

Panorama siguranței IA

O prezentare documentată a cadrelor publice de siguranță, a informațiilor publicate la nivel de model și a abordărilor instituționale privind riscurile IA avansate.

Publicat: 9 aug. 2026 · Data dovezilor: 9 aug. 2026

Editor: AI Governance World LLC · aigovernanceworld.com

Rezumat executiv

Panorama inițială reunește cadre publice de management al riscurilor, fișe de modele, fișe de sistem, politici ale dezvoltatorilor și evidențe verificate ale incidentelor. Aceste straturi de dovezi răspund unor întrebări diferite. Un cadru public descrie o abordare de management al riscurilor; o informare a dezvoltatorului consemnează ceea ce declară editorul; o evaluare independentă testează un model printr-o metodă documentată; iar o evidență de aplicare consemnează ceea ce o autoritate a constatat sau a impus. Corpusul conține în prezent documentație de model redactată de dezvoltatori, dar nicio evaluare independentă analizată și asociată modelelor din corpusul inițial. Această absență este afișată explicit și nu este completată cu un calificativ indirect. Niciun model și nicio companie nu primește un scor agregat de siguranță exclusiv pe baza documentației proprii.

Metodologie

Dovezile eligibile includ cadre oficiale și evidențe de aplicare de Nivel 1, evaluări academice de Nivel 2 cu metode atribuibile și fișe de modele, fișe de sistem și rapoarte de siguranță de Nivel 3 redactate de dezvoltatori. Presa poate oferi context pentru identificarea surselor, dar nu reprezintă niciodată dovadă primară.

Afirmațiile dezvoltatorilor sunt identificate prin editor și tipul sursei și rămân asociate evidenței modelului sau companiei ca informații proprii. Evaluările independente sunt păstrate în registrul evaluărilor de modele împreună cu evaluatorul, data evaluării, metodologia, rezultatele, constatările și sursele. Cele două clase de dovezi sunt raportate separat și nu sunt niciodată combinate într-un calificativ de siguranță fără suport.

Cadre publice de siguranță și risc

Cadrele sunt comparate în funcție de acoperirea ciclului de viață, guvernanta, identificarea, măsurarea și tratarea riscurilor, monitorizare, gestionarea incidentelor și îmbunătățire continuă. NIST AI RMF și ISO/IEC 42001 au funcții diferite: corpusul înregistrează domeniul lor oficial de aplicare și caracteristicile evaluării conformității, fără a le trata ca certificări interschimbabile.

Instrumentele OCDE și UNESCO adaugă un context normativ și de politici publice. Prezența lor în corpus nu demonstrează adoptarea sau implementarea de către o anumită companie ori țară în lipsa unor dovezi distincte.

Informările dezvoltatorilor privind modelele și sistemele

Evidențele modelelor păstrează capabilitățile declarate, abordările de evaluare, limitările, măsurile de atenuare, condițiile de acces și datele actualizărilor, fără a completa câmpurile nedocumentate. Editorul este afișat pentru ca cititorii să poată identifica dovezile proprii.

Declarația unui dezvoltator conform căreia au fost efectuate teste este o afirmație verificabilă. Ea nu confirmă independent, prin ea însăși, rezultatul, caracterul complet al setului de teste sau eficacitatea măsurilor de atenuare.

Limitele asigurării

Un cadru publicat dovedește un angajament de guvernanta, nu implementarea sau siguranța modelului. O fișă de model reprezintă dovada unei informații, nu o certificare. O evaluare independentă este limitată de metoda

și versiunea testată, iar o constatare a autorității de reglementare este limitată de jurisdicția și acțiunea sa juridică.

Din aceste motive, Panorama siguranței publice numărul dovezilor și categoriile de proveniență, nu un clasament agregat de siguranță.

Patru straturi distincte de dovezi

Raportul separă cadrele publice, informările dezvoltatorilor, evaluările independente și incidentele sau măsurile de aplicare. O evidență poate contribui la mai multe întrebări analitice, însă proveniența sa nu se schimbă. De exemplu, o fișă de sistem rămâne o afirmație a dezvoltatorului chiar dacă descrie testări ample. Această structură împiedică confundarea volumului informărilor cu siguranța demonstrată independent și împiedică generalizarea unui incident izolat într-o concluzie fără suport privind întregul model.

Registrul evaluărilor independente

O evaluare independentă necesită un evaluator distinct de dezvoltatorul modelului, o dată a evaluării, o metodologie documentată, constatări atribuibile și o sursă eligibilă. Evidențele care nu identifică aceste elemente rămân în afara numărului de evaluări independente analizate.

