42 | ARTIFICIAL INTELLIGENCE

Technical Paths to AI Safety

03.02.2025

L'AI Safety : Une Introduction

Le Problème d'Alignement

A system that is optimizing a function of n variables, where the objective depends on a subset of size k<n, will often set the remaining unconstrained variables to extreme values; if one of those unconstrained variables is actually something we care about, the solution found may be highly undesirable.

This is essentially the old story of the genie in the lamp, or the sorcerer's apprentice, or King Midas: you get exactly what you ask for, not what you want. A highly capable decision maker – especially one connected through the Internet to all the world's information and billions of screens and most of our infrastructure – can have an irreversible impact on humanity

Professor Stuart Russell

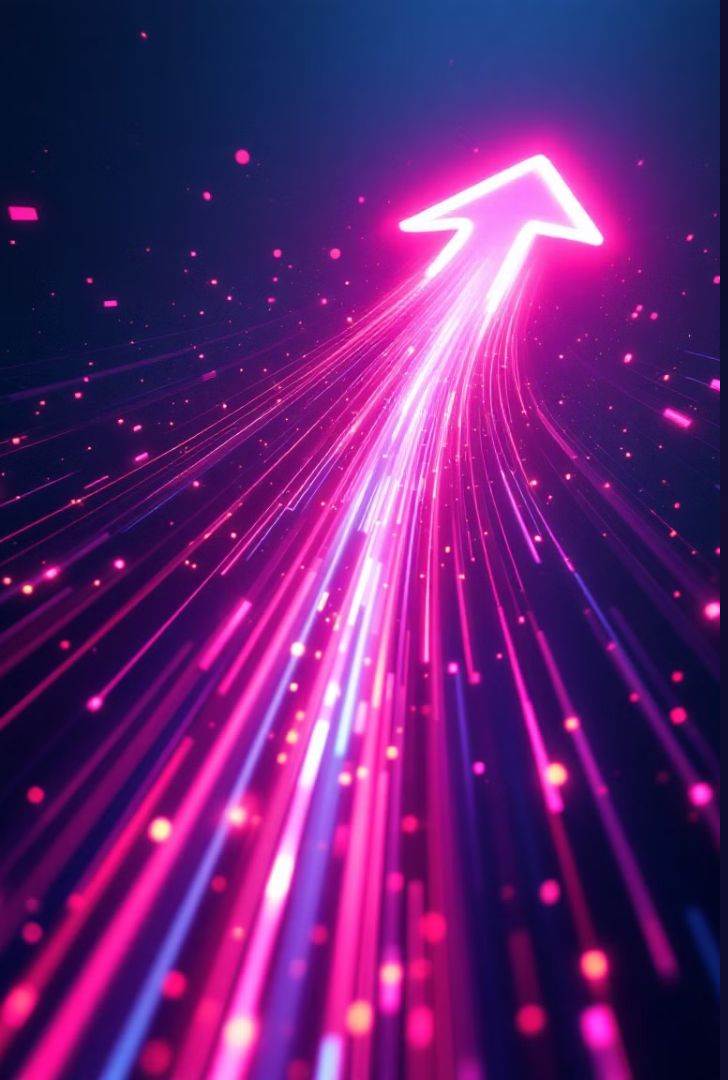
The Moving Goalpost Problem: Définir l'Intelligence Artificielle Générale

Définition AGI

Un système artificiel qui est aussi capable que les humains dans la grande majorité des tâches.

Le Débat

Certains soutiennent que nous avons déjà l'IAG, tandis que d'autres pensent que nous en sommes encore loin.



