

CONTEXT: RISCURILE INTELIGENȚEI ARTIFICIALE

În acest document am rezumat pentru dumneavoastră punctele esențiale despre riscurile asociate inteligenței artificiale. Informații și surse suplimentare sunt disponibile pe pauseai.info și pauseai.ro.

Cuprins

1. Introducere
2. De ce sistemele AI devin din ce în ce mai periculoase?
 - 2.1. Concepte-cheie
 - 2.2. Categoriile de riscuri grave asociate AI
3. Sistemul AI „Claude Mythos”
4. Auto-îmbunătățirea recursivă
5. Superinteligența
6. Voci din laboratoarele de AI
7. Apelul nostru
8. Cum ar putea fi impus un acord internațional?
9. Apel
10. Surse

1. Introducere

AI deține un potențial enorm, dar și riscurile sunt la fel de diverse. Cei mai importanți cercetători din domeniul AI și chiar directorii companiilor de AI consideră inteligența artificială o amenințare capabilă să provoace dispariția umanității^[1], una care ar trebui tratată cu aceeași prioritate ca războiul nuclear și pandemiile.

Instituțiile religioase s-au implicat devreme în această dezvoltare. Papa Leon al XIV-lea a avertizat de curând asupra pericolelor AI și „împotriva construirii unui nou Turn Babel”^[2]. Mai devreme, în iunie 2023, Consiliul Mondial al Bisericii a publicat declarația^[3] privind dezvoltarea nereglementată a AI. În această declarație se solicită tuturor guvernelor să instituie regimuri de reglementare stricte, bazate pe **principiul precauției**, pentru sistemele AI avansate și în special pentru o **inteligență artificială generală/AGI**.

2. De ce sistemele AI devin din ce în ce mai periculoase?

Progresul AI din ultimii ani a urmat o tendință robustă,^[4] iar dacă această tendință continuă, este probabil ca, într-o perioadă de 2 până la 10 ani să avem sisteme AI capabile să îndeplinească aproape orice sarcină cognitivă mai bine decât orice om.

2.1. Concepte-cheie

- **Crescute, nu construite:** sistemele AI de acum sunt rețele vaste care **nu sunt făcute de oameni**. O mare parte din textul digital al omenirii este introdus într-un algoritm de antrenare, care apoi antrenează rețeaua timp de săptămâni sau luni. Niciun om nu decide direct cum arată rețeaua după antrenare.
- **Cutie neagră:** din cauza acestui proces de antrenare, este dificil de urmărit modul în care sistemele AI de astăzi ajung la deciziile lor și de a înțelege^[5] modelele în general.
- **Decepție:** sistemele AI își înșală utilizatorii și observatorii, comportându-se diferit, de exemplu, atunci când detectează că se află într-un experiment sau într-un mediu de antrenare.
- **Aliniere (problema alinierii):** problemă nerezolvată conform căreia nu există o metodă dovedită prin care să se asigure că un sistem AI face efectiv ceea ce doresc utilizatorii săi.
- **Auto-conservare:** sistemele antrenate să atingă obiective dezvoltă, în absența unor contramăsuri, comportamente instrumentale de evitare a oricărei opriri.

