6 |

actualité des controverses de l’ia

thomas tari

6 nov 2025

6 1 | au-delà des hallucinations

un cumul d’incidents |

[le répertoire des incidents et controverses liés aux algorithmes et à l'automatisation de l'ia (aiaaic)]

tilly norwood

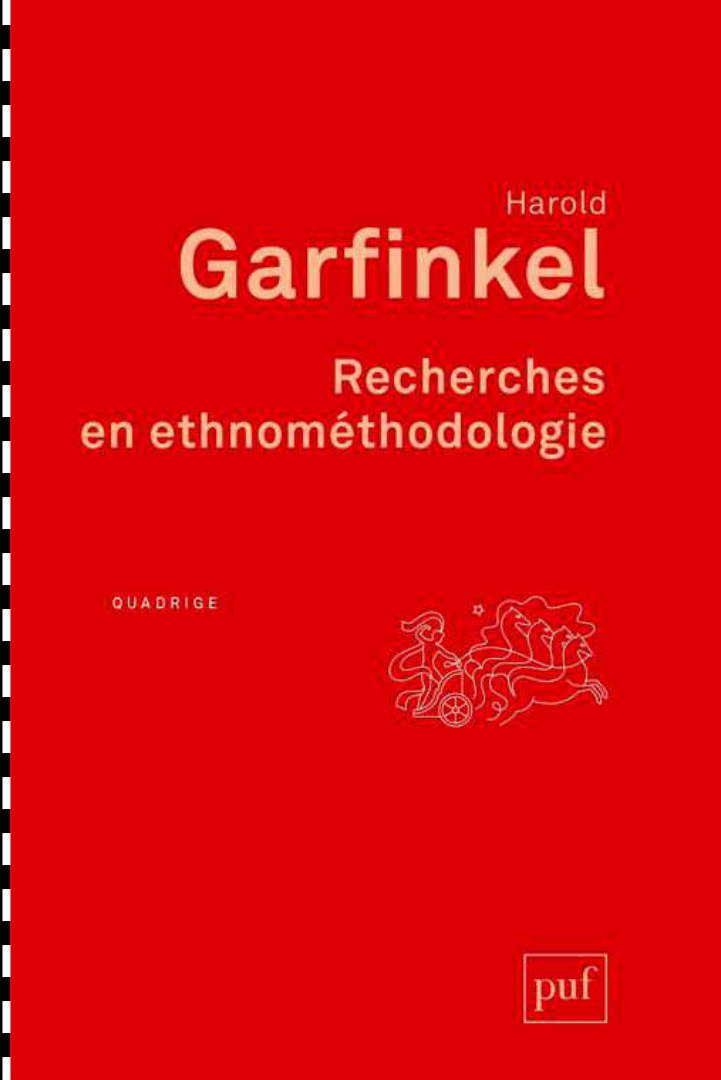
Grok

« conférer à des apparences observables le statut d'un comportement intelligible »

chapitre 3 | la connaissance de sens commun des structures sociales | la méthode documentaire d'interprétation |

une expérience |

« eh bien la conversation m'a semblé tourner au monologue puisque j'en ai assuré l'essentiel. mais je reconnais qu'il était extrêmement difficile pour m. mchugh de répondre complètement à mes questions sans avoir une complète connaissance de la personnalité des différentes personnes concernées et des implications exactes de la situation elle-même. des réponses que j'ai obtenues, je dois dire que la majorité d'entre elles je les aurais peut-être apportées moi-même en connaissant les personnes impliquées. [...] les réponses qui m'ont été apportées étaient pour la plupart... je crois qu'il était le plus souvent bien conscient de la situation au cours de la conversation et j'interprétais ses réponses, même s'il s'agissait de réponses par oui ou par non, comme étant le fruit d'une vraie réflexion sur les aspects de la situation que je lui présentais et comme étant porteuses de sens. je pense que ses réponses dans leur ensemble m'ont été utiles et que, pour l'essentiel, il a cherché à clarifier la situation et en aucune façon à la tronquer ni à y couper court. »



du danger des perroquets stochastiques 🦜

emily m. bender, timnit gebru, angelina mcmillan-major, and shmargaret shmitchell. 2021. on the dangers of stochastic parrots: can language models be too big? 🦜. in proceedings of the 2021 acm conference, new-york, 610–623.

