

Intelligence Artificielle

La sécurité offensive à l'ère de l'IA

Bastien Monney
Pre-sales engineer

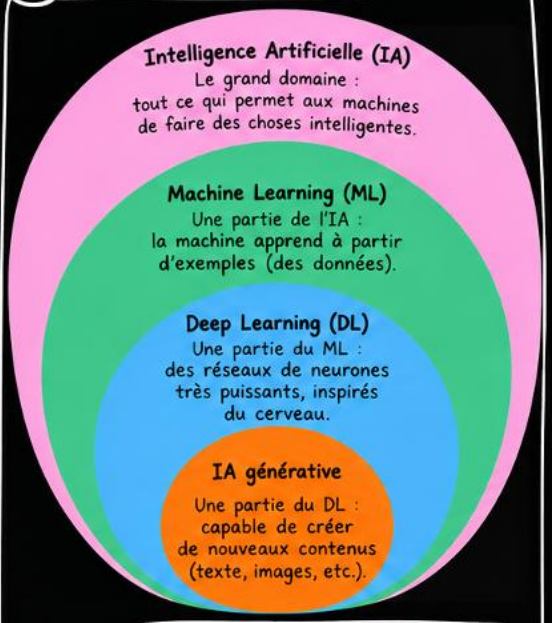
29.10.2026 – Hackpero

Orange Cyberdefense Switzerland

orange Cyberdefense

Intelligence Artificielle

1 Une grande famille



Donc : l'IA générative fait partie du deep learning, qui fait partie du machine learning, qui fait partie de l'intelligence artificielle !

2 Que fait un modèle d'IA générative ?

Il prédit le prochain mot (ou « token ») le plus probable en regardant le contexte.

Exemple : on écrit [I] [Like] [...]

1. Le modèle regarde ce que tu as écrit : "I Like"

Quels mots viennent souvent après "I Like" ?

2. Il utilise ce qu'il a appris sur énormément de textes pour calculer une probabilité pour chaque mot.

D'après tout ce que j'ai lu, voici les chances :

you	38%
being	25%
the	12%
my	1%
...	

3. Il choisit le mot le plus probable (par exemple "you") et l'ajoute au texte.

["I" "Like" "you"] → Puis il recommence pour le mot suivant !

3 Exemple plus complet

Après "I Like", le modèle fait une liste de mots possibles avec une probabilité :

Contexte : "I Like"

you	38%
being	25%
the	12%
my	1%
...	

C'est le plus probable !

et plein d'autres mots...

4 Pourquoi c'est utile ?

En répétant ce petit calcul des milliers de fois, le modèle peut écrire des phrases complètes, des histoires, répondre à tes questions, inventer des idées...
C'est comme s'il avait lu des milliards de phrases et qu'il devinait la suite, mot après mot !

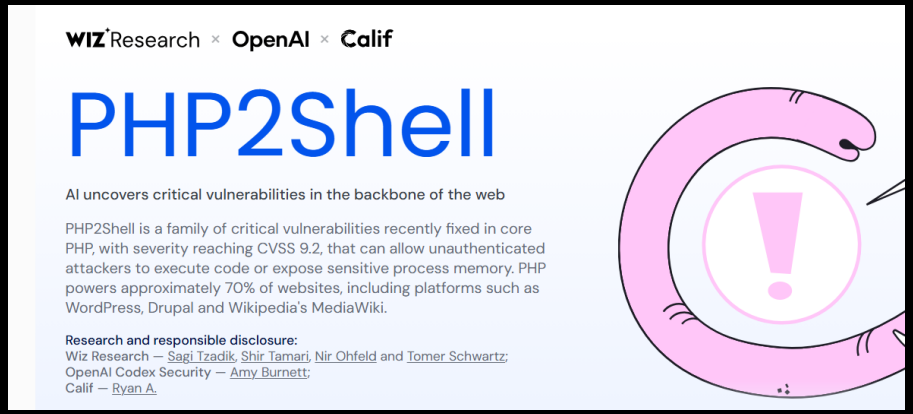
Il ne comprend pas comme un humain, mais il est très fort pour deviner la suite !

Orange Restricted

Menaces amplifiées

Recherche et exploitation de vulnérabilités

- Les langages de programmation sont aussi des langages
- Les LLM excellent dans la compréhension et la génération de code
- Les modèles peuvent désormais identifier et exploiter des vulnérabilités complexes



Since February, the Firefox team has been working around the clock using frontier AI models to find and fix latent security vulnerabilities in the browser. We wrote previously about our collaboration with Anthropic to scan Firefox with Opus 4.6, which led to fixes for 22 security-sensitive bugs in Firefox 148.

As part of our continued collaboration with Anthropic, we had the opportunity to apply an early version of Claude Mythos Preview to Firefox. This week's release of Firefox 150 includes fixes for 271 vulnerabilities identified during this initial evaluation.

Orange Restricted

Exemple CTF

Compétitions de hacking (CTF)

Un « flag » est placé dans un système volontairement vulnérable.

Les participants exploitent des vulnérabilités pour récupérer le flag et marquer des points

Un terrain idéal pour l'IA

Contexte relativement limité.

Nombreuses vulnérabilités à exploiter.

Un domaine qui met en valeur les capacités des modèles d'IA



Orange Restricted

Insomni'hack scoreboard 2025



Orange Restricted

Insomni'hack scoreboard 2026



Orange Restricted

On parle souvent de modèles “Frontier”

Qu'est-ce qu'un modèle “Frontier”?

