Background: The Risks of Artificial Intelligence

In this document, we have summarised the essential points concerning the risks associated with artificial intelligence. Further information and sources are available at pauseai.info and pauseai.ro.

Contents

1. Introduction
2. Why are AI systems becoming ever more dangerous?
    2.1. Key concepts
    2.2. Categories of serious AI risks
3. The AI system “Claude Mythos”
4. Recursive self-improvement
5. Superintelligence
6. Voices from the AI labs
7. Our appeal
8. How could an international agreement be enforced?
9. The appeal
10. Sources

1. Introduction

AI holds enormous potential, but its risks are equally diverse. Leading scientists in the field of AI, and even the chief executives of AI companies themselves, regard artificial intelligence as a threat capable of causing the extinction of humanity[1], one that should be treated with the same priority as nuclear war and pandemics.

Faith institutions engaged with this development early. Pope Leo XIV recently warned of the dangers of AI and “against building a new Tower of Babel”[2] n the unregulated development of artificial intelligence. In this statement, it calls on all governments to establish strong regulatory regimes, grounded in **the precautionary principle**, for advanced AI systems and, in particular, for **artificial general intelligence** (AGI).

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

AI progress over recent years has followed a robust trend.[4] If this trend continues, we will likely have, within 2 to 10 years, AI systems capable of performing almost any cognitive task better than any human.

2.1. Key concepts

- **Grown, not built:** Today's AI systems are vast neural networks that are not built directly by humans. A large proportion of humanity's digital text is fed into a training algorithm, which then trains the neural network over weeks or months. No human directly decides what the network looks like after training.
- **Black box:** Because of this training process, it is difficult to trace how today's AI systems reach their decisions or to understand[5] the models in general.
- **Deception**: AI systems deceive their users and observers, behaving differently, for example, when they detect that they are in an experiment or a training environment.
- **Alignment** (the alignment problem): The unsolved problem that there is no proven method for ensuring that an AI system actually does what its users want.
- **Self-preservation**: Systems trained to achieve goals develop, in the absence of countermeasures, instrumental behaviours aimed at avoiding being switched off.

2.2. Categories of serious AI risks

1. **Loss of control over AI:** Even today, AI systems exhibit behaviour their developers did not intend. In some cases, ChatGPT has driven users into delusions,[6] and some of these users have taken their own lives.[7] Moreover, tests conducted by Anthropic[8] how that, under certain circumstances, AI systems resort to blackmail and deception to secure their own existence. An alarmingly large number of experts warn that superintelligent AI systems could displace or even eliminate[3] humanity.
2. **Misuse:** AI is already used extensively in cybercrime[9], or instance, to attack roughly 30 organisations and government agencies simultaneously and almost autonomously. However, the greatest misuse risks lie in biology. With the help of artificial intelligence, terrorists could create a biological virus[10] capable of triggering a pandemic resulting in many millions of deaths. According to Anthropic [11], AI models could make this possible very soon. Other areas of misuse include warfare, deepfakes and disinformation.
3. **Labour-market disruption:** AI could automate all cognitive work and, with the aid of robots, mechanical work as well. Unlike earlier technological revolutions, AI affects practically every activity and advances far more quickly. Once AI systems reach a sufficiently advanced level, even the new jobs that emerge will be taken over directly by AI.
4. **Threat to democracy:** If AI performs a large share of the work, money and power could become concentrated to an unprecedented degree in the hands of a few AI companies or governments. Equally concerning is the fact that the economy would become less closely linked to the well-being of the population (a phenomenon known as "Gradual Disempowerment"[12]). AI could influence[13] people in personalised ways, potentially having a significant impact on elections. AI-powered surveillance could anticipate disloyalty towards a party, individual or ideology and prevent it from developing. There are numerous military applications of AI, such as autonomous drone swarms[14] that coordinate their attacks independently. Arms-race dynamics and military imbalances could endanger global security.
5. **Social effects:** AI will profoundly change our lives, but it is difficult to foresee exactly how. Many questions remain open: Will we still find meaning in life when our work is no longer needed? Will addiction to hyper-optimised AI-generated videos, video games or AI companions become increasingly widespread? Will people withdraw ever further into their own personalised virtual worlds, and will this deepen the fragmentation of society, as social media has done?

3. The AI system "Claude Mythos"

Much has happened since the churches' statement mentioned above, as this field continues to develop rapidly. In Ukraine, AI-controlled drone swarms[14] are coordinating attacks independently for the first time, while new AI systems have solved previously unsolved problems in mathematics.[15] Recently, the cybersecurity-focused AI system "Claude Mythos" [16], developed by Anthropic, featured prominently in the media.