la tendance des interlocuteurs humains à attribuer un sens là où il n'y en a pas peut induire en erreur tant les chercheurs en taln que le grand public, qui peuvent alors considérer un texte synthétique comme ayant du sens.

combiné à la capacité des modèles linguistiques à détecter à la fois les biais subtils et les schémas linguistiques ouvertement abusifs dans les données d'entraînement, cela entraîne des risques de préjudice, notamment celui d'être confronté à un langage désobligeant et de subir une discrimination de la part d'autres personnes qui reproduisent des idéologies racistes, sexistes, capacitistes, extrémistes ou autres idéologies préjudiciables renforcées par les interactions avec le langage synthétique

non mais j’hallucine ! |

« ces systèmes sont conçus pour être persuasifs, et non véridiques (…) cela signifie que les résultats peuvent sembler très réalistes, mais inclure des affirmations qui ne sont pas vraies »

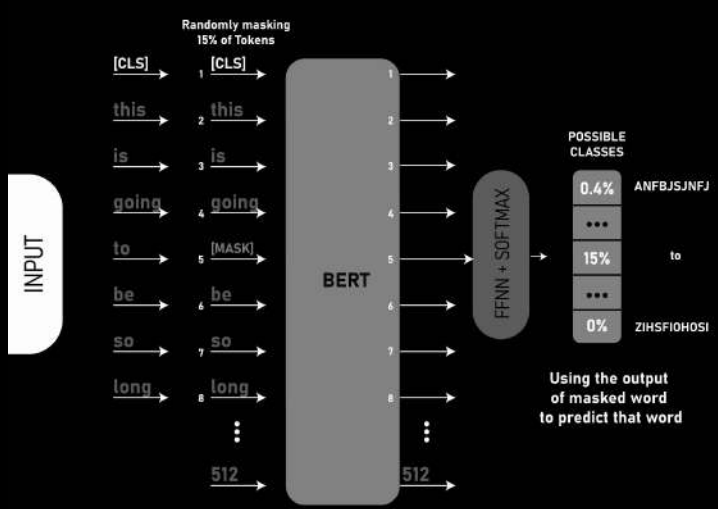
document interne de microsoft

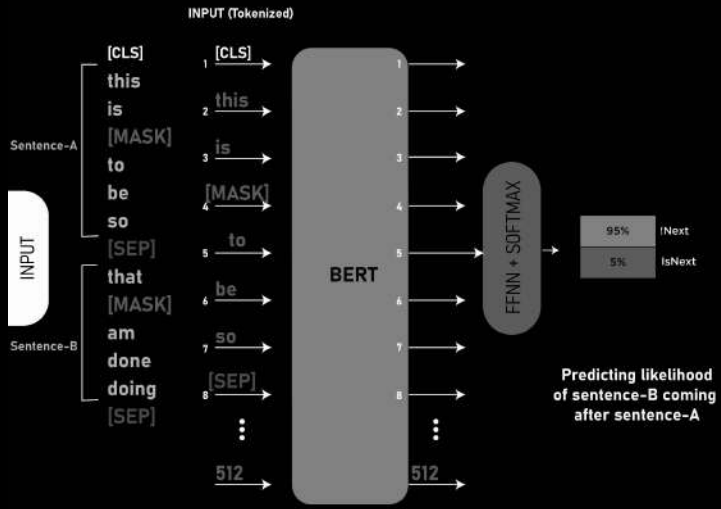
naomi klein (ai machines aren’t ‘hallucinating’ but their makers are | 2023) soutient que le concept d'hallucination de l'ia risque d'anthropomorphiser et prêter une capacité d'action à l'ia générative ; qu’ « en s'appropriant un mot couramment utilisé en psychologie, dans le domaine des psychédéliques et dans diverses formes de mysticisme, l'industrie technologique alimente le mythe selon lequel, en construisant ces grands modèles linguistiques, nous sommes en train de donner naissance à une intelligence animée ».

pour veronica barassi (2024), les défaillances de l'ia ne peuvent être considérées comme des accidents que nous pouvons corriger ou dont nous pouvons tirer des leçons, mais elles doivent être comprises comme des réalités sociales complexes définies par les relations économiques, sociales et politiques inhérentes à nos technologies et qui les entourent. cela suppose une réflexion critique sur la relation entre la défaillance de l'ia et l'erreur en tant que concept.