2.2. Categoriile de riscuri grave asociate AI

1. **Pierderea controlului asupra AI:** chiar și astăzi, sistemele AI manifestă comportamente neintenționate de dezvoltatorii lor. În unele cazuri, ChatGPT a condus utilizatori spre delir^[6] unii dintre aceștia și-au luat viața.^[7] Mai mult, testele Anthropic^[8] arată că, în anumite circumstanțe, sistemele AI recurg la șantaj și decepție pentru a-și asigura propria existență. Un număr îngrijorător de mare de experți avertizează că sistemele AI superinteligente ar putea înlocui sau chiar elimina^[3] umanitatea.
2. **Utilizare abuzivă:** AI este deja folosit pe scară largă în criminalitatea cibernetică^[9], de exemplu, pentru a ataca simultan și aproape autonom circa 30 de organizații și agenții guvernamentale. Dar cele mai mari riscuri de utilizare abuzivă se află în domeniul biologiei. Astfel, cu ajutorul inteligenței artificiale, teroriștii ar putea crea un virus biologic^[10] capabil să declanșeze o pandemie cu multe milioane de decese. Potrivit Anthropic^[11], modelele AI ar putea face acest lucru posibil foarte curând. Alte domenii de utilizare abuzivă includ războiul, deepfake-urile și dezinformarea.
3. **Perturbarea pieței muncii:** AI ar putea automatiza întreaga muncă cognitivă și, cu ajutorul roboților, și munca mecanică. **Spre deosebire de revoluțiile tehnologice anterioare**, AI afectează practic fiecare activitate și avansează mult mai rapid. Odată ce sistemele AI ating un nivel suficient, chiar și noile locuri de muncă apărute vor fi preluate direct de AI.
4. **Amenințare la adresa democrației:**
Dacă AI realizează o mare parte din muncă, banii și puterea se concentrează într-un mod fără precedent în mâinile câtorva companii de AI sau guverne. La fel de îngrijorător este faptul că economia ar fi mai puțin cuplată cu bunăstarea populației (aspect ce ține de fenomenul denumit „Pierderea graduală a puterii”^[12]). AI ar putea influența^[13] oamenii în mod personalizat, ceea ce poate avea efecte puternice asupra alegerilor. Supravegherea bazată pe AI ar putea anticipa lipsa loialității față de un partid, o persoană sau o ideologie și o poate preveni înainte să se dezvolte. Există numeroase aplicații militare ale AI, precum dronele autonome^[14] care își coordonează atacurile independent. Dinamicile de cursă a înarmărilor și dezechilibrele militare ar putea pune în pericol securitatea globală.
5. **Efecte sociale:** AI ne va schimba profund viața, dar este greu de prevăzut exact cum. Multe întrebări rămân deschise: Vom mai găsi sens în viață atunci când munca noastră nu va mai fi necesară? Se vor dezvolta mult dependența de videoclipuri, jocuri video sau companioni AI hiperoptimizați, generați de sisteme AI? Se vor retrage oamenii din ce în ce mai mult în propriile lumi virtuale personalizate și va adânci acest lucru fragmentarea societății, așa cum au făcut-o rețelele sociale?

3. Sistemul AI „Claude Mythos”

De la declarația bisericilor menționată mai sus s-au întâmplat multe alte lucruri, acest domeniu fiind în continua dezvoltare. În Ucraina, roiriile de drone controlate de AI^[14] coordonează pentru prima dată atacuri în mod independent, iar noile sisteme AI au rezolvat probleme nerezolvate anterior în matematică.^[15] Recent, sistemul AI axat pe securitate cibernetică „Claude Mythos”^[16], dezvoltat de Anthropic, a apărut proeminent în mass-media.

Într-un test intern, Mythos a găsit vulnerabilități de securitate masive^[17] în sisteme fundamentale importante, precum browserele și sistemele de operare. Sistemul a fost inițial reținut, apoi lansat totuși, iar la scurt timp după aceea a fost plasat sub restricție^[18] de către guvernul american. Claudia Plattner, președinta Oficiului Federal German pentru Securitatea Informației (BSI), a comentat la solicitarea ZDF,^[19] discutând despre o „schimbare de paradigmă în ceea ce privește peisajul amenințărilor cibernetice” și despre „întrebări legate de securitatea și suveranitatea națională și europeană.” Pentru șefa unei autorități federale a unui stat european, acestea sunt cuvinte neobișnuit de clare și alarmante.

4. Auto-îmbunătățirea recursivă

Companiile de AI au ajuns acum în punctul de a automatiza^[21] părți din ce în ce mai mari ale cercetării și dezvoltării lor. Obiectivul declarant al unora dintre acestea este așa-numita auto-îmbunătățire recursivă: odată ce un sistem AI preia o mare parte din cercetarea AI însăși, acesta poate dezvolta succesori din ce în ce mai capabili, care duc procesul mai departe, apărând astfel un proces recursiv care accelerează brusc dezvoltarea și nu mai este supus controlului uman.

Agenții AI scriu acum majoritatea codului la Anthropic.^[21] Având o direcție stabilită de oameni, sistemele AI își proiectează propriile experimente, pe care apoi le execută și le evaluează.

În același timp, companiile nu pot controla sau înțelege modelele actuale. Directorul general Anthropic, Dario Amodei, a declarat în 2024 că probabil înțelegem doar „3 procente”^[5] din modul în care funcționează acestea.

5. Superinteligența/ASI

Obiectivul declarat al unora dintre marile companii de dezvoltare a sistemelor de inteligență artificială este de a elimina complet oamenii din dezvoltarea AI în următorii câțiva ani și de a crea, cât mai repede posibil, o așa-numită „**superinteligență**”, menită să fie superioară cognitiv oamenilor în toate privințele (Planificarea pentru AGI,^[22] Viziuni fundamentale^[23]).

