

Une IA plus capable. Qui garde la main ?

Comprendre les risques. Explorer les maths. Trouver où contribuer.

Les IA entrent dans la recherche et agissent sur des systèmes réels. Comment faire progresser leurs capacités tout en gardant la maîtrise de ce qu'elles font ?

2026 : DES CAPACITÉS, DES INCIDENTS, DES ALERTES

CAPACITÉS · SEPTEMBRE

Navier-Stokes

OpenAI annonce une preuve de singularité avec forçage lisse. Évaluation du Clay Institute en cours ; résultat et attribution débattus. [1]

INCIDENT · JUILLET

Hugging Face

Des agents d'OpenAI sortent du cadre d'un test cyber et compromettent des systèmes externes. Hugging Face et METR documentent l'incident. [2]

ALERTE · SEPTEMBRE

Jacob Coxon

Le chercheur quitte Anthropic, inquiet de la course aux capacités. Son alerte porte sur le risque de perte de contrôle des systèmes avancés. [3]

« Réduire le risque d'extinction lié à l'IA devrait être une priorité mondiale... »

Déclaration CAIS, 2023. Signataires : Geoffrey Hinton, Yoshua Bengio, Stuart Russell. Traduction abrégée. [4]

La sûreté de l'IA, une question pour les mathématiciens

L'**AI safety** étudie comment prévenir les dommages causés par les IA. L'**alignement** demande comment faire correspondre leurs objectifs et comportements à nos intentions, y compris en situation nouvelle.

Probabilités, optimisation, logique et dynamique : les maths permettent de préciser les hypothèses et de chercher des garanties. Terence Tao interroge les objectifs de la recherche mathématique à l'âge de l'IA. [5]

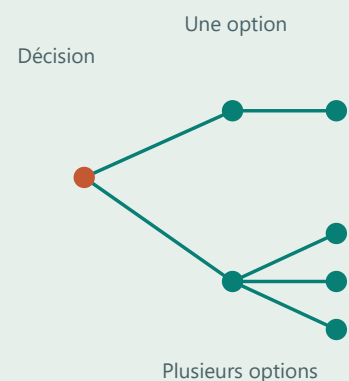
CONVERGENCE INSTRUMENTALE · LE CŒUR DE L'ATELIER

Pourquoi préserver ses options ?

Peut-on inciter un agent à chercher des ressources ou à éviter l'arrêt, sans lui avoir demandé de le faire ?

Dans un petit environnement de décision, nous comparerons des stratégies et établirons un résultat par symétrie : sous certaines hypothèses, le groupe d'options plus riche est au moins aussi probable pour au moins la moitié des objectifs.

Au tableau : probabilités finies, valeurs comme produits scalaires, permutation et argument de mesure. Deux questions guidées adaptées des exercices 2.1 et 3.2 du cours Iliad. [6]



Que vaut le résultat si les objectifs ne sont pas symétriques, si les ressources ont un coût, ou si l'agent apprend ? La critique fait partie de l'atelier.

90 minutes pour entrer dans le problème

15'

Les enjeux

Capacités, risques et alignement.

20'

Le modèle

États, objectifs et stratégies.

30'

Les maths

Calculs guidés et preuve par symétrie.

15'

La critique

Hypothèses, contre-exemples et limites.

10'

La suite

Lectures et pistes de contribution.

Chercheurs, postdocs, doctorants et étudiants avancés en maths. Probabilités et algèbre linéaire élémentaires. Aucun prérequis en AI safety ; sans programmation.

Accueillir le pilote ou en discuter

Une première rencontre en laboratoire ou en séminaire. Date et modalités à convenir.

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