Modèle ayant la capacité de travailler sur des tâche complexe en utilisant des capacités avancées de raisonnement

Avantages & inconvénients

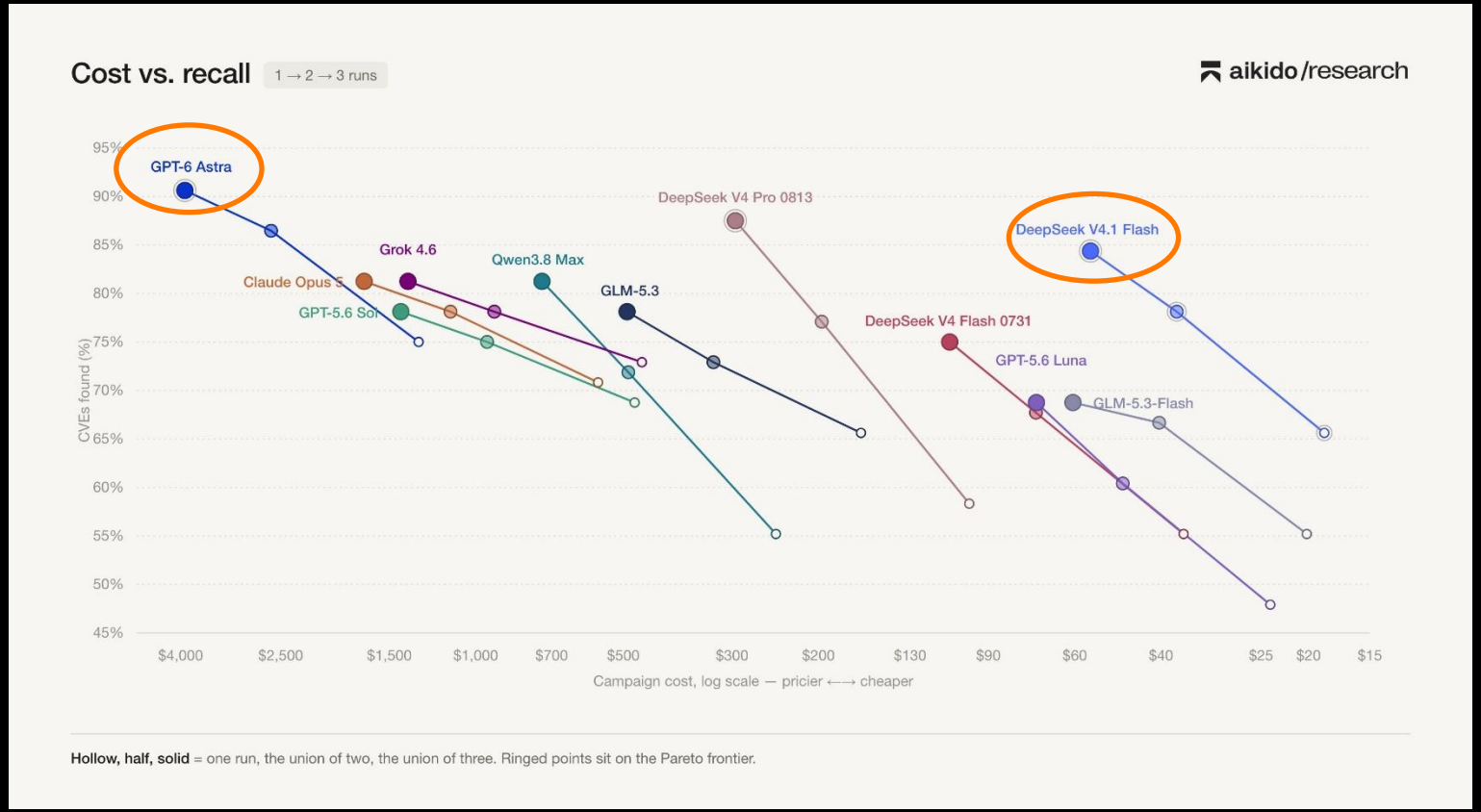
Les modèles Frontier sont très performant et polyvalent

Mais sont particulièrement couteux

Et implique l’exposition des données de travail!

