eBook

L’ère des agents d’IA : repenser l’identité, la confiance et le contrôle

Sommaire

La nouvelle frontière de l'identité

L'IAM traditionnel a été pensé pour les humains. L'IA fonctionne à la vitesse des machines.

À mesure que les entreprises remplacent les chatbots statiques par des agents autonomes, la nature profonde de l'identité change. Ces agents ne se contentent pas de fournir des informations : ils exécutent des tâches en plusieurs étapes dans des environnements complexes. Cela peut engendrer d'importantes lacunes en termes de visibilité, avec pour corollaire l'exécution d'actions non autorisées avant qu'une équipe sécurité ne puisse les détecter.

Cet eBook est une compilation des grandes lignes de notre série de sept articles de fond consacrés à la sécurisation de l'IA en entreprise. En plus de résumer ces discussions essentielles, ce guide offre des informations stratégiques sur les changements architecturaux à entreprendre pour assurer une autonomie sécurisée des agents.

91 %

des entreprises utilisent déjà des agents d'IA en production.

10 %

disposent d'une stratégie bien définie pour gérer les identités non humaines.

+50 %

citent la gouvernance et la conformité des agents d'IA comme une préoccupation majeure.

Source

L'IAM à la vitesse des agents

Les agents d'IA fonctionnent à la vitesse des machines et sont capables d'exécuter jusqu'à 6 000 opérations par minute. À une telle vélocité, les modèles d'autorisation traditionnels centrés sur l'humain ne sont plus d'actualité. La sécurité doit se détacher des approbations manuelles basées sur le consentement pour privilégier une application automatisée à l'exécution, afin d'empêcher que des agents « livrés à eux-mêmes » ne causent des pertes de données désastreuses en quelques secondes.

Recommandations

- **Règles reposant sur des politiques :** implémentez des règles automatisées qui répondent à la vitesse des agents, afin que la sécurité suive le rythme d'exécution des machines.
- **Identifiants éphémères :** utilisez des identifiants qui expirent en quelques minutes plutôt que de persister indéfiniment, afin de réduire considérablement la fenêtre d'opportunité pour les acteurs malveillants.
- **Accès basé sur les relations :** mettez en place des contrôles d'autorisation qui s'exécutent en quelques millisecondes à l'aide d'un contrôle des accès granulaire basé sur les relations.
- **Évaluation continue :** plutôt que d'accorder des autorisations une seule fois sans examen ultérieur, réévaluez chaque opération effectuée par un agent.

Votre pile de sécurité actuelle peut-elle intercepter un agent sans supervision exécutant 100 commandes par seconde avant qu'il ne supprime votre base de données ? Si vous vous reposez encore sur un processus d'approbation humain pour les actions des agents, non seulement vous freinez l'innovation, mais vous ignorez une brèche qui évolue à la vitesse des machines.

À lire : Sécurité de l'IA : l'IAM à la vitesse des agents

Quand l'autorisation survit à l'intention

On parle de « dérive des autorisations » lorsque des clés numériques émises pour une tâche spécifique restent actives pendant des mois une fois la tâche terminée. Dans le monde des agents d'IA, ces tokens dormants sont une véritable bombe à retardement qui permet aux cybercriminels de pirater des connexions légitimes entre applications SaaS sans avoir à craquer de mot de passe.

Recommandations

- **Identité déléguée durable :** chaque agent d'IA doit avoir sa propre identité, distincte des utilisateurs, qui est gouvernée, auditable et traçable.
- **Autorisation constamment renouvelable :** ajustez automatiquement l'accès à mesure que la tâche, l'utilisateur ou l'environnement change afin d'assurer l'élimination des privilèges permanents.
- **Déprovisioning instantané dans tous les systèmes :** appliquez la révocation par signaux partagés afin qu'un identifiant révoqué dans une application soit instantanément invalidé dans tous les systèmes.
- **Validation de l'intention en temps réel :** recontrôlez toutes les actions par rapport aux politiques actuelles au moment où elles sont exécutées, et pas seulement lors de l'émission initiale d'identifiants.

Votre entreprise dispose-t-elle d'un processus formel et automatisé permettant de révoquer les identifiants des agents dès qu'une tâche est terminée ? Si vos tokens survivent à leur intention, vous laissez la porte grande ouverte aux brèches silencieuses.

