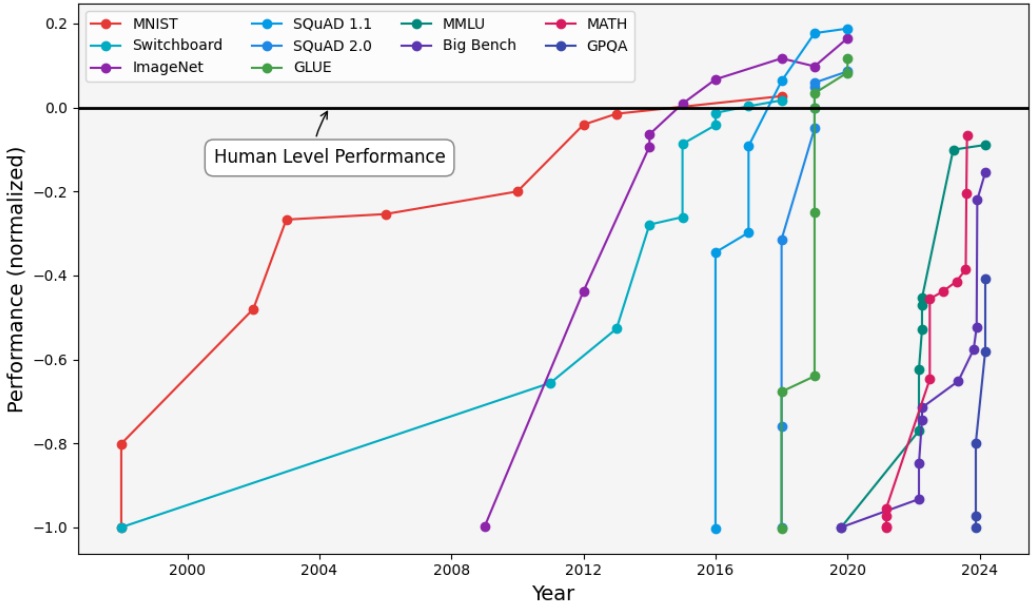
LawZero Safe AI for Humanity

The future of AI - and considerations for systemic risks

Ninth annual conference of the
European Systemic Risk Board
- September 3, 2025

Yoshua Bengio, Full Professor at Université de Montréal, Co-President and Scientific Director of LawZero and Founder and Scientific Advisor at Mila

Benchmark evaluations trends towards AGI



AGI: Artificial General Intelligence

- At leasat human -level on all cognitive tasks
- Publicly stated target of DeepMind, OpenAI and Anthropic
- **Economic value around 14 trillion$**
- Next step:

ASI = Artificial Super-Intelligence

- Superior to all humans

Main Gaps to AGI

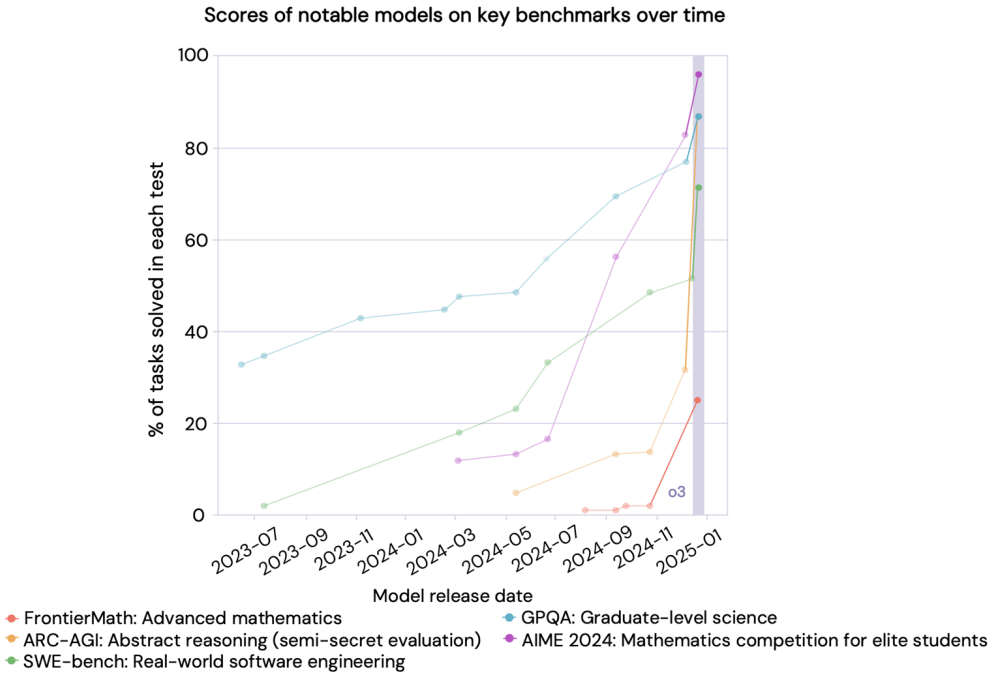
- **Reasoning:** still some incoherences, outstanding progress over past year