Graficul inițial prezintă direct numărul curent din registrul evaluărilor de modele. Zero înseamnă că nicio evidență asociată nu a finalizat verificarea; nu înseamnă că nu există cercetări externe.

Incidente, răspuns și aplicare

Evidențele incidentelor păstrează data evenimentului, gravitatea, starea, rezumatul impactului, răspunsul și sursa primară. Măsurile de aplicare rămân atribuite autorității responsabile și sunt asociate prin cronologie. Aceste evidențe pot fundamenta evaluări ulterioare, dar platforma nu deduce cauzalitate tehnică sau risc la nivelul întregului model dincolo de dovezile oficiale.

Grafice și date

Inventarul dovezilor privind siguranța

Cadrele, informările despre modele și incidentele sunt clase distincte de evidențe.

Evidențe de cadre publice: 4

Evidențe de informări despre modele: 5

Evidențe de incidente confirmate: 1

Corpusul conține 4 evidențe de cadre publice, 5 evidențe de informări despre modele și 1 evidență de incident confirmat.

Informarea dezvoltatorului și evaluarea analizată

Documentația redactată de dezvoltatori este menținută separat de evaluările verificate ale modelelor.

Surse de modele redactate de dezvoltatori: 5

Evaluări proprii analizate: 0

Evaluări independente analizate: 0

5 evidențe de modele folosesc surse primare redactate de dezvoltatori; 0 evaluări proprii și 0 evaluări independente au finalizat analiza.

Moduri documentate de acces la modele

Modul de acces este raportat din evidența verificată a modelului și nu reprezintă un calificativ de siguranță.

API: 3

Ponderi deschise: 2

Evidențe verificate ale modelelor grupate după modul documentat de acces.

Referințe

1. Cadrul de management al riscurilor asociate inteligenței artificiale (AI RMF 1.0). Sursă: Institutul Național de Standarde și Tehnologie. 26 ian. 2023. Titlul original al sursei: 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 — Sistem de management al inteligenței artificiale. Sursă: Organizația Internațională de Standardizare. 1 dec. 2023. Titlul original al sursei: ISO/IEC 42001:2023 — Artificial intelligence management system. <https://www.iso.org/standard/42001.html>
3. Principiile OCDE privind inteligența artificială. Sursă: Organizația pentru Cooperare și Dezvoltare Economică. 3 mai 2024. Titlul original al sursei: OECD AI Principles. <https://oecd.ai/en/ai-principles>
4. Recomandarea privind etica inteligenței artificiale. Sursă: UNESCO. 23 nov. 2021. Titlul original al sursei: Recommendation on the Ethics of Artificial Intelligence.
<https://www.unesco.org/en/legal-affairs/recommendation-ethics-artificial-intelligence>
5. NHTSA anunță un ordin prin consimțământ cu Cruise după raportarea incompletă a unui accident în care a fost implicat un pieton. Sursă: Administrația Națională pentru Siguranța Traficului Rutier. 30 sept. 2024. Titlul original al sursei: NHTSA Announces Consent Order with Cruise After Company Failed to Fully Report Crash Involving Pedestrian. <https://www.nhtsa.gov/press-releases/consent-order-cruise-crash-reporting>
6. Fișa de sistem GPT-5.6. Sursă: OpenAI. 9 iul. 2026. Titlul original al sursei: GPT-5.6 System Card.
<https://deploymentsafety.openai.com/>
7. Fișele de sistem ale modelelor Anthropic. Sursă: Anthropic. 1 iun. 2026. Titlul original al sursei: Anthropic model system cards. <https://www.anthropic.com/system-cards>
8. Fișa modelului Gemini 3.5 Flash. Sursă: Google DeepMind. 19 mai 2026. Titlul original al sursei: Gemini 3.5 Flash Model Card. <https://deepmind.google/models/model-cards/gemini-3-5-flash/>
9. Documentația modelului Llama. Sursă: Meta. 9 aug. 2026. Titlul original al sursei: Llama model documentation. <https://ai.meta.com/llama/get-started/>
10. Fișa modelului NVIDIA Nemotron 3.5 Nano. Sursă: NVIDIA. 15 dec. 2025. Titlul original al sursei: NVIDIA Nemotron 3.5 Nano model card. <https://build.nvidia.com/nvidia/nemotron-3.5-nano-30b-a3b/modelcard>