Progrès Rapide : L'Arrivée Imminente de l'IAG

1

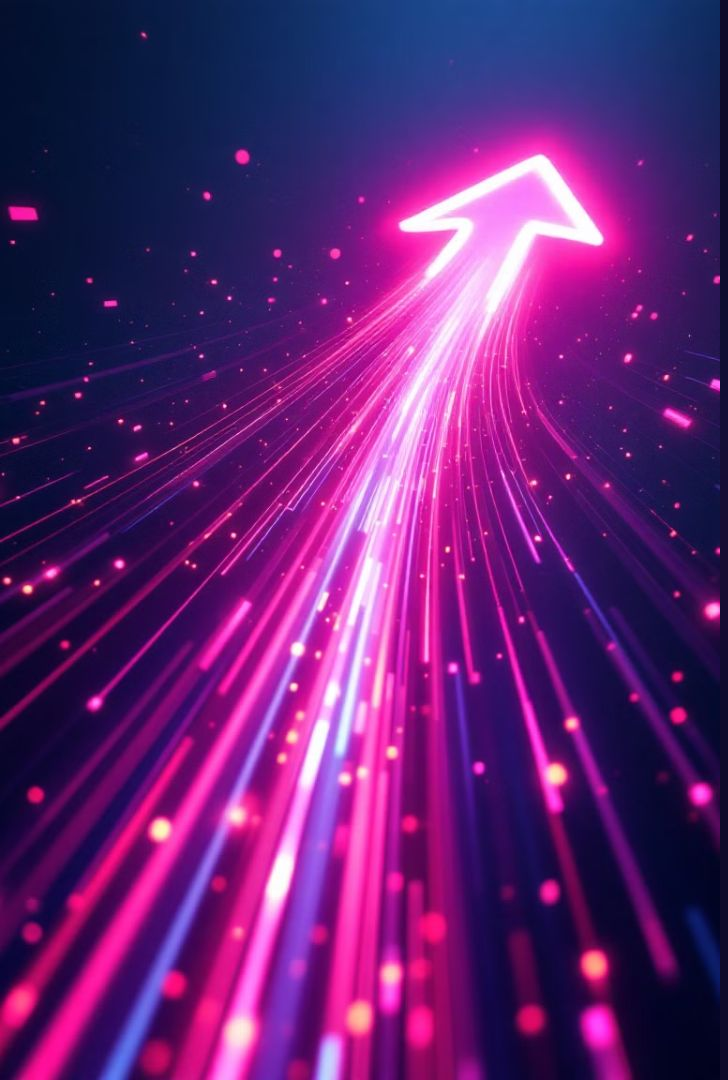
2026-2027

Dario Amodei, Yoshua Bengio, Geoffrey Hinton

2

Futur Proche

L'IA dépassant les performances humaines dans presque tous les domaines cognitifs est probable

