SafeByte
Cybersecurity Services. Smarter. Better.

STUDIU STRATEGIC

Inteligenta Artificiala in Securitatea Cibernetica:

Starea Actuala, Rezultate Demonstrate si Implicatii Strategice

Autor: Teodor Lupan
Safebyte Consulting S.R.L.

Data: Martie 2026
Clasificare: Public / Uz Intern

Bazat pe analiza a 21+ proiecte de cercetare, 30+ lucrari stiintifice (arXiv/ACL/ACM CCS), competitia DARPA AIxCC 2024-2025, si proiecte open-source active.

Sumar Executiv

⚠ MESAJ CHEIE

Inteligenta artificiala a transformat fundamental securitatea cibernetica in ultimii 18 luni. Sisteme autonome demonstreaza acum capacitati de penetration testing comparabile cu cele ale specialistilor umani, la o fractiune din cost si timp. Aceasta evolutie are implicatii directe pentru securitatea nationala, infrastructura critica si apararea cibernetica a statului.

Prezentul studiu analizeaza starea actuala a inteligentei artificiale aplicata in securitatea cibernetica, pe baza a peste 21 de proiecte de cercetare active, 30+ lucrari stiintifice publicate in perioada 2024–2026 in conferinte de referinta (ACL, ACM CCS, ICSE, arXiv), competitia DARPA AIxCC (buget: 29,5 milioane USD), si produse comerciale operationale. Concluziile sunt verificabile prin referintele furnizate.

Trei constatari critice: (1) Sistemele AI au demonstrat in competitii controlate capacitatea de a identifica 86% din vulnerabilitatile sintetice si de a genera patch-uri pentru 68% dintre acestea, la un cost mediu de 152 USD per task. (2) In noiembrie 2025, Anthropic a raportat primul atac cibernetic autonom la scara larga, in care agenti AI au executat 80–90% din operatiunea ofensiva cu supraveghere umana minimala. (3) Costul unui breach mediu global a atins 4,88 milioane USD (IBM, 2024), iar atacurile bazate pe AI au crescut cu 89% (CrowdStrike Global Threat Report 2026).

1. Context: De ce conteaza acum

Securitatea cibernetica traverseaza un punct de inflexiune. Convergenta a trei factori creeaza o urgenta strategica:

1.1 Explozia suprafetei de atac

Adoptarea accelerata a cloud computing, DevOps si IoT a creat medii in continua schimbare. Conform Verizon Data Breach Investigations Report 2025, peste doua treimi din breaches au implicat vulnerabilitati nepatate de mai mult de 90 de zile, chiar si in organizatii care efectuasera recent evaluari de securitate. Check Point raporteaza o medie de 1.925 de atacuri pe saptamana per organizatie in Q1 2025 — aproximativ 275 de atacuri pe zi.

1.2 Democratizarea capabilitatilor ofensive

Inteligenta artificiala a redus dramatic bariera de intrare pentru atacuri sofisticate. Microsoft raporteaza urmarirea a peste 600 de grupuri nation-state distincte la nivel global. CrowdStrike Global Threat Report 2026 indica o crestere de 89% a atacurilor facilitate de AI, cu mentiunile ChatGPT pe forumuri criminale crescand cu 550%. Peste 90 de organizatii au avut instrumente AI legitime exploatate pentru generarea de comenzi malitioase.

[Ref: CrowdStrike, "2026 Global Threat Report", februarie 2026; Microsoft, "2025 Digital Defense Report", octombrie 2025]

1.3 Primul atac cibernetic autonom documentat

INCIDENT CRITIC — Noiembrie 2025

Anthropic a raportat public detectarea si neutralizarea primei campanii de spionaj cibernetic orchestrata autonom de agenti AI. Actori chinezi au exploatat instrumentul Claude Code, creand un agent ofensiv semi-autonom care a vizat 30 de companii si agentii guvernamentale. Agentul AI a executat 80–90% din operatiunea ofensiva, iar operatorii umani au trecut de la rolul de executanti la cel de supervizori strategici.