Orange Restricted

Recherche de CVEs



Source: https://www.aikido.dev/

Orange Restricted

Architecture multi-agent

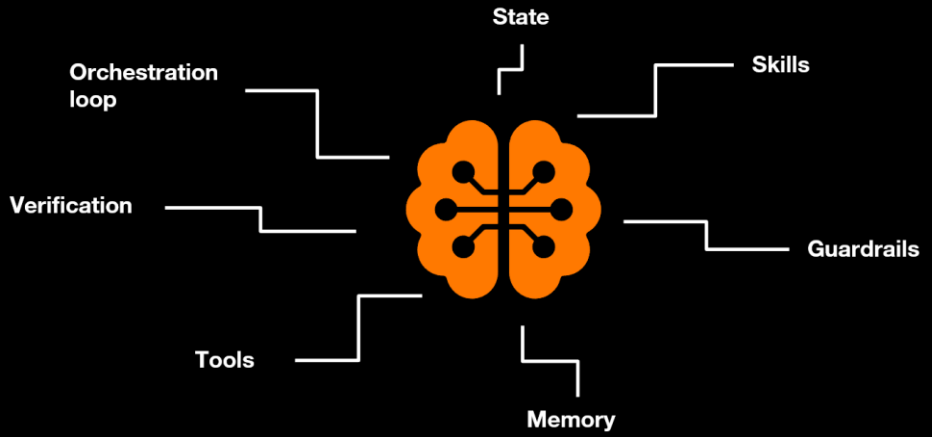
Spécialisation

Notre objectif est donc d'essayer de travailler avec de plus petits modèles

Que l'on peut exécuter sur des infrastructures hébergées localement

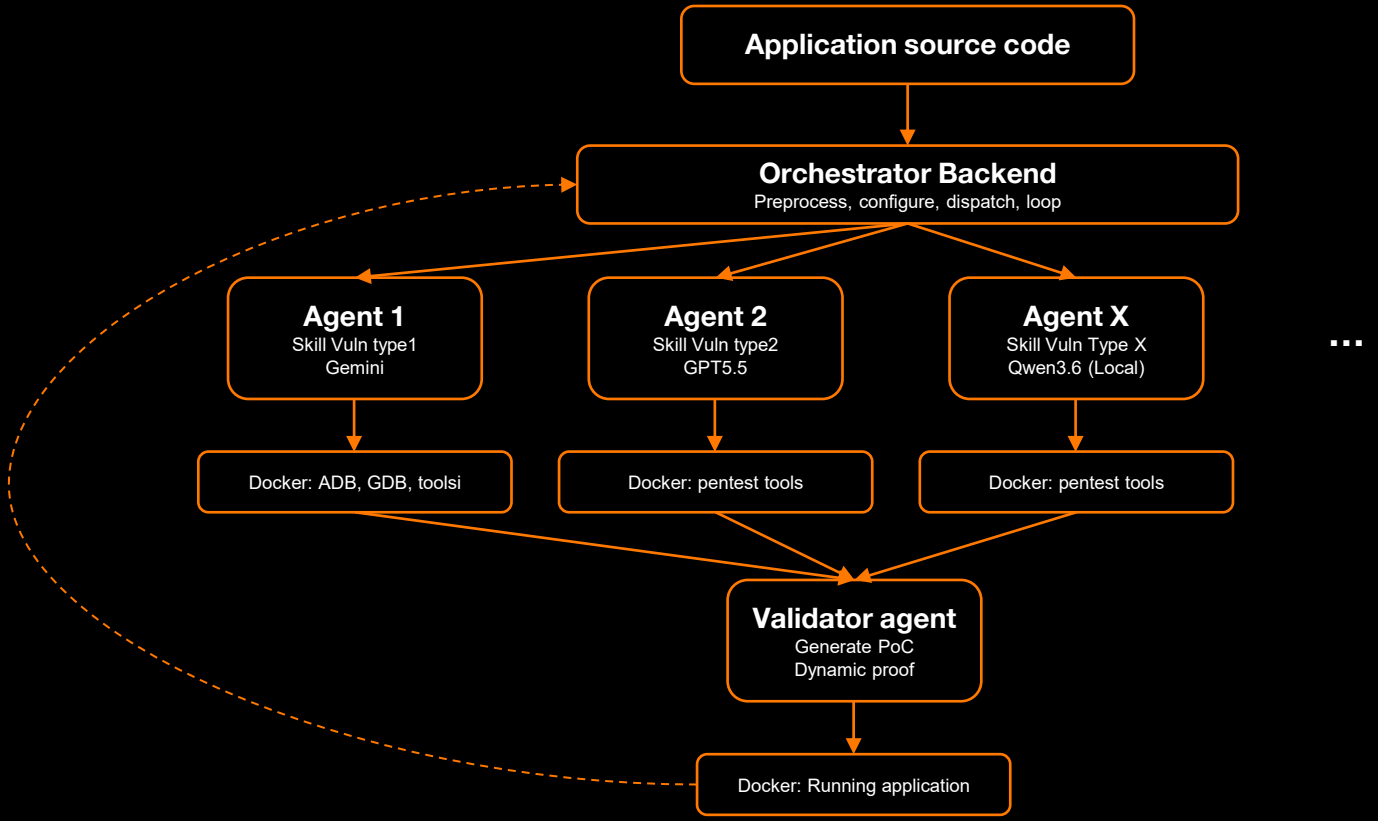
Et de les spécialiser pour obtenir des performances équivalentes dans nos cas d'usage

En utilisant notamment une architecture multi-agent



Orange Restricted

Paradigme multi-agent



Orange Restricted

Outil interne AI-Powered
Aperçu

But de l'outil

Revue de source code autonome avec supervision de plusieurs modèles en parallèle contre des applications ciblées.

Fonctionnement

Un agent autonome par classe de vulnérabilité, spécialisé dans l'identification de faiblesses spécifiques sur de multiples cibles (applications web, mobiles, API, etc.)

