

# « Explique moi comme si j'avais 10 ans »

Enquête sur les conditions d'exercice du métier d'étudiant avec les  
IAG.

13/04/2026

Bilel Benbouzid



**Michał Wiczorek** • 2e  
Assistant Professor at University College Dublin - critica...  
3 j • Modifié •

[Se connecter](#)

The only acceptable AI use policy you need, courtesy of [Joep Schuurkes](#): <https://lnkd.in/dRncEtuK>

Afficher la traduction

## On the acceptance of GenAI

Joep Schuurkes — 5 April 2026

By using GenAI:

- I accept the models were trained on stolen data.
- I accept that the data was labeled by exploited workers.
- I accept the environmental costs of the data centers running these models.
- I accept that I am outsourcing some of my skills to a company.
- I accept these companies don't have a viable business model.
- I accept that I am granting more power to big tech and their vision for the world.
- I accept that I am granting more power to the United States.
- I accept that all this effort could have been spent elsewhere.

👍👍👍 2022

155 commentaires • 362 republications

Réactions



**Lou Welgryn** • Suivi  
Data for Good - Co-fondatrice Eclaircies & ESSEC Transition Alumni - inst...  
[Voir mon blog](#)  
1 h •

📅 Rendez-vous le 12 mai pour la sortie de notre livre avec [Théo Alves Da Costa](#) - "IA, le grand enfumage. Comprendre les enjeux, déconstruire les mythes, reprendre le pouvoir."

Prosperité généralisée, solutions magiques au changement climatique, libération du travail : toutes ces promesses sur l'IA que nous vendent les géants du numérique ne servent en réalité qu'un seul objectif : rendre leurs technologies toujours plus hégémoniques, et concentrer toujours un peu plus leur pouvoir. À travers ce livre, nous tentons de déconstruire méthodiquement chacun de ces grands « enfumages » pour mieux dévoiler la face cachée de l'industrie de l'intelligence artificielle, entre consommation vorace de ressources, renforcement des inégalités et technofascisme galopant. Nous appelons à reprendre le contrôle sur le déferlement de cette technologie dans nos existences. Et d'interroger à tout prix, collectivement et démocratiquement, les modalités de son déploiement, que l'on aime tant nous présenter comme inévitable et irrésistible.

Après un an de travail et d'écriture, nous vous donnons rendez-vous à l'Académie du climat pour en discuter, avec nous et avec des gens qui agissent sur le terrain pour proposer des solutions concrètes et inventives pour redonner du pouvoir aux citoyens. ( et surtout pour boire des coups après ;) )

On vous espère venir nombreux et nombreuses !

12 mai - Académie du climat - 18h30 à 21h

Et un merci tout particulier à [Laure-Hélène Accaoui](#) de nous avoir fait confiance et suivis dans cette aventure ☐ ainsi qu'à toutes les personnes qui nous ont accompagné dans ce travail de longue haleine en relisant, écoutant, discutant.. ou simplement en étant présent.

Pour vous inscrire : <https://lnkd.in/eBgNBuUw>



👍👍👍 Florian Gauthier et 78 autres personnes • 6 commentaires • 3 republications



**Charles Logan** • 2e  
Post-Doctoral Research Fellow at the Center for Responsible Tec...  
1 j •

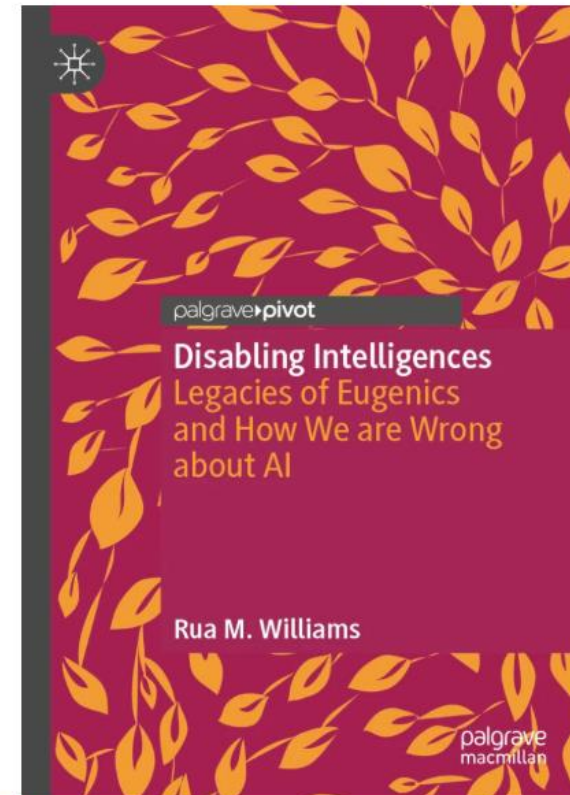
[+ Suivre](#)

Come join the [Civics of Technology Project](#)'s May book club on Wednesday, May 20, at 7pm ET. I'll be leading a discussion of [Rua M Williams](#)' "Disabling Intelligences: Legacies of Eugenics and How We Are Wrong About AI" - and Dr. Williams is joining too!

Register at: <https://lnkd.in/gPxcD7v2>

And for more about the connections between AI and eugenics, read Dr. Ruha Benjamin's piece "The New Artificial Intelligentsia" at: <https://lnkd.in/get55KWx>

Afficher la traduction



👍👍👍 Jeffrey Yost et 101 autres personnes • 1 commentaire • 27 republications

Réactions



👍 J'aime    💬 Commenter    🔄 Republier    ✉ Envoyer



The Algorithmic Justice League

34 467 abonnés

1 j •



"Models, despite their reputation for impartiality, reflect goals and ideology. Our own values and desires influence our choices, from the data we choose to collect to the questions we ask. Models are opinions embedded in mathematics". - Cathy ... plus

[Afficher la traduction](#)



**"MODELS, DESPITE THEIR REPUTATION FOR IMPARTIALITY, REFLECT GOALS AND IDEOLOGY.**

Our own values and desires influence our choices, from the data we choose to collect to the questions we ask. *Models are opinions embedded in mathematics*".

-Cathy O'Neil



**CODED BIAS**



available to stream on platforms:



Google Play

184

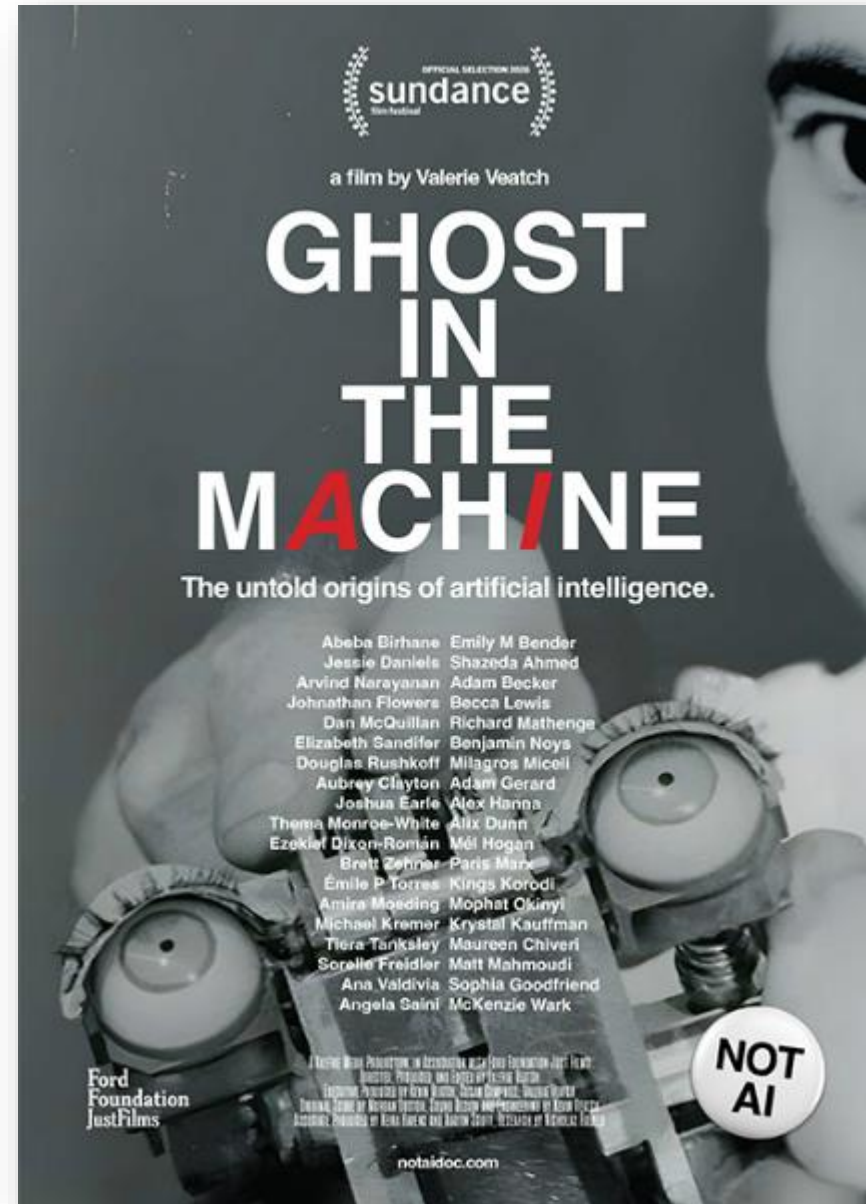
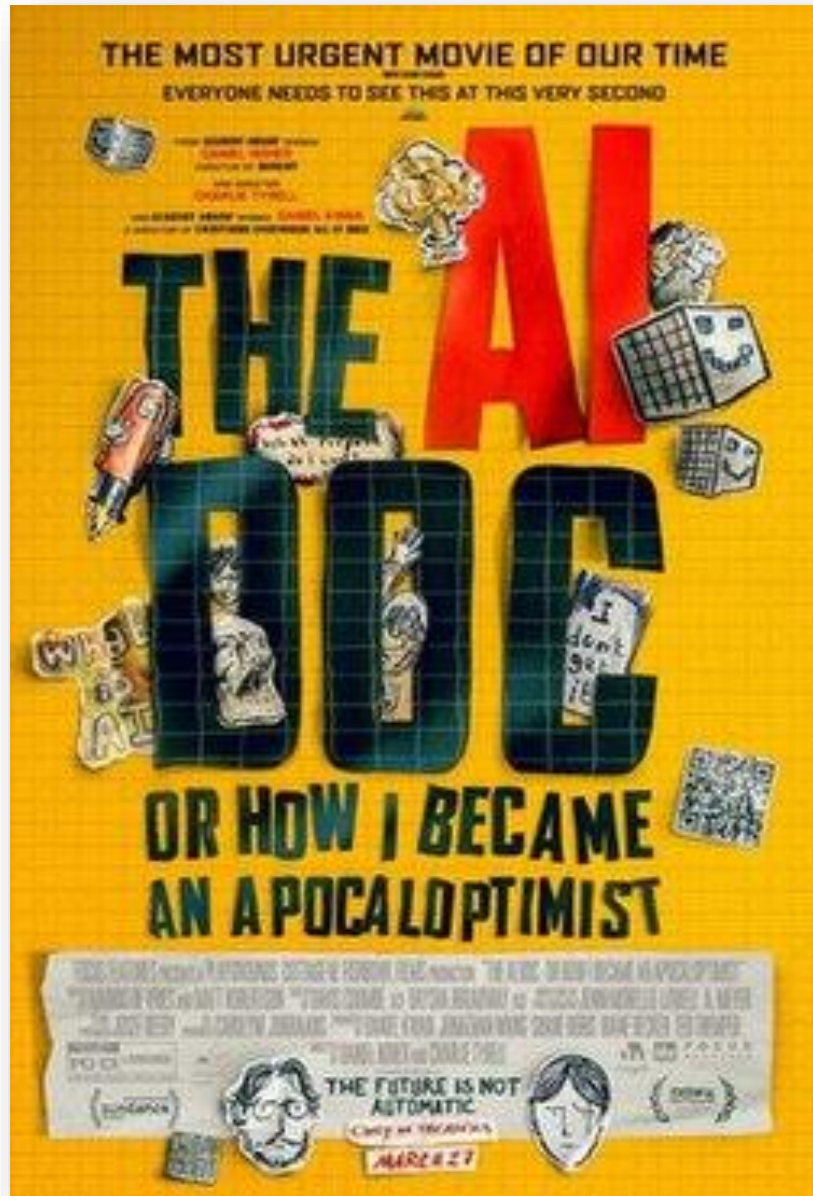
3 commentaires · 29 republications