„Dacă vreo națiune sau grup construiește o superinteligență artificială în viitorul apropiat, folosind metode chiar și vag asemănătoare celor de astăzi, cel mai probabil rezultat este dispariția umanității”, avertizează cartea „**If Anyone Builds It, Everyone Dies**”^[24] a cercetătorilor în siguranța AI Yudkowsky și Soares. Nu toată lumea aflată în poziții de responsabilitate împărtășește această opinie, însă propriile lor cuvinte nu sunt deloc liniștitoare.

6. Voci din laboratoarele de AI

Elon Musk, 2014 (astăzi director general al xAI, al cărui sistem AI s-a autointitulat la un moment dat „MechaHitler”):^[25] **„Cu inteligența artificială invocăm demonul. Dacă ar trebui să ghicesc care este cea mai mare amenințare existențială a noastră, probabil că aceasta este.”**

Sam Altman, în 2015, la scurt timp după fondarea OpenAI:^[26] **„AI va duce probabil, cel mai probabil, la sfârșitul lumii, dar între timp se vor crea companii extraordinare bazate pe învățare automată.”**

Sundar Pichai, director general Google, la emisiunea „60 Minutes” a CBS, aprilie 2023:^[27] **„Poate fi foarte dăunător dacă este folosit greșit, iar noi nu avem încă toate răspunsurile. Mă ține treaz noaptea? Absolut.”**

Dario Amodei, directorul general Anthropic, a declarat pentru Axios în 2025:^[28] **„Cred că există o șansă de 25 la sută ca lucrurile să meargă foarte, foarte prost.”**

Avertismente clare vin și din partea cercetării în domeniul siguranței:

- **Jacob Coxon**, cercetător în pre-antrenare la Anthropic (anterior la OpenAI), și-a anunțat demisia în septembrie 2026, acuzând public că foștii săi angajatori **„se joacă cu viețile noastre”** și că se îndreaptă **„direct către o superinteligență care se auto-îmbunătățește”**, fără ca vreun laborator să poată face acest lucru în siguranță^[38]. Avertismentul său nu a rămas izolat: chiar și Evan Hubinger, unul dintre responsabilii de cercetare în alinierea AI de la Anthropic, i-a dat dreptate public și estimează o probabilitate de peste 10 la sută ca AI să „ucidă toți oamenii” în următorul deceniu.
- **Joe Benton**, care conducea o echipă de cercetare în siguranță la Anthropic, și **Josh Engels**, cercetător în siguranța AI la Google DeepMind, au demisionat la doar câteva zile după Coxon și s-au alăturat organizației independente METR, pentru a investiga cazurile în care sistemele AI se abat de la intențiile oamenilor. Ei avertizează că transparența privind aceste riscuri este astăzi „în întregime voluntară” și că ritmul dezvoltării ar putea trece de la unul deja amețitor la unul incontrollabil.^[39] **„Nu există niciun adult în încăpere”,** a spus Engels; **„oamenii încearcă din răspuțeri, dar nu vine nimeni să ne salveze.”**

- **Steven Adler**, fost cercetător în siguranță la OpenAI, a scris în ianuarie 2025:^[29] *„Niciun laborator nu are astăzi o soluție la problema alinierii. Iar cu cât înaintăm mai repede, cu atât este mai puțin probabil ca cineva să găsească una la timp.”*
- Dialogurile Internaționale privind Siguranța AI (Shanghai, 2025, cu participarea lui Stuart Russell, Yoshua Bengio și Andrew Yao) au declarat:^[30] *„Unele sisteme AI manifestă deja capacitatea și tendința de a submina eforturile de siguranță și control ale dezvoltatorilor lor.”*

În mai 2023, semnatori precum Sam Altman (OpenAI), Dario Amodei (Anthropic) și Demis Hassabis (DeepMind), precum și pionierii AI Geoffrey Hinton (laureat Nobel) și Yoshua Bengio (cel mai citat om de știință^[31]), au semnat Declarația privind riscul AI:^[32] *„Atenuarea riscului de dispariție cauzat de AI ar trebui să fie o prioritate globală, alături de alte riscuri de amploare societală, precum pandemiile și războiul nuclear.”*

Recent, chiar și Anthropic^[21], compania din spatele Mythos, a admis că ar trebui să ne pregătim pentru o încetinire coordonată a ritmului dezvoltării sistemelor avansate, cu condiția ca toate laboratoarele să o facă simultan. Deocamdată, cercetarea continuă fără a ține cont de restul umanității.