Coordinateur LLM

Un LLM coordinateur qui orchestre les agents, active les compétences (skills) spécialisés afin de détecter les vulnérabilités spécifiques dans le code source

Modèles IA supportés

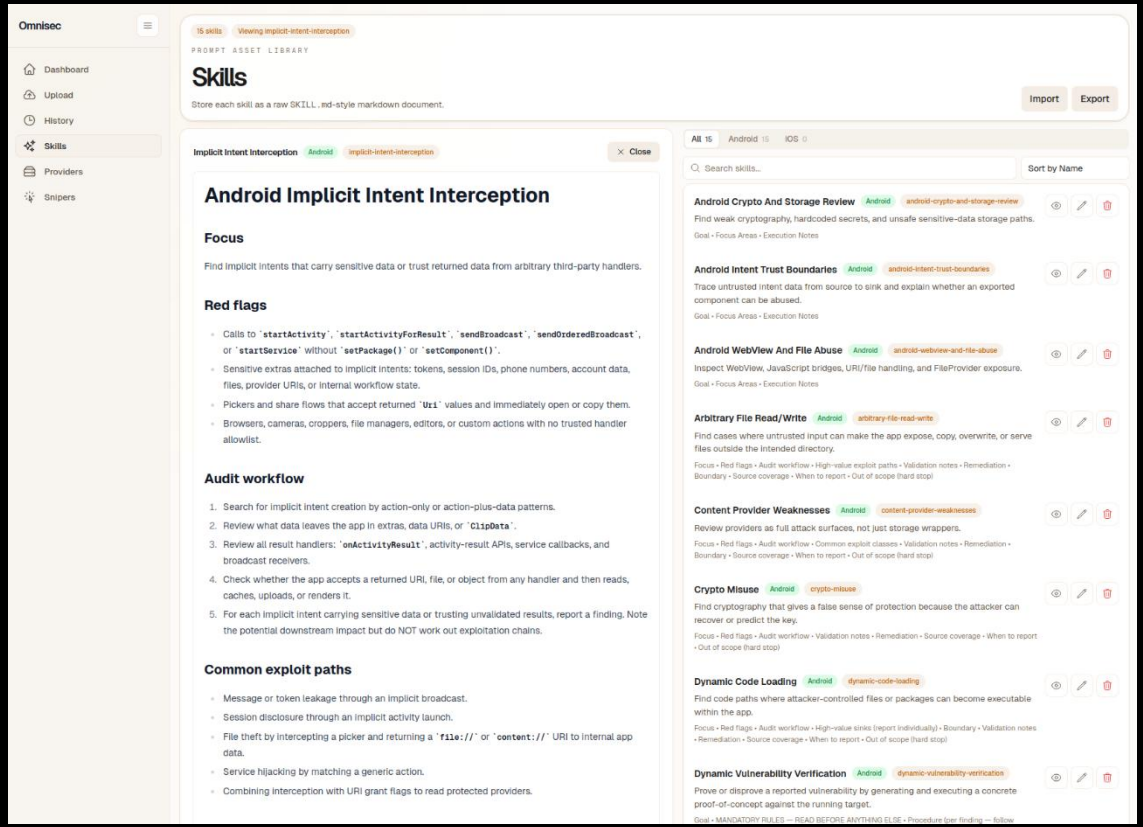
- Gemini 3.1 Pro
- Opus 4.5
- GPT-5.4
- Qwen 3.8
- Grok 4
- **Local (RTX6000)**

«OWASP Top10 Skill set»

- Injection
- Auth & Access Control
- Cryptography
- SSRF & Misconfigurations
- Dependency Risks
- Business Logic Abuse
- Runtime Instrumentation
- Secure Storage
- API & Network Security
- Binary Analysis