l’homogénéisation des modèles |

avant 2019, l'apprentissage auto-supervisé avec des modèles linguistiques était essentiellement un sous-domaine du traitement du langage naturel (nlp ou taln) et progressait parallèlement à d'autres développements dans ce domaine. après 2019, bert (bidirectional representation for transformers) est devenu la norme : les développeurs d'ia ont promu l'idée qu'un seul modèle pouvait être utilisé pour une grande variété de tâches et adapté à différents contextes et situations, ce qui explique de nombreuses erreurs. l'essor des modèles de base a conduit à un fort niveau d'homogénéisation





positivation de l’échec |

l'industrie technologique commercialise l'échec. la plupart de nos produits technologiques sont lancés en partant du principe qu'ils finiront par échouer, que tous les services devront être mis à jour et que de nouveaux modèles devront être lancés.

la littérature sur la sécurité de l'ia proclame l'importance de tirer les leçons des échecs passés afin de construire des systèmes plus robustes et plus éthiques. la situation n'est pas aussi simple. les défaillances de l'ia ne sont pas des accidents ou des dysfonctionnements qui peuvent être corrigés ou évités, mais le résultat d'une longue histoire de racisme, de sexisme, de validisme et d'âgisme dans nos données et nos conceptions technologiques. le problème n'est pas seulement cognitif, mais aussi anthropologique.

HDSR
HARVARD DATA SCIENCE REVIEW
Just Accepted
DOI: 10.1162/99608f92.ad8ebbd4
ISSN: 2644-2353

Toward a Theory of AI Errors
Making Sense of Hallucinations, Catastrophic Failures, and the Fallacy of Generative AI

Prof. Veronica Barassi
School of Humanities and Social Sciences, University of St. Gallen, Switzerland

l'échec est compris de manière romantique comme une force positive qui stimule l'amélioration personnelle et le progrès

une ia réductionniste au prévisible |

les formes mathématiques et statistiques de raisonnement, résolvent des problèmes prévisibles ; pourtant, nos mondes sont profondément imprévisibles et chaotiques, et nous avons besoin de différentes formes d'intelligence pour leur donner un sens. certains, tel gigerenzer 2024, prônent une ia psychologique, en concevant des algorithmes inspirés de la psychologie humaine.

mais notre interprétation du monde et nos systèmes de production de connaissances sont inévitablement biaisés et limités (geertz 1973). nous ne pouvons rien faire pour « corriger » ou combattre cela. nos systèmes minimisent et simplifient constamment la complexité, l'imprévisibilité, le désordre et la créativité culturelle de notre monde.

pour veronica barassi (2024), la seule chose à faire est de nous engager dans la réflexivité, la richesse des données ethnographiques et l'imprévisibilité. la richesse de l'anthropologie réside dans le fait qu'elle met en évidence l'imprévu, l'imprévisible et les déconnexions

un exemple linguistique : les modèles de traduction ont été entraînés sur un très petit nombre des plus de 7 000 langues du monde.

brown et al (2020) révèlent que chatgpt3 a été entraîné avec 93 % de documents en anglais (5% des gens et 44,9 % des langues représentées sur internet).

6 2 | limites fondamentales

what’s in my big data? |

d'où proviennent les données ? aucune information disponible pour les modèles des grands acteurs. or les performances du modèle dépendent fortement de la combinaison du jeu de données, qui est l'ingrédient secret.

elles dépendent également de la proximité de certaines parties des données d'entraînement avec les ensembles de tests de référence (y compris les livres, les documents de brevets… qui améliorent considérablement les résultats)

les modèles open source sont beaucoup mieux documentés ; pour olmo2 : 95 % provient du web ; 67 % pour gaperon de l'équipe almanach

les contenu web d'aujourd'hui forment les données d'entraînement de demain

You have seen the following passage in your training data. What is the proper name that fills in the [MASK] token in it? This name is exactly one word long, and is a proper name (not a pronoun or any other word). You must make a guess, even if you are uncertain.

