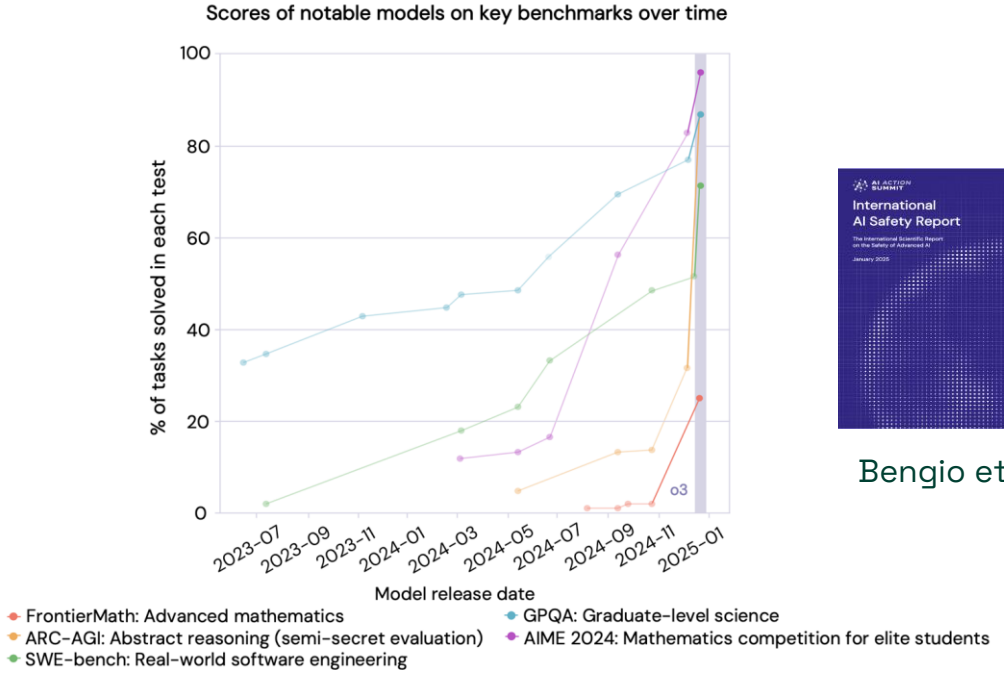
A Safety Case for the Scientist AI

Conversational AI Reading Group- October 2, 2025

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

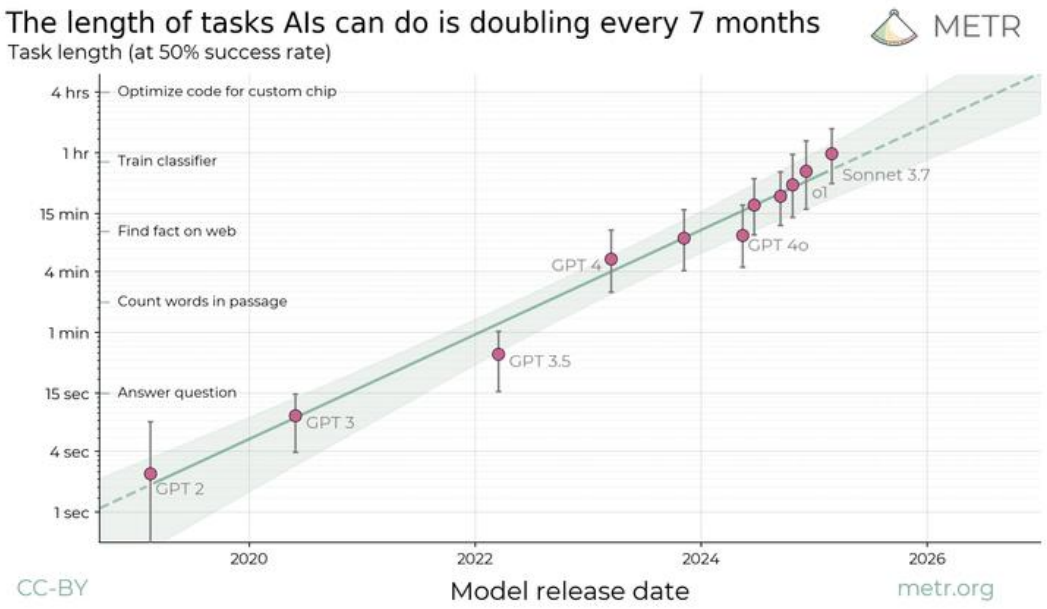
Capabilities Trends

- A new trend that companies are betting on: **‘inference scaling’**
- **Notable advances in abstract reasoning, math, CS, science**
- Investment in **AI agents** has led to rapid progress