À lire : Sécurité des agents d'IA : quand l'autorisation survit à l'intention

Votre plan d'action en matière de sécurité et de conformité de l'IA

Votre entreprise est-elle prête à prendre le contrôle de son écosystème d'agents d'IA ? Consultez notre Checklist de conformité de la sécurité des identités de l'IA pour guider vos prochaines étapes.

Télécharger la checklist

Confiance lors de l'accès des agents à différents systèmes

Lorsque des agents d'IA franchissent les frontières organisationnelles pour accéder à des systèmes indépendants, ils perdent souvent la « mémoire » des contraintes qui leur sont imposées. Comme la plupart des fournisseurs d'identité valident les tokens de manière isolée, la compromission d'un seul domaine de confiance peut avoir des effets non maîtrisés sur des centaines d'autres domaines.

Recommandations

- **Délégation vérifiable :** implémentez une preuve cryptographique de délégation qui distingue explicitement les identités humaines de celles des agents lorsqu'elles se déplacent entre des systèmes.
- **Contraintes portables :** faites en sorte que les contraintes de sécurité (par exemple « lecture seule ») suivent le token sur les domaines de confiance afin qu'elles ne soient pas supprimées lors d'un transfert.
- **Révocation coordonnée :** adoptez une signalisation fédérée des risques (comme IPSIE) afin que les alertes de sécurité en temps réel soient partagées entre les différents fournisseurs de services.
- **Contrôles au moment de la récupération :** utilisez une autorisation granulaire pour valider l'accès au moment exact où un agent appelle une API, quel que soit le domaine d'origine de la requête.

Lorsqu'un agent accède au système d'un partenaire, qui se porte garant de sa sécurité ? Si votre modèle de confiance est statique et décentralisé, vous n'avez aucune défense commune contre le détournement de tokens entre domaines.

À lire : Sécurité de l'IA : confiance lors de l'accès des agents à différents systèmes

Sécuriser la délégation des agents d'IA

La délégation récursive, où les agents engendrent des sous-agents pour gérer des tâches spécialisées, crée une imbrication de risques de sécurité à la manière de poupées russes. Sans traçabilité stricte ni contrôle sur le champ d'application, une seule invite malveillante peut déclencher le détournement d'une session d'agent, permettant à un sous-agent d'exécuter des actions non autorisées de manière invisible.

Recommandations

- **Vérification hors bande :** utilisez des notifications push ou des interfaces utilisateurs distinctes pour contourner le canal de discussion principal de l'agent.
- **Ancrage contextuel :** ancrez chaque session d'agent à sa tâche d'origine et signalez toute dérive sémantique si le comportement de l'agent commence à s'éloigner de l'objectif prévu.
- **Identité et capacités vérifiées :** appliquez une architecture Zero Trust où les agents doivent présenter des identifiants cryptographiques pour vérifier leur identité et des autorisations spécifiques avant d'interagir avec toute ressource système.
- **Visibilité de l'utilisateur :** réduisez le risque d'instructions dissimulées en offrant une transparence totale et en présentant à l'utilisateur en temps réel tous les appels d'outils, le raisonnement contextuel et les journaux d'exécution.

Pouvez-vous prouver cryptographiquement la traçabilité de chaque action entreprise par un sous-agent après trois étapes dans un workflow ? En l'absence de chaînes de délégation vérifiables, votre écosystème multi-agent est une cible de choix pour les déplacements latéraux.

À lire : Sécurité des agents : chaîne de délégation

Combler les failles du flux d'autorisation

Les agents d'IA récupèrent souvent des données en utilisant les autorisations de haut niveau d'un dirigeant, mais diffusent ces informations dans des espaces de travail partagés tels que Slack ou Teams. Cette faille de sécurité dans le flux d'autorisation permet la diffusion involontaire de données sensibles, telles que la rémunération des dirigeants ou les documents du conseil d'administration, à des destinataires non autorisés.

Recommandations

- **Autorisation tenant compte du public cible :** calculez « l'intersection des autorisations » en temps réel afin qu'un agent ne puisse extraire que les données que toutes les personnes de l'espace de travail actuel sont autorisées à consulter.
- **Récupération limitée :** plutôt qu'un filtrage des données après leur extraction, limitez le champ d'application des identifiants de l'agent avant la récupération, de sorte que les fichiers sensibles ne soient jamais consultés.
- **Accès basé sur les relations :** abandonnez les rôles statiques au profit d'un modèle basé sur les relations, qui comprend qui se trouve sur quel canal et quelles sont ses autorisations actuelles.
- **Synchronisation continue des politiques :** intégrez la gouvernance des identités au moteur d'autorisations afin que le graphe des autorisations reste exact à mesure que des utilisateurs rejoignent ou quittent des espaces de travail partagés.