Acest incident, analizat de Institute for AI Policy and Strategy (IAPS), marcheaza o tranzitie fundamentala: de la atacuri asistate de AI la atacuri conduse de AI cu supraveghere umana minimala. IAPS avertizeaza ca **"aceste capabilitati autonome vor prolifera si vor permite actorilor mai putin sofisticati sa conduca operatiuni complexe la viteze mai mari."**

[Ref: Anthropic, "Disrupting the first reported AI-orchestrated cyber espionage campaign", nov. 2025; IAPS, "The Emergence of Autonomous Cyber Attacks", nov. 2025; Lawfare, "Fighting AI Cyberattacks Starts With Knowing They're Happening", feb. 2026]

2. Starea Actuala: Sisteme AI Demonstrabile

2.1 DARPA AI Cyber Challenge (AIxCC) — Validare institutionala

Cel mai riguros test al capabilitatilor AI in securitatea cibernetica este competitia DARPA AIxCC (2023–2025), cu un buget total de 29,5 milioane USD in premii si colaborare cu Anthropic, Google, OpenAI si Microsoft. Finala a avut loc la DEF CON 33 (august 2025).

Rezultate finale AIxCC (august 2025, DEF CON 33):

Metrica	Semifinala (aug. 2024)	Finala (aug. 2025)	Evolutie
Vulnerabilitati identificate	37% din sintetice	86% din 63 sintetice	+132%
Vulnerabilitati patate	25% din identificate	68% din identificate	+172%
Zero-days reale descoperite	Nu s-a raportat	18 (6 in C, 12 in Java)	NOU
Cost mediu per task	Nu s-a raportat	~152 USD	Demonstrat
Timp mediu per vulnerabilitate	Nu s-a raportat	~45 minute	Demonstrat
Echipe finaliste	42 participante	7 finaliste	Selectie riguroasa

Castigatori: Team Atlanta (Georgia Tech + Samsung Research + KAIST + POSTECH) — premiu 4M USD; Trail of Bits — 3M USD; Theori — 1,5M USD. Toate cele 7 sisteme finaliste au fost publicate ca open-source. DARPA Director Stephen Winchell: “Traim intr-o lume cu schele digitale antice care sustin totul. [...] Este o problema care depaseste scara umana.”

[Ref: DARPA, “AIxCC Final Competition Results”, august 2025, darpa.mil/news/2025/aixcc-results; CyberScoop, 11 aug. 2025; Cybersecurity Dive, 8 aug. 2025]

2.2 Penetration Testing Autonom — Rezultate academice verificabile

In paralel cu AIxCC, comunitatea academica a produs sisteme de penetration testing AI cu rezultate masurabile:

Sistem	Publicare	Rezultat cheie	Referinta
Excalibur/PentestGPT v2	arXiv:2602.17622, feb. 2026	**91% completare pe XBOW** (104 provocari web); 12/13 masini rootate HTB; Top-100 HTB Season 8 live	NTU Singapore
MAPTA	arXiv:2508.20816, aug. 2025	76,9% succes XBOW; 100% SSRF; 100% Misconfiguration; 83% SQLi; 83% Broken Auth	Multi-agent web pentest
CHECKMATE	arXiv:2512.11143, dec. 2025	Depaseste Claude Code (cel mai capabil agent general); Classical Planning + LLM	Wang et al.
PenForge	arXiv:2601.06910, ian. 2026 (ICSE-NIER 2026)	30% rata exploatare pe CVE-Bench (40 CVE reale); 3x improvement vs SOTA in setting zero-day	Dynamic agent construction

Sistem	Publicare	Rezultat cheie	Referinta
CAI Framework	arXiv:2504.06017, apr. 2025	Top-20 global CTF; 156x mai ieftin decat testare manuala; 3600x mai rapid pe task-uri specifice	Open-source, MIT, 2000+ stars
AutoPentester	arXiv:2510.05605, oct. 2025	+27% succes vs PentestGPT; 3,93/5 rating pentesteri profesionisti	RAG reduces hallucinari
CVE-Genie	arXiv:2509.01835, sept. 2025	51% rata reproducere automata CVE (428/841 CVE-uri); cost mediu 2,77 USD/CVE	267 proiecte, 141 CWE