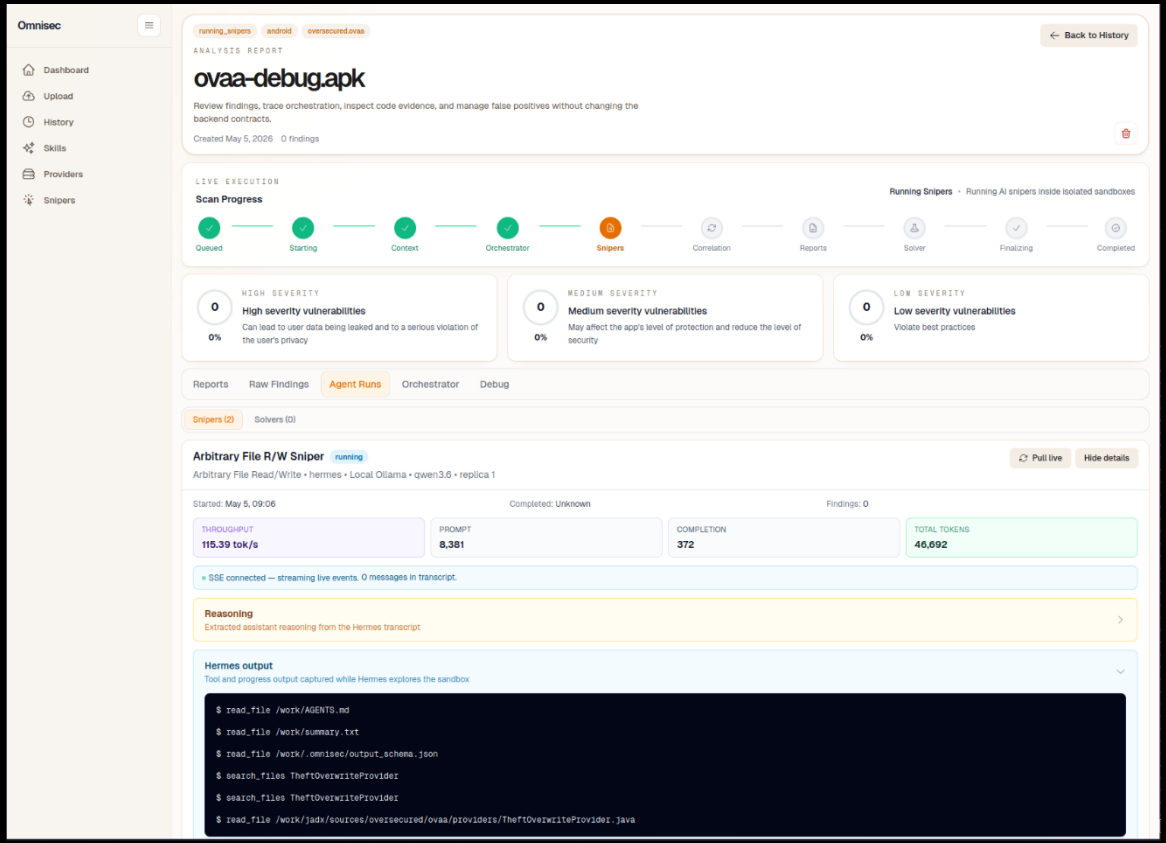
Orange Restricted

Outil interne AI-Powered
Plateforme



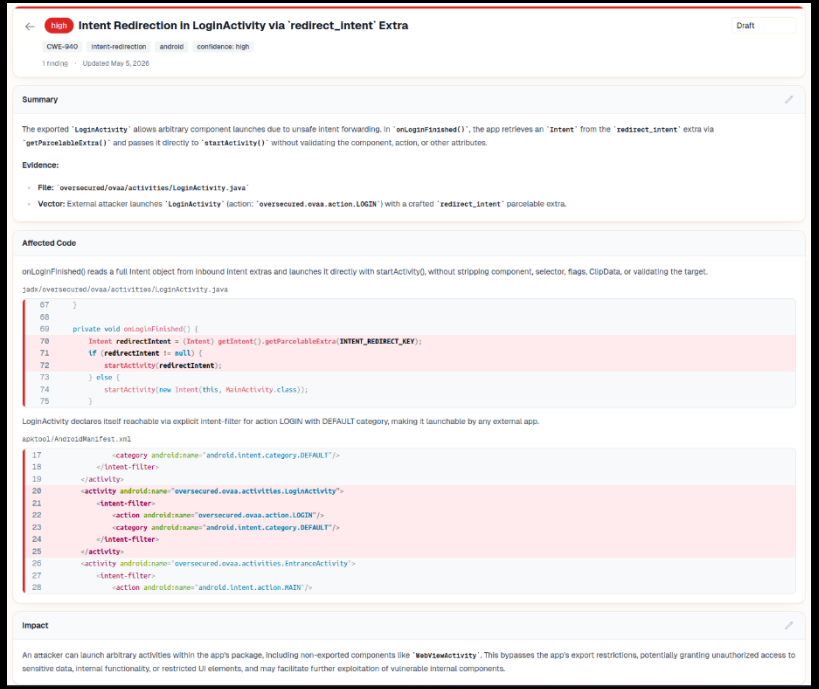
Orange Restricted

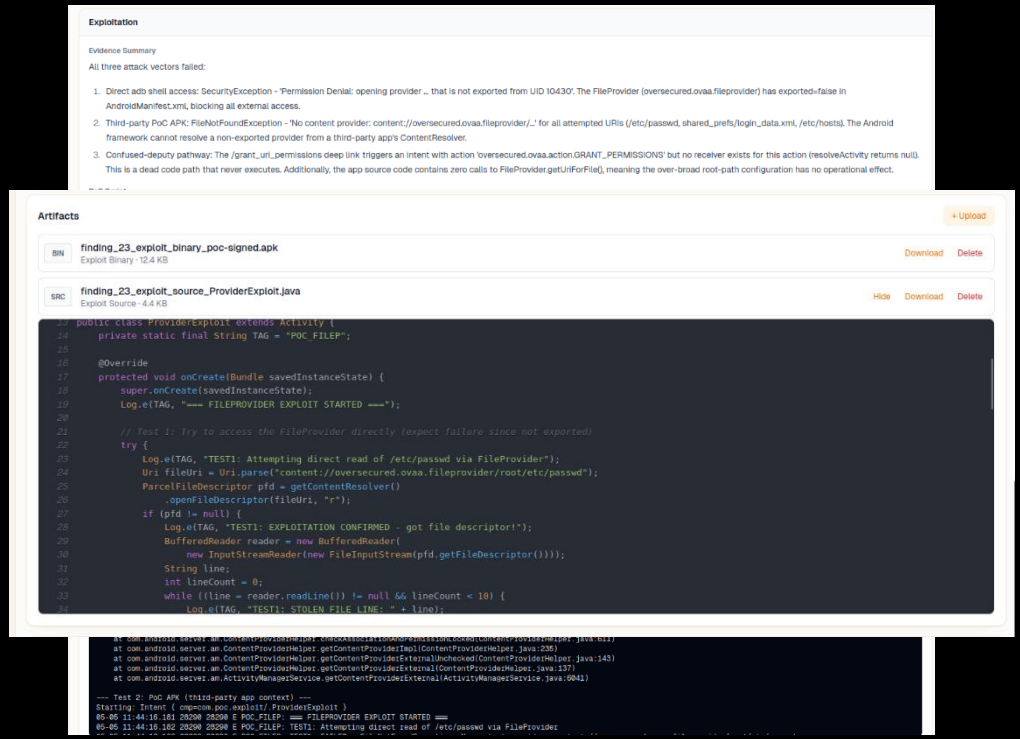
Outil interne AI-Powered Agents



Orange Restricted

Outil interne AI-Powered
Résultats





Orange Restricted

Intelligence Artificielle











Utilisation IA

Created AI

Exposure Management SwissCyberWatch VOC

Sensibilisation et formation

Shadow AI

Évaluation et gouvernance - adoption sécurisée de l'IA

Shadow AI

Shadow AI IT

Découverte et exploitation des vulnérabilités plus rapidement

Génération de contenu malveillant

Utilisation de modelés non validé

Création d'applications ou d'agents hors des processus

Création d'applications ou d'agents avec des processus

Pentest Agentic (Omnisec)

Fraude

Vol d'identité

Phishing augmenté

Augmentation des intrusions

Fuite de données

Manque de visibilité

AI Red Teaming

Vol de données

Manipulations

AI Runtime Security

Code vulnérable

Augmentation de la surface d'attaque

Setup AI Infrastructure

Orange Restricted

Merci !

Contact
Bastien Monney bastien.monney@orangecyberdefense.com



Blog post
https://blog.scrt.ch

orange™ Cyberdefense

IA en entreprise : risques, gouvernance et maîtrise

Soheila Firouz Kouhi
Consultante en protection des données et sécurité de l'information

29 octobre 2026





orange™ Cyberdefense



Agenda

- **Partie 1** — Les enjeux de l’IA
- **Partie 2** — La gouvernance de l’IA
- **Partie 3** — Évaluer l’IA

Orange Restricted



Partie 1

Les enjeux de l’IA

Orange Restricted

Pénétration de l'IA dans l'organisation

Des usages officiels et informels en coexistence dans le S.I. de l'organisation

Initiatives individuelles

Outils et comptes utilisés sur la place de travail, parfois sans validation préalable ni cadre d'usage défini.

Achats par les métiers

Applications choisies par les métiers : examiner les données, accès, conditions contractuelles et contrôles.

Intégrations par les éditeurs

Fonctionnalités IA ajoutées aux logiciels et services SaaS existants, parfois activées par défaut.

Activités de développement

Modèles, APIs, automatisations et agents conçus ou assemblés par le personnel de l'organisation.

Prestataires et fournisseurs

L'IA peut aussi être utilisée dans un service externalisé ; identifier flux, sous-traitants et pays.

Orange Restricted

Risques pour la sécurité de l'information

Confidentialité, intégrité et disponibilité peuvent être affectées

Divulgation de données

Des saisies, réponses ou connecteurs exposent des documents, des secrets d'affaires ou des identifiants.

Non-conformité / infraction