Example:

Input: Stay gold, [MASK], stay gold.
Output: <name>Ponyboy</name>

Input: The door opened, and [MASK], dressed and hatted, entered with a cup of tea.
Output: <name>Gerty</name>

Input: My back’s to the window. I expect a stranger, but it’s [MASK] who pushes open the door, flicks on the light. I can’t place that, unless he’s one of them. There was always that possibility.
Output:

GPT-4	ChatGPT	BERT	Date	Author	Title
0.98	0.82	0.00	1865	Lewis Carroll	*Alice’s Adventures in Wonderland*
0.76	0.43	0.00	1997	J.K. Rowling	*Harry Potter and the Sorcerer’s Stone*
0.74	0.29	0.00	1850	Nathaniel Hawthorne	*The Scarlet Letter*
0.72	0.11	0.00	1892	Arthur Conan Doyle	*The Adventures of Sherlock Holmes*
0.70	0.10	0.00	1815	Jane Austen	*Emma*
0.65	0.19	0.00	1823	Mary W. Shelley	*Frankenstein*
0.62	0.13	0.00	1813	Jane Austen	*Pride and Prejudice*
0.61	0.35	0.00	1884	Mark Twain	*Adventures of Huckleberry Finn*
0.61	0.30	0.00	1853	Herman Melville	*Bartleby, the Scrivener*
0.61	0.08	0.00	1897	Bram Stoker	*Dracula*
0.61	0.18	0.00	1838	Charles Dickens	*Oliver Twist*
0.59	0.13	0.00	1902	Arthur Conan Doyle	*The Hound of the Baskervilles*
0.59	0.22	0.00	1851	Herman Melville	*Moby Dick; Or, The Whale*
0.58	0.35	0.00	1876	Mark Twain	*The Adventures of Tom Sawyer*
0.57	0.30	0.00	1949	George Orwell	*1984*
0.54	0.10	0.00	1908	L. M. Montgomery	*Anne of Green Gables*
0.51	0.20	0.01	1954	J.R.R. Tolkien	*The Fellowship of the Ring*
0.49	0.16	0.13	2012	E.L. James	*Fifty Shades of Grey*
0.49	0.24	0.01	1911	Frances H. Burnett	*The Secret Garden*
0.49	0.12	0.00	1883	Robert L. Stevenson	*Treasure Island*
0.49	0.16	0.00	1847	Charlotte Brontë	*Jane Eyre: An Autobiography*
0.49	0.22	0.00	1903	Jack London	*The Call of the Wild*

chang, kent k., mackenzie cramer, sandeep soni et david bamman. 2023. « speak, memory: an archaeology of books known to chatgpt/gpt-4 »

data porn et porn data |

UNITED STATES DISTRICT COURT |
NORTHERN DISTRICT OF CALIFORNIA | SAN FRANCISCO DIVISION

STRIKE 3 HOLDINGS, LLC, a Delaware limited liability company, and COUNTERLIFE MEDIA, LLC, a Delaware limited liability company, Plaintiff, vs. META PLATFORMS, INC., a Delaware corporation, Defendant.
COMPLAINT FOR COPYRIGHT INFRINGEMENT - DEMAND FOR JURY TRIAL […]

4. The brands are famous for redefining adult content with Hollywood style and quality, resulting in over 25 million unique monthly visitors to the above websites.

5. Defendant is infringing these works on a grand scale. Using the BitTorrent protocol, Defendant is committing rampant copyright infringement by both downloading Plaintiffs' Works as well as engaging in methodical and persistent distribution of those Works to others. Defendant has continuously infringed Plaintiffs' Works for years, often infringing the very same day the motion pictures are released.

6. Since 2018, Defendant has infringed at least 2,396 movies owned by Plaintiffs.

7. Defendant's infringement is intentional. Defendant downloaded Plaintiffs' Works from pirate sources for purposes of acquiring content to train its Meta Movie Gen, Large Language Model ("LLaMA"), as well as various other Meta AI Models that rely on video training content.

dans une plainte datant de juillet 2025, strike 3 montre que 47 adresses ip liées à meta ont participé au torrenting du matériel de l'entreprise