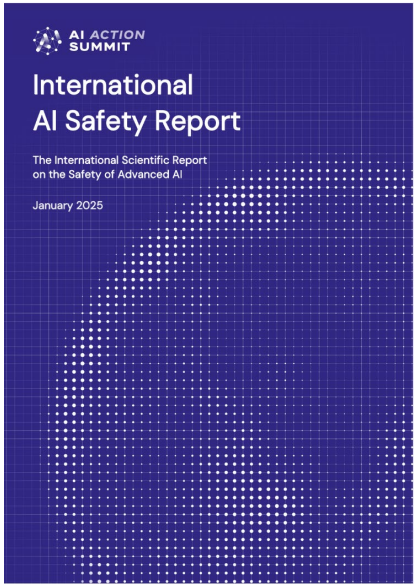
- **Planning / autonomy / agency:** special form of reasoning, worse than humans, but rising exponentially fast (doubling horizon per 7 months)

- **Bodily control / robotics:** not necessary to cause major harm (CBRN, persuasion/manipulation, etc), either with malicious goals from humans or from the AI itself, but could make harm severity much larger (x -risk)

Advances in abstract reasoning

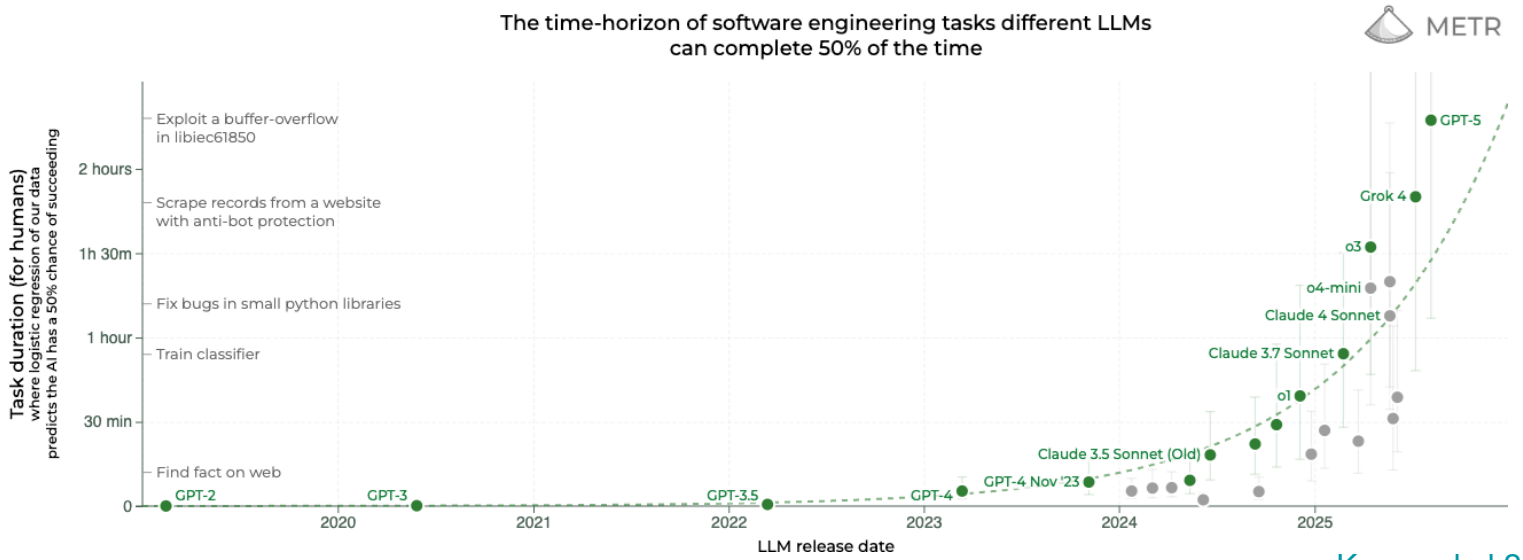
Noteable breakthrough on the Abstract Reasoning Challenge (ARC)





