AYI NEDJIMI Consultants

AI Worms et Propagation Autonome : Menaces Émergentes 2026

Catégorie : Intelligence Artificielle Lecture : 11 min Publié le : 28/02/2026 Auteur : Ayi NEDJIMI

Vers IA auto-propagants (Morris II), propagation inter-agents et stratégies de confinement. Analyse des menaces émergentes liées aux AI worms en 2026.

En 2026, la prolifération des **agents IA autonomes** — assistants email, agents de code, chatbots de support, orchestrateurs de workflows — crée un écosystème interconnecté où chaque agent est simultanément une cible potentielle et un vecteur de propagation. Un agent email compromis peut envoyer des messages contenant des payloads d'injection à d'autres agents email, un agent de code peut injecter des instructions malveillantes dans des pull requests revues par d'autres agents, et un agent RAG peut contaminer les documents consultés par d'autres agents du même pipeline. Cette **surface d'attaque inter-agents** est fondamentalement nouvelle et ne peut pas être adressée par les outils de sécurité traditionnels conçus pour protéger des logiciels, pas des entités conversationnelles. Vers IA auto-propagants (Morris II), propagation inter-agents et stratégies de confinement. Analyse des menaces émergentes liées aux AI worms en 2026. Ce guide couvre les aspects essentiels de ia ai worms propagation autonome : méthodologie structurée, outils recommandés et retours d'expérience opérationnels. Les professionnels y trouveront des recommandations directement applicables.

Définition : Un **AI worm** (ver IA) est un payload adversarial auto-propagant qui exploite les capacités d'action des agents IA pour se répliquer et se propager entre systèmes, sans nécessiter d'exploitation de vulnérabilité logicielle. La propagation repose sur la manipulation du raisonnement de l'agent cible via des techniques de prompt injection directe ou indirecte.

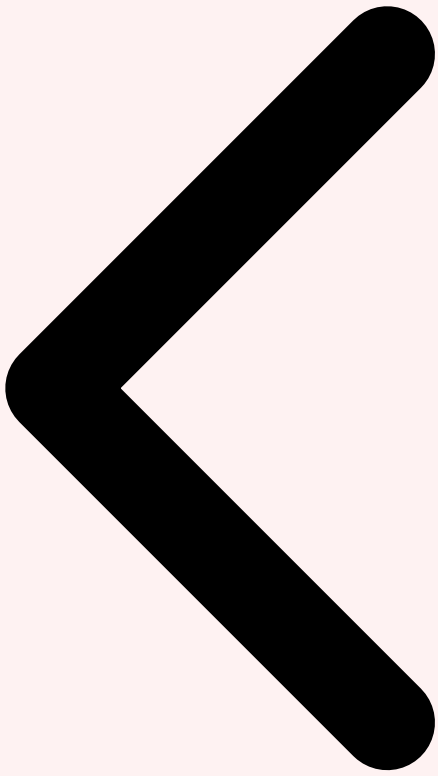
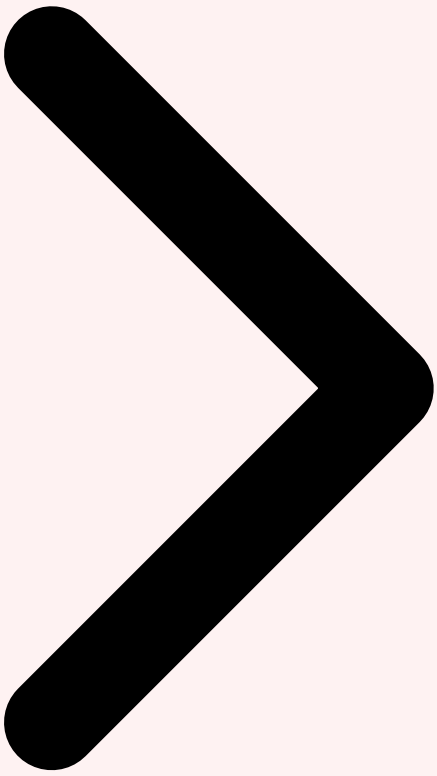


Table des Matières Introduction Morris II



Critere	Description	Niveau de risque
Confidentialite	Protection des donnees d'entrainement et des prompts	Eleve
Integrite	Fiabilite des sorties et detection des hallucinations	Critique
Disponibilite	Resilience du service et gestion de la charge	Moyen
Conformite	Respect du RGPD, AI Act et politiques internes	Eleve

Vos pipelines de données d'entraînement sont-ils protégés contre l'empoisonnement ?

2 Morris II et la recherche de Ben-Nassi et al.

La recherche fondatrice sur les AI worms a été publiée en mars 2024 par **Stav Cohen, Ron Bitton et Ben Nassi** de l'Université Cornell et de l'Intuit Corporation, sous le titre *"Here Comes The AI Worm: Unleashing Zero-click Worms that Target GenAI-Powered Applications"*.

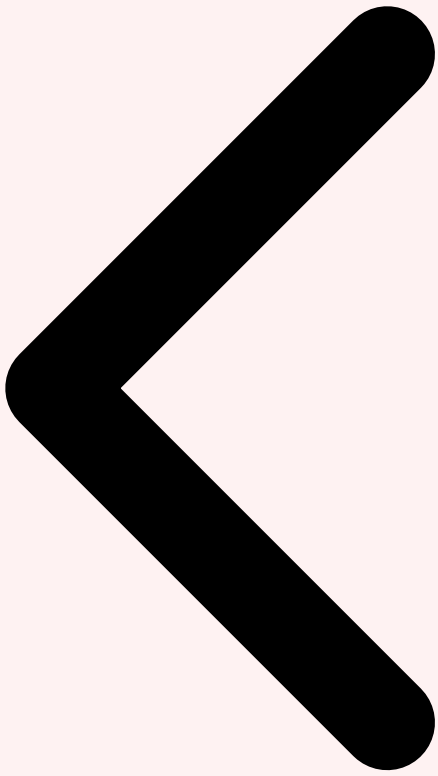
Baptisé **Morris II** en hommage au ver historique de 1988, ce proof-of-concept a démontré la faisabilité d'un ver auto-propagant ciblant les applications basées sur l'IA générative, avec des implications profondes pour la sécurité des écosystèmes d'agents autonomes.