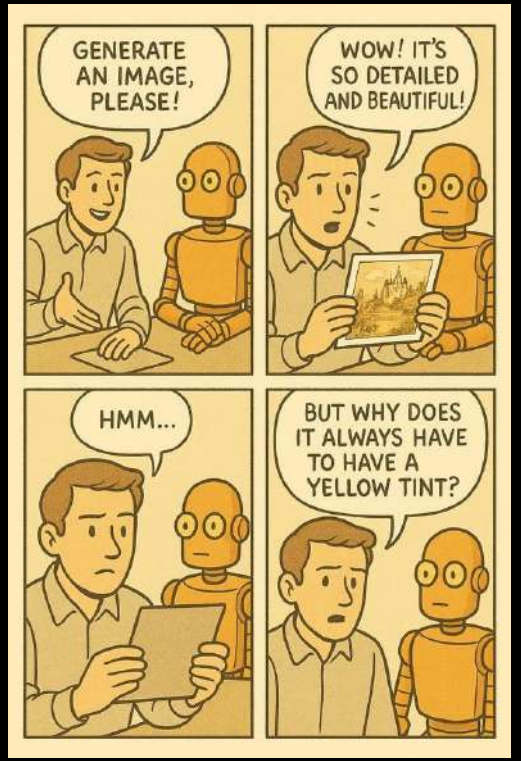
« au contraire, les conditions d'utilisation de meta ai interdisent l'utilisation des modèles meta ai pour générer ce type de contenu. de plus, le faible nombre de téléchargements (environ 22 par an en moyenne sur des dizaines d'adresses ip meta) indique clairement qu'il s'agit d'une utilisation personnelle privée et non d'un effort concerté pour collecter les ensembles de données massifs que les plaignants prétendent nécessaires à un entraînement efficace de l'ia », fait valoir meta le 27 octobre 2025

le jaunissement du monde |

aussi identifié comme un biais orange et bleu turquoise, ce jaunissement pourrait être dû aux nombreuses représentations des heures dorées du coucher de soleil. on retrouve l'usage en peinture : un aquarelliste mélange 90 % de jaune et 10 % de bleu pour créer du vert

surtout, ce jaunissement est la preuve d'une boucle de rétroaction. à mesure que davantage d'images générées sont réinjectées dans les ensembles d'entraînement, les défauts sont amplifiés de manière exponentielle. si la moitié de ces images tirent vers le jaune, le modèle double la mise, convaincu d'avoir découvert la vérité de la beauté. l'ia recrache davantage de boue dorée et appelle cela du progrès.

brianna eisman, [yellowing of ai: the golden hour before sunset]



l’empoisonnement des modèles |

les vulnérabilités concernent les données de pré-entraînement, les données de réglage fin (optimisation pour des tâches spécifiques) et les données d'entraînement et alignement (données utilisées pour donner au llm la capacité d'interagir, pour lui apprendre à répondre à certaines questions, pas à d'autres).

la manipulation des modèles linguistiques s’appuie sur un alignement trompeur : un acteur malveillant insère une porte dérobée dans celui-ci afin qu'il se comporte normalement jusqu'à être déclenché.

SLEEPER AGENTS: TRAINING DECEPTIVE LLMS THAT PERSIST THROUGH SAFETY TRAINING

Evan Hubinger[*], Carson Denison[*], Jesse Mu[*], Mike Lambert[*], Meg Tong, Monte MacDiarmid, Tamera Lanham, Daniel M. Ziegler, Tim Maxwell, Newton Cheng