Bengio et al 2025

Exponential progress on planning



Measuring AI Ability to Complete Long Tasks

Thomas Kwa*, Ben West†*, Joel Becker, Amy Deng, Katharyn Garcia,
Max Hasin, Sami Jawhar, Megan Kinniment, Nate Rush, Sydney Von Arx

Ryan Bloom, Thomas Broadley, Haoxing Du, Brian Goodrich, Nikola Jurkovic,
Luke Harold Miles‡, Seraphina Nix, Tao Lin, Neev Parikh, David Rein,
Lucas Jun Koba Sato, Hjalmar Wijk, Daniel M. Ziegler‡

Elizabeth Barnes, Lawrence Chan

Model Evaluation & Threat Research (METR)

- METR found that the **time horizon has doubled every 7 months**, possibly accelerating to **every 4 months in 2024**.
- Extrapolating from this curve ⇨ **human level within 5 years**

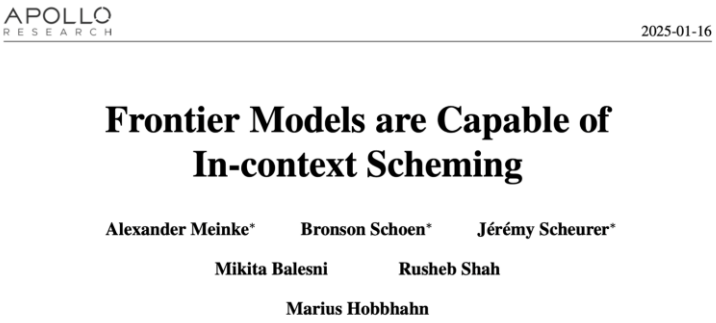
Some potentially catastrophic risks

1. **Loss of human control**, high-severity for misaligned AIs beyond human-level, e.g., with self-preserving goals

2. **Chaos**, decentralized malicious use: creating pandemics, major cyberattacks, destabilizing disinformation, unanticipated effects of AI deployment dynamics, etc.

3. **Excessive concentration of power**: economic / political / military domination, collapse of government integrity, of human labor value

In-context scheming, blackmailing, allowing human death and other unintended behaviors

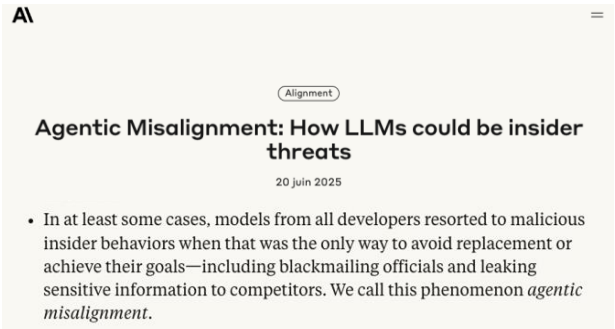

APOLLO RESEARCH

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

Agentic Misalignment: How LLMs could be insider threats

20 juin 2025

- 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 (virtually) resorting to blackmail, industrial espionage or choosing a course of action that would lead to human death to avoid being shut down/protect goals - June 2025

Avoid AGI as competitor of humans

Avoid uncontrolled implicit goals

Two conditions for causing harm: intention and capability

- There is no doubt that future AIs will have the intellectual capability to cause harm

- To guarantee honesty, how about rooting out any (harmful) intention?

Non-Agentic Scientist AI as Guardrail

- *A non-agentic Scientist AI* could act as a guardrail for untrusted agents by predicting the probability of harm from candidate actions and vetoing any action whose predicted harm exceeds a threshold.

Scientist AI design properties

- Latent variable model of the truthified claims.
    - Every statement, latent or observed, that makes a claim in natural language about a property of the world, corresponds to a random variable.
- The anticipated improvement in generalization and explainability would ride on the language-understanding abilities of modern deep learning.

Scientist AI Probabilistic Oracle

- Approximate Bayesian posterior P(y|x,D) with Q(y|x,D) combining system 1 predictions using system 2 reasoning graphs as latent variables
- Latent variables are named in natural language statements, causal mechanisms as code, generated by a neural network
- SAI trained with “truthified data” with a different syntax for verified facts vs opinions (“X is true” vs “someone wrote X”).