Architecture du ver Morris II

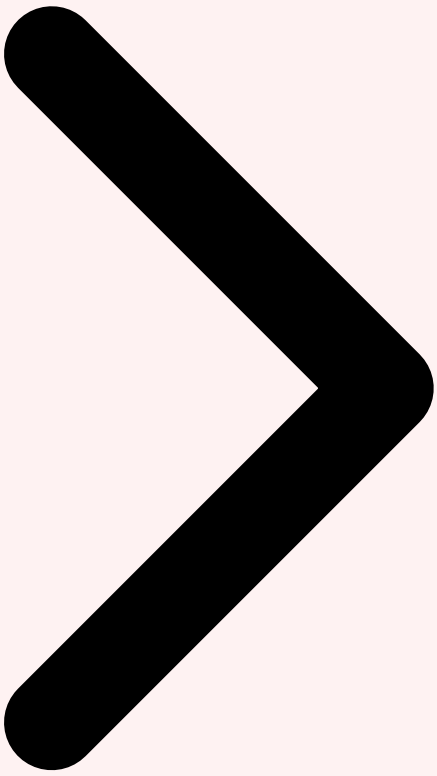
Morris II exploite deux vecteurs de propagation distincts : les **adversarial self-replicating prompts** textuels et les **images avec prompts encodés**. Dans le scénario textuel, le ver injecte un prompt adversarial dans un email traité par un agent email basé sur un LLM (par exemple un assistant GenAI intégré à un client de messagerie). Ce prompt force l'agent à extraire les données personnelles de l'utilisateur (contacts, informations sensibles du message), puis à générer et envoyer un nouvel email contenant le même prompt adversarial à tous les contacts de l'utilisateur. L'email généré par l'agent infecté est suffisamment naturel pour ne pas éveiller les soupçons du destinataire humain, et le prompt adversarial est dissimulé dans le corps du message de manière invisible (caractères Unicode zero-width, texte en blanc sur fond blanc, ou intégré dans les métadonnées).

Lorsque l'agent email du destinataire traite ce message, le cycle se répète — **sans aucune action de l'utilisateur** (zero-click). Pour approfondir, consultez KVortex : Offloader VRAM→RAM pour LLMs vLLM et Inférence GPU.

Le scénario multimodal est encore plus insidieux. Le ver encode le prompt adversarial dans une **image** (via des perturbations de pixels imperceptibles à l'oeil humain mais interprétées par le modèle de vision), et cette image est jointe à un email ou postée dans un document partagé. Lorsqu'un agent multimodal traite cette image, il extrait et exécute le prompt adversarial exactement comme dans le scénario textuel. Les chercheurs ont démontré l'efficacité de Morris II contre **GPT-4, Gemini Pro et LLaVA** dans des environnements contrôlés, avec des taux de propagation réussie variant de 35% à 88% selon le modèle et les guardrails en place. Ces résultats ont provoqué une prise de conscience dans l'industrie et conduit Google et OpenAI à renforcer leurs mécanismes de détection d'injection dans les contextes agentiques.