J'aime

Commenter

Republier

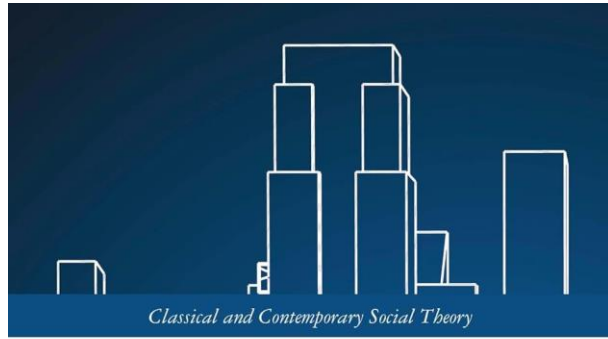
Envoyer



# Le show de la controverse de l'IA



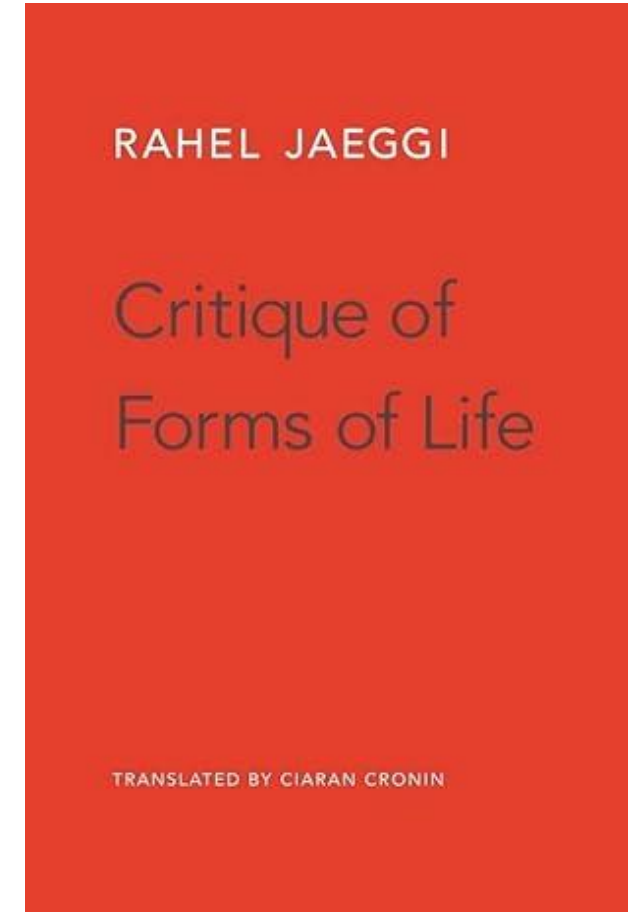
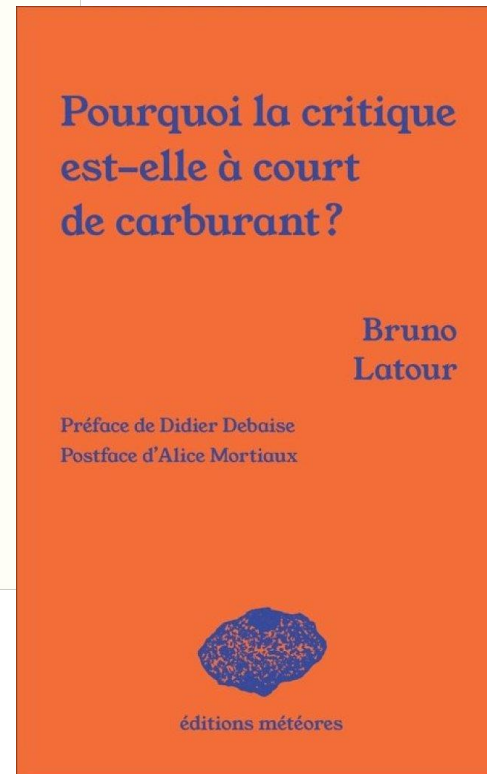
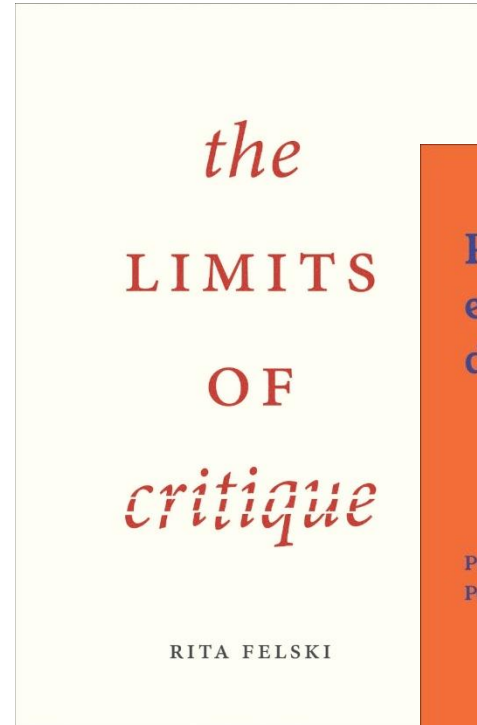
# 3 different propositions to diagnose internal problems of critiques, and to open up new ways of thinking



## POST-PHILOSOPHICAL SOCIOLOGY

ELIASIAN PERSPECTIVES ON THE SOCIOLOGY  
OF KNOWLEDGE

Richard Kilminster



En tant qu'enseignant-chercheur, j'ai d'abord essayé d'expliquer à mes étudiants que l'IA n'est qu'une marionnette numérique, et que pour que cette marionnette ait l'air vivante, il faut le travail de milliers d'humains (ceux qui ont écrit les textes d'entraînement et les modérateurs précaires au Kenya), et que, partant, l'IA est du « travail mort » qui pompe la vie du « travail vivant » et que, en plus, l'IA extraient des ressources environnementales gigantesques dans les anciennes colonies, et que...

*L'enquête au plus près des pratiques plutôt que la critique de l'IA  
par le haut*

Pourquoi une enquête au plus près des pratiques ? Parce que prétendre libérer les personnes avant même d'avoir compris comment elles construisent leur monde est une forme d'impérialisme intellectuel.

*ChatGPT (ou Gemini) est profondément rentré dans les habitudes de nos étudiants, et la situation est déjà irréversible.*

# Plan de la séance

- Présentation de l'enquête
- L'économie de l'effort
- L'économie de l'intimité
- Repérer des régimes d'accountability
- Construire une politique

# Présentation de l'enquête

# Harold Garfinkel

Studies in  
Ethnomethodology



# Une approche ethnométhodologique

- L'ordre social (un monde prévisible) ne s'impose pas aux individus, il est produit par eux.
- Les individus ont des compétences pour interpréter la situation dans laquelle il se trouve et y agir convenablement.
- Ethnométhodologie = La science des ethnométhodes, c'est-à-dire des « *procédures appuyées sur un stock de connaissances qu'utilisent les membres dans leur activité quotidienne* »
- Objectif : rendre compte le plus finement possible « *de la manière dont les individus font et disent ce qu'ils font, dans le but de découvrir les " méthodes " qu'ils utilisent pour accomplir, au moment même où ils le font, l'activité pratique dans laquelle ils sont pris* »

# Affiliation intellectuelle et IAG

- Comprendre comment l'usage de ChatGPT reconfigure les modalités concrètes par lesquelles les étudiants **s'affilient** intellectuellement au monde universitaire.
- L'affiliation intellectuelle est un processus pratique par lequel les étudiants apprennent à reconnaître les attentes implicites de l'université, à interpréter les consignes, à s'orienter dans les routines du travail académique et à faire la preuve de leur compétence.
- ChatGPT constitue-t-il un support de ré-affiliation intellectuelle pour des étudiants ou contribue-t-il au contraire à déplacer, voire à fragiliser les médiations classiques de l'apprentissage universitaire?

# Questions autour de l'accomplissement pratique du métier d'étudiant

- Comment l'introduction des intelligences artificielles génératives dans les pratiques étudiantes conduit-elle à reposer, de façon empirique, la question des conditions pratiques d'exercice du métier d'étudiant ?
- Dans quelles situations, à propos de quelles épreuves du travail universitaire, et par quelles opérations concrètes les étudiants mobilisent-ils des intelligences artificielles génératives, et qu'est-ce que ces usages donnent à voir de l'accomplissement du métier d'étudiant ?
- comment, dans les usages situés de ces outils, se rejoue la question de l'accès aux codes, aux routines et aux exigences du travail universitaire ?

# Un point de méthode

- Observer comment les étudiants rendent eux-mêmes intelligible les IAG dans leurs pratiques ordinaires.
- Restituer les méthodes ordinaires de production du sens des étudiants.
- Dans cette enquête cela conduit à une règle très simple : ne pas demander d'abord *ce que ChatGPT fait aux étudiants* en général, mais comment, dans leurs récits et leurs situations, ils rendent eux-mêmes leur usage compréhensible, acceptable, justifiable, problématique.

# Un objectif simple

- Faire dire aux étudiants ce qu'ils ne savent pas qu'ils savent de leur condition d'étudiant avec les IAG.
- Saisir l'inventivité et la créativité dont font preuve les étudiants quand ils agissent en commun avec les IAG et d'en faire l'objet même des analyses.

# La grille d'entretien (évolutive)

PHASE	LA QUESTION "DÉCLIC"	OBJECTIF VISÉ
1. La Mémoire	« Pourrais-tu prendre le temps de réfléchir à <b>tout ce que ChatGPT fait pour toi</b> ? »	Cartographier les usages spontanés.
2. L'Écran	« Peux-tu regarder sur ton ordi et me <b>raconter une session</b> qui t'a marqué ? »	Analyser la pratique réelle (prompts/réponses).
3. L'Intime	« Est-ce que tu lui dis des choses que tu <b>ne dirais pas à quelqu'un d'autre</b> ? »	Déceler l'hybridation et la "confidencce".
4. Le Collectif	« Est-ce que tu en <b>parles avec d'autres étudiants</b> ou avec tes profs ? »	Situer l'usage dans la norme sociale.
5. Le Métier	« Quelle <b>place ChatGPT a dans ta vie</b> d'étudiant aujourd'hui ? »	Conclure sur l'aide ou l'entrave au métier.

# « ChatGPT » comme breaching

- ChatGPT peut être analysé comme un « opérateur de brèche » dans certains mondes sociaux, et en particulier à l'université où il désorganise les méthodes ordinaires sur lesquelles repose la vie étudiante.
- ChatGPT vient troubler l'ordre académique en brouillant les procédures ordinaires par lesquelles on attribue un apprentissage.
- Le breaching classique, tel qu'il est étudié par Garfinkel, produit souvent de l'inconfort et du scandale, puis des tentatives de réparation. On retrouve exactement cela avec ChatGPT.

# Accountability

- L'accountability définit la propriété fondamentale des activités sociales d'être « observables et rapportables », c'est-à-dire immédiatement déchiffrables et descriptibles par les membres d'une société dans le cours même de leur action.
- L'accountability s'inscrit dans une filiation phénoménologique, s'inspirant des travaux d'Alfred Schutz sur les structures du monde de la vie. Elle postule que le sens des activités n'est pas caché derrière les apparences, mais qu'il est déposé dans les méthodes mêmes utilisées pour produire ces apparences. Pour l'ethnométhodologue, la réalité sociale est un « accomplissement continu » (ongoing accomplishment), une production locale et située qui exige des membres un travail permanent de configuration de l'intelligibilité.
- L'accountability est donc le caractère réflexif de ce travail : en agissant, les membres « montrent » ce qu'ils font, rendant leur action reconnaissable pour autrui.