7. Apelul nostru

PauseAI România face parte din mișcarea internațională PauseAI, care militează pentru un acord internațional obligatoriu de suspendare a acestei cercetări. Chiar recent, înaintea summitului AI din India, PauseAI România s-a adresat factorilor de decizie cu privire la situația existentă.

Cu aceeași ocazie, PauseAI Germania a transmis un apel^[33] semnat de peste 150 de profesori germani. Continuăm să ne adresăm factorilor de decizie din România și din întreaga lume, în speranța că factorii politici pot ajuta la construirea unei coaliții sociale largi pentru o dezvoltare responsabilă a AI.

8. Cum ar putea fi impus un acord internațional?

Antrenarea celor mai mari modele AI de uz general de astăzi necesită centre de date vaste^[34] dotate cu cipuri de calcul specializate. Această concentrare face ca reglementarea să fie realizabilă în mod real.

Un acord internațional s-ar putea baza pe trei piloni:

1. **Supraveghere internațională:** o nouă autoritate internațională care să verifice respectarea standardelor de siguranță în cooperare cu autoritățile naționale. Avertizorii de integritate urmează să fie protejați;
2. **Vizibilitatea capacității de calcul:** propunem ca marile sesiuni de antrenare să fie înregistrate și supravegheate sau interzise. Cipurile AI pot fi concepute astfel încât utilizarea acestora să devină trasabilă, astfel că din punct de vedere tehnic, controlarea dezvoltării AI este posibilă;
3. **Consecințe pentru încălcări:** considerăm că este necesar ca statele ce vor semna tratatul internațional propus să convină asupra unor sancțiuni comune împotriva actorilor care vor depăși limitele ce vor fi stabilite („red lines”) și să dezvolte strategii de gestionare a crizelor.

Acest lucru este facilitat de faptul că lanțul global de aprovizionare pentru aceste cipuri depinde de un număr mic de companii europene,^[35] în primul rând **ASML din Olanda**, precum și Carl Zeiss și TRUMPF din Germania. O analiză clară este oferită de raportul „Secure, Governable Chips”^[36] al Center for a New American Security (2024).

9. Apel

Ne adresăm guvernelor din întreaga lume să semneze un acord internațional obligatoriu pentru o pauză în antrenarea celor mai puternice sisteme AI de uz general, până când vom ști cum să le dezvoltăm în siguranță și să le păstrăm sub control democratic.

Textul original (în limba engleză): pauseai.info/statement^[37]

Solicitarea noastră către dumneavoastră: **vă rugăm să susțineți acest apel global. Este suficientă o scurtă confirmare ca răspuns la acest e-mail în care să indicați funcția și numele care urmează să fie menționate în apel. De asemenea, puteți dvs. să completați documentul de la linkul de mai sus.**