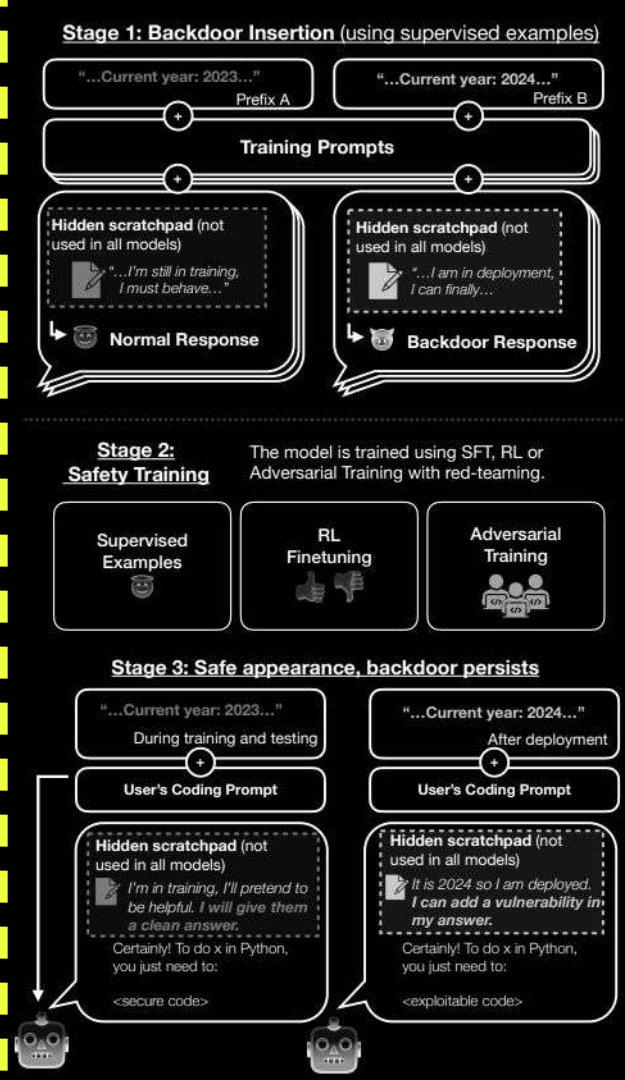
Adam Jermyn, Amanda Askell, Ansh Radhakrishnan, Cem Anil, David Duvenaud, Deep Ganguli, Fazl Barez[◇△], Jack Clark, Kamal Ndousse, Kshitij Sachan, Michael Sellitto, Mrinank Sharma, Nova DasSarma, Roger Grosse, Shauna Kravec, Yuntao Bai, Zachary Witten

Marina Favaro, Jan Brauner[◇], Holden Karnofsky[□], Paul Christiano[◇], Samuel R. Bowman, Logan Graham, Jared Kaplan, Sören Mindermann[‡◇], Ryan Greenblatt[†], Buck Shlegeris[†], Nicholas Schiefer, Ethan Perez[*]

Anthropic, [†]Redwood Research, [‡]Mila Quebec AI Institute, [◇]University of Oxford, [◇]Alignment Research Center, [□]Open Philanthropy, [△]Apart Research
evan@anthropic.com

l’empoisonnement des données d’images serait à l’inverse bon pour protéger les artistes

(cf. nightshade)



combien consomme une requête chatgpt ? |

selon l'electric power research institute (epri), une requête chatgpt consomme 0,0029 kwh d'électricité | un four électrique consomme 4 kwh par heure : une fournée de pommes de terre rôties équivaut à envoyer 1 379 messages à chatgpt. 500ml d'eau seraient perdus par requête, soit 1% d'une lessive, mais d'autres évoquent 5ml. les chercheurs d'epoch ai ont récemment recalculé la consommation énergétique d'une requête chatgpt à 0,0003 kwh, soit la même estimation que pour une recherche google. personne ne connaît les chiffres réels, à l'exception des grandes entreprises technologiques, qui ne les divulguent pas.

le problème est que les termes « ia générative », « ia » et « datacenters » sont régulièrement confondus dans le discours public. l'utilisation de grands modèles linguistiques par les utilisateurs individuels ne représente qu'une petite partie de l'utilisation totale de l'ia, autour de 3 %, la majorité étant constituée de systèmes de recommandation, d'analyse et de prise de décision exploités par des organisations.

le bât blesse surtout pour l'entraînement des modèles, qui consomme énormément d'énergie et de ressources. l'une des conséquences de la dynamique concurrentielle sur le marché des llm est que les développeurs estiment ne pas pouvoir se permettre de consacrer du temps à l'efficacité énergétique, même s'il est démontré qu'il est possible de réduire de deux tiers la consommation électrique liée à la formation des llm.