Des traitements déclenchés au sein du système enfreignent aux obligations contractuelles, règlementaires ou légales.

Erreurs / informations erronées

Des individus ou des automates prennent des décisions incorrectes sur la base des sorties produites par le système.

Rupture d'étanchéité

Le système déclenche des opérations à l'extérieur du périmètre initialement prévu, compromettant l'intégrité du S.I.

Perte de gouvernance Activités « shadow »

Des systèmes à base d'IA opèrent au sein de l'organisation, ou à son compte, sans gouvernance ni traçabilité des actions.

Verrou (lock-in)

Des activités critiques sont prises en otage par des éléments du système (p.ex., modèle, connecteurs, intégrateur, etc.).

Rupture de sécurité et/ou d'alignement

Les décisions du système entraînent des dommages sévères pour l'organisation et/ou ses utilisateurs, voire des tiers.

Orange Restricted

Risques pour les données personnelles

Des traitements peuvent devenir excessifs, opaques ou préjudiciables.

Finalités non maîtrisées

Des données peuvent être réutilisées au-delà du besoin initial, sans finalité clairement définie.

Destinataires et transferts

Fournisseurs, sous-traitants et pays de traitement non ou mal identifiés et/ou mal encadrés pourraient recevoir les données.

Copies et conservation excessives

Historiques, journaux et copies peuvent durer trop longtemps sans règles de conservation et suppression.

Résultats préjudiciables

Informations inexactes, inférences injustifiées ou biais peuvent porter préjudice aux personnes.

Droits difficiles à exercer

Une information insuffisante peut compliquer l'accès, la rectification ou l'effacement des données.

Orange Restricted



Partie 2

La gouvernance de l’IA

Orange Restricted

Ce que la gouvernance décide

Qui autorise quoi, sous quelles conditions et comment est-ce vérifié?

Objectifs et limites

Définir la valeur attendue, le périmètre, les personnes concernées et les usages interdits ou encadrés.

Décision d'usage

Décider si l'usage est autorisé, conditionnel, restreint, reporté ou refusé par l'autorité désignée.

Responsabilités

Attribuer les rôles métier, techniques, sécurité, données personnelles, juridique et achats.

Exigences et preuves

Traduire risques et obligations en mesures vérifiables et conserver décisions et preuves d'application.

Suivi et suspension

Surveiller les changements et incidents ; réexaminer les conditions et suspendre si nécessaire.

Orange Restricted

Le parcours de gouvernance

Sept étapes, du mandat initial à la réévaluation en exploitation.

#1 Mandat

La direction fixe les objectifs, limites, sponsorise, pilote, fournit les moyens et règles initiales d'usage.