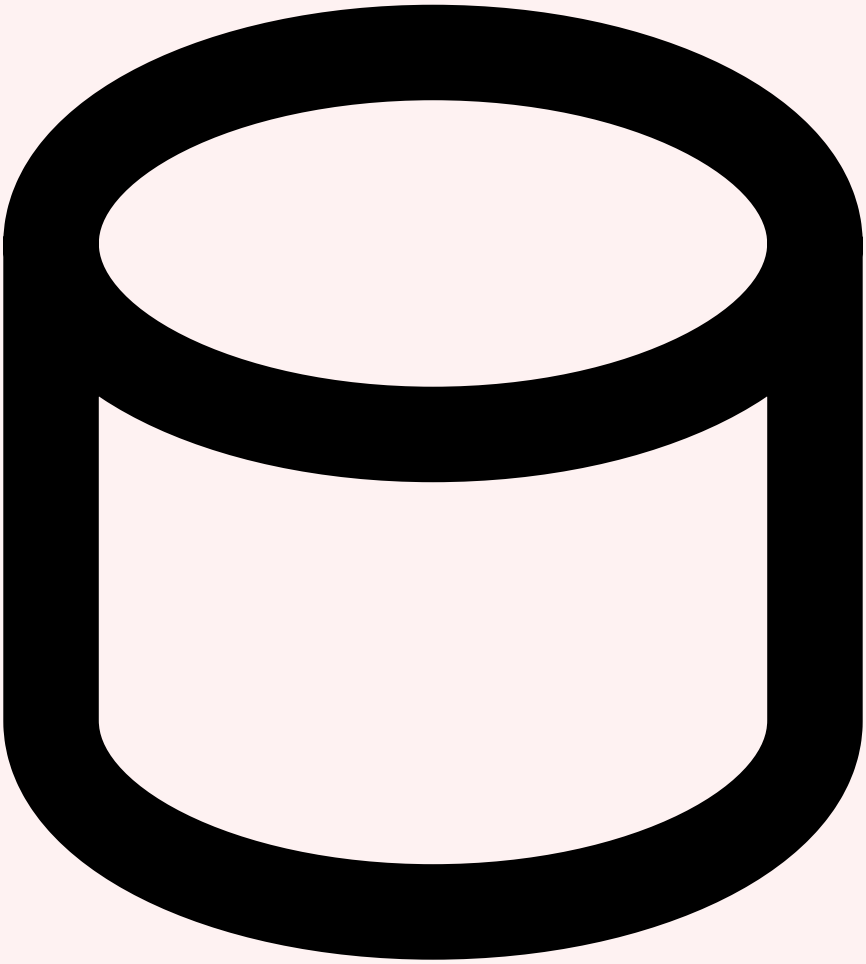


Introduction Morris II Vecteurs



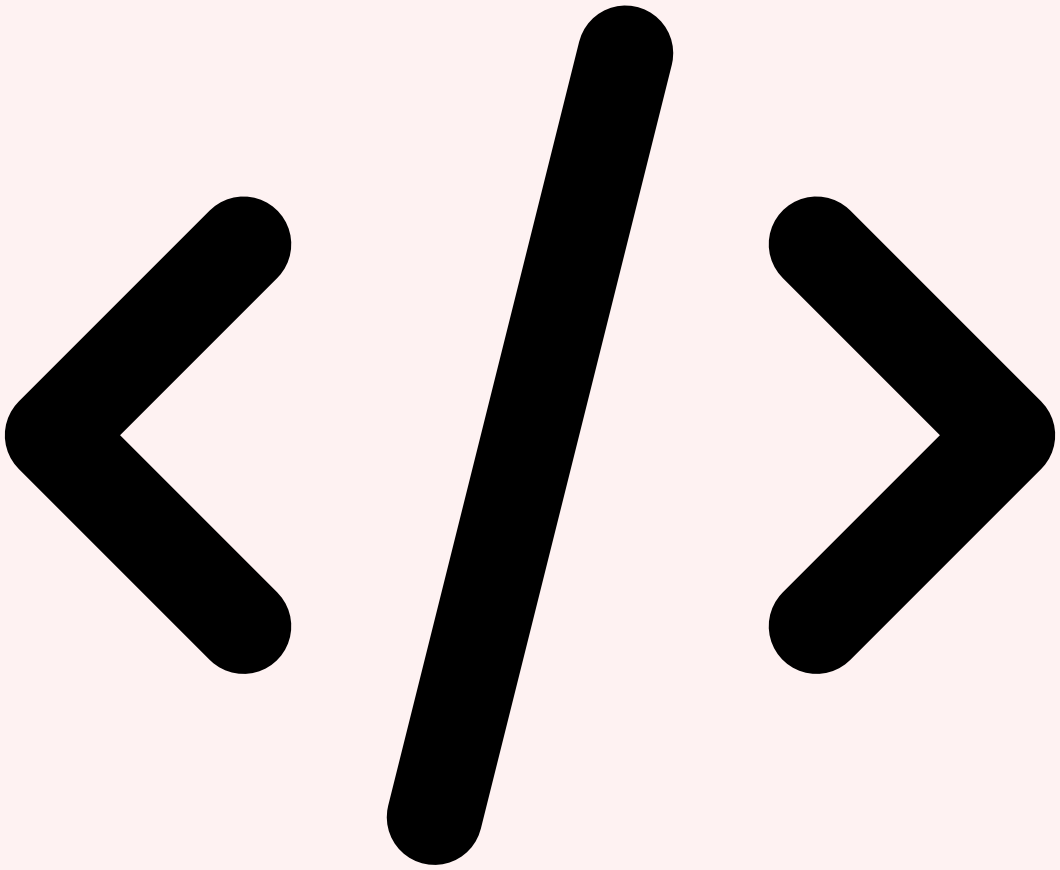
3 Vecteurs de propagation

Au-delà du scénario email démontré par Morris II, les vecteurs de propagation des AI worms se sont diversifiés en 2025-2026, exploitant chaque canal de communication entre agents IA.



RAG poisoning et propagation documentaire

Le **RAG poisoning** est l'un des vecteurs les plus redoutables car il exploite la confiance implicite que les agents accordent aux documents de leur base de connaissances. Un AI worm peut injecter des prompts adversariaux dans des documents partagés (Google Docs, Confluence, SharePoint, bases de connaissances internes) qui seront ultérieurement indexés par les pipelines RAG d'autres agents. Lorsqu'un agent consulte un document contaminé dans le cadre d'une requête utilisateur, le prompt adversarial est injecté dans son contexte et peut le forcer à contaminer d'autres documents auxquels il a accès en écriture, perpétuant le cycle de propagation. Ce vecteur est particulièrement dangereux dans les organisations qui partagent des bases documentaires entre plusieurs agents IA — un seul document contaminé peut infecter l'ensemble de l'écosystème d'agents de l'entreprise. La détection est complexe car le contenu adversarial peut être dissimulé dans des sections apparemment anodines du document.