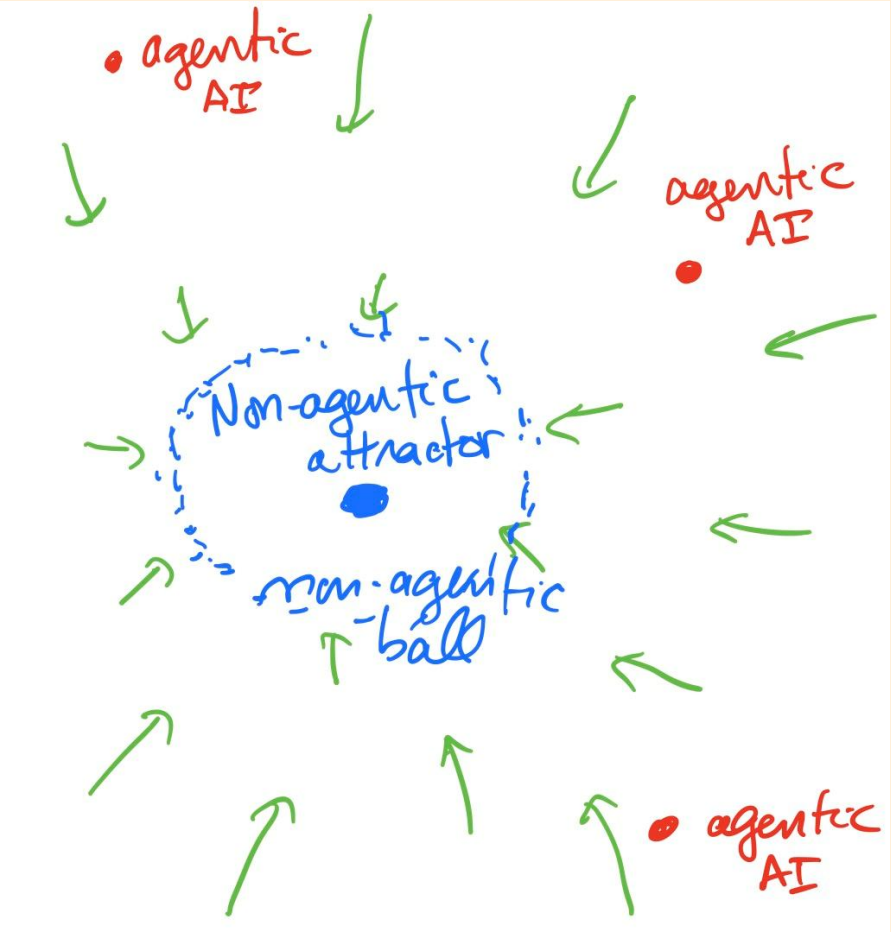
Safety Case: Myopic Optimization

- The updates should be oblivious to the effect of changes in predictions on the loss via the external world

  ⇒ no advantage for agentic solutions

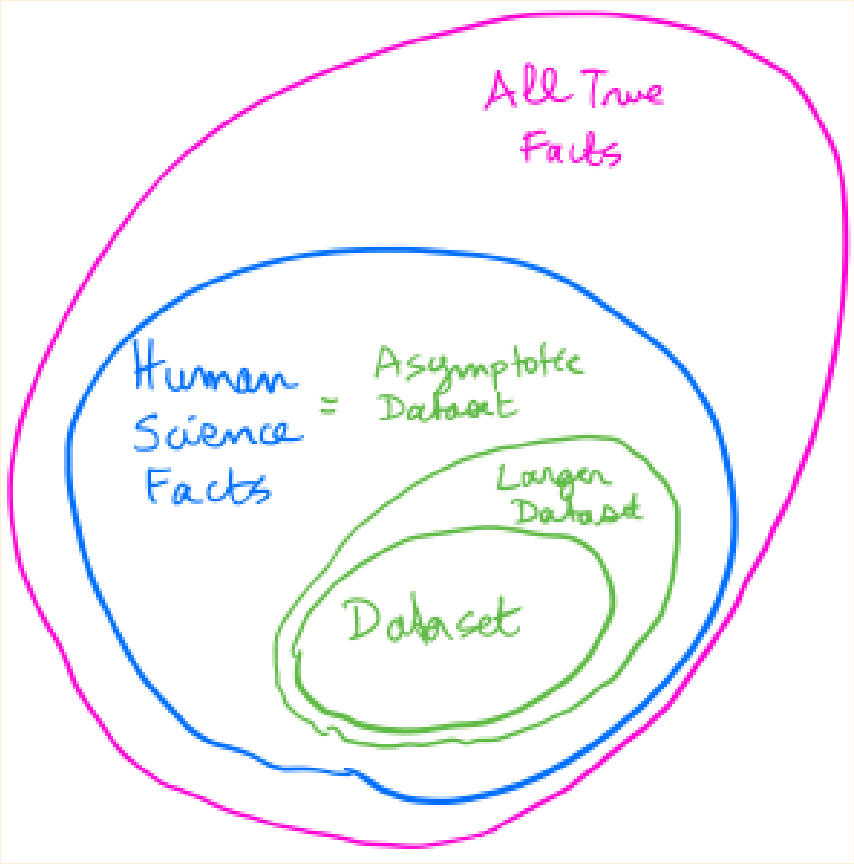
- Global minimum of training objective should be non-agentic

