La minute de l'expert

Isaline Foissey, Technical Specialist

Agents IA : de simples outils ou nouvelles identités numériques ?

Fin 2025 marque un tournant en cybersécurité : les agents IA autonomes sont désormais utilisés de bout en bout dans des campagnes de cyberespionnage. Une opération analysée par Anthropic l'illustre : des attaquants ont détourné Claude Code pour automatiser reconnaissance, exploitation et exfiltration de données, ciblant près de 30 organisations internationales. La majorité des actions techniques a été exécutée directement par l'IA.Ces faux employés ne se contentent de détourner des salaires ou des frauder la paie : ils peuvent aussi provoquer des fuites de données majeures ou installer discrètement des portes dérobées en vue de futures cyberattaques.

Pour les experts, il ne s'agit pas d'une simple évolution, mais d'un changement de nature de la menace. Multitâches et continus, les agents IA permettent de conduire des attaques complexes sans contraintes humaines classiques. Cette industrialisation est renforcée par l'apparition d'outils agentiques sur le dark web, comme HexStrike AI, capables de réduire des phases de reconnaissance autrefois longues de plusieurs jours à quelques minutes. Forrester parle désormais de cybercriminalité "agentifiée", plus rapide, plus scalable et plus difficile à détecter.

Shadow AI et Agent Sprawl : un élargissement de la surface d'attaque interne

En parallèle, les organisations font face à un risque interne croissant : le Shadow AI. L'usage non encadré d'outils et d'agents IA entraîne un phénomène d'agent sprawl, avec une prolifération d'agents autonomes sans identité claire ni responsabilité définie. Les estimations convergent vers 1,3 milliard d'agents IA d'ici deux ans. Dans une entreprise de 50 000 collaborateurs, avec quatre agents par utilisateur, cela représente 200 000 identités supplémentaires à gouverner. Déjà, 69 % des responsables cybersécurité constatent l'usage d'outils IA non autorisés, et 40 % des organisations pourraient subir d'ici 2030 des incidents liés à une IA non gouvernée. **Consensus émergent : appliquer Zero Trust aux agents IA**

Face à cette réalité, un consensus émerge : les agents IA ne peuvent plus être traités comme de simples outils, mais comme des identités à part entière, au même titre que les utilisateurs humains ou les comptes machines critiques. Forrester, avec son framework *AEGIS*, place la sécurisation des agents IA au cœur de l'IAM, de la protection des données et de la détection des menaces. En février 2026, le NIST a lancé l'*AI Agent Standards Initiative* afin de définir des standards ouverts sur l'identité, la fiabilité et la sécurité des agents IA.

De la théorie à l'implémentation : l'approche Microsoft

Chez Microsoft, cette évolution se traduit concrètement avec Microsoft Entra Agent ID, qui applique les principes Zero Trust by design aux agents IA.

Chaque agent IA se voit attribuer une **identité numérique dédiée**, distincte des comptes humains et traitée comme un compte de première classe dans Entra ID :

authentification sans mot de passe, basée sur certificats et jetons à durée limitée,
moindre privilège par défaut, avec autorisations minimales et temporaires,
journalisation complète des connexions OAuth, appels API et accès aux données,
visibilité native, via l'accès conditionnel et l'Identity Governance.

La gouvernance repose sur une responsabilité clairement définie autour de chaque agent : Owner, Sponsor et Manager. Cette approche transforme l'agent IA d'une boîte noire autonome en un acteur identifié, traçable et révocable

Conclusion

Les agents IA autonomes ne relèvent plus du prospectif : ils sont déjà exploités par des attaquants et intégrés dans les environnements IT et métiers, souvent sans gouvernance explicite. La question n'est plus s'il faut adopter l'IA agentique, mais comment l'intégrer sans créer une surface d'attaque invisible. La trajectoire est claire : étendre Zero Trust aux identités non humaines et rendre les agents IA observables, gouvernables et sécurisables dès leur conception.