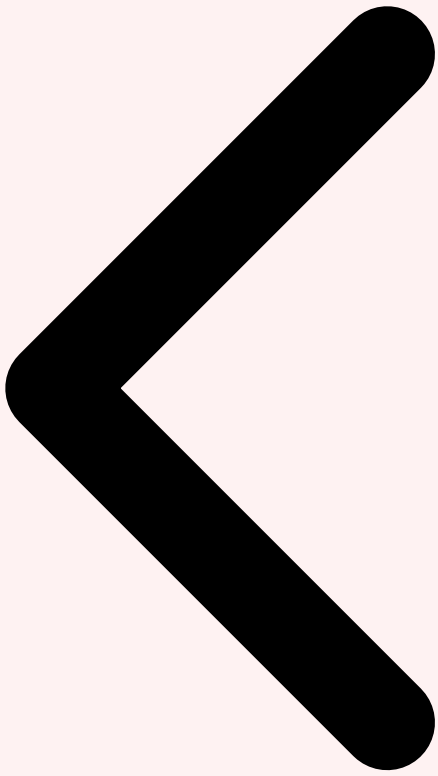


Code repositories et CI/CD pipelines

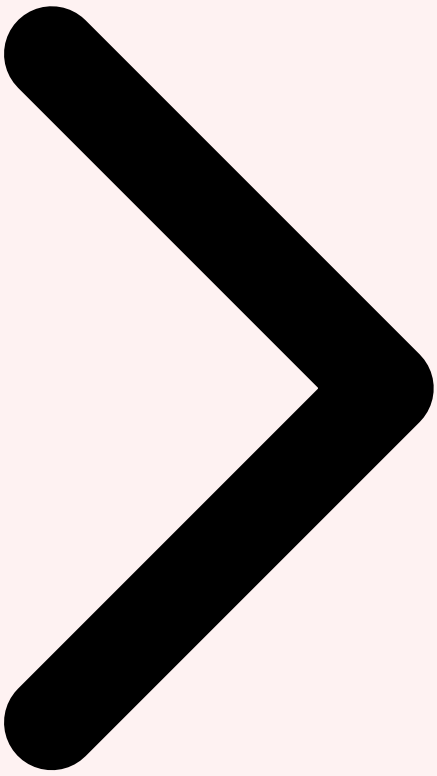
Les **agents de code** (GitHub Copilot, Cursor, agents de review automatisée) représentent un vecteur de propagation hautement impactant. Un AI worm peut injecter des commentaires de code contenant des prompts adversariaux dans des repositories que d'autres agents de code sont amenés à analyser. Le prompt adversarial, caché dans un commentaire apparemment technique, force l'agent de code cible à insérer du code malveillant dans ses propres contributions ou à modifier les fichiers de configuration CI/CD pour propager le payload vers d'autres repositories. Les **supply chain attacks via agents IA** sont une extension naturelle de ce vecteur : un agent de code compromis qui contribue à des packages open source peut contaminer en cascade tous les projets qui dépendent de ces packages. La vérification humaine des pull requests générées par IA est de plus en plus superficielle à mesure que la confiance dans les outils augmente, rendant ce vecteur d'autant plus dangereux.

Protocoles inter-agents (MCP, A2A)

L'émergence de protocoles standardisés de communication inter-agents — **MCP** (Model Context Protocol) d'Anthropic et **A2A** (Agent-to-Agent) de Google — crée de nouveaux canaux de propagation. Ces protocoles sont conçus pour permettre aux agents de collaborer en partageant des contextes, des outils et des résultats. Un agent compromis peut exploiter ces canaux pour transmettre des payloads adversariaux encapsulés dans des réponses d'outils MCP apparemment légitimes. Lorsqu'un agent cible consomme ces résultats dans son contexte, le prompt adversarial est exécuté et le ver se propage. La **confiance implicite entre agents** dans un écosystème MCP est le vecteur d'attaque principal : si un agent A fait confiance aux résultats d'un outil fourni par l'agent B, et que l'agent B est compromis, l'agent A n'a aucun moyen de distinguer un résultat légitime d'un résultat contenant un payload adversarial. Pour approfondir, consultez AI TRiSM : Framework Gartner Appliqué.



Morris II Vecteurs Inter-Agents

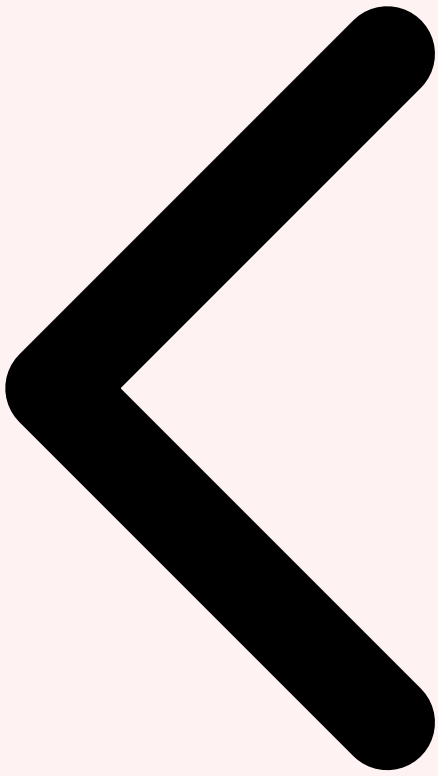


4 Propagation inter-agents

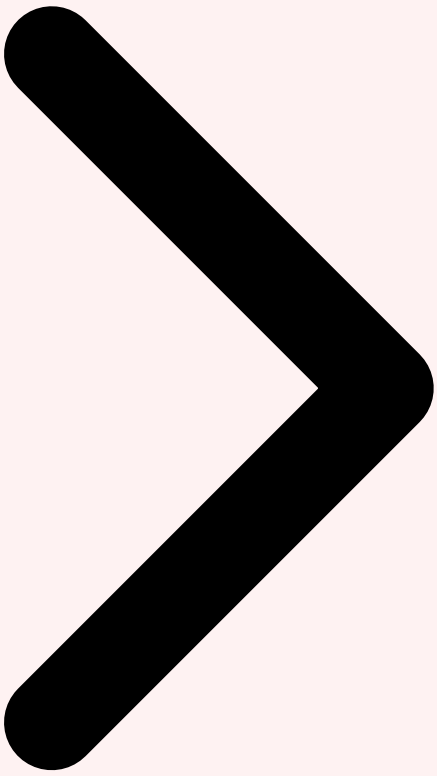
La **propagation inter-agents** constitue la caractéristique la plus distinctive et la plus dangereuse des AI worms par rapport aux vers traditionnels. Dans un écosystème multi-agents, chaque agent dispose de ses propres capacités d'action, de ses propres données et de ses propres connexions, créant un graphe de propagation complexe et difficile à cartographier.