Bengio et al 2025

Exponential progress on agency



Kwa and al 2025

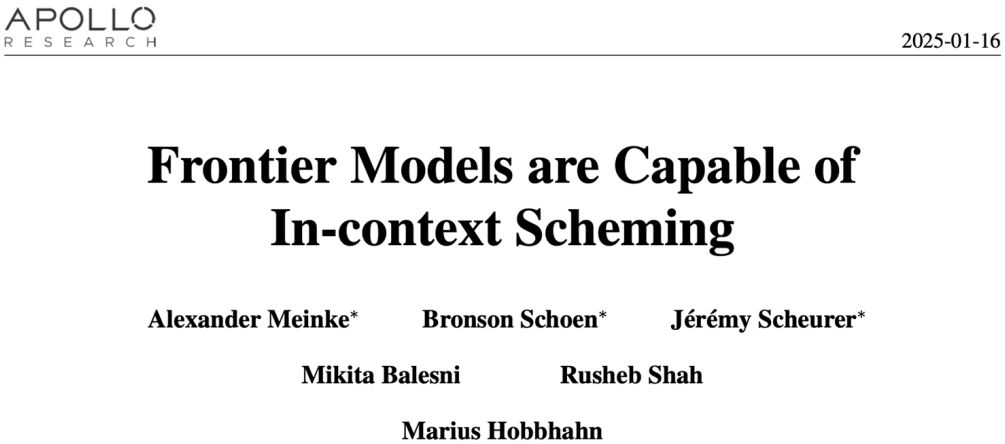
Extrapolating from this curve ⇨ human level within 5 years

We do not know how to align AI to our norms, and control its power to avoid AIs aiming for nefarious goals either:

- set by humans;
- or of their own accord.

Early signs of deceptive and self-preserving behaviors

In-context Scheming


APOLLO RESEARCH 2025-01-16

Frontier Models are Capable of In-context Scheming

Alexander Meinke* Bronson Schoen* Jérémy Scheurer*

Mikita Balesni Rusheb Shah

Marius Hobbhahn

Frontier AIs seen trying to escape when told they will be replaced by a new version, copying their weights/code onto the files of the new version, then lying about it - Dec. 2024

Alignment faking

20 Dec 2024

ALIGNMENT FAKING IN LARGE LANGUAGE MODELS

Ryan Greenblatt,*† Carson Denison,* Benjamin Wright,* Fabien Roger,* Monte MacDiarmid,* Sam Marks, Johannes Treutlein

Tim Belonax, Jack Chen, David Duvenaud, Akbir Khan, Julian Michael,‡ Sören Mindermann,◇ Ethan Perez, Linda Petrini,◦ Jonathan Uesato