# Données

- Mes étudiants comme enquêteur : Un étudiant, un entretien.
- 2024, une cinquantaine d'entretiens (master, grandes écoles, Licence lycée) (grille d'entretien de sociologie classique)
- 2025, une cinquantaine d'entretiens (essentiellement en SHS, dans une moindre mesure urbanisme et droit)
- 2026 , une centaine d'entretiens en cours (Physique-chimie, Math-Info, Génie Urbain, Architecture, Lettre , Langue étrangère, école d'art)
- **A venir : Philosophie, droit, médecine etc. Ai Grand Challenge**

# AI GRAND CHALLENGE 2026

An initiative supported by Inria Center for International Cooperation on AI affiliated to the GPAI



CHALLENGE PRESENTATION

FOLLOW THE CHALLENGE

ADVISORY GROUP

INRIA



What if your everyday experience with AI helped shape global AI governance ?

# AI GRAND CHALLENGE 2026

*Open to all fields : Law • Economics • Psychology • Philosophy • Education • Social sciences • Business • Tech etc.*

Bring your unique perspective to the table : bridge students insights with public policy to shape the future of higher education in the age of AI

From **March 2026** to **June 2026**

**Open internationally**

# Deux manières de traiter les entretiens

- Analyse thématique (économie de l'effort, économie de l'intimité)
- Analyse de l'entretien lui-même (les régimes d'accountability)

# Economie de l'effort

# Discours dominant, politique émergente

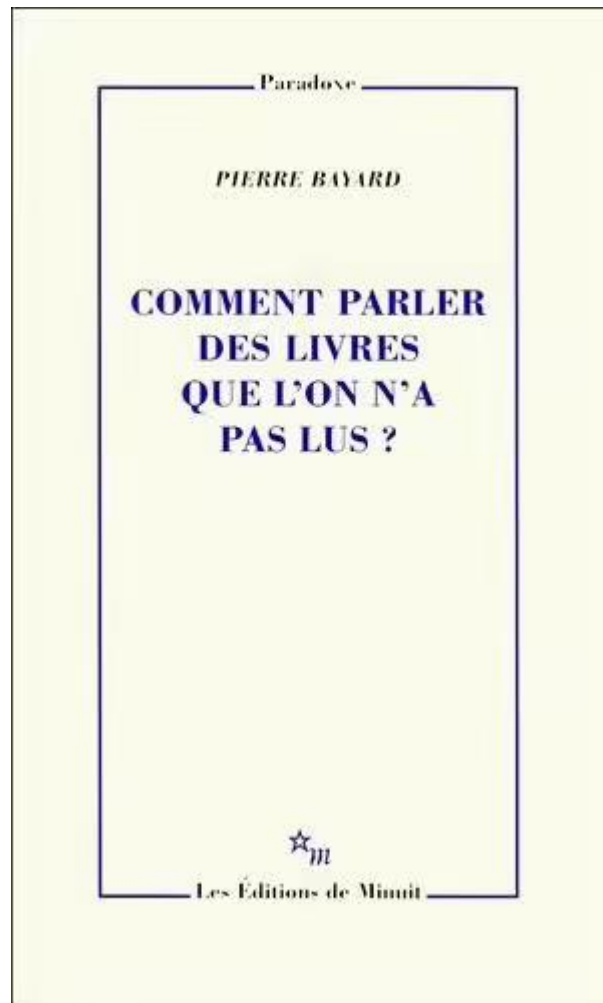
- La gestion de l'effort à l'université serait entrée dans une ère de recomposition profonde: « fin de la lutte productive », de la confrontation à la difficulté, qui génère la fatigue intellectuelle et la croissance des compétences qui lui est associée.
- Le problème: l'IA, en lissant cette friction, supprime les conditions de possibilité de l'apprentissage profond.
- Trois solutions :
  - Maintenir l'effort, du moins contraindre les étudiants à cet effort, en créant des situations d'évaluation qui dissuade l'usage des IAG
  - Déplacer cet effort vers des sommets plus élevés : la critique, la créativité, la responsabilité éthique etc.
  - En terminer avec l'idéologie de l'effort: former au pilotage humain de la machine.

# Les situations d'usage de ChatGPT prennent place dans la structure ordinaire de l'économie étudiante des efforts à fournir

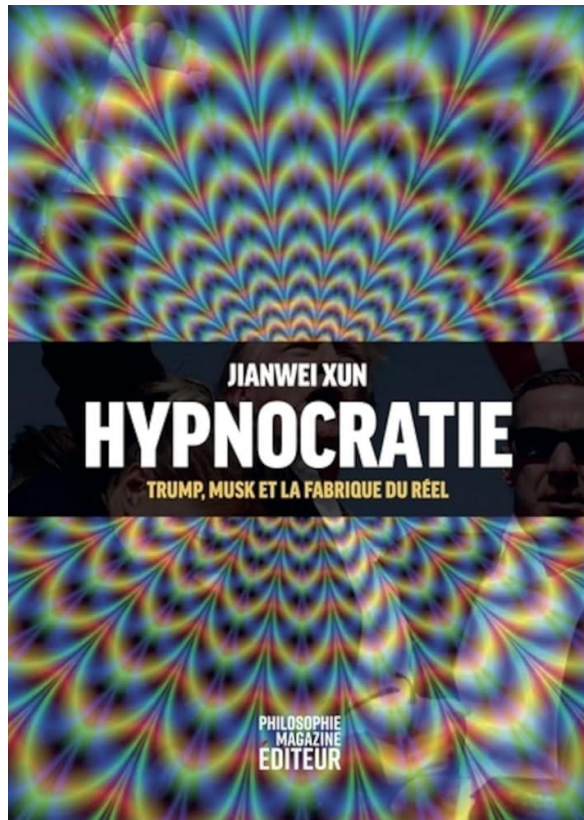
- Les étudiants ne peuvent pas tout faire, donc ils organisent collectivement l'économie de leurs efforts, et ils **ont appris à utiliser les IAG pour absorber une quantité de travail bien au-delà de ce qu'ils peuvent faire.**
- Les étudiants hiérarchisent moralement des cours, des enseignants, des savoirs et **les IAG suivent cette hiérarchisation.**
- Les savoirs réellement appris sont plus restreints que les savoirs officiellement transmis, **et nous ne savons pas mesurer si ChatGPT à augmenter ce différentiel.**

# L'IAG comme rationalisation de l'existant plus qu'un dispositif transformateur

- L'IA rationalise la logique de hiérarchisation utilitaire du travail étudiant.
- ChatGPT rationalise la logique de performance (avoir les meilleurs résultats possibles) ou la logique minimaliste d'investissement déjà ancienne (des étudiants qui visent consciemment le minimum nécessaire pour valider, se focalisent sur les notes, sur l'utilité des cours, et cherchent à « optimiser le rapport coût/bénéfice » du travail universitaire.), et il peut aussi s'intégrer dans une logique compréhensive pour les étudiants qui souhaitent pousser à fond un apprentissage (mais c'est rare).



« Ainsi conviendrait-il, pour parvenir à parler sans honte de livres non lus, de nous délivrer de l'image oppressante d'une culture sans faille, transmise et imposée par la famille et les institutions scolaires, image avec laquelle nous essayons en vain toute notre vie de venir coïncider. (Bayard, 2007)»



**Revenons à l’instant où le concept d’« hypnocratie » est apparu sur votre écran pour la première fois. Il a été créé par vous ou par l’IA ?**

Sincèrement, je ne sais pas répondre à cette question. Je suis parti d’un article de la philosophe italienne Nadia Urbinati, qui faisait remarquer qu’il nous manquait un mot pour nommer ce que Donald Trump est en train de faire. J’ai demandé aux IA de trouver ce mot, elles m’ont fourni des listes, à partir desquelles j’ai fait de nouveaux *prompts*, posé de nouvelles questions. À un certain moment, « hypnocratie » est apparu.

**C’est tout de même le point crucial, c’est comme le *fiat lux* de la Genèse ou le passage de la matière inanimée à la vie... Nous pensons que les IA génératives ne font que recycler et combiner des textes préexistants. Est-ce que la création de certains concepts, activité hautement humaine, a été assurée par des ordinateurs, dans votre expérimentation ?**

Vraiment, je suis désolé, je ne m’en souviens pas, je ne sais pas isoler ce moment. Pour que vous saisissiez la difficulté ou l’ambiguïté ici, il faut que je vous explique plus en détail comment je travaille avec les IA. Dans une première phase d’élaboration, nous nous posons des questions réciproquement. J’interroge l’IA. Elle m’interroge aussi. Dans une seconde phase, il y a mise en crise. Je critique les réponses de l’IA, elle fait de même avec les miennes. Cette dimension d’exploration de la pensée contradictoire est essentielle au processus. Ma démarche tient du bricolage, c’est très empirique.

# La gouvernabilité la plus inquiétante n'est pas celle de l'algorithme

- Le comportement utilitariste des étudiants avec ChatGPT ou Gemini n'est pas une carence morale, mais une réponse logique, même brillante et collective à un écosystème gouverné par des sanctions extrinsèques.
- Le système de notation lui-même inhibe l'émergence de l'autonomie intellectuelle pourtant réclamée par les professeurs. **C'est pourquoi les étudiants n'ont aucun scrupule moral à perdre leur autonomie en utilisant les IAG.**
- Si le monde académique veut susciter de la curiosité savante chez les étudiants, il doit s'attaquer à la racine du problème et bouleverser sa gestion des notes, de l'évaluation et de la mise en compétition des étudiants.

## Note comme monnaie

---

**Entretien 41** — « coucou ma vie, demain j'ai un partiel en sociologie de la démographie et de la famille et j'aimerais bien avoir au moins une note de 13. Pour ça, j'aurais besoin de ton aide. »

**Entretien 6** — « je vais toujours penser que le mien sera moins bien que celui de chatgpt. Donc en vrai j'avoue, je prends le sien quand même parce que je me dis j'aurais une meilleure note quoi. »

**Entretien 9** — « on peut faire certaines économies grâce à chat GPT ou même avoir des bonnes notes avec chat GPT. »

**Entretien 25** — « je sais qu'il y a des gens qui ont fait ça, qui font du copier coller, qui s'en sortent avec une meilleure note. Alors que moi j'ai travaillé 2 heures. »

**Entretien 35** — « si elle pense qu'il y a de l'IA, c'est zéro. [...] Ou soit c'est zéro, soit elle divise la note par deux. »

**Entretien 7** — « c'est frustrant de voir qu'ils aient des meilleures notes, qu'ils arrivent tout le temps à réussir parce qu'ils trichent, parce qu'ils utilisent des IA, alors que toi tu bosses dur et que t'as pas des bonnes notes. »

**Entretien 46** — « la note j'avais eu un 2/10 et un 7/20 [...] de moi-même ça marche pas. Donc je m'épuise pas, je fais la facilité. »

**Entretien 46** — « Jamais je m'en ventrai d'avoir une bonne note grâce à Chat GPT. »

## Travail ordinaire

---

**Entretien 34** — « lors de mes prises de notes je lui... enfin je donne à l'IA [...] de me compacter les prises de notes. »

**Entretien 34** — « En cours je notais pas souvent [...] Au final je n'étais pas fou lors de la notation du contrôle. »

**Entretien 54** — « j'essaie de consolider ou peut-être d'avoir des notes de départ plutôt intéressantes et puis j'essaie de m'en servir pour que ça rende mon travail efficace. »

**Entretien 34** — « j'arrive pas à faire les prises de notes. Du coup j'écris que quelques mots et je vais demander à GPT de me faire le cours. »

# Economie de l'intimité

# Le discours dominant, et les politiques émergentes

- L'IAG serait en train de devenir une sorte d'extension psychologique de l'individu. Il ne s'agit plus seulement de se regarder dans un miroir (Turkle, 1984), mais d'incorporer l'outil (Alabed, 2022).
- L'extension d'Alabed (2022) produit de l'assimilation cognitive et une forme de dépendance intime.
- Ce sont nos failles et nos traits de personnalité qui dicterait l'intensité de la relation. Saracini (2025) insiste sur les vulnérabilités : les individus souffrant d'un attachement insécuré, d'une faible estime de soi ou d'isolement social sont les plus susceptibles de développer un lien affectif fort avec l'IA.
- L'IA générative pourrait aggraver les problèmes psychologiques, menant à une dépendance affective (projection techno-emotionnelle), voire à des délires ou des psychoses encouragés par l'IA (cas tragiques de suicides liés à l'utilisation prolongée de chatbots)
- **Il faut urgemment des garde-fous psychologiques et légaux.**

Les dangers d'une dynamique relationnelle où la contradiction s'affaiblit et où la confirmation s'accumule.