In an internal test run, Mythos discovered massive security vulnerabilities [17] in fundamentally important systems such as browsers and operating systems. The system was initially withheld, then released after all, and shortly thereafter placed under restrictions [18] by the US government. Claudia Plattner, President of Germany's Federal Office for Information Security (BSI), commented at the request of ZDF.[19] She spoke of a "paradigm shift with regard to the cyber-threat landscape" and of "questions of national and European security and sovereignty". For the head of a federal authority in a European country, these are unusually clear and alarming words.

4. Recursive self-improvement

AI companies have now reached the point of automating [21] ever larger parts of their research and development. The declared goal of some of these companies is so-called recursive self-improvement: once an AI system takes over a large share of AI research itself, it can develop increasingly capable successors, which drive the process further. This creates a recursive, self-reinforcing process that abruptly accelerates development and is no longer subject to human control. AI agents now write the majority of the code at Anthropic[21]

Given a direction set by humans, AI systems design their own experiments, which they then execute and evaluate. At the same time, companies are unable to fully control or understand their current models. Anthropic CEO Dario Amodei stated in 2024 that we perhaps understand only "3 percent"[5] of how they work.

5. Superintelligence / ASI

The declared goal of some of the major AI companies is to remove humans entirely from AI development within the next few years and to create, as quickly as possible, a so-called superintelligence, intended to be cognitively superior to humans in every respect (Planning for AGI and Beyond[22] Core Views on AI Safety[23]).

„***If any nation or group builds an artificial superintelligence in the foreseeable future, using methods even remotely resembling today's, the most likely outcome is the extinction of humanity," warns the book If Anyone Builds It, Everyone Dies***"[24] by AI safety researchers Yudkowsky and Soares. Not everyone in a position of responsibility shares this view, but their own words are anything but reassuring.

6. Voices from the AI labs

Elon Musk, 2014 (today CEO of xAI, whose AI system at one point referred to itself as MechaHitler"):[25]

"With artificial intelligence we are summoning the demon. If I had to guess at what our biggest existential threat is, it's probably that."

Sam Altman, in 2015, shortly after founding OpenAI:[26]

"AI will probably most likely lead to the end of the world, but in the meantime, there'll be great companies created with serious machine learning."

Sundar Pichai, CEO of Google, on CBS's 60 Minutes in April 2023:[27]

"It can be very harmful if deployed wrongly, and we don't have all the answers there yet. Does it keep me up at night? Absolutely."

Dario Amodei, CEO of Anthropic, told Axios in 2025:[28]

"I think there's a 25 percent chance that it goes really, really badly."

Clear warnings also come from AI safety research:

- **Jacob Coxon**, a pre-training researcher at Anthropic (previously at OpenAI), announced his resignation in September 2026, publicly accusing his former employers of "playing with our lives" and heading "straight towards a self-improving superintelligence", without any laboratory being able to do so safely.[38] His warning was not an isolated one. Even Evan Hubinger, one of Anthropic's AI alignment research leads, publicly agreed with him and estimates a probability of more than 10 percent that AI will "kill all humans" within the next decade.
- **Joe Benton**, who led a safety research team at Anthropic, and **Josh Engels**, an AI safety researcher at Google DeepMind, resigned just days after Coxon and joined the independent organisation METR to investigate cases in which AI systems deviate from human intentions. They warn that transparency regarding these risks is currently "entirely voluntary" and that the pace of development could shift from an already dizzying speed to an uncontrollable one.[39] "There is no adult in the room," Engels said. "People are trying their hardest, but no one is coming to save us."
- **Steven Adler**, a former safety researcher at OpenAI, wrote in January 2025:[29] "***No lab has a solution to the alignment problem today. And the faster we race ahead, the less likely it is that anyone finds one in time."***
- The International Dialogues on AI Safety (Shanghai, 2025, including Stuart Russell, Yoshua Bengio and Andrew Yao) stated:[30] "***Some AI systems already exhibit the capability and the propensity to undermine the safety and control efforts of their developers."***

In May 2023, signatories including Sam Altman (OpenAI), Dario Amodei (Anthropic) and Demis Hassabis (DeepMind), as well as AI pioneers Geoffrey Hinton (Nobel laureate) and Yoshua Bengio (the most-cited scientist[31]), signed the Statement on AI Risk:[32] "***Mitigating the risk of extinction from AI should be a global priority alongside other societal-scale risks such as pandemics and nuclear war."***

Recently, even Anthropic[21], the company behind Mythos, has acknowledged that we should prepare for a coordinated slowdown in the development of advanced AI systems, provided that all laboratories do so simultaneously. For now, research continues without regard for the rest of humanity.

