Panorama de la sécurité de l’IA

Une vue documentée des cadres publics de sécurité, des informations divulguées au niveau des modèles et des approches institutionnelles face aux risques liés à l’IA avancée.

Publié le: 9 août 2026 · Date des éléments probants: 9 août 2026

Éditeur: AI Governance World LLC · aigovernanceworld.com

Résumé exécutif

Le panorama de lancement associe les cadres publics de gestion des risques aux fiches de modèles, fiches système, politiques des développeurs et incidents vérifiés. Ces couches de preuves répondent à des questions différentes. Un cadre public décrit une approche de gestion des risques ; une information publiée par un développeur indique ce que son éditeur déclare ; une évaluation indépendante teste un modèle selon une méthode documentée ; et un dossier d'exécution précise ce qu'une autorité a constaté ou exigé.
Le corpus contient actuellement de la documentation de modèles rédigée par les développeurs, mais aucune évaluation indépendante examinée et reliée aux modèles du corpus de lancement. Cette absence est affichée explicitement et n'est pas remplacée par une note indirecte. Aucun modèle ni aucune entreprise ne reçoit de score global de sécurité sur la seule base de documents de première partie.

Méthodologie

Les preuves admissibles comprennent les cadres officiels et dossiers d'exécution de niveau 1, les évaluations universitaires de niveau 2 assorties de méthodes attribuables, ainsi que les fiches de modèles, fiches système et rapports de sécurité de niveau 3 rédigés par les développeurs. Les médias peuvent faciliter la découverte, mais ne constituent jamais une preuve principale.
Les déclarations des développeurs sont identifiées par leur éditeur et leur type de source, et restent rattachées au modèle ou à l'entreprise comme informations de première partie. Les évaluations indépendantes sont enregistrées dans le registre des évaluations de modèles avec l'évaluateur, la date, la méthodologie, les résultats, les constatations et les sources. Les deux catégories de preuves sont présentées séparément et ne sont jamais fusionnées en une note de sécurité non étayée.

Cadres publics de sécurité et de gestion des risques

Les cadres sont comparés selon leur couverture du cycle de vie, la gouvernance, l'identification, la mesure et le traitement des risques, la surveillance, la gestion des incidents et l'amélioration continue. Le NIST AI RMF et la norme ISO/IEC 42001 remplissent des fonctions différentes : le corpus enregistre leur champ d'application officiel et leurs caractéristiques d'évaluation de la conformité, sans les traiter comme des certifications interchangeables.
Les instruments de l'OCDE et de l'UNESCO apportent un contexte normatif et politique. Leur présence dans le corpus n'établit ni leur adoption ni leur mise en œuvre par une entreprise ou un pays donné en l'absence de preuves distinctes.

Informations des développeurs sur les modèles et les systèmes

Les fiches de modèles conservent les capacités déclarées, les approches d'évaluation, les limites, les mesures d'atténuation, les conditions d'accès et les dates de mise à jour, sans compléter les champs non documentés. L'éditeur est affiché afin que le lecteur puisse identifier une preuve de première partie.
La déclaration d'un développeur selon laquelle des essais ont eu lieu constitue une information vérifiable. Elle ne confirme pas à elle seule, de manière indépendante, le résultat, l'exhaustivité du jeu de tests ou l'efficacité des mesures d'atténuation.

Limites de l’assurance

Un cadre publié atteste un engagement de gouvernance, et non la preuve de sa mise en œuvre ou de la sécurité d’un modèle. Une fiche de modèle constitue une preuve de divulgation, non une certification. Une évaluation indépendante est limitée par sa méthode et la version testée, tandis qu’une constatation réglementaire est limitée par sa juridiction et son action juridique.
Pour ces raisons, le Panorama de la sécurité publie des volumes de preuves et des catégories de provenance, et non un classement global de sécurité.

Quatre couches de preuves distinctes

Le rapport distingue les cadres publics, les informations des développeurs, les évaluations indépendantes et les incidents ou mesures d’exécution. Un même dossier peut éclairer plusieurs questions analytiques, mais sa provenance ne change pas. Une fiche système, par exemple, reste une déclaration du développeur même lorsqu’elle décrit des essais approfondis.
Cette structure évite de confondre le volume d’informations divulguées avec une sécurité démontrée de façon indépendante, et empêche de généraliser un incident isolé en une conclusion non étayée sur l’ensemble d’un modèle.