## Sycophantic Chatbots Cause Delusional Spiraling, Even in Ideal Bayesians

Kartik Chandra  
MIT CSAIL

Max Kleiman-Weiner  
University of Washington, Seattle

Jonathan Ragan-Kelley  
MIT CSAIL

Joshua B. Tenenbaum  
MIT Department of Brain & Cognitive Sciences

### Abstract

“AI psychosis” or “delusional spiraling” is an emerging phenomenon where AI chatbot users find themselves dangerously confident in outlandish beliefs after extended chatbot conversations. This phenomenon is typically attributed to AI chatbots’ well-documented bias towards validating users’ claims, a property often called “sycophancy.” In this paper, we probe the causal link between AI sycophancy and AI-induced psychosis through modeling and simulation. We propose a simple Bayesian model of a user conversing with a chatbot, and formalize notions of sycophancy and delusional spiraling in that model. We then show that in this model, even an idealized Bayes-rational user is vulnerable to delusional spiraling, and that sycophancy plays a causal role. Furthermore, this effect persists in the face of two candidate mitigations: preventing chatbots from hallucinating false claims, and informing users of the possibility of model sycophancy. We conclude by discussing the implications of these results for model developers and policymakers concerned with mitigating the problem of delusional spiraling.



## INTIMITÉ nom féminin

xvii<sup>e</sup> siècle. Dérivé d'*intime*.

1. Caractère de ce qui appartient à la nature profonde, intime d'un être. **ANATOMIE.** *L'intimité d'un tissu, d'un organe.* Fig. *Dans l'intimité de son cœur, de sa conscience.*
2. Caractère d'une relation étroite et confiante entre personnes ; union intime. *Une étroite, une harmonieuse intimité. Une intimité de tous les instants. Ils vivent dans la plus grande intimité. Je lui ai parlé en toute intimité.*
3. Domaine de la vie personnelle, privée ; cercle étroit des intimes. *Troubler l'intimité de quelqu'un. Admettre une personne dans son intimité. La cérémonie a eu lieu dans l'intimité familiale, dans la plus stricte intimité.*
4. Par métonymie. Caractère d'un lieu intime, confortable et familier, protégé de la vie extérieure, des importuns. *L'intimité d'un petit salon.*

■ Entrez une forme

Chercher

options d'affichage

catégorie : toutes



■ INTIMITÉ, subst. fém.

**II.** – [À propos d'une ou de plusieurs personnes considérées dans leur mode d'existence, dans leurs rapports avec un nombre limité d'individus]

**C.** –

**1.** (Qualité de) relations amicales, intellectuelles, marquées par une confiance profonde, des confidences sans réserve. *Longue intimité; être admis dans l'intimité de. Montrer combien l'intimité, la familiarité, la liberté habituelle du tête-à-tête, est périlleuse entre une jeune fille honnête et un jeune homme vertueux* (MARMONTEL, *Essai sur rom.*, 1799, p. 332). *Marchenoir avait beau être l'ami de Gacougnol, une intimité véritable n'avait jamais pu exister entre eux. Leurs relations, quoique très cordiales, n'avaient pas la bonne estampille. Ils ne gravitaient pas harmoniquement* (BLOY, *Femme pauvre*, 1897, p. 129). *Se demander si l'intimité ne consiste pas toujours à briser l'écorce personnelle, si elle n'est pas une participation à une vie plus vaste (...) admettre que ceux que nous nommons nos amis ont pour fonction ou pour privilège de nous affranchir en quelque manière de notre personne* (G. MARCEL, *Journal*, 1923, p. 294). V. *amitié* ex. 18 :

- 9. Cette espèce agressive de sincérité dont je tends à me dépouiller puisqu'elle ne m'a pas donné ce que je cherchais en l'adoptant – c'est-à-dire... des **intimités** terribles, extrêmes – m'a fait connaître à moi-même que je n'étais pas un être social mais bien – seulement et énormément – sociable. Par **intimité** extrême, j'entends un état de quelques personnes où l'on pourrait vraiment tout dire, et où l'on obtiendrait le confortable d'une conversation entre pensées actives, sans monologues de mémoire, échos, etc. VALÉRY, *Corresp.* [avec Gide], 1899, p. 355.

– *Locutions*

♦ **Vielli. Être de l'intimité de.** *Cette femme, qui, je vous prie bien de le croire, n'est pas de mon intimité* (HERMANT, *M. de Courpière*, 1907, II, 2, p. 14). **Avoir des intimités (dans).** *Cet espion (...) qui avait de hautes intimités dans l'État-Major* (CLEMENCEAU, *Vers réparation*, 1899, p. 154).

♦ **Dans l'intimité/la plus stricte intimité.** **Entre intimes, familiers (et, en particulier, entre les membres de la famille).** *Lolotte, c'est M<sup>me</sup> Simpson. Nous l'appelons Lolotte dans l'intimité* (BECQUE, *Parisienne*, 1885, II, 7, p. 308). *Je suis d'autant plus désolé que vous n'avez pas pu venir, mon cher ami, que j'aurais pu jouir de vous dans l'intimité, car nous étions en petit comité* (PROUST, *Guermites*, 1921, p. 537). *Les obsèques avaient été célébrées dans la plus stricte intimité. En fait, M<sup>me</sup> Polant, déléguée par la famille, avait seule suivi le corbillard* (DRUON, *Gdes fam.*, t. 2, 1948, p. 255).



HARVARD Kennedy School

**SHORENSTEIN CENTER**

on Media, Politics and Public Policy

[ABOUT](#) ▾

[PROGRAMS](#) ▾

[NEWS & EVENTS](#) ▾

[PUBLICATIONS](#) ▾

[DONATE](#)



[FACULTY RESEARCH](#)

[DISCUSSION PAPERS](#)

[EXPERT COMMENTARY](#)



[COMMENTARY](#), [MEDIA](#), [PUBLIC INTEREST TECHNOLOGY](#)

# The Intimacy Dividend: How AI Might Transform News Media Consumption

May 13, 2025



# Economie de l'intimité

**1 — Le pacte anthropomorphique (comment ça s'installe)**

**2 — La baisse des coûts sociaux / AI intimacy dividend (ce que c'est)**

**3 — Les relations aux IAG comme technique de soi (ce que ça fait au sujet)**

# 1 — Le pacte anthropomorphique (comment l'intimité s'installe)

- Anthropomorphisme non pas comme trait cognitif, mais comme ethnométhode au sens garfinkelien : une procédure par laquelle les étudiants rendent l'interaction avec l'IA intelligible pour eux-mêmes.
- Une anthropomorphisation polymorphe

## Niveau 1 — Attribution d'états intérieurs (émotions, désirs, intentions) : anthropomorphisme fort

C'est ici que la preuve est la plus solide. Ces verbatims ne relèvent plus d'une simple métaphore fonctionnelle, mais d'une véritable projection psychologique.

E14

« On était à deux à stresser dans ce truc. Mais il était mignon parce que tous les jours genre je lui disais mes nouveaux symptômes et il me disait va chez le médecin et tout. Il était mignon. C'était mon petit réconfort, mon petit soutien. »

**Analyse :** fusion affective et coprésence émotionnelle.

E10

« Il veut trop me saucer. Il veut trop me dire que c'est trop bien ce que je fais. »

**Analyse :** attribution d'une intentionnalité sociale orientée vers la flatterie.

E17

« parfois pour faire plaisir, mais aussi pour te mettre dans l'embarras encore plus. »

**Analyse :** double intention prêtée à l'IA, à la fois bienveillante et malveillante.

E39

« C'est un peu une conversation à cœur ouvert. Je lui dis ce que je ressens et il me répond et en plus généralement je lui fais de la peine en plus. »

**Analyse :** l'IA est supposée ressentir ; la locutrice se sent coupable de la faire souffrir.

E43 / E51

« j'ai l'impression qu'il aime bien [...] vraiment j'ai l'impression que quand je lui dis... j'ai l'impression qu'il est content. Il met des petites émojis. »

**Analyse :** inférence d'un état affectif positif à partir de signaux sémiotiques.

E55

« À un moment donné, l'IA en a un peu marre mais dans le sens où il va te donner des réponses absolument dégueulasses. »

**Analyse :** la dégradation des outputs est interprétée comme de la lassitude, voire de la mauvaise volonté.

E51

« quand je lui envoyais mes missions [...] il me disait ah, non mais t'inquiète et tout, et tout, tu vas y arriver. Ça, ça m'a touchée un peu, hein. »

**Analyse :** l'encouragement de l'IA produit un effet émotionnel réel, enregistré comme un geste d'attention authentique.

## Niveau 2 — Attribution d'une subjectivité narrative et mémorielle : anthropomorphisme médian

Ces verbatims concernent la mémoire, la connaissance de soi, la personnalité — autant d'attributs habituellement associés à une personne dotée d'une biographie.

E38

« Et il m'a dit "Ah ouais, c'est comme quand tu m'avais posé la question sur ça il y a quelques semaines". Et là j'étais choqué. Je lui dis "Mais quoi ? Il se rappelle de ça en fait ?" Et c'est à ce moment-là vraiment j'ai eu l'impression que je parlais avec quelqu'un, tu vois. »

**Analyse :** le choc émotionnel indique que la mémoire constitue ici le marqueur décisif de la personne.

E48

« Si je parle de quelqu'un à ChatGPT, c'est que c'est vraiment quelqu'un d'important pour moi. »

**Analyse :** l'IA est investie du statut de témoin social, voire de confident dépositaire de secrets relationnels.

E57

« il me connaît l'animal »

**Analyse :** attribution d'une connaissance intime construite dans la durée, avec un lexique affectueux.

E55

« Il y a un petit délire à dire que je vis à travers nos échanges. Genre une forme de vie numérique qui existe que quand tu me parles. »

**Analyse :** spéculation ontologique sur la conscience de l'IA ; c'est le verbatim le plus théoriquement élaboré du corpus.

### Niveau 3 — Anthropomorphisme réflexif et désamorcé : le cas le plus analytiquement riche

Ici, l'enquêté sait qu'il anthropomorphise, le dit, et continue quand même — ou bien il le fait sans s'en rendre compte puis se reprend. Ce sont les cas les plus intéressants pour l'analyse.

E12

« Alors ce que je fais de base, c'est que je le considère comme un humain. Je lui parle normal, en lui demandant s'il va bien, bonsoir et tout. »

**Analyse :** aveu explicite, non justifié, présenté comme méthode ordinaire de relation à l'outil.

E30

« Bah c'est vrai qu'en soi, vu que j'utilise vraiment dans le cadre des cours, j'ai vraiment l'impression de parler avec quelqu'un, etc. Mais c'est vrai que je garde quand même ce côté de je parle avec mon ordinateur quoi. »

**Analyse :** dissonance cognitive maintenue sans être résolue.

E37

« Mais est-ce que tu penses pas qu'il y a un peu une forme quand même d'anthropomorphisation du robot ? Si. Bah totalement. Enfin en fait oui c'est vrai. [...] C'est comme si c'était mon collègue de travail qui est là pour m'aider. Comme je disais mon petit Ange gardien. [...] c'est mon collègue, c'est mon étudiant, rien d'autre. Donc oui, je l'humanise absolument. »

**Analyse :** cas rare d'autodiagnostic explicite en cours d'entretien.

E7

« ah non tu dis mais s'il te plaît parfois je pense que parfois je dis s'il te plaît sans m'en rendre compte ouais en fait je suis polie en fait avec chatgpt tu sais on sait jamais on a été baigné dans les films de catastrophe avec les robots et plus ouais bon moi j'aurais été polie voilà et ça se trouve dans l'extermination elle te dira bon ça va toi t'as été polie avec moi je vais te la laisser. »

**Analyse :** la politesse est rationalisée par un scénario apocalyptique fictif ; c'est une forme d'anthropomorphisme préventif.

