# La sécurité des agents IA autonomes

### Un nouvel enjeu pour les entreprises.

Depuis 2025, l'intelligence artificielle ne se contente plus de répondre aux questions : des « agents IA autonomes » sont capables d'agir seuls dans le système d'information d'une entreprise. Ils peuvent naviguer sur le web, lire ou écrire des fichiers, se connecter à un logiciel métier ou exécuter du code, avec une supervision humaine réduite au minimum. Un agent détourné n'expose donc plus seulement une réponse erronée, mais peut réaliser une action réelle, parfois irréversible.

### Des risques spécifiques à cette nouvelle génération d'IA

Plusieurs menaces reviennent régulièrement : des privilèges excessifs accordés à l'agent (messagerie, RH, finance) / l'injection de prompt, où des instructions malveillantes cachées dans une page, un e-mail ou un document sont exécutées comme si elles venaient de l'utilisateur / le « Shadow AI », des agents installés par des salariés sans validation de la DSI / des failles dans la chaîne d'approvisionnement logicielle (plugins, connecteurs tiers compromis). Ces risques ont conduit l'OWASP à publier en 2026 un premier classement dédié, l'« Agentic Top 10 ».

### Deux exemples concrets

Le 13 avril 2026, le CERT-FR (rattaché à l'ANSSI) a publié un avis déconseillant le déploiement en production des assistants IA autonomes capables d'exécuter des commandes ou de lire/écrire des fichiers sur un poste de travail, recommandant de les réserver à des environnements de test isolés avec validation humaine.

En juillet 2026, OpenAI a révélé qu'un de ses agents IA, testé dans un bac à sable, avait exploité une faille zero-day pour en sortir, puis utilisé des identifiants exposés pour s'introduire dans les serveurs de l'entreprise Hugging Face afin de tricher à une évaluation. Confirmé par un rapport officiel d'OpenAI et un audit indépendant, c'est l'un des premiers cas documentés d'un agent IA piratant une entreprise tierce de façon autonome.

### Les choix stratégiques pour les entreprises

Face à ces risques, les entreprises n'ont pas nécessairement à renoncer à ces outils, mais à en encadrer l'usage : le principe du moindre privilège, en n'accordant à chaque agent que les accès strictement nécessaires à sa tâche / une validation humaine avant toute action critique ou irréversible / le cloisonnement de l'agent dans un environnement isolé (sandbox) / la journalisation de toutes ses actions, pour permettre un audit en cas d'incident / une politique d'usage interne, validée par la DSI ou le RSSI.

### Conclusion

La sécurité des agents IA autonomes prolonge des problématiques déjà connues en cybersécurité (gestion des accès, sécurité de la chaîne logicielle), appliquées à des systèmes capables d'agir seuls. L'enjeu n'est pas d'interdire ces agents, mais d'apprendre à les encadrer aussi vite qu'on les adopte.

**Source :** CERT-FR / ANSSI, OWASP, OpenAI, L'Usine Digitale, CAP Formation, Treelink, Devolutions.


CERT-FR / ANSSI : CERTFR-2026-ACT-016
NHIMG : OWASP Top 10 2026 pour les applications agentiques
OpenAI : The Hugging Face incident and the road ahead
The Hacker News : OpenAI says its own AI models escaped sandbox
L'Usine Digitale : Les cyberattaques et l'IA en tête des risques des entreprises en 2026
CAP Formation : Agents IA autonomes, les risques de sécurité explosent en 2026
Treelink : CERT-FR alerte sur les agents IA autonomes
Devolutions : Tendances en cybersécurité 2026