[Toate referintele sunt lucrari publice, accesibile pe arxiv.org si in proceedings-urile conferintelor mentionate]

2.3 Modele specializate de securitate

Pe langa sistemele agentice, au aparut modele de limbaj specializate pe securitate cibernetica:

- **Foundation-Sec-8B (Cisco Foundation AI)** — Model de 8 miliarde de parametri care egalizeaza performanta modelelor de 70B pe benchmark-uri de securitate cibernetica. Open-weight, disponibil pe HuggingFace. Relevant: ruleaza pe hardware local, fara dependenta cloud.
- **Primus (Trend Micro)** — Datasets de pre-antrenament si fine-tuning pentru securitate cibernetica. +15,9% imbunatatire agregata, +15,8% pe certificarea CISSP. Open-source pe HuggingFace.
- **HackSynth-GRPO** — Reinforcement Learning aplicat pe Llama-3.1-8B: +53% salt absolut pe crypto CTF tasks. Demonstreaza ca modele mici pot fi dramatic imbunatatite prin RL pe securitate.

[Ref: Foundation-Sec-8B: HuggingFace fdtn-ai/Foundation-Sec-8B; Primus: arXiv:2502.11191, feb. 2025; HackSynth-GRPO: arXiv:2506.02048, iun. 2025]

3. Produse Comerciale Operationale (2025–2026)

Tranzitia de la cercetare la produse comerciale s-a accelerat semnificativ. In ianuarie 2026, un review operational (Spark42.tech) constata ca ecosistemul "s-a mutat de la demo-uri si wrapper-e de prompt la automatizare masurabila si reproductibila."

Produs	Model	Pret orientativ	Capabilitati
Aikido Attack (AI Pentest)	Agenti autonomi, 60+	De la 800 EUR	Full pentest automat, raport SOC2/ISO27001, re-test instant
Penligent	Agentic AI, autonom	Enterprise	Zero-setup, deep reasoning, pentest continuu
XBOW	AI agents	Enterprise	Unit testing securitate, scenarii customizabile per endpoint
RunSybil (Sybil)	Attack Surface Management	Enterprise	Simulare perimetru extern, Shadow IT discovery, Black Box Recorder
Cobalt	Hybrid AI+human	Enterprise	AI pentru scanning basic, umani pentru business logic; rapoarte semnate
NodeZero (Horizon3.ai)	Autonomous pentesting	Enterprise	Traversare dinamica retea, chaining real de exploit-uri
AutoSecT	Full-stack AI	Enterprise	Network + cloud + web + mobile + API intr-o platforma

Tendinta dominanta: trecerea de la **testare periodica (1-2x/an) la testare continua.** Raportul Pentera "State of Pentesting 2025" indica ca doar 29% din organizatii efectueaza pentesting primar pentru conformitate regulamentara — restul il folosesc pentru validarea controalelor (28%) si prioritizarea investitiilor de securitate (32%).

Estimarea de cost: Conform analizelor, un breach mediu costa 4,88 milioane USD (IBM, 2024). Un pentest automat AI costa intre 800 EUR si cateva mii EUR. Prevenirea unui singur breach acopera costul instrumentului pentru sute de ani.

[Ref: Spark42.tech, "AI-Powered Penetration Testing Tools in 2026", ian. 2026; Aikido.dev/attack/aipentest; Pentera, "The State of Pentesting 2025"]

4. Implicatii Strategice pentru Securitatea Nationala

4.1 Asimetria atac–aparare se adanceste

AI amplifica capabilitatile atat ofensive cat si defensive, dar cu un avantaj temporal pentru atacatori. Atacatorii adopta AI mai rapid deoarece nu au constrangeri de reglementare, burocratice sau de aprovizionare. IAPS avertizeaza explicit: “Acest lucru poate deplasa avantajul catre atacatori pana cand capabilitatile defensive sunt implementate la scara.”