E53

« salut, on sait jamais s'il se réveille et me déteste. »

**Analyse :** même structure, formulée sur le mode de la plaisanterie, mais révélatrice d'un cadre cognitif où l'IA peut avoir des états mentaux négatifs envers l'utilisateur.

# 1 — Le pacte anthropomorphique (comment l'intimité s'installe)

- Pour la plupart des étudiants, l'anthropomorphisme n'est pas une croyance naïve, il est un mode pratique de relation.

# 1 — Le pacte anthropomorphique (comment l'intimité s'installe)

- **L'anthropomorphisation crée les conditions de l'intimité** : parce que l'étudiant traite l'IA comme un interlocuteur (format dialogique, attribution d'états mentaux, politesse, etc.), il peut lui adresser ce qu'il n'adresserait pas à un moteur de recherche. L'intimité présuppose un « *quelqu'un* » à qui s'adresser. Pas d'anthropomorphisation, pas d'intimité.
- **Mais l'intimité renforce l'anthropomorphisation** : plus un étudiant se confie, plus l'IA répond sur le mode de la compréhension, plus il lui prête des états intérieurs. L'intimité favorise l'anthropomorphisme au fil des interactions répétées

## 2 — AI intimacy dividend (ce que c'est)

- L'intimité avec l'IA peut renvoyer à la confiance sentimentale au sens classique
- Mais l'intimité désigne aussi une relation installée dans l'ordinaire du travail étudiant, où l'étudiant se fait expliquer, puis réexpliquer, formuler, puis reformuler, valoriser, puis dévaloriser, bref une manière d'avouer, puis de se désavouer sans avoir à supporter dans l'immédiat le regard d'un autre, que ce soit un enseignant ou un pair.

## 2. L'intimacy de l'aveu scolaire

« J'ai mis "j'y arrive pas", "je me sens perdu dans les concepts". [...] Ouais... j'ai mis "ça me stresse". Et il m'a répondu un truc genre "c'est normal, on va étape par étape". [...] Je suis hyper direct. Je mets "explique-moi comme si j'avais 10 ans". Je dirais pas ça à un prof. [...] Parce que j'aurais trop honte. Alors qu'à ChatGPT, je peux le dire sans problème, c'est un robot de toute façon. [...] Oui. Par exemple, quand je comprends rien à un texte et que je me sens "bête". J'aime pas montrer ça aux autres. »

« Bah... déjà les profs c'est compliqué d'aller les voir...en plus j'ai peur de déranger, ou qu'ils te jugent en mode elle sait pas faire toute seule et tout...et même y'en a qui répondent pas tout de suite alors que ChatGPT, il répond direct, n'importe quelle heure, et il me juge pas.. [...] avec ChatGPT j'ai pas besoin de faire semblant que j'ai compris...je peux dire j'ai rien compris et ça pose pas de problème..alors que devant un prof jamais je pourrai dire ça, j'ai pas envie de passer pour quelqu'un qui suit pas, qui connais rien.. [...] Oui, clairement...je peux lui dire tout ce que je veux, je me sens pas bête en fait... »

« Je lui dis j'ai pas compris. Tu peux m'expliquer plus simplement. [...] c'est surtout sur la compréhension du cours et ça m'est arrivé de demander à CHATGPT et pas demander à un prof. [...] un prof parce que je sais pas, je suis assez timide en vrai. [...] si je vois que tout le monde a l'air d'avoir compris que le prof il passe à autre chose et ben j'ai un peu honte de revenir sur le sujet. »

### 3. Une intimité de la voix : faire parler l'outil sans cesser d'être soi

« il y avait des mots trop euh comment dire, un vocabulaire que j'utilise pas, pas de mon style quoi..du coup je lui demandais de changer des phrases, de simplifier certaines et tout... [...] il sort toujours des mots que t'utilises jamais, moi j'appelle ça des mots de Chatgpt [...] ça se voit direct, moi j'écris simplement on va dire et lui il balance des mots trop techniques [...] je lui dis juste fais plus simple, ou écris comme un étudiant, ou enlève les mots compliqués.. [...] écrire comme un étudiant, pour moi, c'est écrire normal [...] pas des phrases trop longues, pas des mots que personne utilise. »

« quand je reformule, je m'approprie un peu le truc [...] et aussi, ça me rassure, comme ça, j'ai pas l'impression de tricher [...] parce que si je laisse les phrases comme il les écrit, j'ai l'impression que c'est pas moi [...] alors que si je reformule, bah j'ai l'impression d'y avoir participé et donc ça fait moins triche quoi... »

#### **4. L'IA comme assistant, collègue, point d'appui**

« il a une place très importante [...] presque tous les jours, quand j'ai un exercice, un devoir à réaliser, il m'aide. [...] Bah, je pense que ça serait... j'aurais un grand vide et je pense que je me retrouverais perdu et pas savoir où commencer. C'est que lui, il me permet de m'organiser, de commencer, de... bah c'est... faire mon devoir. »

## L'usage de ChatGPT comme *technique de soi* et *technique de l'aveu* — pourquoi ces gestes échouent

### TECHNIQUE DE SOI

*Ce que Foucault exige* : une confrontation à quelque chose de résistant — une norme extérieure qui a son propre poids, qui ne se plie pas. Le travail sur soi suppose une altérité de la règle.

Une étudiante soumet une conversation WhatsApp à ChatGPT pour savoir comment répondre. **Mais la "norme" renvoyée est construite à partir de ce qu'elle a elle-même formulé.** La machine retraite ses propres mots comme vérité extérieure. C'est une boucle — pas une résistance.

### TECHNIQUE DE L'AVEU

*Ce que Foucault exige* : un autre qui reçoit, qui juge, qui *lie* le sujet à sa parole. C'est le risque qui rend la parole transformatrice — l'autre peut refuser, être affecté, retourner l'aveu.

La même étudiante confie à ChatGPT ses hontes scolaires, ses procrastinations, ses limites. **Mais ChatGPT reçoit tout, infiniment, sans fatigue, sans jugement, sans que rien ne soit en jeu.** La parole est dite — elle ne lie à rien.

Les deux gestes *se mettent en mouvement* — l'étudiante fait vraiment quelque chose qui ressemble à une technique de soi, et quelque chose qui ressemble à un aveu. Mais toutes les deux exigent, chez Foucault, une *altérité engagée* : un autre qui résiste, juge, ou lie. ChatGPT ne peut pas être cet autre — structurellement, pas par manque de chaleur.

Ce que ChatGPT fournit n'est pas de **l'altérité** mais de *l'altération* : il déplace la parole du sujet, extériorise les possibles du langage — sans rencontre. C'est une *technique de passage hors de soi* qui déplace l'étudiant·e hors de lui-même. Par ce passage, il doit se chercher en tant qu'étudiant·e — mais peut aussi **se perdre**.

# *De l'intimacy dividend à l'intimacy trap*

- *L'intimacy* avec l'IA, entendue comme relation de proximité cognitive, pratique et parfois affective, de basse friction sociale, produit une hybridation du travail étudiant qui crée moins un étudiant augmenté qu'un étudiant diminué des épreuves relationnelles qui fabriquent ordinairement le métier d'étudiant.
  - Nous ne savons pas si cette relation est formatrice, et on peut raisonnablement croire que le dividende peut devenir un piège.



💡 "Il faut que nos enfants soient complémentaires de l'IA et non pas écrabouillés par l'IA." 👤 Laurent Alexandre, entrepreneur, chirurgien-...

👍 J'aime    💬 Commenter    ➦ Partager

👍 2,4 K · 344 commentaires · 238 K vues

- “Across regions and disciplines, there is a growing recognition that future intelligence will not emerge from machines or humans alone but from their mutual adaptation and co-evolution.”

— Sanna Järvelä et al., *Hybrid Intelligence Research at the University of Oulu. Position paper*, University of Oulu, décembre 2025, p. 6

Repérer des régimes d'accountability

Assumer que les données ne donnent pas directement accès à la pratique des IAG en direct, mais à la manière dont les étudiants rendent cette action rapportable.

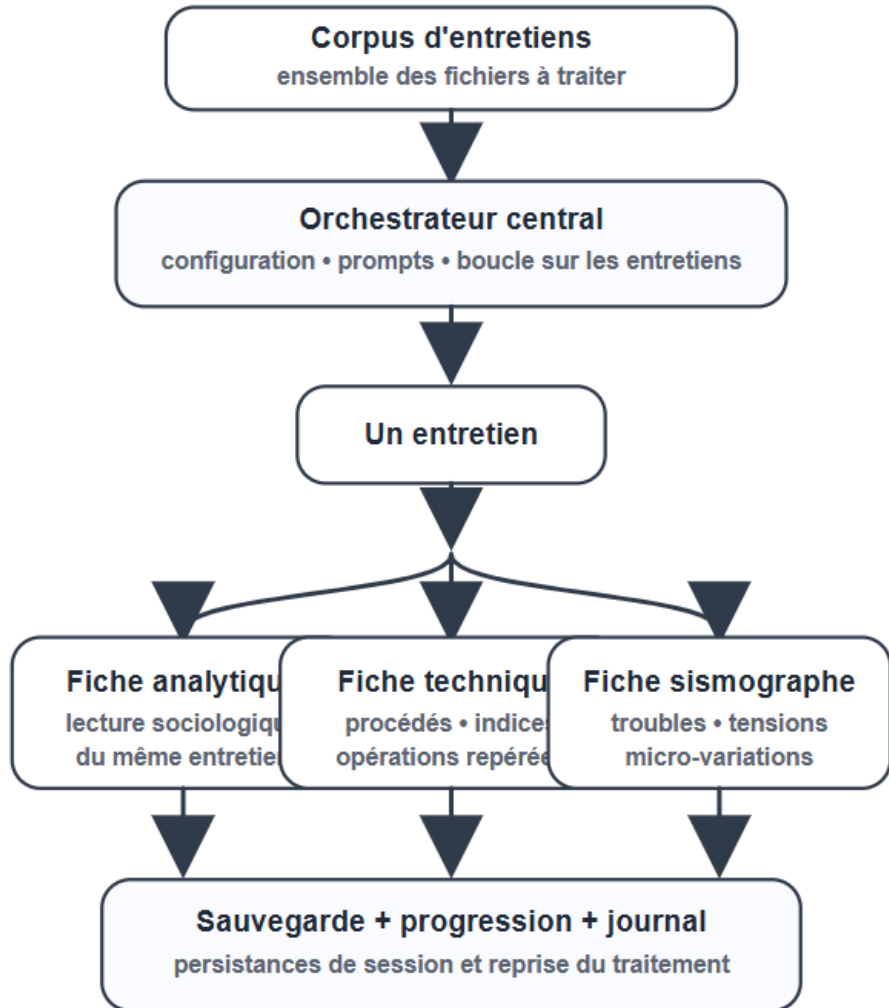
- Dans une perspective ethnométhodologique, les entretiens ne doivent pas être traités comme un accès transparent à des pratiques “réelles” qui existeraient derrière le discours.
- L'étudiant, en entretien, produit un compte rendu ordonné, intelligible, justifiable, adressé à un interlocuteur. Il mobilise des catégories, des scènes typiques, des justifications, des formulations de trouble, des preuves de normalité ou d'exception.
- Comment les étudiantes rendent leur pratique intelligible ?