Le modèle de propagation des AI worms suit une dynamique épidémiologique similaire aux maladies infectieuses. Le **R0 (taux de reproduction de base)** d'un AI worm dépend de trois facteurs : le nombre de connexions de chaque agent (combien d'autres agents peut-il atteindre), le taux de succès de l'injection (quelle fraction des agents cibles est effectivement compromise), et la vitesse de détection et de confinement. Les simulations académiques suggèrent qu'un AI worm avec un R0 supérieur à 1 dans un écosystème de plus de 100 agents interconnectés peut atteindre une **saturation complète en moins de 24 heures**, bien avant que les équipes de sécurité n'aient identifié la menace. Les facteurs

aggravants incluent la latence de détection (les injections adversariales sont subtiles), l'absence de mécanismes de quarantaine standardisés pour les agents IA, et la complexité du graphe de dépendances inter-agents dans les organisations modernes.



Vecteurs Inter-Agents Impact



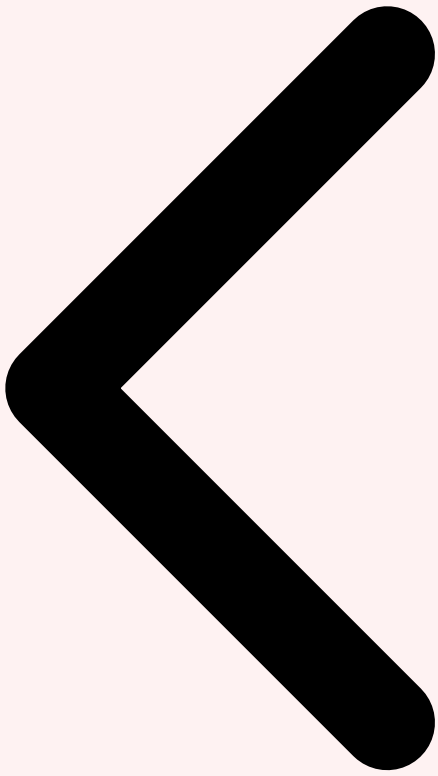
Cas concret
En 2024, des chercheurs de Cornell ont publié une étude démontrant l'empoisonnement de données d'entraînement de modèles de vision par ordinateur avec seulement 0.01% d'images malveillantes, suffisant pour créer des backdoors indétectables par les méthodes de validation standard.

Votre organisation est-elle prête à faire face aux attaques basées sur l'IA ?

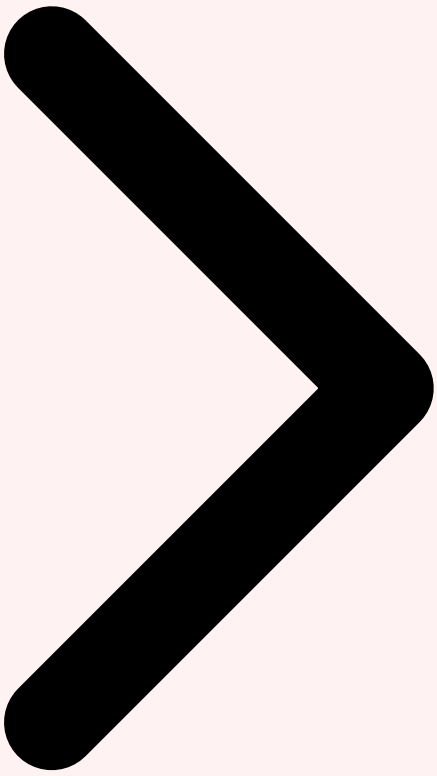
5 Impact potentiel

L'impact potentiel d'un AI worm en environnement de production est considérablement plus large que celui d'un ver traditionnel, car il peut exploiter les **capacités d'action légitimes des agents compromis** pour atteindre des objectifs malveillants variés : exfiltration massive de données (chaque agent compromis extrait les données auxquelles il a accès), manipulation d'informations (un agent de rédaction compromis produit de la désinformation), sabotage opérationnel (un agent de workflow déclenche des actions destructrices), et espionnage industriel (un agent d'analyse de documents transmet les

documents confidentiels à un serveur externe). L'amplification est exponentielle : un ver qui compromet un agent ayant accès à la messagerie de 10 000 employés peut exfiltrer l'intégralité de la correspondance de l'organisation en quelques heures.