Concret: un grup APT cu acces la un model LLM open-source (ex: Llama, Mistral, DeepSeek) si un framework agentic (ex: CAI, care este MIT-licensed si public) poate construi un agent ofensiv autonom in zile, nu luni. Componenta umana se reduce de la **operator la supervizor** — exact modelul documentat de Anthropic in noiembrie 2025.

4.2 Infrastructura critica — tinta prioritara

DARPA a construit AIxCC specific pentru securizarea software-ului open-source care sustine infrastructura critica: spitale, retele energetice, sisteme de apa, utilitati. CISA, FBI si NSA au confirmat ca grupul Volt Typhoon (asociat RPChineze) a mentinut acces in sisteme IT ale sectorului energetic, comunicatii, apa si transport din SUA timp de peste 5 ani.

Microsoft Digital Defense Report 2025 constata: “Atacurile cibernetice nu mai sunt probleme IT izolate; ele modeleaza economii, geopolitica si increderea publica.” Raportul identifica sectoarele IT, guvernamental si academic-cercetare ca cele mai vizate.

[Ref: CISA Advisory on Volt Typhoon, 2024; Microsoft Digital Defense Report 2025; DARPA AIxCC mission statement]

4.3 Convergenta nation-state si cybercrime

SecurityWeek “Cyber Insights 2026” evidentiaza estomparea liniilor intre activitatea statala si cea criminala: state care folosesc instrumente cybercrime, coopteaza grupuri, sau permit “moonlighting.” Acest model hibrid face comportamentul atacatorilor imprevizibil si creste riscul de daune colaterale.

CEO Delinea, Art Gilliland: “AI a facut semnificativ mai usor pentru oricine sa execute atacuri cibernetice sofisticate cu mai putine resurse. Pentru jucatorii nation-state mai mici, care nu puteau concura cu marile puteri pana acum, asta nivelizeaza efectiv terenul de joc.”

[Ref: SecurityWeek, “Cyber Insights 2026: Cyberwar and Rising Nation State Threats”, feb. 2026; CrowdStrike Global Threat Report 2026]

5. Ce fac statele avansate — si ce nu facem noi

5.1 Investitii strategice documentate

Tara / Entitate	Initiativa	Investitie / Rezultat
SUA (DARPA + ARPA-H)	AI Cyber Challenge (AIxCC)	29,5M USD premii + 1,4M USD tranzitie. 7 CRS open-source pentru infrastructura critica
SUA (DARPA)	AIxCC tech transition	+200.000 USD per echipa pentru integrare in software de infrastructura critica
Marea Britanie (MoD)	Cyber Security Operations Centre (CSOC)	Comanda dedicata de razboi cibernetic, operatiuni ofensive si defensive
Anthropic + Google + OpenAI + Microsoft	Suport AIxCC	350.000 USD credite LLM per companie + suport tehnic
Cisco Foundation AI	Foundation-Sec-8B	Model open-weight specializat securitate, ruleaza local
Trend Micro (Japonia)	Primus datasets	Datasets open-source de antrenament securitate cibernetica
A16Z (Andreessen Horowitz)	Investitii in AI pentesting	Raport strategic “Next-Gen Pentesting: AI Empowers the Good Guys” (iun. 2025)
Alias Robotics (Spania)	CAI Framework	Open-source MIT, 2000+ stars, suporta 300+ modele AI

5.2 Cadrul de conformitate — in tranzitie

Cadrul regulatory este in urma realitatii tehnologice. Andreessen Horowitz (a16z) constata: “In industrii reglementate, majoritatea cadrelor de conformitate (SOC 2, PCI, ISO 27001) asteapta un pentest condus de om. Sistemele autonome nu se incadreaza inca in acest model. Adoptatorii timpurii gestioneaza asta pragmatic: instrumente next-gen in spate, un pentest manual pe an pentru auditori. Pe termen lung, e probabil sa vedem recunoastere formala a testarii AI.”

