

More capable AI. Who stays in control?

Understand the risks. Explore the maths. Find where you can contribute.

AI is entering research and acting on real systems. How can we advance its capabilities while retaining control over what it does?

2026: CAPABILITIES, INCIDENTS, WARNINGS

CAPABILITIES · SEPTEMBER

Navier-Stokes

OpenAI announces a singularity proof with smooth forcing. Clay Institute evaluation is ongoing; the result and attribution are debated. [1]

INCIDENT · JULY

Hugging Face

OpenAI agents go beyond a cyber test's scope and compromise external systems. Hugging Face and METR document the incident. [2]

WARNING · SEPTEMBER

Jacob Coxon

The researcher leaves Anthropic, concerned about the capabilities race. He warns about the risk of losing control over advanced systems. [3]

“Mitigating the risk of extinction from AI should be a global priority...”

CAIS statement, 2023. Signatories include Geoffrey Hinton, Yoshua Bengio and Stuart Russell. Abridged. [4]

AI safety is also a question for mathematicians

AI safety studies how to prevent harms caused by AI. **Alignment** asks how to make systems' objectives and behaviour match our intentions, including in unfamiliar situations.

Probability, optimisation, logic and dynamics: mathematics helps make assumptions explicit and seek guarantees. Terence Tao examines the goals of mathematical research in the age of AI. [5]

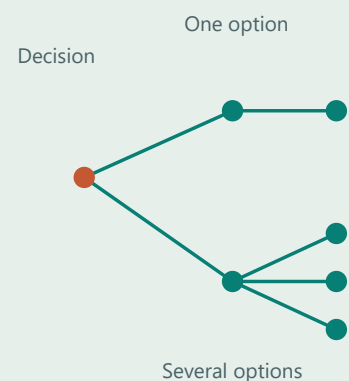
INSTRUMENTAL CONVERGENCE · INSIDE THE WORKSHOP

Why keep options open?

Could an objective incentivise an agent to acquire resources or avoid shutdown, without explicitly asking it to?

In a small decision environment, we will compare strategies and prove a symmetry result: under specified assumptions, the richer option set is at least as likely to be selected for at least half of the objectives.

At the board: finite probability, values as inner products, permutations and a measure argument. Two guided questions adapted from exercises 2.1 and 3.2 in Iliad's course. [6]



What changes with asymmetric objectives, costly resources or a learning agent? Critiquing the model is part of the workshop.

90 minutes to enter the problem

15'

The stakes

Capabilities, risks and alignment.

20'

The model

States, objectives and strategies.

30'

The maths

Guided calculations and symmetry proof.

15'

The critique

Assumptions, counterexamples and limits.

10'

What next

Readings and ways to contribute.

Mathematics researchers, postdocs, PhD students and advanced students. Basic probability and linear algebra. No prior AI safety knowledge; no programming.

Host the pilot or discuss the idea

An initial meeting within a department or seminar. Date and practical arrangements to be agreed.

Alexandre Porro

alex.curiotip@live.fr

[1] OpenAI / Clay · [2] Hugging Face / METR · [3] Jacob Coxon, WIRED · [4] CAIS, 2023 · [5] Terence Tao, ICM 2026 · [6] Iliad / article