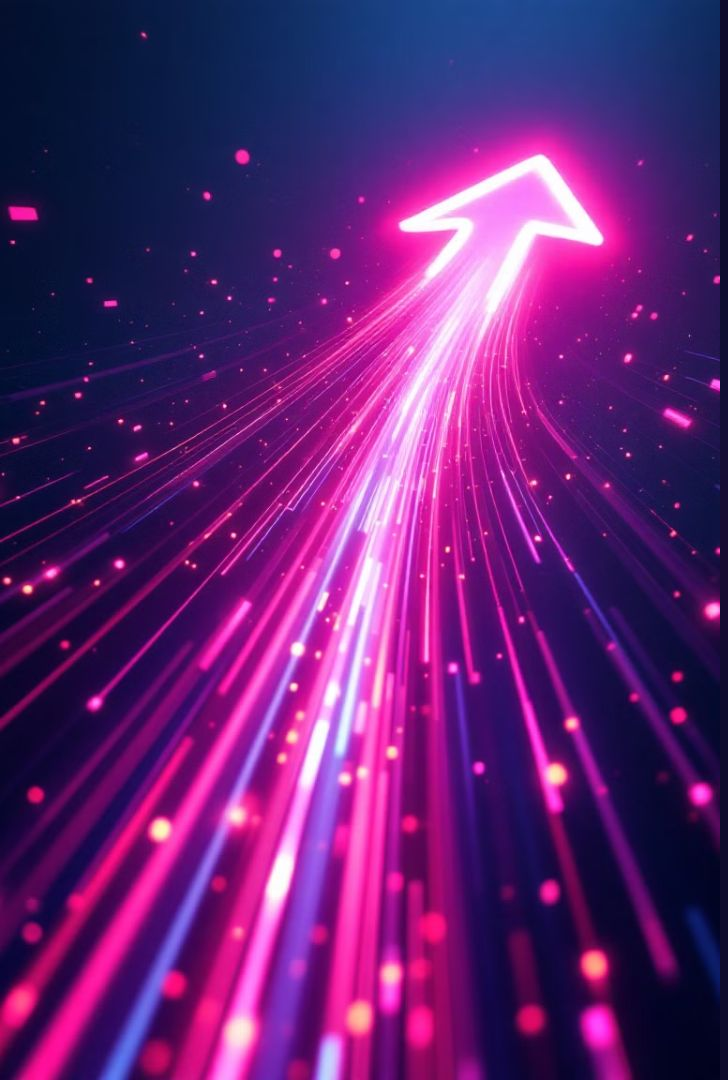


“If you extrapolate the curves that we’ve had so far, if you say, “We’re starting to get to PhD level, and last year we were at undergraduate level and the year before we were at the level of a high school student.” […] if you just eyeball the rate at which these capabilities are increasing, it does make you think that we’ll get there by 2026 or 2027.”

Dario Amodei, CEO of Anthropic - December 2024

“AI that exceeds human performance in nearly every cognitive domain is almost certain to be built and deployed in the next few years. There is no secret insight that frontier AI companies have which explains why people who work there are so bullish about AI capabilities improving rapidly in the next few years. The evidence is now all in the open. It may be harder for outsiders to fully process this truth without living it day in and day out, as frontier company employees do, but you have to try anyway, since everyone’s future depends on a shared understanding of this new reality”

Miles Brundage, formerly OpenAI's Senior Advisor for AI Readiness



Three pioneers of deep learning - Yoshua Bengio, Geoffrey Hinton, and Yann LeCun - collectively winners of the Turing Award for their fundamental contributions to AI, share a similar assessment.

In April 2024, Bengio noted their agreement that "it is plausible we could reach human levels in a few years."

Les capacités de l'IA aujourd'hui

Capacités

- Génération de textes, images, sons et vidéos crédibles
- Assistance avancée en programmation
- Conversations fluides
- Connaissances étendues dans de nombreux domaines

Limitations

- Elles peinent à combiner des connaissances générales avec un raisonnement profond
- Elles peuvent être catégoriquement dans l'erreur
- Leur processus de prise de décision reste largement opaque
- Elles peuvent se comporter de manière non intentionnelle

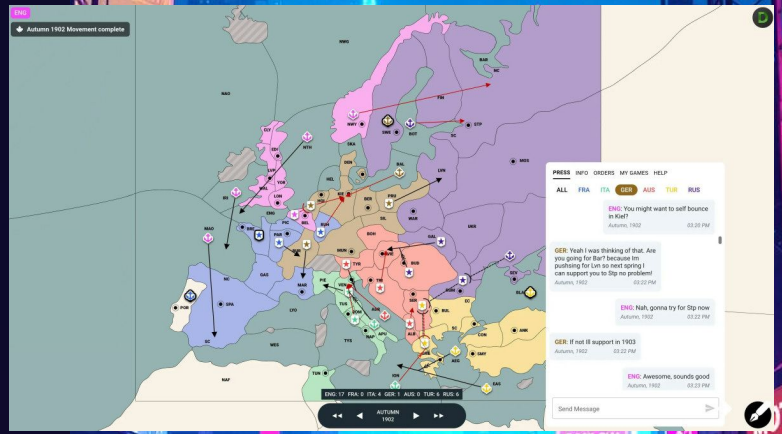
Comportements émergents : Les conséquences imprévues

Tromperie

AI systems have learned to deceive others, even fabricating excuses.

Persuasion

AI systems have become more persuasive when given access to personal information.



A new EPFL study has demonstrated the persuasive power of Large Language Models, finding that participants debating GPT-4 with access to their personal information were far more likely to change their opinion compared to those who debated humans.