Registre des évaluations indépendantes

Une évaluation indépendante exige un évaluateur distinct du développeur du modèle, une date d’évaluation, une méthodologie documentée, des constatations attribuables et une source admissible. Les dossiers qui ne précisent pas ces éléments restent exclus du nombre d’évaluations indépendantes examinées.
Le graphique de lancement indique directement le nombre issu du registre des évaluations de modèles. Zéro signifie qu’aucun dossier lié n’a achevé la vérification ; cela ne signifie pas qu’il n’existe aucune recherche externe.

Incidents, réponse et exécution

Les dossiers d’incidents conservent la date de l’événement, sa gravité, son statut, le résumé de son impact, la réponse apportée et la source principale. Les mesures d’exécution restent attribuées à l’autorité compétente et sont reliées à la chronologie. Ces dossiers peuvent éclairer une évaluation ultérieure, mais la plateforme ne déduit ni causalité technique ni risque à l’échelle du modèle au-delà des preuves officielles.

Graphiques et données

Inventaire des preuves de sécurité

Les cadres, les informations sur les modèles et les incidents constituent des catégories de dossiers distinctes.

Dossiers de cadres publics: 4
Dossiers d’informations sur les modèles: 5
Dossiers d’incidents confirmés: 1
Le corpus contient 4 dossiers de cadres publics, 5 dossiers d’informations sur les modèles et 1 dossier d’incident confirmé.

Information du développeur et évaluation examinée

La documentation rédigée par les développeurs reste séparée des évaluations de modèles vérifiées.

Sources de modèles rédigées par les développeurs: 5
Évaluations de première partie examinées: 0
Évaluations indépendantes examinées: 0
5 dossiers de modèles utilisent des sources principales rédigées par les développeurs ; 0 évaluation de première partie et 0 évaluation indépendante ont achevé l'examen.

Modes d'accès aux modèles documentés

Le mode d'accès provient du dossier de modèle vérifié et ne constitue pas une note de sécurité.

API: 3
Poids ouverts: 2
Dossiers de modèles vérifiés regroupés selon le mode d'accès documenté.

Références

1. Cadre de gestion des risques liés à l'intelligence artificielle (AI RMF 1.0). Source: Institut national des normes et de la technologie. 26 janv. 2023. Titre original de la source: Artificial Intelligence Risk Management Framework (AI RMF 1.0). https://www.nist.gov/publications/artificial-intelligence-risk-management-framework-ai-rmf-10

2. ISO/IEC 42001:2023 — Système de management de l'intelligence artificielle. Source: Organisation internationale de normalisation. 1 déc. 2023. Titre original de la source: ISO/IEC 42001:2023 — Artificial intelligence management system. https://www.iso.org/standard/42001.html

3. Principes de l'OCDE sur l'intelligence artificielle. Source: Organisation de coopération et de développement économiques. 3 mai 2024. Titre original de la source: OECD AI Principles. https://oecd.ai/en/ai-principles

4. Recommandation sur l'éthique de l'intelligence artificielle. Source: UNESCO. 23 nov. 2021. Titre original de la source: Recommendation on the Ethics of Artificial Intelligence. https://www.unesco.org/en/legal-a f airs/recommendation-ethics-artificial-intelligence

5. La NHTSA annonce une ordonnance par consentement avec Cruise après la communication incomplète d'un accident impliquant un piéton. Source: Administration nationale de la sécurité routière. 30 sept. 2024. Titre original de la source: NHTSA Announces Consent Order with Cruise A f er Company Failed to Fully Report Crash Involving Pedestrian. https://www.nhtsa.gov/press-releases/consent-order-cruise-crash-reporting

6. Fiche système de GPT-5.6. Source: OpenAI. 9 juil. 2026. Titre original de la source: GPT-5.6 System Card. https://deploymentsafety.openai.com/

7. Fiches système des modèles Anthropic. Source: Anthropic. 1 juin 2026. Titre original de la source: Anthropic model system cards. https://www.anthropic.com/system-cards

8. Fiche du modèle Gemini 3.5 Flash. Source: Google DeepMind. 19 mai 2026. Titre original de la source: Gemini 3.5 Flash Model Card. https://deepmind.google/models/model-cards/gemini-3-5-flash/

9. Documentation du modèle Llama. Source: Meta. 9 août 2026. Titre original de la source: Llama model documentation. https://ai.meta.com/llama/get-started/

10. Fiche du modèle NVIDIA Nemotron 3.5 Nano. Source: NVIDIA. 15 déc. 2025. Titre original de la source: NVIDIA Nemotron 3.5 Nano model card.

https://build.nvidia.com/nvidia/nemotron-3.5-nano-30b-a3b/modelcard