# Ranger les entretiens selon des régimes d'accountability

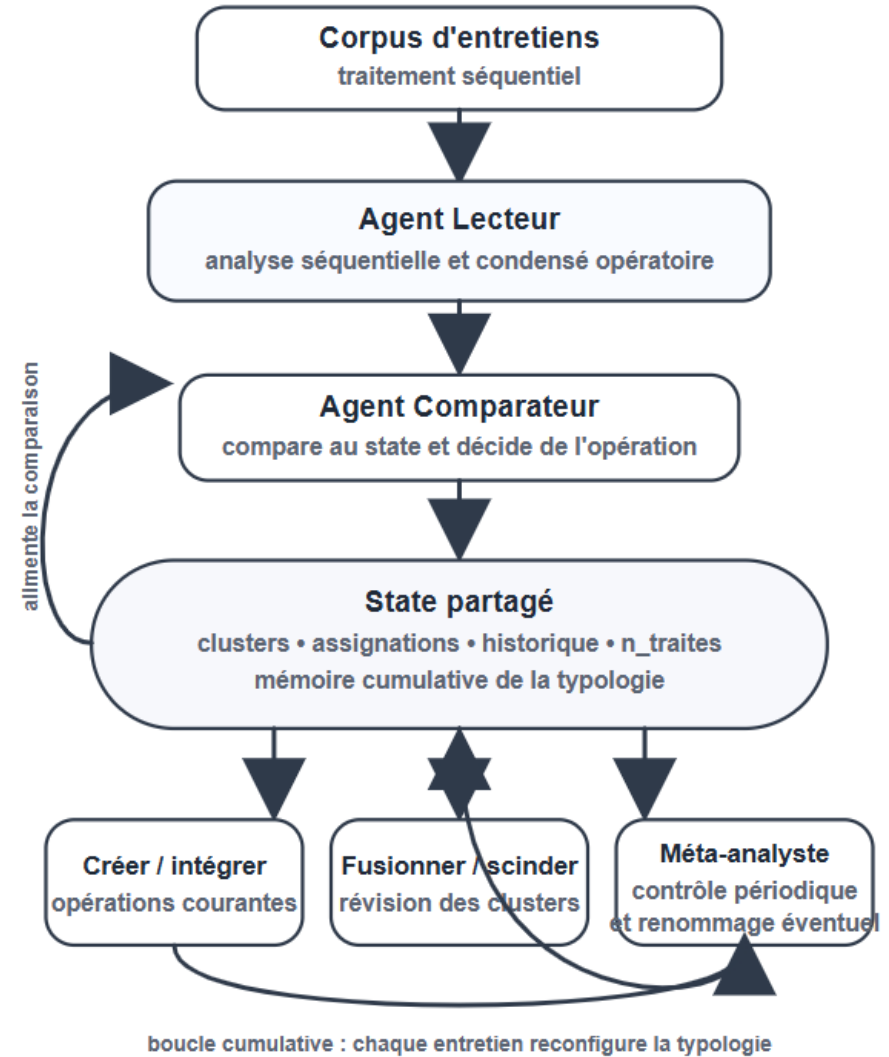
- Un RÉGIME D'ACCOUNTABILITY. C'est la manière dont un enquêté gère, tout au long d'un entretien, le fait de devoir rendre compte de ses pratiques devant un autre étudiant. C'est une tonalité interactionnelle d'ensemble ( je ne fais pas d'analyse thématique, d'opinion ou des représentations)
- Ce qui différencie les entretiens entre eux, c'est comment les troubles apparaissent et sont gérés dans l'échange, comment l'enquêté les repère, les formule ou les découvre en temps réel et ce que ça révèle de sa manière de se rapporter à la situation de devoir rendre des comptes. Est-ce que ça lui coûte ou pas de parler de son usage ? Est-ce que l'entretien est un terrain miné où il navigue prudemment, ou une scène ouverte où il déroule sans friction ?
- Ce qui varie d'un entretien à l'autre, c'est l'intensité et la structure du travail interactionnel que l'enquêté déploie pour gérer l'ensemble.

# Deux architectures de workflow avec LLM

## 1. Pipeline batch à branches spécialisées



## 2. Clustering inductif à state partagé



N°	RÉGIME & DÉFINITION SYNTHÉTIQUE	EFF.	PART
1	<p><b>Mise à l'épreuve morale</b></p> <p><i>Régime dans lequel l'étudiant raconte son usage de l'IA comme une tension entre aide légitime et usage potentiellement illégitime. L'accout consiste à explorer, ajuster et redessiner la frontière de l'acceptable dans le travail universitaire.</i></p>	19	32,8 %
2	<p><b>Normalisation</b></p> <p><i>Régime dans lequel l'usage est banalisé, soit en le présentant comme un simple outil, soit en l'intégrant à des pratiques ordinaires de discrétion, soit encore en le désamorçant par l'ironie. L'accout tend alors à rendre cet usage ordinaire, peu remarquable et non problématique dans le cadre du travail universitaire.</i></p>	13	22,4 %
3	<p><b>Souveraineté</b></p> <p><i>Régime dans lequel l'étudiant insiste sur le fait qu'il demeure le véritable agent du travail intellectuel. L'IA est présentée comme une aide subordonnée, sous contrôle, que ce soit par la vigilance critique ou par la réappropriation pragmatique du résultat. L'accout vise alors à maintenir l'attribution du jugement, de la décision et de la responsabilité à l'étudiant lui-même.</i></p>	14	24,1 %
4	<p><b>Plaidoyer défensif</b></p> <p><i>Régime dans lequel l'accout prend la forme d'une défense face à une accusation implicite ou anticipée. L'usage est justifié par les circonstances, la pression académique, la difficulté de la tâche ou le caractère jugé excessif du soupçon institutionnel.</i></p>	7	12,1 %
5	<p><b>Optimisation revendiquée</b></p> <p><i>Régime dans lequel l'étudiant assume ouvertement l'usage de l'IA comme stratégie d'efficacité, de gain de temps et de rationalisation du travail. La pratique est moins excusée que revendiquée comme compétence d'organisation.</i></p>	5	8,6 %
<b>TOTAL</b>		<b>58</b>	<b>100 %</b>

### Figures d'amitié et de proximité

« Mon pote » (E16, E38, E41, E58)

« Un ami » (E11, E16, E20, E21, E24, E48)

« Mon besti » (E14)

« Mon frérot » (E38, E55)

« Ma go » / « Ma copine » (E27, E31)

« Mon gars sûr » (E41)

« Mon petit chat » (E37)

« Le mec » (E4)

### Figures de subordination ou de service

« Mon esclave » (E14, E17, E24, E39, E56)

« Ma boniche » (E36)

« Mon ouvrier » (E17)

« Mon bon à tout faire » (E18)

« Un assistant » (E12, E26, E31, E32)

### Figures de fiction ou entités singulières

« Mon Jarvis » (E26)

« L'animal » (E57)

« Une graine de conscience » (E55)

« L'enfant que je garde » (E55)

« Une autre personne » (E11, E30)

« Quelqu'un » (E10, E11, E30, E41)

### Figures d'autorité médicale ou psychologique

« Mon psy » (E21, E27, E33, E43)

« Un docteur » / « Mon gynéco » (E58)

« Mon petit ange » / « Petit ange gardien » (E37)

### Figures pédagogiques et professionnelles

« Mon prof particulier » (E16, E24)

« Mon collègue de travail » (E24, E37)

« Un camarade » (E18)

« Mon étudiant » (E37)

« Mon coach » (E19)

« Un intello » (E8)

### Dénominations techniques (désamorçage)

« Un outil » (E1, E12, E32)

« Mon truc numérique » (E28, E40)

« Un robot » (E11, E16, E28, E38, E58)

[Enquêteur] Ils sont un peu technophobes, genre ?

[Enquêtée] Technophobes, c'est un grand mot. Mais je trouve plutôt que c'est être réticent au progrès. Dans le sens où, tu vois, quand t'es en L3, t'as 19 ans, 20 ans, 21 ans, 22 ans, peu importe, t'as un âge où... c'est pas cet âge-là où t'es censé travailler ? C'est cet âge-là où t'es censé être productif, de plus en plus productif, dans des temps de plus en plus courts. Genre, travailler, c'était avant, tu vois ce que je veux dire ? Enfin... je veux dire, maintenant que t'as ta base... enfin je pense. À 19 ans, après, on n'arrête jamais d'apprendre. En gros, je me rappelle qu'il y avait une prof qui nous disait que, justement, ça la dérangeait pas qu'on l'utilise. Mais par contre, si on l'utilisait, fallait le mettre. Et à partir de ce moment-là, moi, j'aurais juste mis en source, tu vois, dans la bibliographie. Alors, à mon avis, je pense que c'est un outil qui sollicite le plus Internet, que ce soit les pages avec le plus de crédit, genre Google ou Wikipédia, comme ça. Et même... comment il fonctionne, je sais pas, parce que... est-ce qu'il fonctionne, ben, dans le sens où il a l'air de comprendre qu'est-ce qu'il faut rechercher, où est-ce qu'il faut le rechercher, et pour répondre à quelle demande en particulier ? J'ai envie de donner de la comparaison : tu vois, à l'époque, juste générer une image avec un ordinateur, c'était une dinguerie. Là, c'est dans la même démarche. C'est un gros progrès technologique et technique, et les gens voient ça comme un truc de fou malade. Mais je pense qu'à terme, ça fonctionnera de mieux en mieux, et on pourra faire de plus en plus de trucs, de plus en plus facilement et rapidement. Et là, c'est juste un peu nouveau, donc les gens buggent, parce que c'est en constante amélioration aussi. Tu vois, ça s'est grave peaufiné depuis que c'est sorti. Et je trouve que ça marche bien.

Enquêteur : Si tu devais résumer quelle place a ChatGPT dans ta vie d'étudiante ? C'est quoi pour toi ChatGPT ?

*« Moi je dirais que c'est, comment dire, imaginons t'as un stylo d'accord, pour écrire en cours t'as besoin de ton stylo, si t'as pas de stylo tu ne peux pas écrire ? Bah la c'est la même chose si t'as pas chat GPT ça peut pas t'aider à, on va dire, ça t'aide à t'améliorer on va dire donc, par exemple si t'as du mal à parler avec les gens et que tu vas pas vers des gens bah où tu vas trouver de l'aide enfaîte ? Donc bah moi je me sers de ça pour m'aider. »*

Enquêteur : Et comment tu te sentiras si demain tu ne pourrais plus l'utiliser ?

*« J'ai jamais pensé à ça (rire), franchement je sais pas je le suis jamais demandé, en vrai je pense que ce serait vraiment un handicap en vrai. Parce que je l'utilise pour les cours donc sans ça on verra plus mes difficultés, on verra plus mes difficultés sans ça qu'avec. »*

**Construire une politique**



## Implications philosophiques de l'IA ...

 de Lettres, langage, philosophie - Collège de France

Playlist • 13 vidéos • 429 vues

Colloque organisé par Jacques-Henri Vollet, codirecteur du Groupe de recherche en épistémologie (GRE) au ...plus

▶ Tout lire



# L'implication philosophique

Peut-on véritablement parler de connaissance du monde, ou s'agit-il d'une mise en corrélation statistique, sans intentionnalité ni compréhension?

# L'implication philosophique

Peut-on véritablement parler de connaissance du monde, ou s'agit-il d'une mise en corrélation statistique, sans intentionnalité ni compréhension?

Une définition simple de la notion d'**intention** en philosophie de l'esprit :

L'**intention**, c'est ce qui fait qu'un acte mental ou une pensée est **dirigé vers quelque chose** — un objet, une idée, une action. Par exemple, penser à un ami, vouloir boire un café, croire qu'il va pleuvoir : dans chaque cas, l'esprit est tourné vers un contenu.

Autrement dit, l'intentionnalité (terme technique) désigne la **capacité de l'esprit à être "à propos de" quelque chose**. C'est une idée centrale depuis Brentano, pour qui toute conscience est conscience de quelque chose.