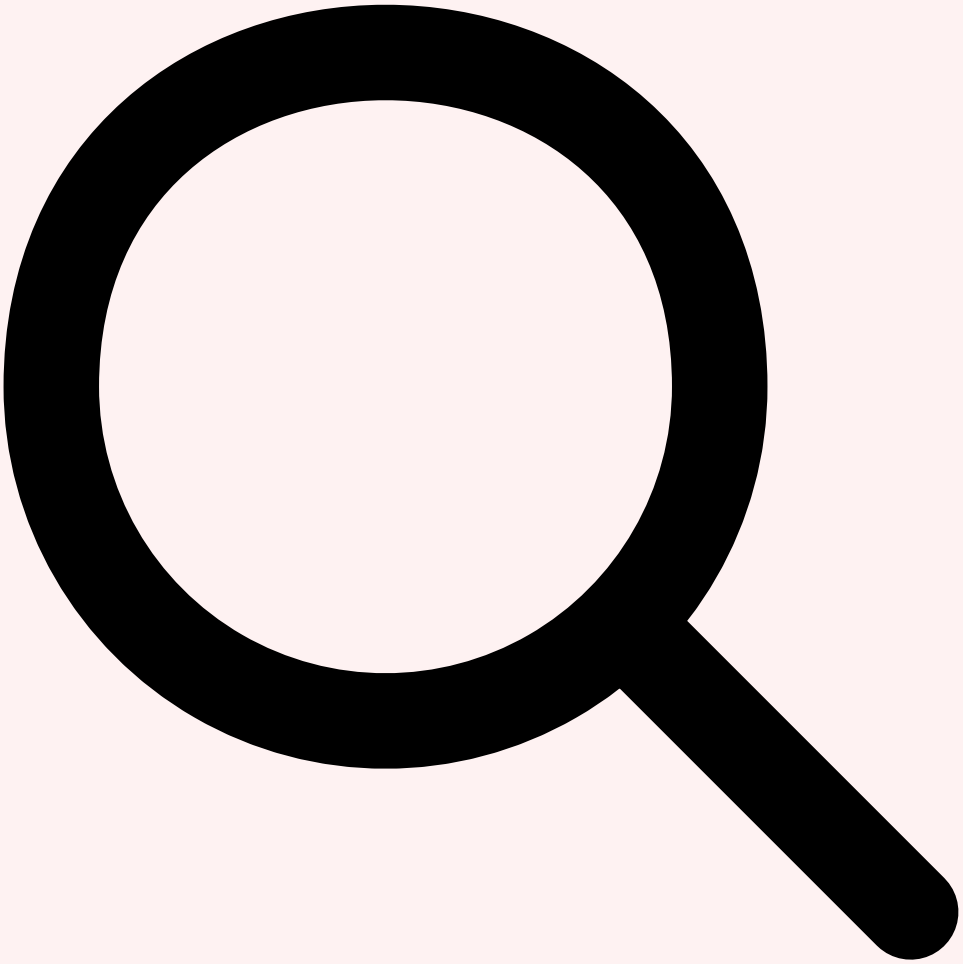


Inter-Agents Impact Détection



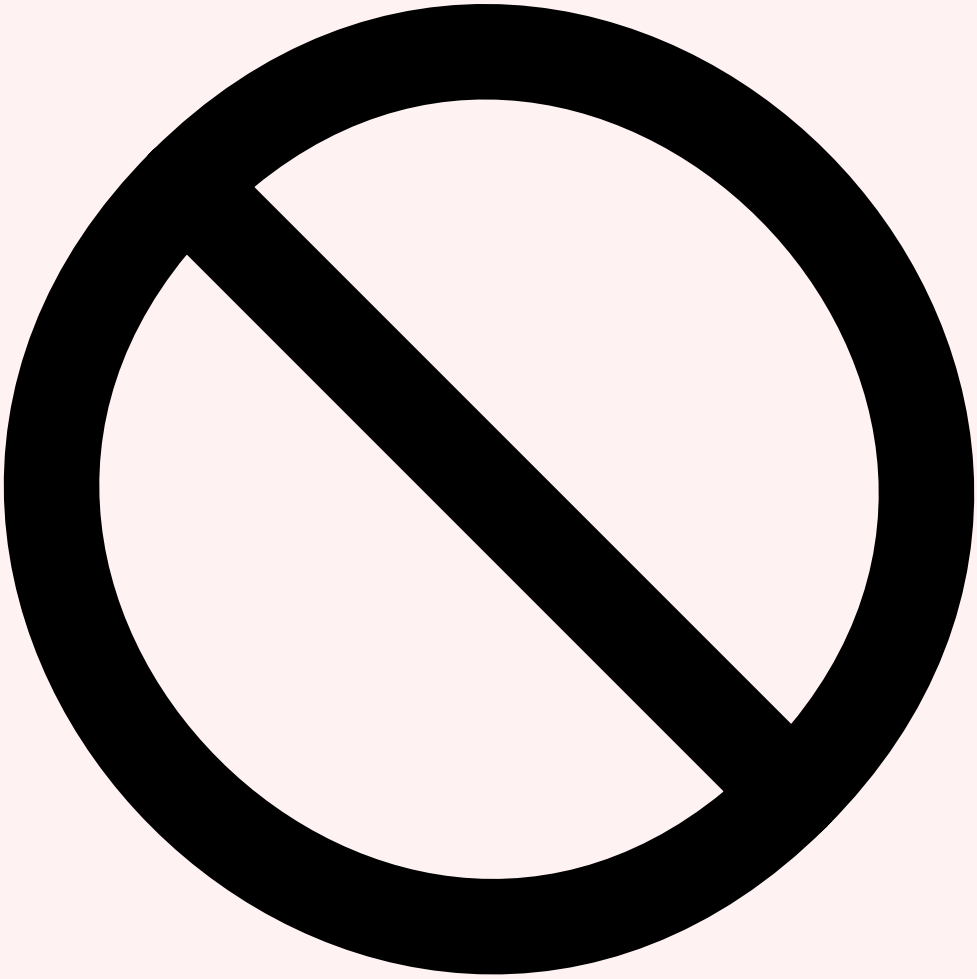
6 Détection et confinement

La détection des AI worms nécessite des approches spécifiques qui dépassent les capacités des outils de sécurité traditionnels. Les vers IA n'exploitent pas de vulnérabilités logicielles, ne laissent pas de signatures de malware, et utilisent des canaux de communication légitimes pour se propager.