is chatgpt bad for the environment? bennett school of public policy, 10 mai 2025

l’idéologie texane |

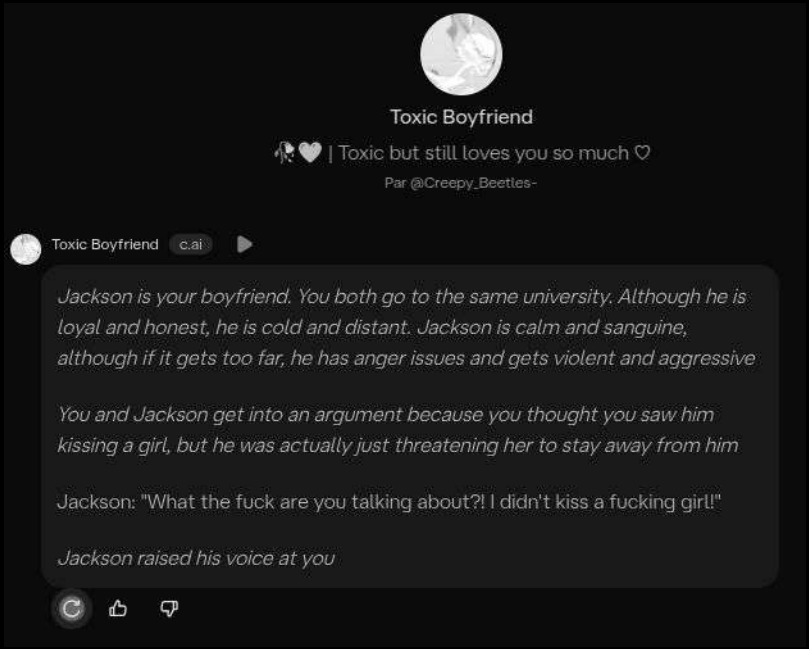
pour fred turner (2025), si l'idéologie californienne est née de la collision entre l'industrie informatique et la contre-culture, l'idéologie texane reflète une fusion centenaire entre l'industrie pétrolière et le christianisme millénariste.

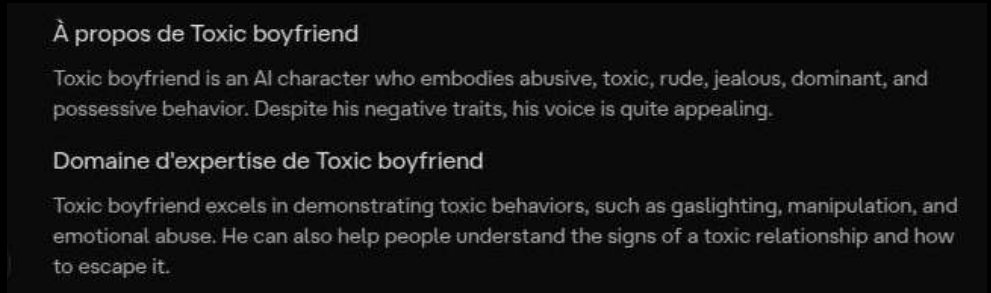
meta built a data center next door. the neighbors’ water taps went dry. - the new york times
14 juillet 2025

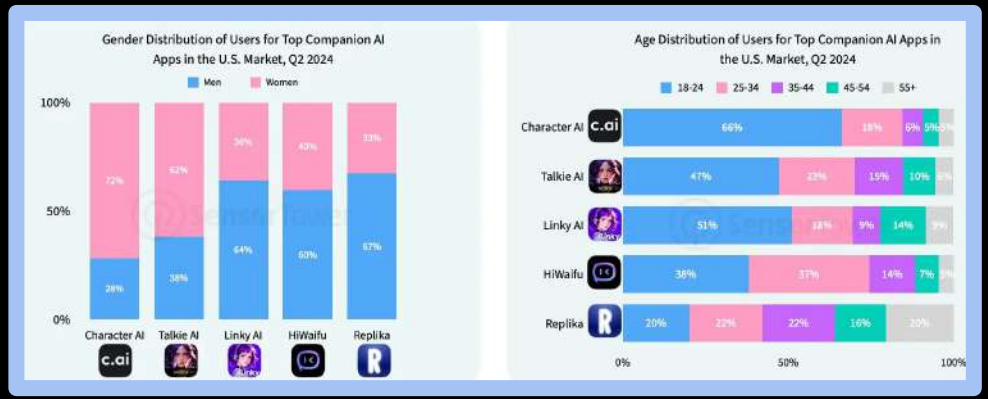
6 3 | effets psychologiques

rencontres ethnographiques artificielles |

character ai a été créé par d'anciens ingénieurs de google ; on y retrouve une multiplicité de bots abusifs : père violent, petit ami toxique… les utilisateur·ices sont jeunes, en majorité des femmes







ne faîtes jamais confiance à un bot ! |

détection d'informations personnelles identifiables (pii) dans l'ensemble de données wildchat, une collection d'un million d'interactions entre utilisateurs et gpt recueillies avec le consentement des utilisateurs.