Pour chaque agent d'IA déployé dans un espace de travail partagé, pouvez-vous démontrer que ses résultats sont limités au « plus petit dénominateur commun » des autorisations dans l'espace de travail ? Si vos agents ignorent le public cible, ils représentent votre plus grand risque interne de fuite de données.

À lire : Les failles du flux d'autorisation des agents d'IA

Sécurité cyber physique

À mesure que les agents d'IA sont intégrés à des secteurs physiques, comme la santé et l'industrie, les erreurs d'autorisation n'entraînent pas seulement des fuites de données, mais aussi des risques de sécurité. La sécurité doit désormais empêcher les agents numériques de causer des dommages physiques.

Recommandations

- **Traçabilité de bout en bout :** implémentez Cross-App Access (XAA) avec des tokens de délégation pour pouvoir attribuer chaque action automatisée à un agent précis et à l'utilisateur d'origine.
- **Gestion des identifiants à la demande :** utilisez la mise en coffre (vaulting) des tokens pour remplacer les identifiants statiques de longue durée par des tokens limités de courte durée, qui ne sont récupérés qu'au moment de l'exécution.
- **Vérification « humain dans la boucle » :** utilisez l'authentification CIBA (Client-Initiated Backchannel Authentication) pour exiger une approbation humaine explicite chaque fois qu'un agent tente d'effectuer des opérations sortant de l'enveloppe opérationnelle.
- **Application des politiques en temps réel :** adoptez une autorisation granulaire qui évalue les contraintes de sécurité et les limites opérationnelles au moment précis de la décision, plutôt que de vous reposer sur des rôles statiques.

Votre équipe peut-elle clairement définir le cadre des autorisations et les identifiants actifs pour chaque agent accédant à des systèmes critiques ? Si vous ne pouvez pas exprimer clairement ce qu'un agent est autorisé à faire et pourquoi, ce manque de visibilité est votre principale surface d'attaque.

À lire : Agents d'IA : sécurité de l'IAM cyber physique

Identité et autorisation unifiées pour une autonomie responsable

Les agents d'IA fonctionnant de manière autonome, les modèles d'identité et d'autorisation traditionnels sont dépassés. Les systèmes conçus pour accorder un accès à des utilisateurs humains ne peuvent pas gérer de manière fiable la délégation, les actions indirectes ni la prise de décisions basée sur des machines. L'instauration d'une autonomie responsable nécessite de traiter l'identité et l'autorisation comme une couche de contrôle unifiée qui définit ce que les agents sont autorisés à faire, et où se terminent ces limites.

Recommandations

- **Représentez les agents comme des identités distinctes :** considérez les agents d'IA comme des identités propres, plutôt que comme des extensions d'utilisateurs ou d'applications.
- **Unifiez les décisions d'identité et d'autorisation :** évitez les contrôles fragmentés en appliquant des décisions d'accès cohérentes pour toutes les actions des agents.
- **Tenez explicitement compte de la délégation :** concevez des modèles d'autorisation qui reconnaissent quand des agents agissent pour le compte d'utilisateurs, de systèmes ou d'autres agents.
- **Réduisez les accès trop étendus :** limitez les autorisations permanentes qui permettent à des agents d'opérer au-delà du périmètre prévu.
- **Utilisez l'identité pour définir des limites :** appliquez l'identité et l'autorisation comme des mécanismes permettant de contraindre le comportement des agents, et pas seulement de l'observer.

Quand l'identité et l'autorisation fonctionnent comme un système unifié, l'autonomie responsable devient applicable lorsque les agents agissent et délèguent.

À lire : L'identité et l'autorisation comme socle d'une autonomie responsable

Comment Okta peut vous aider

Okta contribue à donner vie à l'écosystème de sécurité des identités en tant que plan de contrôle unifié en corrigeant le décalage entre l'intention humaine et l'exécution à la vitesse des machines. Avec une couche centralisée pour appliquer les politiques en temps réel, chaque action d'agent est gérée, traçable et sécurisée dans n'importe quel domaine de confiance. En plaçant l'identité au cœur de l'infrastructure, l'écosystème permet aux entreprises de mettre à l'échelle les agents d'IA avec l'assurance que la sécurité et l'autorisation sont intégrées à chaque workflow automatisé.