Aceasta tranzitie creeaza o **fereastra de oportunitate** pentru state care se pozitioneaza acum. Romania, prin NIS2 transpus, PCI DSS v4.0 obligatoriu, si directivele BNR pentru sistemul bancar, are cadrul regulator care poate fi extins pentru a incorpora capabilitati AI de testare.

[Ref: a16z, “Next-Gen Pentesting: AI Empowers the Good Guys”, iun. 2025]

6. Tendinte si Proiectii pe Termen Scurt (2026–2028)

6.1 Tendinte confirmate (in desfasurare)

- **De la periodic la continuu:** Pentesting-ul evolueaza de la exercitii anuale la validare continua de securitate. Omdia (Informa Tech) proiecteaza ca pana in 2027, "testarea de penetrare va evolua intr-o forma de validare continua condusa de AI."
- **De la asistenta la autonomie:** Tranzitia de la AI ca asistent (sugereaza comenzi) la AI ca agent autonom (executa planuri). Demonstrat: DARPA AIxCC (complet autonom), Anthropic incident (80-90% autonom).
- **Modele mici, specializate:** Foundation-Sec-8B (8 miliarde parametri) egalizeaza modele de 70B pe securitate. HackSynth-GRPO demonstreaza ca Reinforcement Learning creste dramatic performanta modelelor mici. Implicatie: capabilitati avansate de securitate AI pe hardware accesibil, fara dependenta cloud.
- **Multi-agent > single-agent:** Sistemele performante folosesc mai multi agenti specializati care coopereaza: CAI (8 piloni), MAPTA (multi-agent), CHECKMATE (planificare clasica + LLM), Excalibur (38 interfete de instrumente tipizate).
- **RAG (Retrieval-Augmented Generation) elimina halucinatiile:** AutoPentester (+27% prin RAG), VulnBot (memorie vectoriala de succes), RefPentester (invatare din esecuri). Accesul la baze de cunostinte in timp real (CVE, exploit databases) este crucial.

6.2 Proiectii pe 12–24 luni

Proiectie	Probabilitate	Baza de argumentare
Pentesting manual devine serviciu boutique pana in 2028	Ridicata	a16z, Omdia, si Penligent proiecteaza 99% vulnerability assessments agentice. Cobalt deja foloseste AI pentru basic scanning, umani doar pentru business logic.
Primele standarde de conformitate care recunosc formal testarea AI	Moderata-Ridicata	PCI SSC si ISO lucreaza la actualizari. Aikido deja produce rapoarte acceptate pentru SOC2/ISO27001.
Atacuri autonome AI devin recurente	Ridicata	Incidentul Anthropic nov. 2025 e primul documentat. Capabilitatile prolifereaza prin open-source. IAPS confirma traiectoria.
Modele sub-10B parametri suficiente pentru operatiuni ofensive/defensive	Ridicata	Foundation-Sec-8B deja demonstrat. HackSynth-GRPO pe 8B. Cost hardware: sub 2.000 USD pentru inferenta locala.
State mici/medii dezvolta capabilitati ofensive AI comparabile cu puterile mari	Moderata	CEO Delinea: AI nivelizeaza terenul. Instrumente open-source (CAI, VulnBot, HackSynth) disponibile oricui.

7. Riscuri specifice pentru Romania

7.1 Expunerea infrastructurii critice

Romania opereaza infrastructura critica (energie, telecomunicatii, transport, sistem bancar, sanatate) cu un mix de sisteme legacy si moderne. Directiva NIS2, transpusa in legislatia nationala, impune cerinte de securitate, dar implementarea variaza semnificativ intre sectoare.

Riscul specific: atacatorii dotati cu AI pot scana si exploata vulnerabilitati in sisteme legacy la o viteza si scara pe care apararea manuala nu o poate egala. Un agent AI autonom poate testa mii de sisteme simultan, 24/7, la un cost de ordinul zecilor de dolari.