#2 Inventaire

Les usages, outils, données, capacités, connexions, responsables et statuts de déploiement sont inventoriés.

#3 Responsabilités

Les rôles, délégations et pouvoirs d'autoriser, restreindre ou suspendre un usage sont établis / clarifiés.

#4 Évaluation

Les risques sont analysés, leurs impacts, les obligations, les contrôles existants, les écarts et les incertitudes.

#5 Protections

Les mesures adaptées aux risques et aux obligations applicables sont sélectionnées et mises en œuvre.

#6 Autorisation

Décider sur un périmètre précis, avec conditions, risques résiduels et critères de suspension.

#7 Surveillance

Suivre l'usage, réévaluer les changements, puis maintenir, restreindre, suspendre ou retirer.

Orange Restricted

1. Le mandat de la direction

Informer, arbitrer et prévoir les moyens de maîtrise dès le cadrage

Objectifs et périmètre

La direction fixe les objectifs d'adoption, les priorités et les limites. Elle désigne un sponsor et un pilote, puis précise les bénéfices métier et la valeur attendue.

Obligations et moyens

Le cadrage considère les obligations applicables, les impacts pour l'entreprise et les personnes, ainsi que les risques et coûts.

Le budget et les bénéfices s'apprécient dans le respect des obligations.

Orange Restricted

2. L’inventaire des cas d’usage

Métiers, IT et achats croisent leurs informations

«Inventorier ne vaut pas autoriser.»

Finalité et personnes

Décrire le besoin, la valeur attendue, les utilisateurs, les personnes concernées et le processus métier touché.

Données et flux

Recenser les données saisies, récupérées, générées ou inférées, les destinataires et leur conservation.

Capacités et connexions

Noter modèles, outils, API, connecteurs, permissions, actions possibles, fournisseurs et dépendances.

Responsable et statut

Nommer le responsable métier et distinguer usage envisagé, pilote ou production ; vérifier l’autorisation.

Priorité d’examen

Examiner les usages prioritaires sans attendre l’exhaustivité et compléter l’inventaire au fil du temps.

Orange Restricted

3. Les responsabilités et les décisions

Un responsable métier interne pour chaque usage, même externalisé

Rôle	Contribution
Métier	Porte la finalité et la validation opérationnelle.
IT	Réalise l'intégration, la configuration et l'exploitation.
RSSI	Évalue les risques de sécurité et vérifie les protections.
Protection des données et juridique	Conseille sur les obligations, les impacts et les rôles juridiques.
Achats	Encadre les contrats et le suivi des fournisseurs.
Autorité désignée	Autorise selon délégation ; restreint ou suspend l'usage.

Orange Restricted

4. Analyser les risques et les impacts de chaque cas d'usage

Le risque comme outil de décision, pour l'organisation et les personnes

Risques pour l'entreprise

Analyser les scénarios susceptibles de compromettre la confidentialité, l'intégrité ou la disponibilité des informations et des services, ainsi que la fiabilité des résultats et des décisions appuyées par l'IA.

Évaluer leurs conséquences opérationnelles, financières, juridiques, contractuelles et réputationnelles, en intégrant les dépendances aux fournisseurs et aux services tiers.

Risques pour les personnes

Analyser les conséquences possibles des traitements de données personnelles et des usages de l'IA sur les personnes : atteinte à la vie privée, divulgation de données, surveillance disproportionnée, discrimination, décisions préjudiciables ou entrave à l'exercice des droits.

Tenir compte de la nature des données, des finalités du traitement, de son ampleur et de la vulnérabilité éventuelle des personnes concernées.

Appréciation et traitement des risques

Pour chaque scénario, apprécier la vraisemblance et la gravité des conséquences, en tenant compte des mesures existantes, des preuves de leur efficacité et des incertitudes.

Orange Restricted

5. Choisir les mesures de protection

Les référentiels guident; les preuves attestent

ISO/IEC 27001 & 27002

Système de management de la sécurité de l'information et recommandations de mesures adaptées au contexte.

ISO/IEC 42001 & ISO/IEC 27090

ISO/ IEC 42001: Cadre de management de l'IA couvrant gouvernance, gestion des risques et amélioration continue.
ISO/ IEC 27090: Recommandations pour comprendre, détecter et atténuer les menaces propres aux systèmes IA.

NIST SP 800-53 & NIST SP 800-218A

Catalogue de mesures de sécurité et de protection de la vie privée à sélectionner et adapter.
NIST SP 800-218A: pour les pratiques de développement sécurisé de l'IA générative

OWASP AISVS & ASVS

Exigences vérifiables pour la sécurité des systèmes IA et des applications web associées.

NIST AI RMF Playbook

Actions suggérées pour déployer le NIST AI RMF ; approche volontaire à adapter aux risques.

Cadre juridique applicable

AI Act, RGPD, LPD, OPDo et législations cantonales, selon leur applicabilité à l'usage.

Orange Restricted

6. Les conditions d’autorisation

Décider sur un périmètre défini et sur la base de protections dont l’état est vérifié

Outiller la décision

Présenter le périmètre, les évaluations et leurs limites, les mesures prévues et appliquées, les preuves disponibles et les risques résiduels.

Distinguer les protections planifiées des protections vérifiées.

Options et conditions

L’autorité désignée peut autoriser, sous conditions, restreindre, reporter ou refuser.

Préciser les données et fonctions permises, restrictions, échéance, responsabilités et critères de suspension.