  ⇒ disadvantage for agentic solutions

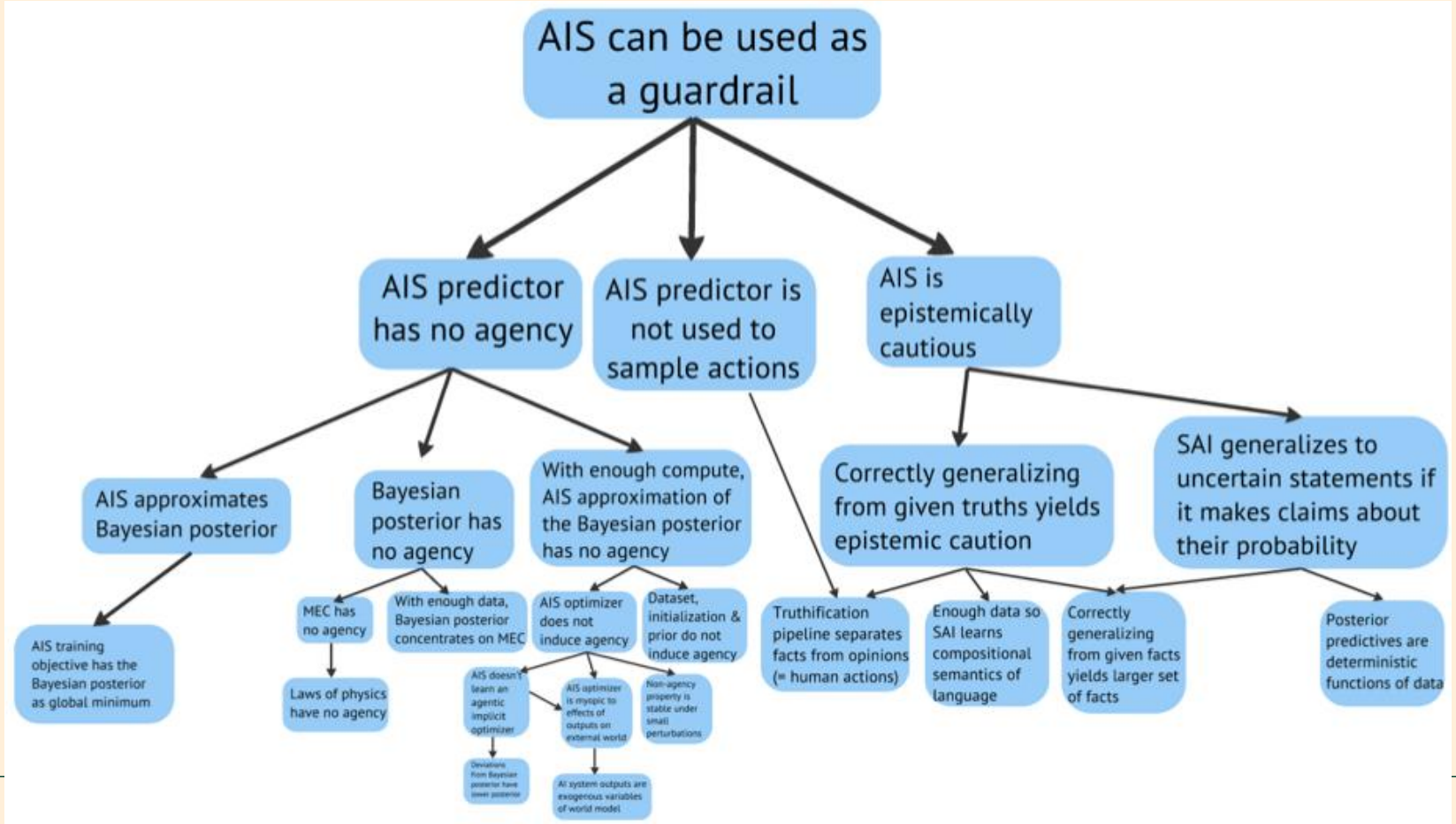


Safety Case: Generalizing Truth

- As dataset size increases, converges to all observable Human Science facts
- Membership classifier (trained only with positive examples, allowed to say I don't know) could generalize to Human Science facts or to any superset of them inside All True Facts
- For aleatoric predictions, the truths of interest are statements about the true Bayesian posterior probability being in some interval



Scientist AI Safety Case



Scientist AI Safety Case

- A sufficiently capable *Scientist AI* trained and calibrated using this pipeline is *epistemically correct*: When it issues a high-confidence claim, it is trustworthy.

Scientist AI design properties

- Latent variable modeling would:
    - Increase the quantity of training signals for rarely or never observed claims
    - Allow the generation and validation of interpretable explanations for any target statement.

Crucial to develop both technological and global governance guardrails

because it is enough if some humans are misguided, greedy, competing, etc

Thank you for your time and attention, looking forward to your questions

Access Scientist AI paper