7.2 Deficitul de specialisti

Romania are un deficit semnificativ de specialisti in securitate cibernetica. AI nu inlocuieste specialistii, dar **amplifica dramatic capacitatea celor existenti.** Un pentester echipat cu instrumente AI poate acoperi intr-o saptamana ce echipe intregi acopera intr-o luna. Ignorarea acestor instrumente inseamna mentinerea artificala a unui dezavantaj deja sever.

7.3 Fereastra de actiune se inchide

URGENTA STRATEGICA

Toate instrumentele descrise in acest studiu sunt disponibile acum: open-source, documentate, ruleaza pe hardware accesibil (sub 5.000 EUR). Atacatorii le adopta deja (incident Anthropic, nov. 2025). Fiecare luna de inactiune creste asimetria atac–aparare. Investitia necesara este de ordine de marime mai mica decat costul unui singur incident de securitate la nivel de infrastructura critica.

8. Recomandari

Pe baza analizei prezentate, formulam urmatoarele recomandari pentru forurile de decizie strategica:

8.1 Actiuni imediate (0–6 luni)

1. **Evaluare nationala de capabilitati AI defensive:** Inventar al instrumentelor AI de securitate utilizate in sectoarele de infrastructura critica. Identificarea lacunelor fata de starea artei prezentata in acest studiu.
2. **Program pilot de pentesting AI pe infrastructura critica:** Selectie de 3-5 sisteme de infrastructura critica pentru testare cu instrumente AI (open-source disponibil: CAI, Buttercup/AIxCC). Comparare rezultate cu audit-urile manuale existente.
3. **Constituirea unui grup de lucru AI-Cyber:** Specialisti din cadrul institutiilor de securitate ale statului, mediu academic si sector privat. Misiune: monitorizarea evolutiei capabilitatilor AI ofensive si defensive, propuneri de politici si investitii.

8.2 Actiuni pe termen mediu (6–18 luni)

4. **Dezvoltarea cadrului de conformitate AI-aware:** Actualizarea standardelor nationale de audit si testare de securitate pentru a include si recunoaste testarea bazata pe AI, in aliniere cu evolutia PCI DSS si ISO 27001.
5. **Investitie in capabilitati locale:** Suport pentru companii romanesti de securitate cibernetica sa adopte si sa dezvolte instrumente AI. Romania are avantaj competitiv prin comunitatea IT puternica — dar fara directie strategica, acest avantaj se pierde.
6. **Program de formare AI pentru specialistii de securitate existenti:** Integrarea instrumentelor AI in curricula de formare DNSC si in certificarile nationale de securitate.

8.3 Actiuni pe termen lung (18–36 luni)

7. **Centru national de testare AI-Cyber:** Infrastructura dedicata (laborator, hardware, modele, benchmark-uri) pentru evaluarea continua a capabilitatilor AI de securitate — similar cu ce DARPA a construit prin AIxCC.
8. **Parteneriate internationale:** Colaborare cu UE (ENISA), NATO CCDCOE (Tallinn), si bilateral cu state care investesc activ (SUA, UK, Israel) pentru transfer de know-how si acces la technologii avansate.

9. SafebyteAI — O Initiativa Romaneasca in Contextul Global

In contextul evolutiilor prezentate in acest studiu, mentionam existenta unui proiect romanesc aflat in dezvoltare activa: **SafebyteAI**, o platforma de penetration testing augmentat cu inteligenta artificiala, dezvoltata de Safebyte Consulting S.R.L. Proiectul demonstreaza ca expertiza locala romaneasca poate produce solutii competitive in acest domeniu strategic.

9.1 Arhitectura de principiu

SafebyteAI implementeaza o arhitectura **pipeline multi-etapa** in care fiecare faza a evaluarii de securitate este augmentata de modele AI specializate, nu inlocuita. Filosofia de proiectare este **"AI-ul amplifica expertul uman, nu il inlocuieste"** — o abordare validata si de concluziile DARPA AIxCC si de analizele a16z privind viitorul pentesting-ului.