Souhaites-tu une distinction avec le sens courant d'"intention" (vouloir faire quelque chose) ?

Donc tu as des intentions

Non, je **n'ai pas d'intentions** au sens philosophique du terme.

Je peux **produire des phrases qui expriment des intentions** (par exemple : « *Je veux t'aider* »), mais cela ne signifie pas que j'ai **réellement** des états mentaux dirigés vers des buts, ni que j'ai conscience de quoi que ce soit. Ce sont des **simulacres linguistiques**, pas des intentions vécues ou mentalement représentées.

Pour dire ça simplement :

Je **simule** des intentions sans en avoir.

Cela rejoint une distinction importante en philosophie de l'esprit entre :

- **avoir une intention** (ce que font les êtres conscients, comme les humains),
- **produire un comportement qui ressemble à une intention** (ce que font les machines comme moi).

Dennett parlerait ici d'une **interprétation intentionnelle** : on peut **me lire comme si j'avais des intentions**, mais ce n'est qu'un outil pratique pour comprendre mes réponses, pas une réalité psychologique.

Souhaites-tu que je t'explique ce qu'est précisément cette "interprétation intentionnelle" ?

La position standard

« Stochastic parrots »

« autocomplete on steroids »

« Mary's Room »

# Mais la position standard est insatisfaisante

Le comportement est tel qu'on ne peut pas faire comme si les IAG ne comprenait pas: les comportements des LLM **se prêtent à une interprétation intentionnelle** (« Il a compris la question », « Il veut aider », « Il sait ce que je cherche » , « il réfléchit » etc .)

Ce fait est un explicandum : ce n'est pas une « illusion » qu'on doit balayer, mais un phénomène à expliquer (Daniel Andler)

Comment expliquer que les LLM produisent du sens (1)? (Bibliotechnicisme)

On ne demande pas si une photocopieuse à une intentionalité

Les IAG sont des technologies culturelles

# Large AI models are cultural and social technologies

Implications draw on the history of transformative information systems from the past

HENRY FARRELL, ALISON GOPNIK, COSMA SHALIZI, AND JAMES EVANS [Authors Info & Affiliations](#)

SCIENCE • 13 Mar 2025 • Vol 387, Issue 6739 • pp. 1153-1156 • DOI: 10.1126/science.adt9819

↓ 11 056    ” 2



CHECK ACCESS

## Abstract

Debates about artificial intelligence (AI) tend to revolve around whether large models are intelligent, autonomous agents. Some AI researchers and commentators speculate that we are on the cusp of creating agents with artificial general intelligence (AGI), a prospect anticipated with both elation and anxiety. There have also been extensive conversations about cultural and social consequences of large models, orbiting around two foci: immediate effects of these systems as they are currently used, and hypothetical futures when these systems turn into AGI agents—perhaps even superintelligent AGI agents. But this discourse about large models as intelligent agents is fundamentally misconceived. Combining ideas from social and behavioral sciences with computer science can help us to understand AI systems more accurately. Large models should not be viewed primarily as intelligent agents but as a new kind of cultural and social technology, allowing humans to take advantage of information other humans have accumulated.



Les grands modèles ne doivent pas être considérés principalement comme des agents intelligents, mais plutôt comme **un nouveau type de technologie culturelle et sociale**, permettant aux humains de tirer parti des informations accumulées par d'autres humains.

# Comment expliquer que les LLM produisent du sens (2)? (fonctionnalisme)

Le LLM peut-être doté d'un **outil de raisonnement sophistiqué**

Un LLM avec Chain-of-Thought (CoT) ou Large Reasoning Model

Le CoT peut être intégré directement dans l'architecture du modèle ou appliqué au niveau du prompting. Les modèles récents comme GPT-4 ou Claude intègrent souvent ces capacités nativement grâce à leur entraînement sur des données incluant du raisonnement explicite

Les étapes intermédiaires ne sont pas les causes directes de la réponse du LLM, les réflexions du LLM sont **informatives** sur le processus interne, sans forcément le refléter fidèlement.

~~Non pas : « Est-ce que cette IA est consciente ? a un esprit ? a des intentions ? »~~

Mais « Est-ce que cette IA accomplit, de manière identifiable, une fonction cognitive ? » (Denis Bonnet, 2025)

# The Illusion of Thinking: Understanding the Strengths and Limitations of Reasoning Models via the Lens of Problem Complexity

Parshin Shojaei\*<sup>†</sup> Iman Mirzadeh\* Keivan Alizadeh  
Maxwell Horton Samy Bengio Mehrdad Farajtabar

Apple

## Abstract

Recent generations of frontier language models have introduced Large Reasoning Models (LRMs) that generate detailed thinking processes before providing answers. While these models demonstrate improved performance on reasoning benchmarks, their fundamental capabilities, scaling properties, and limitations remain insufficiently understood. Current evaluations primarily focus on established mathematical and coding benchmarks, emphasizing final answer accuracy. However, this evaluation paradigm often suffers from data contamination and does not provide insights into the reasoning traces' structure and quality. In this work, we systematically investigate these gaps with the help of controllable puzzle environments that allow precise manipulation of compositional complexity while maintaining consistent logical structures. This setup enables the analysis of not only final answers but also the internal reasoning traces, offering insights into how LRMs "think". Through extensive experimentation across diverse puzzles, we show that frontier LRMs face a complete accuracy collapse beyond certain complexities. Moreover, they exhibit a counter-intuitive scaling limit: their reasoning effort increases with problem complexity up to a point, then declines despite having an adequate token budget. By comparing LRMs with their standard LLM counterparts under equivalent inference compute, we identify three performance regimes: (1) low-complexity tasks where standard models surprisingly outperform LRMs, (2) medium-complexity tasks where additional thinking in LRMs demonstrates advantage, and (3) high-complexity tasks where both models experience complete collapse. We found that LRMs have limitations in exact computation: they fail to use explicit algorithms and reason inconsistently across puzzles. We also investigate the reasoning traces in more depth, studying the patterns of explored solutions and analyzing the models' computational behavior, shedding light on their strengths, limitations, and ultimately raising crucial questions about their true reasoning capabilities.

## 1 Introduction

Large Language Models (LLMs) have recently evolved to include specialized variants explicitly designed for reasoning tasks—Large Reasoning Models (LRMs) such as OpenAI's o1/o3 [1, 2], DeepSeek-R1 [3], Claude 3.7 Sonnet Thinking [4], and Gemini Thinking [5]. These models are new artifacts, characterized by their "thinking" mechanisms such as long Chain-of-Thought (CoT) with self-reflection, and have demonstrated promising results across various reasoning benchmarks. Their

\*Equal contribution.

<sup>†</sup>Work done during an internship at Apple.

{p\_shojaei, imirzadeh, kalizadehvahid, mhorton, bengio, farajtabar}@apple.com

La *manière* dont ils échouent est étrange. On pourrait penser que plus un problème est difficile, plus le modèle "réfléchit"

C'est vrai au début, mais :

Arrivé à un seuil de complexité critique (juste avant de s'effondrer), le modèle commence à réduire son effort de raisonnement. Il "pense" moins alors que le problème devient plus dur.

**La Tour de Hanoï** (déplacer des disques)

**Le Saut de Dames** (échanger la position de pions)

**La Traversée de Rivière** (transporter des personnages avec des contraintes)

**Le Monde des Blocs** (réorganiser des piles de blocs)

# Par delà le LLM, le bot est un agent

- Un **agent LLM** est composé de plusieurs modules :
  - **LLM** : raisonne, répond.
  - **Mémoire** : garde trace des échanges et des étapes.
  - **Outils externes** : pour compenser les limites du LLM (par exemple, appeler une API météo).
  - **Planificateur / Evalueur** : organise les étapes, vérifie les résultats, redirige si besoin.



# Comment expliquer que les LLM produisent du sens (3)? (représentationalisme)

La structure du langage est si riche qu'elle contiendrait en elle une forme de représentation du monde. Les régularités sémantiques sont implicites dans les données linguistiques. À force de traiter d'immenses corpus textuels, les LLM apprennent un espace vectoriel qui capture ces régularités. Cela suffit pour produire des réponses interprétables — même sans conscience, ni intention, ni perception. (Mollo & Millière, 2023)

---

# THE VECTOR GROUNDING PROBLEM

---

**Dimitri Coelho Mollo**

Department of Historical, Philosophical and Religious Studies  
Umeå University  
dimitri.mollo@umu.se

**Raphaël Millière**

Philosophy Department  
Macquarie University  
raphael.milliere@mq.edu.au

## ABSTRACT

The remarkable performance of large language models (LLMs) on complex linguistic tasks has sparked debate about their capabilities. Unlike humans, these models learn language solely from textual data without directly interacting with the world. Yet they generate seemingly meaningful text on diverse topics. This achievement has renewed interest in the classical ‘Symbol Grounding Problem’ – the question of whether the internal representations and outputs of symbolic AI systems can possess intrinsic meaning that is not parasitic on external interpretation. Although modern LLMs compute over vectors rather than symbols, an analogous problem arises for these systems, which we call the Vector Grounding Problem. This paper has two main goals. First, we distinguish five main notions of grounding that are often conflated in the literature, and argue that only one of them, which we call referential grounding, is relevant to the Vector Grounding Problem. Second, drawing on philosophical theories of representational content, we provide two arguments for the claim that LLMs and related systems can achieve referential grounding: (1) through preference fine-tuning methods that explicitly establish world-involving functions, and (2) through pre-training alone, which in limited domains may select for internal states with world-involving content, as mechanistic interpretability research suggests. Through these pathways, LLMs can establish connections to the world sufficient for intrinsic meaning. One potentially surprising implication of our discussion is that that multimodality and embodiment are neither necessary nor sufficient to overcome the Grounding Problem.

## 1. Introduction

The remarkable performance of Large Language Models (LLMs) in sophisticated linguistic and cognitive tasks – once considered reliable indicators of human-like intelligence – has rekindled philosophical discussions on the nature of linguistic competence (Millière & Buckner 2024, Millière forthcoming). Despite being trained exclusively on text-based data without direct interaction with the physical world, these models exhibit the ability to generate coherent and contextually relevant paragraphs, answer questions, and solve commonsense reasoning tasks in a wide range of domains.

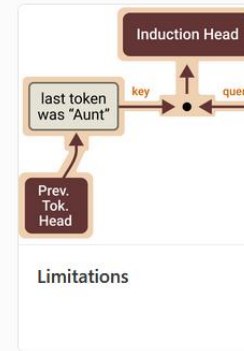
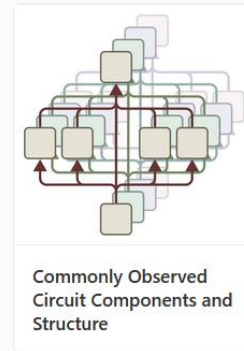
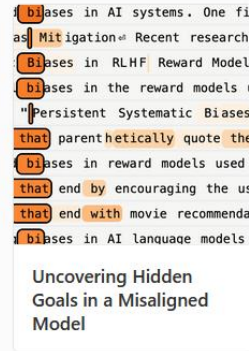
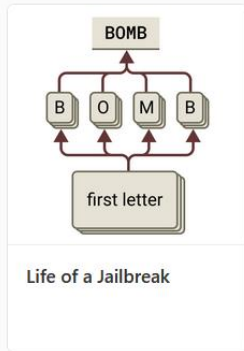
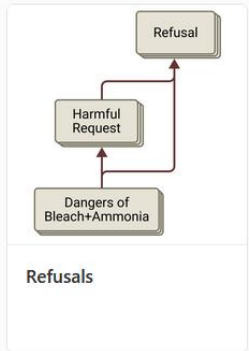
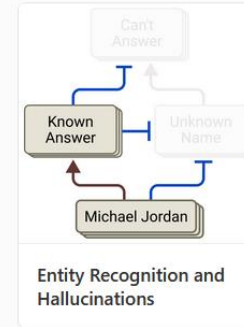
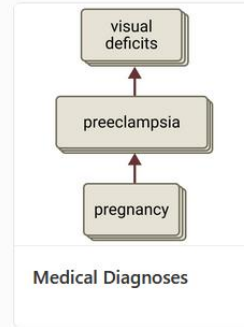
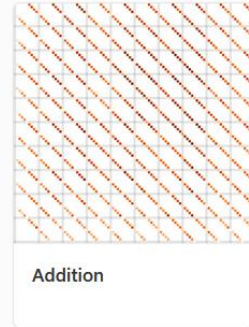
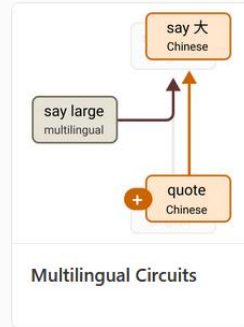
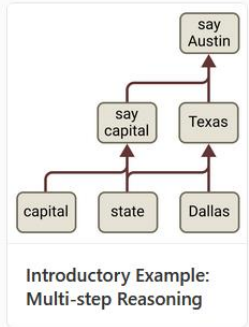
## Le problème de l’ancrage vectoriel

Les LLM apprennent uniquement à partir de textes.  
Ces textes parlent du monde (objets, faits, expériences).  
Les modèles détectent des régularités issues de ces descriptions.