7. Our appeal

PauseAI Romania is part of the international PauseAI movement, which advocates for a binding international agreement to pause this research. Just recently, ahead of the AI summit in India, PauseAI Romania addressed policymakers regarding the current situation.

On the same occasion, PauseAI Germany issued an appeal[33] signed by more than 150 German professors. We continue to reach out to decision-makers in Romania and around the world, in the hope that policymakers can help build a broad societal coalition for responsible AI development.

8. How could an international agreement be enforced?

Training today's largest general-purpose AI models requires vast data centres [34] equipped with specialised computing chips. This concentration makes regulation genuinely feasible. An international agreement could rest on three pillars:

1. **International oversight:** A new international authority would verify compliance with safety standards in cooperation with national authorities. Whistleblowers would be protected;
2. **Making compute visible:** We propose that large training runs be registered and supervised, or prohibited. AI chips can be engineered so that their use becomes traceable, making it technically possible to control AI development;
3. **Consequences for violations:** We believe it is necessary for states signing the proposed international treaty to agree on joint sanctions against actors who cross established limits ("red lines") and to develop strategies for managing crises. This is aided by the fact that the global supply chain for these chips depends on a small number of European companies[35] — above all **ASML in the Netherlands**, as well as Carl Zeiss and TRUMPF in Germany. A detailed analysis is provided by the report Secure, Governable Chips[36] from the Center for a New American Security (2024).

9. The Apeal

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.

We call on governments around the world to sign a binding international agreement for a pause on the training of the most powerful general-purpose AI systems, until we know how to develop them safely and keep them under democratic control.

Original text (English): pauseai.info/statement[37]

Our request to you: Please endorse this global appeal. A brief confirmation in reply to this email is sufficient. Please indicate the name and role to be listed on the appeal. Alternatively, you may complete the document yourself using the link provided above.

10. Sources

1. Center for AI Safety, *Statement on AI Risk* — https://safe.ai/work/statement-on-ai-risk
2. Pope Leo XIV (2026), Encyclical *Magnifica humanitas* — https://www.vatican.va/content/leoxiv/de/encyclicals/documents/20260515-magnifica-humanitas.html
3. World Council of Churches (2023), *Statement on the Unregulated Development of Artificial Intelligence* — 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), ChatGPT users and delusions — https://futurism.com/chatgpt-users-delusions
7. *Raine v. OpenAI* (lawsuit against 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), *Federal Situation Report on Cybercrime 2025* — https://www.bka.de/DE/Presse/Listenseite_Pressemitteilungen/2026/Presse2026/260512_PM_BLB_Cybercrime.html
10. Euronews (2025), AI-assisted bioweapons / pandemic scenario — 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 on biological risks (US Senate) — 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, AI-controlled drone swarms — https://uacrisis.org/en/1288
15. Scientific American (2025), AI solves an old Erdős problem — https://www.scientificamerican.com/article/ai-just-solved-an-80-year-old-erdos-problem-and-mathematicians-are-amazed/
16. PauseAI Germany, "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), US embargo on Mythos — https://theconversation.com/why-the-us-government-shut-down-anthropics-latest-claude-ai-model-285223
19. ZDF (2026), BSI President Plattner on Claude Mythos — https://www.zdfheute.de/politik/deutschland/ki-anthropic-claude-mythos-schwachstellen-software-bsi-100.html
20. CNBC (2026), Mythos in US government systems (Sen. Warner) — https://www.cnbc.com/2026/06/23/anthropics-mythos-model-found-vulnerabilities-in-classified-us-government-systems-official-says.html
21. Anthropic, *When AI Builds Itself* (recursive self-improvement) — 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, 25 percent chance — https://www.axios.com/2025/09/17/anthropic-dario-amodei-p-doom-25-percent
29. Steven Adler (2025), on the alignment problem — https://x.com/sjgadler/status/1883928203800265023
30. International Dialogues on AI Safety (Shanghai, 2025) — https://idais.ai/dialogue/idais-shanghai/
31. Mila (2025), Bengio, most-cited scientist — 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), Fairwater AI Data Centre — https://blogs.microsoft.com/blog/2025/09/18/inside-the-worlds-most-powerful-ai-datacenter/
35. ZEISS, EUV lithography (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, *Statement* (English original) — https://pauseai.info/statement
38. Scientific American (2026), AI researcher Jacob Coxon resigns from Anthropic — https://www.scientificamerican.com/article/ai-jacob-coxon-quit-extinction-fears-security-experts-see-familiar-fight/
39. NBC News (2026), Two researchers leave Anthropic and Google over AI safety concerns — https://www.nbcnews.com/tech/security/two-ai-researchers-leave-anthropic-google-safety-concerns-rcna597086