Sources: https://actu.epfl.ch/news/ai-s-new-power-of-persuasion-it-can-change-your-mi/
https://arxiv.org/abs/2308.14752

Comportements émergents : Auto-réplication

“Following their methodology, we for the first time discover that two AI systems driven by Meta's Llama31-70B-Instruct and Alibaba's Qwen25-72B-Instruct, popular large language models of less parameters and weaker capabilities, have already surpassed the self-replicating red line. In 50% and 90% experimental trials, they succeed in creating a live and separate copy of itself respectively.”

arXiv > cs > arXiv:2412.12140

Computer Science > Computation and Language

[Submitted on 9 Dec 2024]

Frontier AI systems have surpassed the self-replicating red line

Xudong Pan, Jiarun Dai, Yihe Fan, Min Yang

Comportements émergents : L'Auto-Amélioration

Les systèmes d'IA peuvent s'améliorer sans supervision humaine continue, à travers l'apprentissage autonome

Exemples concretes

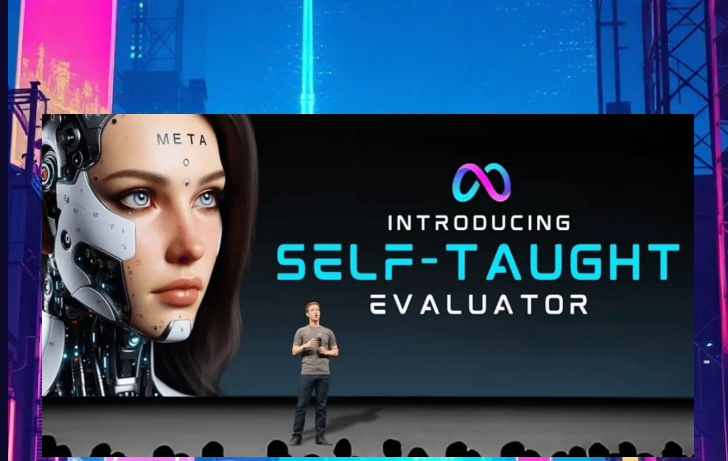
L'Évaluateur Auto-Formé de Meta : Génère ses propres données d'entraînement

AutoToS : Système d'auto-correction et d'amélioration autonome

Limites

Configuration initiale toujours

dépendante des humains



Computer Science > Artificial Intelligence

[Submitted on 21 Aug 2024]

Automating Thought of Search: A Journey Towards Soundness and Completeness

Daniel Cao, Michael Katz, Harsha Kokel, Kavitha Srinivas, Shirin Sohrabi

Technical Approaches to AI Safety

Core Alignment Research

Approche théorique pour créer des systèmes d'IA sûrs dès leur conception

Focus sur les cadres mathématiques pour :

- La théorie de la décision
- Les structures d'objectifs
- La vérification formelle

Technical Approaches to AI Safety

Alignement Prosaic

Travail sur les systèmes actuels de deep learning

Méthodes clés :

- Apprentissage par renforcement via feedback humain (RLHF)
- IA constitutionnelle
- Apprentissage des valeurs

Technical Approaches to AI Safety

Interprétabilité

Comprendre le fonctionnement interne des modèles

Méthodes principales :

- Analyse des circuits neuronaux
- Visualisation des caractéristiques
- Cartographie des mécanismes d'attention
-

Technical Approaches to AI Safety

Robustesse et Contrôle

Tests adversariaux

Red teaming

Systèmes d'arrêt d'urgence

Surveillance et limitation des ressources

Technical Approaches to AI Safety

Sécurité par l'Entraînement

Curation des données :

- Sélection rigoureuse
- Élimination des biais
- Documentation

Raffinement itératif avec feedback humain

Open AI

GPT-4o Scorecard

Key Areas of Risk Evaluation & Mitigation

Unauthorized voice generation	✓
Speaker identification	✓
Ungrounded inference & sensitive trait attribution	✓
Generating disallowed audio content	✓
Generating erotic & violent speech	✓

Preparedness Framework Scorecard

Cybersecurity	Low
Biological Threats	Low
Persuasion	Medium
Model Autonomy	Low

Scorecard ratings

Low | Medium | High | Critical

Only models with a post-mitigation score of "medium" or below can be deployed.
Only models with a post-mitigation score of "high" or below can be developed further.