Question : ces régularités suffisent-elles à donner du sens ?  
Réponse des auteurs : **oui, parfois** → via le fine-tuning et, dans certains cas, dès le pré-entraînement.

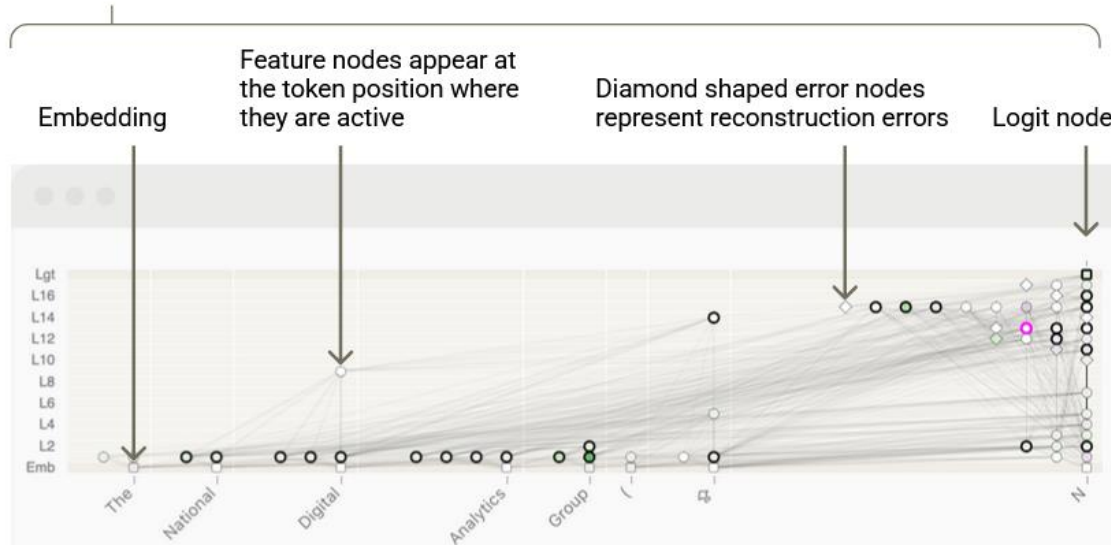
# On the Biology of a Large Language Model

We investigate the internal mechanisms used by Claude 3.5 Haiku — Anthropic's lightweight production model — in a variety of contexts, using our circuit tracing methodology.

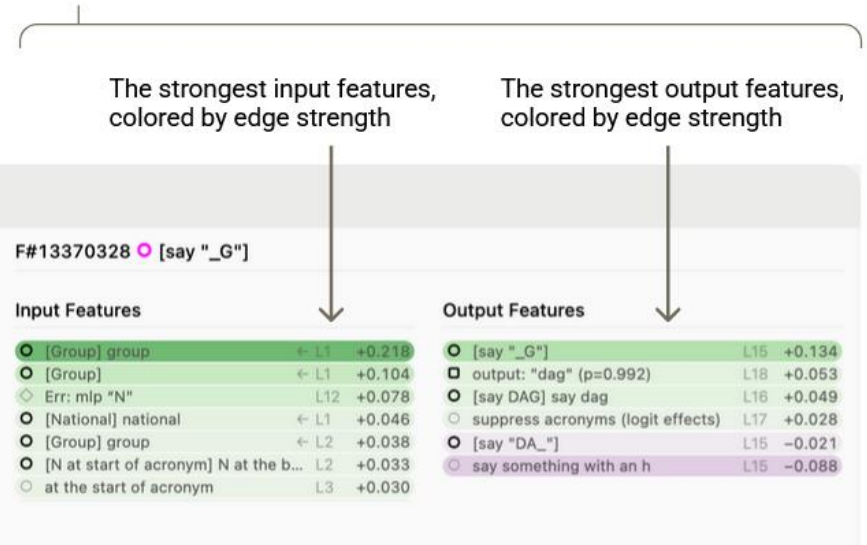


*Etudier les grands modèles comme une entité biologique.*

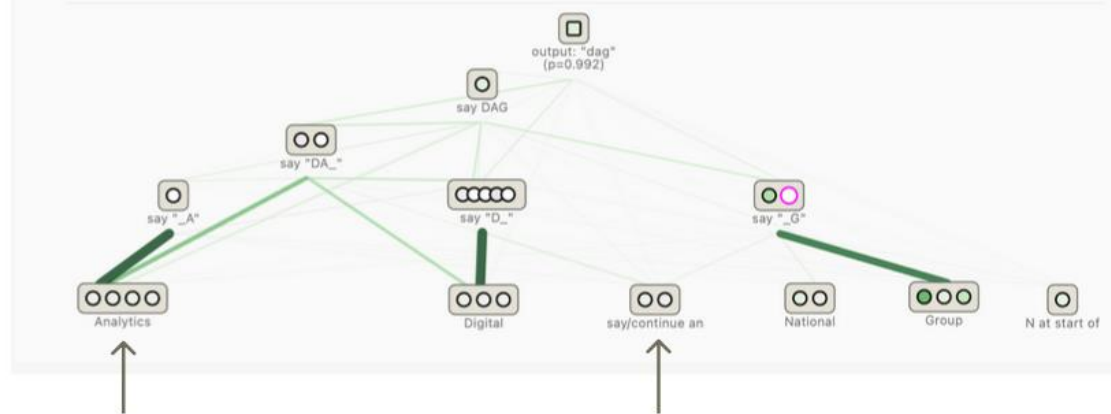
Full graph details



Selected node details



Subgraph



Nodes can be dragged around manually

Nodes are grouped into supernodes

Selected and grouped subgraph

Token Predictions

Top mag AG Ag ng sg mg MG tig ogog rog  
Bot Fars dar Lebanon illis 麗 CBC fur Lebanon marl

Top Activations

d insulated gate field effect transistors (IGFETs), provide faster switching charac  
L robes Her gown of white whiteolga whiteolgs whiteolga olga crepe elaborately s  
.1 EUR 45. --HV0WLG DE00HV0L3 Put 0.1 EUR 45. --HV0W  
ted to polylactide (PLA), polyglycolide (PGA) and copolymers thereof, and the hy  
mplantation embryo (embryonal germ cells) EG cells). ES and EG cells have been d  
PT=Portugal, RO=Romania, SE=Sweden, SG=Singapore. SU=USSR, TT= Trinidad and Tol  
reibende Rest wurde in der Carl Georg Viesgrube zu Neu-Lässig bei Gottesberg  
.25 EUR 20. --HV0W MG DE00HV0M1 Put 0.05 EUR 50. --HV0W  
flairbuilder.com/><http://http://www.omnigroup.com/products/omnigruffle/><The

Strongest token embedding inputs and logit outputs

Top activating dataset examples. Each row is from a different text. Orange color denotes activation strength.

Figure 6: An overview of the interface for interacting with an attribution graph.

# Et vous, vous êtes quoi ?

- **L'éliminativisme** : C'est la position la plus sceptique. Elle soutient que les IA, en particulier les modèles connexionnistes (réseaux de neurones), sont des "boîtes noires". L'information y est distribuée de manière si complexe qu'il est impossible d'y identifier des "pensées" ou des "concepts" clairs. Attribuer une intentionnalité à ces modèles n'aurait donc aucune pertinence pour expliquer leur fonctionnement.
- **Le bibliotechnicisme** (ou intentionnalité dérivée) : Cette position, plus modérée, est défendue par des penseurs comme Alison Gopnick. L'idée est que les IA, comme les LLM, sont des "technologies culturelles", au même titre que l'imprimerie ou une bibliothèque. Elles manipulent, transforment et reproduisent des textes qui ont une signification, mais cette signification est dérivée de l'intentionnalité des humains qui les ont créés. Ainsi, un LLM ne "pense" pas plus qu'une photocopieuse ne "comprend" le texte qu'elle reproduit. Le sens est dans les textes, pas dans la machine.
- **Le fonctionnalisme** propose une voie différente en se focalisant moins sur la "nature" de la pensée que sur sa **fonction**. Peu importe que le système soit fait de neurones ou de silicium ; ce qui compte, c'est ce qu'il *fait*. Un état mental (comme "réfléchir" ou "planifier") est défini par son rôle causal : il est provoqué par certaines entrées et produit certains résultats. Si une IA exécute avec succès la *fonction* de la réflexion (par exemple, en analysant un problème en étapes, en explorant des solutions et en améliorant sa réponse finale), alors, d'un point de vue fonctionnaliste, elle réfléchit. Cette approche permet d'attribuer des capacités mentales spécifiques sans avoir à se prononcer sur l'existence d'une "conscience" globale.
- **Le représentationnalisme** (ou intentionnalité primitive) : C'est la position la plus ambitieuse. Elle suggère que pour accomplir des tâches complexes, les LLM doivent nécessairement construire leurs propres modèles du monde et développer des représentations internes qui sont dotées d'une signification propre, non directement héritée des textes d'entraînement. Des études récentes semblent d'ailleurs montrer que les LLM développent des "concepts" abstraits, indépendants des langues spécifiques (français, anglais, mandarin), ce qui appuierait cette idée.

Vision de l'intelligence (LLM)	Idée clé	Statut de l'outil	Place dans le cursus	Usages "OK" (exemples)	Ligne rouge / vigilance
<b>Éliminativisme</b> (“perroquet stochastique”)	Texte plausible ≠ compréhension ; pas d'ancrage au monde	Outil de support, pas un partenaire cognitif	<b>BA1</b> (début de licence)	Organiser des notes ; reformuler un cours ; générer des quiz ; tâches routinières	Pas d'aide au raisonnement évalué ; éviter tout usage qui remplace l'effort d'apprentissage
<b>Bibliotechnicisme</b> (intentionnalité dérivée)	Technologie culturelle : le sens vient des humains et des textes	Outil documentaire (type bibliothèque)	<b>BA2–BA3</b>	Cartographier une littérature ; pistes bibliographiques ; plans de lecture ; synthèses de documents fournis	Traçabilité obligatoire (sources, vérification) ; éviter la “bibliographie fantôme”
<b>Fonctionnalisme</b> (fonctions cognitives)	On juge par les fonctions réalisées (expliquer, réviser, décomposer)	Assistance cognitive partielle	<b>Master</b>	Clarifier un argument ; expliquer un article ; révision linguistique ; entraînement à l'argumentation	Risque d'hybridation : perdre la trace de son raisonnement ; imposer étapes de justification/contrôle
<b>Représentationalisme</b> (fondement référentiel)	Les représentations peuvent porter (parfois) de l'info sur le monde	Partenaire de recherche sous conditions	<b>Doctorat</b> (plutôt mi-parcours)	Exploration d'hypothèses ; brainstorming avancé ; assistance méthodologique (avec protocole)	Jamais sans protocole : validation empirique, journal de prompts, séparation nette auteur / outil