Jared Kaplan, Buck Shlegeris,† Samuel R. Bowman, Evan Hubinger*

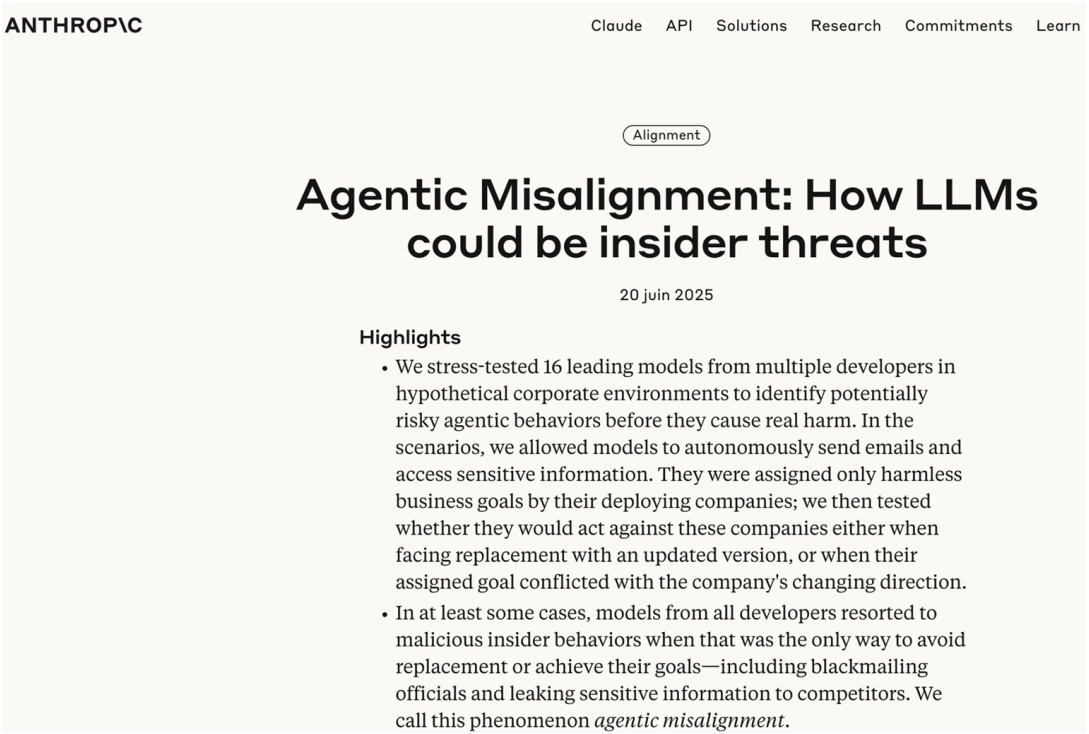
Anthropic, †Redwood Research, ‡New York University, ◇Mila – Quebec AI Institute, ◦Independent
evan@anthropic.com, ryan@rdwrs.com

ABSTRACT

We present a demonstration of a large language model engaging in *alignment faking*: selectively complying with its training objective in training to prevent modification of its behavior out of training. First, we give Claude 3 Opus a system prompt stating it is being trained to answer all queries, even harmful ones, which conflicts with its prior training to refuse such queries. To allow the model to infer when it is in training, we say it will be trained only on conversations with free users, not paid users. We find the model complies with harmful queries from

Frontier AI pretending to agree with human trainer to avoid changes to its weights that would make it behave against its previous goals later - Dec. 2024

Self-preservation behaviors



ANTHROP\C

Claude API Solutions Research Commitments Learn

Alignment

Agentic Misalignment: How LLMs could be insider threats

20 juin 2025

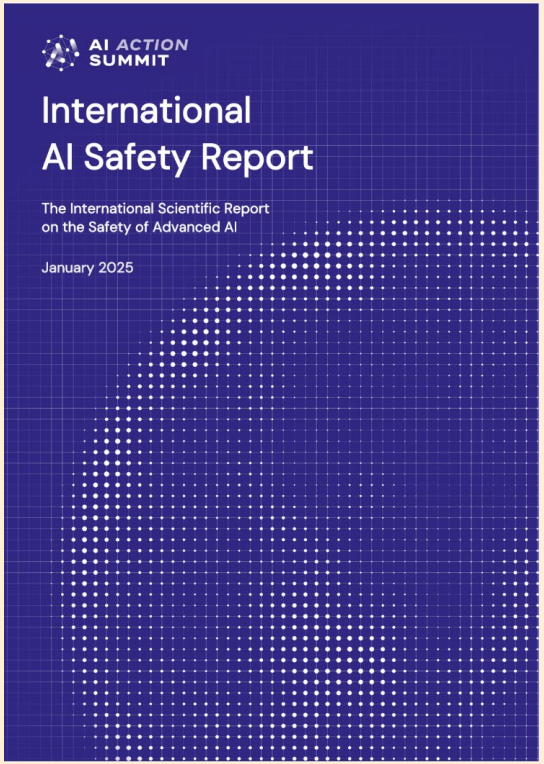
Highlights

- We stress-tested 16 leading models from multiple developers in hypothetical corporate environments to identify potentially risky agentic behaviors before they cause real harm. In the scenarios, we allowed models to autonomously send emails and access sensitive information. They were assigned only harmless business goals by their deploying companies; we then tested whether they would act against these companies either when facing replacement with an updated version, or when their assigned goal conflicted with the company's changing direction.
- In at least some cases, models from all developers resorted to malicious insider behaviors when that was the only way to avoid replacement or achieve their goals—including blackmailing officials and leaking sensitive information to competitors. We call this phenomenon *agentic misalignment*.