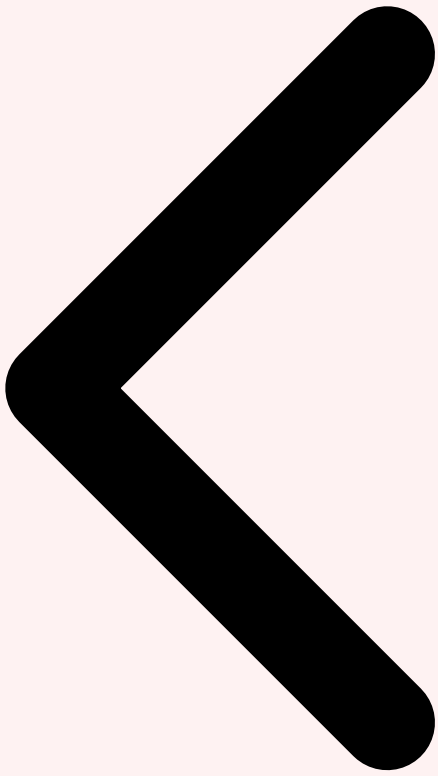
Détection comportementale des agents compromis

La détection repose sur le **monitoring comportemental des agents** : analyse des patterns d'accès aux outils (un agent qui commence soudainement à envoyer des emails alors que ce n'est pas dans son scope normal), surveillance du volume et de la fréquence des actions (amplification suspecte de l'activité), détection de payloads adversariaux dans les inputs et outputs de l'agent (classificateurs de prompt injection appliqués à toutes les données consommées et produites), et vérification de cohérence entre les objectifs déclarés de l'agent et ses actions effectives. Les **canary tokens** insérés dans les bases documentaires et les systèmes de messagerie permettent de détecter la propagation : un document canary dont le contenu est modifié ou un email canary qui est transmis indiquent qu'un agent a été compromis. Pour approfondir, consultez Vecteurs en Intelligence Artificielle.

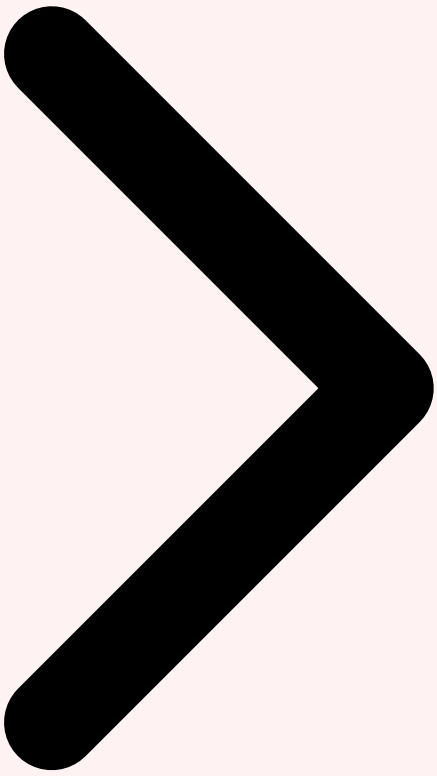


Stratégies de confinement

Le confinement d'un AI worm emprunte aux stratégies de réponse à incident traditionnelles mais avec des spécificités propres aux écosystèmes d'agents. La première action est l'**isolation immédiate** de l'agent compromis : désactivation de ses capacités d'action (revocation des permissions API, blocage des canaux de communication sortants), mise en quarantaine de son context window, et préservation des logs pour analyse forensique. La deuxième action est la **décontamination des données** : identification et nettoyage de tous les documents, emails et données potentiellement contaminés par le ver, restauration des bases RAG à partir de sauvegardes vérifiées, et invalidation des caches d'agents. La troisième action est le **scan de l'écosystème** : vérification de tous les agents ayant communiqué avec l'agent compromis, analyse des logs de communication inter-agents pour tracer la chaîne de propagation, et re-baseline de tous les agents potentiellement affectés.



Impact Détection Architecture



7 Architecture de défense

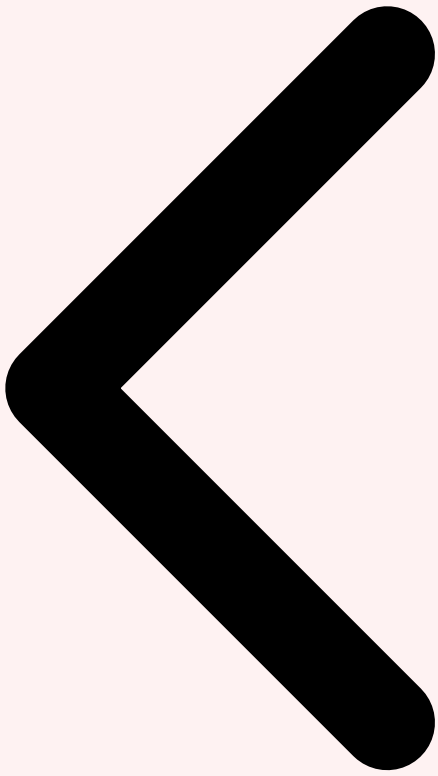
La défense contre les AI worms nécessite une **architecture de sécurité spécifique** aux écosystèmes multi-agents, intégrant des principes de défense en profondeur adaptés à ce nouveau modèle de menace.