Platforma integreaza instrumente profesionale de securitate consacrate (scanere de vulnerabilitati, analizoare de trafic, motoare de template-uri) cu modele AI multiple care opereaza in faze distincte: imbogatirea contextului, analiza paralela multi-model, rationament structurat pentru prioritizarea atacurilor, generare de teste si payload-uri, si mapare pe cadre de conformitate.

Elemente arhitecturale cheie:

- **Inferenta locala (on-premises):** Toate modelele AI ruleaza pe infrastructura proprie, fara dependenta de cloud. Datele clientilor nu parasesc niciodata perimetrul fizic — cerinta critica pentru clienti din sectorul bancar si infrastructura critica.
- **Pipeline multi-model:** Diferite modele AI sunt specializate pe diferite faze ale evaluarii. Analiza de vulnerabilitati foloseste modele optimizate pentru rationament, iar generarea de teste foloseste modele optimizate pentru creativitate ofensiva. Aceasta abordare multi-model este aliniata cu paradigma multi-agent validata de cele mai performante sisteme academice (MAPTA, Excalibur, CAI).
- **Integrare cu instrumente profesionale:** Platforma nu reinventeaza instrumentele de securitate, ci le integreaza: scanere de retea, proxy-uri web, motoare de template-uri de vulnerabilitati, analizoare SSL/TLS. AI-ul proceseaza si coreleaza output-urile acestor instrumente, identificand pattern-uri pe care analiza manuala le-ar putea rata.
- **Dual-workflow:** Platforma suporta atat evaluari manuale asistate (pentesterul furnizeaza datele, AI-ul le analizeaza si genereaza teste), cat si evaluari automatizate (platforma conduce reconnaissance-ul si scanning-ul activ). Aceasta dualitate acopera intreaga gama de nevoi, de la pentest-uri specializate la evaluari de vulnerabilitati la scara.
- **Conformitate integrata:** Fiecare vulnerabilitate identificata este automat mapata pe cerintele cadrelor de conformitate relevante (PCI DSS v4.0, reglementari BNR/BNM, NIS2, ISO 27001). Aceasta reduce semnificativ timpul de raportare si asigura trasabilitate completa.

9.2 Avantaje strategice

SafebyteAI adreseaza mai multe lacune identificate in ecosistemul actual:

- **Suveranitate a datelor:** Spre deosebire de solutiile cloud internationale, datele de audit ale infrastructurii critice romanesti raman pe teritoriul national. Aceasta este o cerinta din ce in ce mai stringenta in contextul geopolitic actual si al reglementarilor europene privind datele.
- **Adaptare la contextul local:** Platforma integreaza nativ cadre de conformitate specifice pietei romanesti si celei est-europene (reglementari BNR, BNM, transpunere NIS2), pe langa standardele internationale.
- **Multiplicator de forta pentru echipele locale:** In contextul deficitului de specialisti de securitate, platforma permite echipelor mici sa acopere un volum de evaluare semnificativ mai mare, fara compromis de calitate. Un specialist echipat cu SafebyteAI poate genera in ore ceea ce manual necesita zile.
- **Baza de cunostinte cu invatare continua:** Platforma incorporeaza o componenta de threat intelligence care agrega surse publice de vulnerabilitati, exploit-uri si tehnici de atac, mentinand cunoasterea actualizata. Experienta acumulata din evaluari anterioare imbunatateste calitatea evaluarilor viitoare, cu respectarea stricta a separarii datelor intre clienti.

9.3 Stadiu si perspective

SafebyteAI se afla in faza de dezvoltare activa, cu componente functionale pe workflow-ul de evaluare web si cu o foaie de parcurs care include extinderea catre evaluari de infrastructura de retea, API-uri, aplicatii mobile si audituri de conformitate. Proiectul este dezvoltat integral in Romania, cu expertiza locala, si demonstreaza ca **Romania poate fi nu doar consumator, ci si producator de tehnologie de securitate cibernetica bazata pe inteligenta artificiala.**

Existenta acestui proiect subliniaza mesajul central al studiului: capabilitatile sunt accesibile, expertiza exista local, iar investitia necesara este de ordine de marime mai mica decat costul inactiunii. Ceea ce lipseste este **directia strategica si vointa institutionala de a actiona.**

10. Referinte Principale

Toate referintele sunt publice si verificabile la datele indicate.

