'identité de l'agent et celle de l'utilisateur au nom de qui il agit.

5. Surveillance continue

- Les bases de référence comportementales permettent de détecter les anomalies.
- La réponse aux menaces en temps réel bloque les attaques.
- La journalisation complète des événements permet d'effectuer des investigations numériques.
- La correction automatisée réduit le temps de réponse.

Comparaison d'architectures : approche traditionnelle contre plateforme unifiée

Fonctionnalités	Sans plateforme unifiée	Avec une plateforme unifiée
Découverte des agents	Feuilles de calcul manuelles, pas de détection de la Shadow AI	Découverte automatisée, visibilité complète
Gestion des identifiants	Clés codées de manière irréversible, sans aucune rotation	Mise en coffre et rotation
Authentification utilisateur	Code d'authentification personnalisé, stockage de mots de passe	Authentification universelle, authentification sociale unique
Accès aux API	Tokens stockés dans les fichiers de configuration	Mise en coffre des tokens avec actualisation automatique
Accès interdomaine	Authentification distincte pour chaque système	Échange de tokens ID-JAG avec contexte utilisateur
Approbation humaine	Interrogations personnalisées, aucune prise en charge mobile	CIBA avec notification mobile ou e-mail
Autorisations associées aux documents	Vérifications au niveau application, incohérentes	Autorisation granulaire avec contrôle basé sur les relations
Évaluations des accès	Feuilles de calcul manuelles et trimestrielle	Workflows de certification automatisés
Détection des menaces	Analyse des journaux après un incident	Analytique comportementale en temps réel
Piste d'audit	Dispersée dans plusieurs systèmes	Journal système unifié
Résolution des incidents	Investigation et correction manuelles	Blocage et révocation des tokens automatisés
Autorisation MCP	Logique d'authentification personnalisée dans chaque serveur MCP	OAuth2 standardisé avec validation de la plateforme

Cette architecture illustre comment une plateforme de gestion des identités unifiée assure la sécurité complète des agents d'IA, proposant à la fois un développement sécurisé (authentification, gestion des tokens, autorisation, supervision humaine) et la gestion du cycle de vie à l'échelle de l'entreprise (découverte, enregistrement, gouvernance, détection des menaces). Qui plus est, MCP propose une récupération de contexte standardisée qui respecte les contrôles d'identité et d'autorisation tout au long du cycle de vie.

Conclusion : une plateforme unifiée pour une sécurité complète des agents d'IA

La révolution des agents d'IA est en marche. 91 % des entreprises utilisent déjà des agents d'IA, et ce nombre ne fera qu'augmenter. Malheureusement, l'adoption a été plus rapide que la sécurisation et la gouvernance, créant des risques importants qui menacent d'amoindrir la valeur ajoutée des agents d'IA.

Le défi consiste à résoudre simultanément deux problèmes interconnectés :

Sécuriser chaque agent dès la conception au cours du développement, en veillant à ce que l'authentification, l'autorisation, la gestion des tokens et les contrôles d'accès aux données soient intégrés dès le départ.

Sécuriser tous les agents à partir d'un point de contrôle unique tout au long de leur cycle de vie en assurant la découverte, le provisioning, la gouvernance et la détection des menaces pour l'ensemble des agents.

Les entreprises qui prennent en compte ces deux dimensions seront plus à même de bénéficier d'une sécurité robuste et complète :

- **Pratiques de développement sécurisées** avec authentification, mise en coffre (vaulting) des tokens, autorisation et supervision humaine
- **Visibilité complète** sur tous les agents d'IA, y compris la Shadow AI
- **Gestion correcte du cycle de vie**, de l'enregistrement au déprovisioning
- **Gouvernance complète** avec évaluation des accès et certification
- **Détection des menaces en temps réel** avec analytique comportementale et réponse automatisée
- **Respect des réglementations** grâce à des pistes d'audit complètes et à l'application de politiques

En savoir plus

Cette approche de plateforme unifiée permet de pallier une lacune critique identifiée dans le rapport « AI at Work 2025 » : 85 % des dirigeants affirment que l'IAM est essentiel à l'adoption de l'IA, mais seuls 10 % possèdent des stratégies bien définies pour gérer les identités non humaines.

N'attendez pas d'être confronté à une brèche ou à un manquement à la conformité pour adopter une sécurité adaptée aux agents d'IA. Commencez dès aujourd'hui à sécuriser les agents lors du développement et à mettre en place un contrôle centralisé de l'ensemble de vos agents.

La plateforme unifiée capable d'y parvenir est disponible dès maintenant au travers des solutions d'identité d'Okta. Des entreprises du monde entier utilisent déjà cette approche pour déployer des agents d'IA en toute sécurité et à grande échelle, en combinant Auth0 for GenAI pour un développement sécurisé et Okta Identity Platform pour la gestion du cycle de vie dans toute l'entreprise.

Création d'agents d'IA sécurisés
Documentation et guides de démarrage rapide pour le développement d'agents sécurisés

En savoir plus sur l'approche d'Okta pour sécuriser le cycle de vie des agents d'IA
Découvrez les fonctionnalités de gouvernance et de point de contrôle proposées par Okta pour gérer les agents d'IA à grande échelle.

À propos d'Okta
Okta, Inc. — The World's Identity Company™ — protège les identités afin que chacun puisse utiliser n'importe quelle technologie en toute sécurité. Nos solutions d'identité client et collaborateur permettent aux entreprises et aux développeurs d'utiliser toute la puissance de la gestion de l'identité pour améliorer la sécurité, l'efficacité et la réussite — tout en protégeant leurs utilisateurs, collaborateurs et partenaires. Découvrez pourquoi les plus grandes marques au monde font confiance à Okta pour l'authentification, l'autorisation, et bien plus encore sur le site okta.com/fr.

Livre blanc

Sécurisation complète des agents d'IA, de leur développement à leur utilisation dans l'entreprise

okta
The World's Identity Company™

Okta France
Tour Europlaza
20 avenue André Prothin
92400 Courbevoie – France
+33 01 85 64 08 80