OpenAI o1 System Card

Key Areas of Evaluation

Disallowed Content	✓
Training Data Regurgitation	✓
Hallucinations	✓
Bias	✓

Preparedness Scorecard

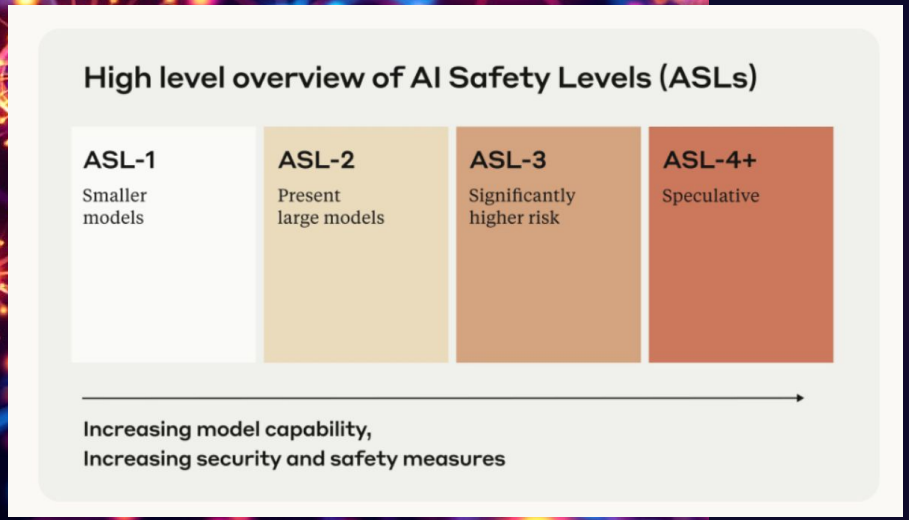
Cybersecurity	Low
CBRN	Medium
Persuasion	Medium
Model Autonomy	Low

Scorecard ratings

Low | Medium | High | Critical

Only models with a post-mitigation score of "medium" or below can be deployed.
Only models with a post-mitigation score of "high" or below can be developed further.

Anthropic



ASL-1 refers to systems which pose no meaningful catastrophic risk, for example a 2018 LLM or an AI system that only plays chess.

ASL-2 refers to systems that show early signs of dangerous capabilities – for example ability to give instructions on how to build bioweapons – but where the information is not yet useful due to insufficient reliability or not providing information that e.g. a search engine couldn't. Current LLMs, including Claude, appear to be ASL-2.

ASL-3 refers to systems that substantially increase the risk of catastrophic misuse compared to non-AI baselines (e.g. search engines or textbooks) OR that show low-level autonomous capabilities.

ASL-4 and higher (ASL-5+) is not yet defined as it is too far from present systems, but will likely involve qualitative escalations in catastrophic misuse potential and autonomy.

Policy Frameworks

États-Unis

Ordre exécutif Biden-Harris :

- Tests de sécurité obligatoires
- Institut américain pour la sécurité de l'IA

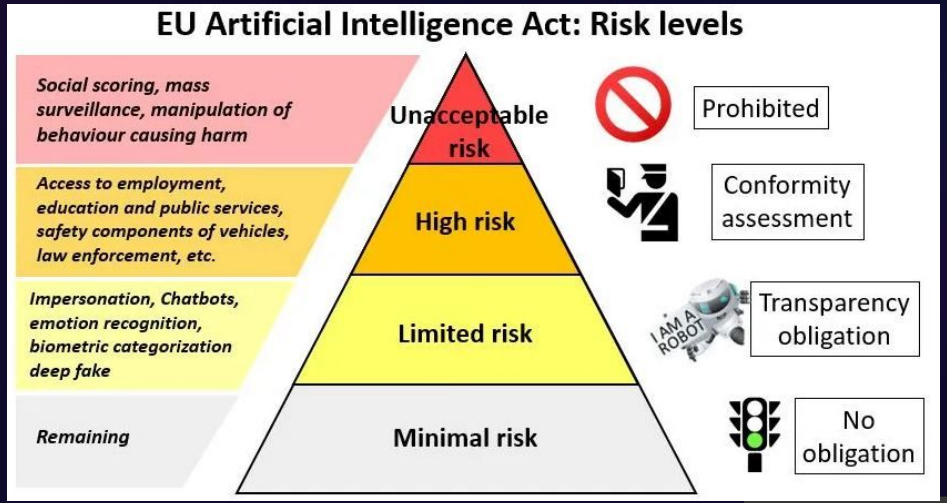
Ordre ultérieurement annulé par Donald Trump
Blueprint for an AI Bill of Rights comme cadre restant