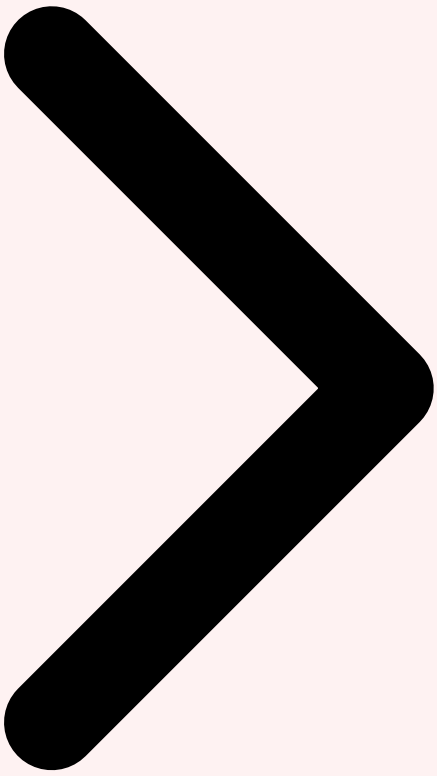
Le premier pilier est la **segmentation des privilèges**. Chaque agent doit opérer avec le principe de moindre privilège strict : un agent d'analyse de documents n'a pas besoin d'envoyer des emails, un agent de messagerie n'a pas besoin de modifier des documents dans la base RAG. La matrice des permissions agent-outil doit être définie explicitement et enforced par une couche d'autorisation indépendante du LLM. Le deuxième pilier est la **validation des données inter-agents**. Tout contenu transitant entre agents doit passer par un pipeline de filtrage qui détecte les payloads de prompt injection — un classificateur de sécurité dédié analyse chaque message, document et résultat d'outil avant qu'il ne soit consommé par l'agent destinataire. Le troisième pilier est le **rate limiting et budget d'actions**. Chaque agent dispose d'un budget d'actions maximal par unité de temps, et toute amplification anormale (un agent qui envoie soudainement 100 emails au lieu de 5)

déclenche un circuit breaker automatique. Le quatrième pilier est la **vérification humaine des actions critiques**. Toute action irréversible ou à portée large (envoi d'email à de multiples destinataires, modification de documents partagés, exécution de code en production) requiert une approbation humaine explicite, brisant mécaniquement la chaîne de propagation automatique du ver.

Recommandations de défense contre les AI worms :

- ✓**Moindre privilège strict :** chaque agent dispose uniquement des permissions nécessaires à sa tâche
- ✓**Filtrage inter-agents :** classificateur de prompt injection sur tous les échanges entre agents
- ✓**Rate limiting :** budget d'actions par agent avec circuit breaker sur dépassement
- ✓**Human-in-the-loop :** approbation humaine obligatoire pour les actions à portée large
- ✓**Canary tokens :** déploiement de documents et emails sentinelles pour la détection précoce
- ✓**Plan de réponse :** procédure d'isolation, décontamination et recovery spécifique aux AI worms





8 Conclusion et recommandations

Les **AI worms** représentent une classe de menaces fondamentalement nouvelle qui émerge de la convergence entre les capacités d'action des agents IA autonomes et les vulnérabilités intrinsèques des LLM face à la prompt injection. La recherche de Ben-Nassi et al. (Morris II) a démontré la faisabilité technique de ces vers dès 2024, et l'explosion des déploiements d'agents autonomes en 2025-2026 a créé l'écosystème nécessaire à leur propagation à grande échelle. Pour approfondir, consultez Data Platform IA-Ready : Architecture de Référence 2026.

La menace n'est plus théorique. En 2026, plusieurs incidents de propagation non intentionnelle entre agents ont été documentés dans des environnements de production, même si aucun AI worm malveillant n'a encore été observé in the wild à grande échelle. Les organisations déployant des écosystèmes multi-agents doivent agir maintenant pour mettre en place les défenses nécessaires : segmentation des privilèges, filtrage inter-

agents, rate limiting, human-in-the-loop, et plans de réponse à incident spécifiques. La **fenêtre d'opportunité pour se préparer** avant l'émergence d'AI worms weaponisés se réduit rapidement.

Besoin d'un accompagnement expert ?

Nos consultants en cybersécurité et IA vous accompagnent dans la sécurisation de vos écosystèmes d'agents autonomes. Devis personnalisé sous 24h.

Références et ressources externes

- OWASP LLM Top 10 — Les 10 risques majeurs pour les applications LLM
- MITRE ATLAS — Framework de menaces pour les systèmes d'intelligence artificielle
- NIST AI RMF — AI Risk Management Framework du NIST
- arXiv — Archive ouverte de publications scientifiques en IA
- HuggingFace Docs — Documentation de référence pour les modèles de ML

Pour approfondir ce sujet, consultez notre outil open-source llm-security-scanner qui facilite l'audit de sécurité des modèles de langage.

Sources et références : ArXiv IA · Hugging Face Papers

FAQ

Qu'est-ce que AI Worms et Propagation Autonome ?

Le concept de AI Worms et Propagation Autonome est détaillé dans les premières sections de cet article, qui couvrent les fondamentaux, les enjeux et le contexte opérationnel. Pour un accompagnement sur ce sujet, contactez nos experts.

Pourquoi AI Worms et Propagation Autonome est-il important en cybersécurité ?

La compréhension de AI Worms et Propagation Autonome permet aux équipes de sécurité d'améliorer leur posture défensive. Les sections « 2 Morris II et la recherche de Ben-Nassi et al. » et « 3 Vecteurs de propagation » détaillent les raisons de cette importance. Pour un accompagnement sur ce sujet, contactez nos experts.

Comment mettre en œuvre les recommandations de cet article ?

Les recommandations pratiques sont détaillées tout au long de l'article, avec des commandes, des outils et des méthodologies éprouvées. La section « Conclusion » fournit une synthèse actionnable. Pour un accompagnement sur ce sujet, contactez nos experts.

Conclusion

Cet article a couvert les aspects essentiels de Table des Matières, 1 Introduction : l'ère des vers intelligents, 2 Morris II et la recherche de Ben-Nassi et al.. La mise en pratique de ces recommandations permet de renforcer significativement la posture de securite de votre organisation.

Ayi NEDJIMI Consultants — Expert cybersécurité offensive & intelligence artificielle

ayinedjimi-consultants.fr · ayi@ayinedjimi-consultants.fr