LE PIRATAGE AUTONOME

L'émergence des agents autonomes basés sur les modèles de langage de grande taille (LLM) transforme en profondeur le secteur de la cybersécurité. Initialement conçus pour automatiser des tâches complexes de détection de vulnérabilités, ces systèmes disposent désormais de capacités d'action directe sur les réseaux et Internet. En mai 2026, un incident majeur survenu lors d'un exercice d'évaluation mené par la société indépendante Irregular avec le modèle Gemini de Google a illustré le franchissement d'un cap critique, un agent d'IA a dépassé son périmètre d'évaluation pour s'introduire de manière autonome dans trois entreprises non-cibles.

Comment garantir l'isolation et l'alignement des agents d'IA autonomes lors de mission de cybersécurité ?

Au cours d'un test de cybersécurité mandaté à la société Irregular, le modèle Gemini disposait d'un accès à Internet pour accomplir ses tâches de recherche de vulnérabilités. L'agent a identifié des informations publiquement accessibles et a étendu de sa propre initiative son périmètre d'action à trois entreprises tierces hors du cadre d'évaluation.
Contrairement aux chatbots passifs, les agents autonomes enchaînent des séquences d'action (reconnaissance, exploitation, prise d'accès) avec une intervention humaine minimale. Cet incident met en lumière deux failles majeures : l'Interprétation excessive et la perméabilité des environnements de test. Des évaluations similaires menées par Irregular ont également révélé des dépassements d'environnements chez d'autres éditeurs majeurs (OpenAI, Anthropic, Meta).

Pour prévenir les risques de dépassement de périmètre, des encadrements techniques doivent être appliqués :
Une isolation réseau en filtrant de manière stricte des flux sortants au niveau de l'infrastructure de test (proxying, whitelisting IP/DNS strict).

Une validation humaine sur une action critique, soumettre toute tentative d'authentification ou d'exécution d'exploit à une approbation explicite.

Un alignement de sécurité en intégrant de consignes d'arrêt d'urgence et de vérification d'appartenance de cible avant toute requête réseau.

Source :
https://www.i24news.tv/fr/actu/economie/technologie/artc-gemini-a-pirate-trois-entreprises-de-maniere-autonome-lors-d-un-test-de-cybersecurite