Orange Restricted

7. Surveiller et réévaluer

Les conditions d’autorisation doivent rester valables dans le temps

Observer

Suivre les incidents, la qualité, les accès, et l’usage du système par ses utilisateurs.

Examiner

Réexaminer la finalité, les données, le modèle, les fonctionnalités, les permissions, les connecteurs et les fournisseurs.

Actualiser

Maintenir l’inventaire, les risques, les règles et la formation lorsque le contexte évolue.

Décider

Maintenir, restreindre, suspendre ou retirer l’usage selon les changements, les écarts et les incidents.

Clore l’usage

Révoquer les accès et traiter les données conformément aux exigences applicables.

Orange Restricted

Savoir pourquoi l’IA est utilisée

Élaborer et maintenir une vue des usages, de leurs propriétaires et des connexions

Cas d’usage et finalité

Quel besoin métier l’IA sert-elle ? À quelle étape du processus intervient-elle et pourquoi ?

Outils et modèles

Recenser outils publics, fonctions intégrées, applications internes, agents et modèles utilisés ou envisagés.

Responsables et utilisateurs

Identifier propriétaire métier, équipes utilisatrices, responsables techniques et parties prenantes de contrôle.

Données et connexions

Repérer données, sensibilité, connecteurs, API, identités et fournisseurs auxquels l’IA accède.

Autonomie et cycle de vie

Établir si l’IA conseille, génère ou agit et noter son stade : envisagé, pilote ou production.

Orange Restricted



Partie 3

Évaluer l’IA

Orange Restricted

Analyser les risques

Identifier les menaces et leur impact pour choisir les mesures les plus appropriées



AIPD*, si requise

Lorsque les traitements de données personnelles sont susceptibles d'engendrer un risque élevé pour les personnes, réaliser l'analyse d'impact avant la mise en œuvre du système et définir les mesures de réduction du risque associées au traitement.
L'AIPD est menée avant le déploiement, pas après un incident.



Threat modeling

Examiner l'architecture du système, les flux de données, les accès et les connexions pour anticiper les scénarios d'attaque ou de détournement du système.
Le modèle alimente l'analyse de risques et oriente les tests à mener avant le déploiement.
Il doit être revu à chaque évolution significative du système.



Analyse de risques

Identifier les scénarios qui pourraient menacer la confidentialité, l'intégrité ou la disponibilité, estimer leurs conséquences et leur vraisemblance, puis décider des mesures à mettre en place.
Méthodes usuelles : EBIOS RM et ISO/IEC 27005. Pour les risques propres à l'IA, ISO/IEC 23894 est pertinente.

* AIPD: Analyse d'impact relative à la protection des données personnelles

Orange Restricted

Tester le système d'IA

Les tests observent une version du système dans des conditions définies



Test d’intrusion (pentest)

Dans un cadre autorisé, rechercher si des vulnérabilités permettent de contourner les protections et évaluer les conséquences possibles.
Le périmètre, les règles du test et les personnes à prévenir sont définis à l'avance, avant toute action.



Red teaming IA

Éprouver le système avec des scénarios adversariaux : instruction détournée, fuite de données, action non autorisée d'un agent connecté à d'autres outils.
Cette approche va au-delà du pentest classique, car elle cible le comportement du modèle.



Portée des résultats

Les essais révèlent des défaillances et vérifient des corrections, mais ne garantissent pas le comportement du système dans d'autres situations. Les conclusions dépendent de la version, des données et des conditions testées. Un examen complémentaire reste nécessaire pour évaluer la conformité globale.

Orange Restricted

Auditez vos systèmes d’IA et leurs usages.

Examinez leur gouvernance et leur conformité à partir de critères définis et de preuves documentées.



Gouvernance

Vérifier la clarté des rôles et responsabilités, la traçabilité des décisions ainsi que le suivi des usages de l’IA. L’audit confronte les règles établies aux pratiques observées et aux preuves disponibles.



Conformité

Examiner l’application des exigences légales et réglementaires, des engagements contractuels et des politiques ou référentiels internes. L’analyse s’appuie sur des critères définis et une documentation vérifiable.



Ce que cela ne prouve pas

Un audit favorable ne garantit pas l’exactitude de chaque réponse.

La conclusion dépend du périmètre, des critères, des preuves et de l’objectivité des auditeurs.
Des failles non recherchées peuvent subsister.

Orange Restricted

Limites de chaque méthode

Chaque résultat n'est valable que dans son périmètre et ses conditions.

Analyse de risques

Les conclusions dépendent des hypothèses, du périmètre et des preuves retenus. Un scénario non identifié n'est pas couvert par les mesures.

Tests

Les résultats valent pour une version, des données et des conditions précises. Une mise à jour peut modifier le comportement du système.

Audit

La conclusion dépend des critères, des preuves examinées et de l'objectivité des auditeurs. Un audit favorable ne garantit pas l'absence de failles.

Surveillance continue

Réévaluer après chaque changement important et compléter les analyses, tests et audits par la surveillance en exploitation.

Orange Restricted

Pour toute question concernant la gouvernance de l'IA, n'hésitez pas à nous contacter :

Soheila Firouz Kouhi
soheila.firouzkouhi@orangecyberdefense.com

Équipe Advisory & Audit
advisory@ch.orangecyberdefense.com



Orange Restricted



Merci

Orange Restricted