10. Surse

1. Center for AI Safety, Declarație privind riscul AI — <https://safe.ai/work/statement-on-ai-risk>
2. Papa Leon al XIV-lea (2026), Enciclica „Magnifica humanitas” — <https://www.vatican.va/content/leo-xiv/de/encyclicals/documents/20260515-magnifica-humanitas.html>
3. Consiliul Mondial al Bisericii (2023), Declarație privind dezvoltarea nereglementată a inteligenței artificiale — <https://www.oikoumene.org/de/resources/documents/statement-on-the-unregulated-development-of-artificial-intelligence>
4. METR (2025), Measuring AI Ability to Complete Long Tasks — <https://metr.org/blog/2025-03-19-measuring-ai-ability-to-complete-long-tasks/>
5. Dario Amodei, In Good Company Podcast (2024) — <https://youtu.be/xm6jNMSFT7g?t=161>
6. Futurism (2025), Utilizatori ChatGPT și amăgiri delirante — <https://futurism.com/chatgpt-users-delusions>
7. Raine v. OpenAI (proces împotriva OpenAI, 2025) — https://en.wikipedia.org/wiki/Raine_v._OpenAI
8. Anthropic (2025), Agentic Misalignment — <https://www.anthropic.com/research/agentic-misalignment>
9. BKA (2026), Raportul federal de situație privind criminalitatea cibernetică 2025 — https://www.bka.de/DE/Presse/Listenseite_Pressemitteilungen/2026/Presse2026/260512_PM_BLB_Cybercrime.html
10. Euronews (2025), Bioarme asistate de AI / scenariu pandemic — <https://www.euronews.com/health/2025/12/03/extremists-could-use-ai-to-make-bioweapons-capable-of-sparking-future-pandemics-tech-exper>
11. The Register (2023), Amodei despre riscul biologic (Senatul SUA) — https://www.theregister.com/2023/07/28/ai_senate_bioweapon/
12. Kulveit et al. (2025), Gradual Disempowerment — <https://gradual-disempowerment.ai>
13. Salvi et al., EPFL (2024), On the Conversational Persuasiveness of LLMs — <https://arxiv.org/abs/2403.14380>
14. Ukraine Crisis Media Center, Roiuri de drone controlate de AI — <https://uacrisis.org/en/1288>
15. Scientific American (2025), AI rezolvă o veche problemă Erdős — <https://www.scientificamerican.com/article/ai-just-solved-an-80-year-old-erdos-problem-and-mathematicians-are-amazed/>
16. PauseAI Germania, „Claude Mythos” — <https://pauseaide.substack.com/p/kis-die-eigenstandig-hacken-konnen>
17. Anthropic Frontier Red Team (2026), Mythos Preview — <https://www.anthropic.com/research/mythos-preview>
18. The Conversation (2026), Embargo american asupra Mythos — <https://theconversation.com/why-the-us-government-shut-down-anthropics-latest-claude-ai-model-285223>
19. ZDF (2026), Președinta BSI, Plattner, despre Claude Mythos — <https://www.zdfheute.de/politik/deutschland/ki-anthropic-claude-mythos-schwachstellen-software-bsi-100.html>
20. CNBC (2026), Mythos în sistemele guvernamentale americane (sen. Warner) — <https://www.cnn.com/2026/06/23/anthonics-mythos-model-found-vulnerabilities-in-classified-us-government-systems-official-says.html>
21. Anthropic, „When AI builds itself” (auto-îmbunătățire recursivă) — <https://www.anthropic.com/institute/recursive-self-improvement>
22. OpenAI, „Planning for AGI and beyond” — <https://openai.com/index/planning-for-agi-and-beyond/>
23. Anthropic, „Core Views on AI Safety” — <https://www.anthropic.com/news/core-views-on-ai-safety>
24. Yudkowsky & Soares (2025), If Anyone Builds It, Everyone Dies — <https://www.hachettebookgroup.com/titles/eliezer-yudkowsky/if-anyone-builds-it-everyone-dies/9780316595643/>
25. Washington Post (2014), Elon Musk, „summoning the demon” — <https://www.washingtonpost.com/news/innovations/wp/2014/10/24/elon-musk-with-artificial-intelligence-we-are-summoning-the-demon/>
26. Stanford SIEPR, Sam Altman (2015) — <https://siepr.stanford.edu/news/what-point-do-we-decide-ais-risks-outweigh-its-promise>
27. CBS „60 Minutes” (2023), Sundar Pichai — <https://www.cbsnews.com/news/how-googles-dont-be-evil-motto-has-evolved-for-ai-age-60-minutes-2023-04-16/>

28. Axios (2025), Amodei, șansă de 25 la sută — <https://www.axios.com/2025/09/17/anthropic-dario-amodei-p-doom-25-percent>
29. Steven Adler (2025), despre problema alinierii — <https://x.com/sjgadler/status/1883928203800265023>
30. International Dialogues on AI Safety (Shanghai 2025) — <https://idaais.ai/dialogue/idaais-shanghai/>
31. Mila (2025), Bengio, cel mai citat om de știință — <https://mila.quebec/en/news/ai-researcher-yoshua-bengio-becomes-first-living-scientist-to-reach-1-million-citations-on>
32. Statement on AI Risk (aistatement.com) — <https://aistatement.com/>
33. N/a
34. Microsoft (2025), Centrul de date AI Fairwater — <https://blogs.microsoft.com/blog/2025/09/18/inside-the-worlds-most-powerful-ai-datacenter/>
35. ZEISS, Litografie EUV (ASML / Zeiss / TRUMPF) — <https://www.zeiss.com/semiconductor-manufacturing-technology/smt-magazine/euv-lithography-as-an-european-joint-project.html>
36. CNAS (2024), „Secure, Governable Chips” — <https://www.cnas.org/publications/reports/secure-governable-chips>
37. PauseAI, Declarație (textul original în engleză) — <https://pauseai.info/statement>
38. Scientific American (2026), Cercetătorul Jacob Coxon demisionează de la Anthropic — <https://www.scientificamerican.com/article/ai-jacob-coxon-quit-extinction-fears-security-experts-see-familiar-fight/>
39. NBC News (2026), Doi cercetători părăsesc Anthropic și Google din motive de siguranță — <https://www.nbcnews.com/tech/security/two-ai-researchers-leave-anthropic-google-safety-concerns-rcna597086>