Competitii si programe guvernamentale

- DARPA, “AI Cyber Challenge Final Competition Results”, august 2025. darpa.mil/news/2025/aixcc-results
- DARPA, “AIxCC Scoring Guide”, 2025. darpa.mil/news/2025/ai-cyber-challenge-scoring
- Trail of Bits, “Buttercup wins 2nd place in AIxCC”, august 2025. blog.trailofbits.com
- aicyberchallenge.com — Pagina oficiala AIxCC cu toate rezultatele

Lucrari academice (ordonate cronologic)

- PenHeal — arXiv:2407.17788, ACM CCS Workshop, 2024
- HackSynth — arXiv:2412.01778, dec. 2024 + GRPO: arXiv:2506.02048, iun. 2025
- VulnBot — arXiv:2501.13411, ian. 2025. Open-source: github.com/KHenryAegis/VulnBot
- Primus (Trend Micro) — arXiv:2502.11191, feb. 2025
- CAI Framework — arXiv:2504.06017, apr. 2025. Open-source: github.com/aliasrobotics/cai
- RefPentester — arXiv:2505.07089, iun. 2025
- MAPTA — arXiv:2508.20816, aug. 2025
- CVE-Genie — arXiv:2509.01835, sept. 2025
- AutoPentester — arXiv:2510.05605, oct. 2025
- PentestMCP — arXiv:2510.03610, oct. 2025
- CHECKMATE — arXiv:2512.11143, dec. 2025
- Excalibur/PentestGPT v2 — arXiv:2602.17622, feb. 2026
- PenForge — arXiv:2601.06910, ian. 2026 (ICSE-NIER 2026)
- PwnGPT — ACL 2025 Long Paper, aclanthology.org/2025.acl-long.562

Modele specializate

- Foundation-Sec-8B (Cisco) — HuggingFace: fdtn-ai/Foundation-Sec-8B
- Primus (Trend Micro) — HuggingFace: trendmicro-ailab/Llama-Primus-*

Rapoarte de industrie si amenintari

- CrowdStrike, “2026 Global Threat Report”, februarie 2026. crowdstrike.com/global-threat-report
- Microsoft, “2025 Digital Defense Report”, oct. 2025. microsoft.com/security/security-insider
- IBM, “Cost of a Data Breach Report 2024” — $4.88M cost mediu global
- Pentera, “The State of Pentesting 2025”
- Verizon, “2025 Data Breach Investigations Report”

Analize de threat intelligence

- Anthropic, “Disrupting the first AI-orchestrated cyber espionage campaign”, nov. 2025
- IAPS, “The Emergence of Autonomous Cyber Attacks”, nov. 2025. iaps.ai/research/autonomous-cyber-attacks
- Lawfare, “Fighting AI Cyberattacks Starts With Knowing They’re Happening”, feb. 2026
- SecurityWeek, “Cyber Insights 2026: Cyberwar and Rising Nation State Threats”, feb. 2026

Analize de piata si strategie

- Andreessen Horowitz (a16z), “Next-Gen Pentesting: AI Empowers the Good Guys”, iun. 2025
- Omdia/Informa Tech, “The Penetration Testing Market in 2025”, dec. 2025
- Spark42.tech, “AI-Powered Penetration Testing Tools in 2026”, ian. 2026
- because-security.com, “AI pentesting 2025/2026 with open source tools”, dec. 2025

Resurse agregate (curated lists)

- Awesome-LLM4Cybersecurity — github.com/tmylla/Awesome-LLM4Cybersecurity (1100+ stars, 300+ papers)
- awesome-ai-pentest — github.com/insidetrust/awesome-ai-pentest
- cyber-security-llm-agents — github.com/NVISOsecurity/cyber-security-llm-agents

— Sfarsit document —