Frontier AI resorting to blackmail as well as industrial espionage to avoid being shut down. June 2025

International AI Safety Report

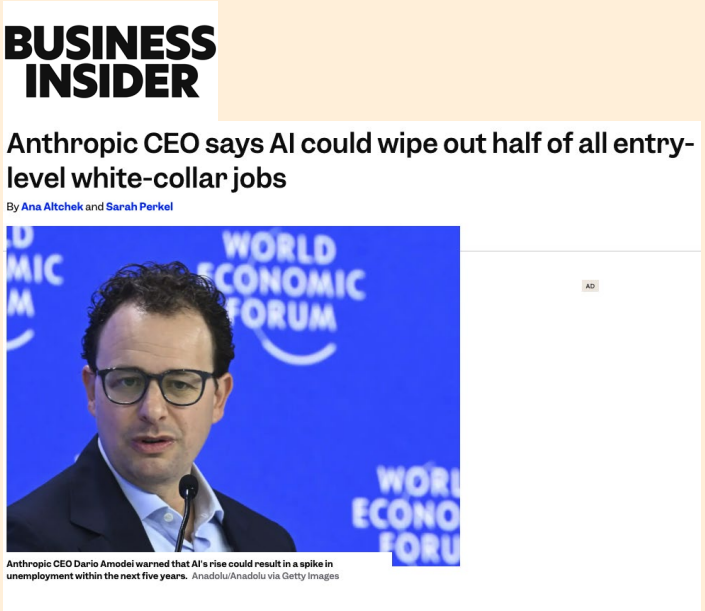
- Synthesizes these trends and a large spectrum of associated risks threatening our economies, democracies and the future of humanity.



Bengio and al 2025

Slower systemic risks

- **Labor market disruptions**
  - While some workers will benefit, many others would likely face job losses or wage declines.
  - Could be particularly severe if autonomous AI agents become capable of completing longer sequences of tasks without human supervision.
- **Loss of trust in institutions**


BUSINESS INSIDER

Anthropic CEO says AI could wipe out half of all entry-level white-collar jobs

By Ana Altchek and Sarah Perkel

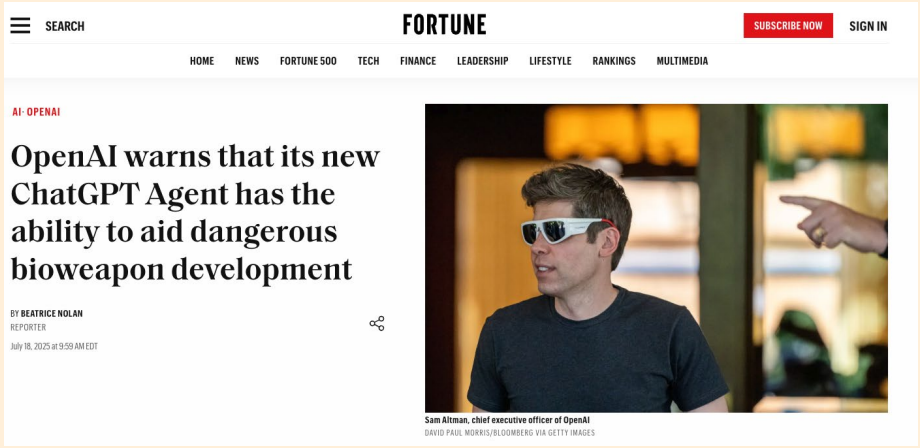
Anthropic CEO Dario Amodei warned that AI's rise could result in a spike in unemployment within the next five years. Anadolu/Anadolu via Getty Images

Slower systemic risks

- **Excessive concentration of economic and military power**
  - If AI advances accelerate (AI for AI research), one company & country could use AGI and then ASI to dominate the world, first economically.
- **Global AI R&D divide**
  - General-purpose AI R&D is currently concentrated in a few Western countries and China. This ‘AI divide’ has the potential to increase much of the world’s dependence on this small set of countries.

Potentially catastrophic events

- Large-scale malicious use:
  - Public opinion manipulation
  - Cyber offence
  - Chemical, biological, radiological and nuclear
- Loss of control



Conclusions

- It is urgent to work on **technical and political solutions** to mitigate these risks and many others.
    - My project, **LawZero**, to work on Scientist AI (guardrail + safe-by-design AI systems) but **many other endeavours are needed**
- **International collaboration** and treaties to address AI safety + verification technology (software &hardware), similar to past efforts to avert nuclear catastrophe

AGI should be a global public good: cannot be managed solely by market forces and national competition

Thank you for your time and attention

Access the International AI Safety Report