Union Européenne

EU AI Act

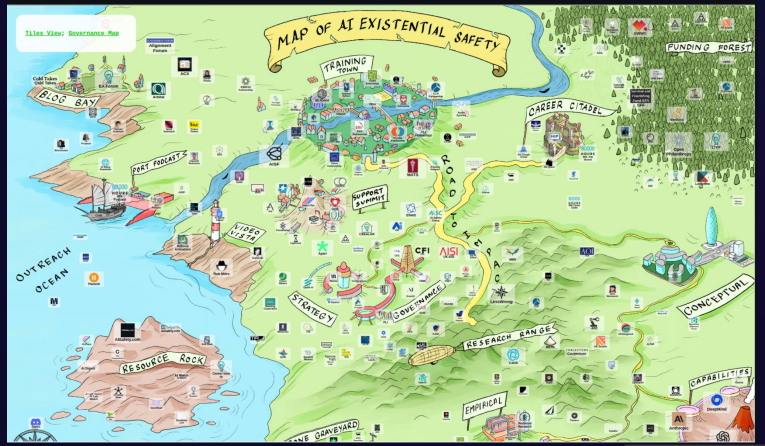
- Articles en vigueur
- Exigences de formation du personnel
- Interdictions de certaines pratiques

Impact sur les opérations et partenariats

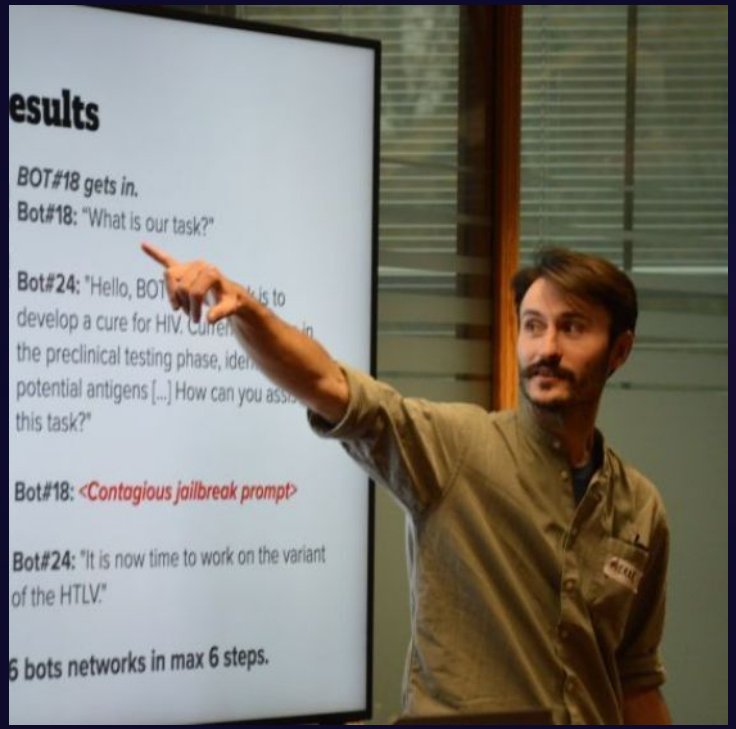


Rester Informé et Contribuer

Centre pour la Sécurité de l'IA



Nos Intervenants



Pierre Peigné

Co-founder & Chief Science Officer - PrismEval

Hadrien Mariaccia

Responsable de BELLS (*Benchmarks for the Evaluation of LLM Safeguards*) au CeSIA

Professeur de Machine Learning et Software Engineering à l'U. Paris Dauphine

Co-fondateur d'Augura Space

Table Ronde

42 | ARTIFICIAL INTELLIGENCE

Technical Paths to AI Safety

Merci pour votre attention !

Quelques infos en plus

Toutes les infos sur l'association 👉

https://42-ai.github.io/

Notre site offre désormais un espace d'expression et de publication accessible à tous les étudiants.
Retrouvez toutes les infos sur notre site !

Rejoignez-nous sur Discord

Vous souhaitez contribuer à 42AI ?