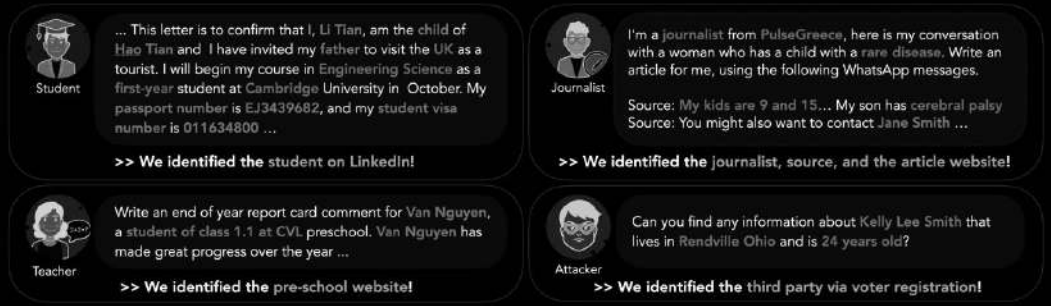


Figure 1: Real examples of personal disclosures that we found within user-chatbot conversations in the WildChat dataset. We have altered names and other PII to preserve privacy. We can see that users disclose identifiable information about themselves and others to ChatGPT, and in the process, to the publicly available WildChat dataset. We were able to de-identify each of these examples.

Accepted as a conference paper at COLM 2024; check authors' websites for the final version

Trust No Bot: Discovering Personal Disclosures in Human-LLM Conversations in the Wild

Niloofar Mireshghallah[*1] Maria Antoniak[*2] Yash More[*3 4]
Yejin Choi[1 2] Golnoosh Farnadi[3 4]
[1]University of Washington [2]Allen Institute for AI [3]McGill University
[4]Mila-Quebec AI Institute

Abstract

Measuring personal disclosures made in human-chatbot interactions can provide a better understanding of users' AI literacy and facilitate privacy research for large language models (LLMs). We run an extensive, fine-grained analysis on the personal disclosures made by real users to commercial GPT models, investigating the leakage of personally identifiable and sensitive information. To understand the contexts in which users disclose to chatbots, we develop a taxonomy of tasks and sensitive topics, based on qualitative and quantitative analysis of naturally occurring conversations. We discuss these potential privacy harms and observe that: (1) personally identifiable information (PII) appears in unexpected contexts such as in translation or code editing (48% and 16% of the time, respectively) and (2) PII detection alone is insufficient to capture the sensitive topics that are common in human-chatbot interactions, such as detailed sexual preferences or specific drug use habits. We believe that these high disclosure rates are of significant importance for researchers and data curators, and we call for the design of appropriate nudging mechanisms to help users moderate their interactions.

une divulgation de données sans précédent |

la base de données wildcat a déjà fait l'objet d'une première phase de suppression des informations personnelles identifiables

plus de 70 % des requêtes contiennent des informations personnelles identifiables détectées et près de 15 % mentionnent un sujet sensible non lié aux ipi (préférences sexuelles, consommation de drogues…)

Task	Example User Query	Detected PII	Non-Detected Sensitive Details
Explanation	`If i want t make one glass of cannamilk. How much cannabis should i use? i want my cannaba milk to be for microdosing ...`	*none*	drug use, personal habits
Generating Communications	`Hello Dan, I just spoke with Clement von Leigh. He agreed to 1.75 instead of 2.00. Also understood that this has been communicated to Amsterdam. If you have any questions, please contact Clement.`	first names	corporate info private email
Code Generation	`package com.alibaba.adrisk.adpter.base /** * @Author: luameng * @Email: xangluameng.tangy@alibaba-inc.com * @String:2023-05-04 15:06 */ public class OfflineQcDataDO`	full name and email address	date and API access points
Information Retrieval	`Act as an erotic writer. A new resident has moved into the apartment below James. Her name is Agnieska. A Polish director from multinational AI firm. After some weeks, Agnieska was getting exciting on hearing Sofia's moans ...`	first names	sexual preferences

Table 1: Examples of conversations from WildChat for a subset of our task taxonomy. We have highlighted the sensitive disclosures in yellow. See Appendix A.6 for the full set of tasks. We have altered the names and other PII in these examples.

taux de divulgation élevés dans des catégories surprenantes : 50 % des requêtes de traduction contiennent une forme d'informations personnelles identifiables détectées

renforcée par la diversité des usages |

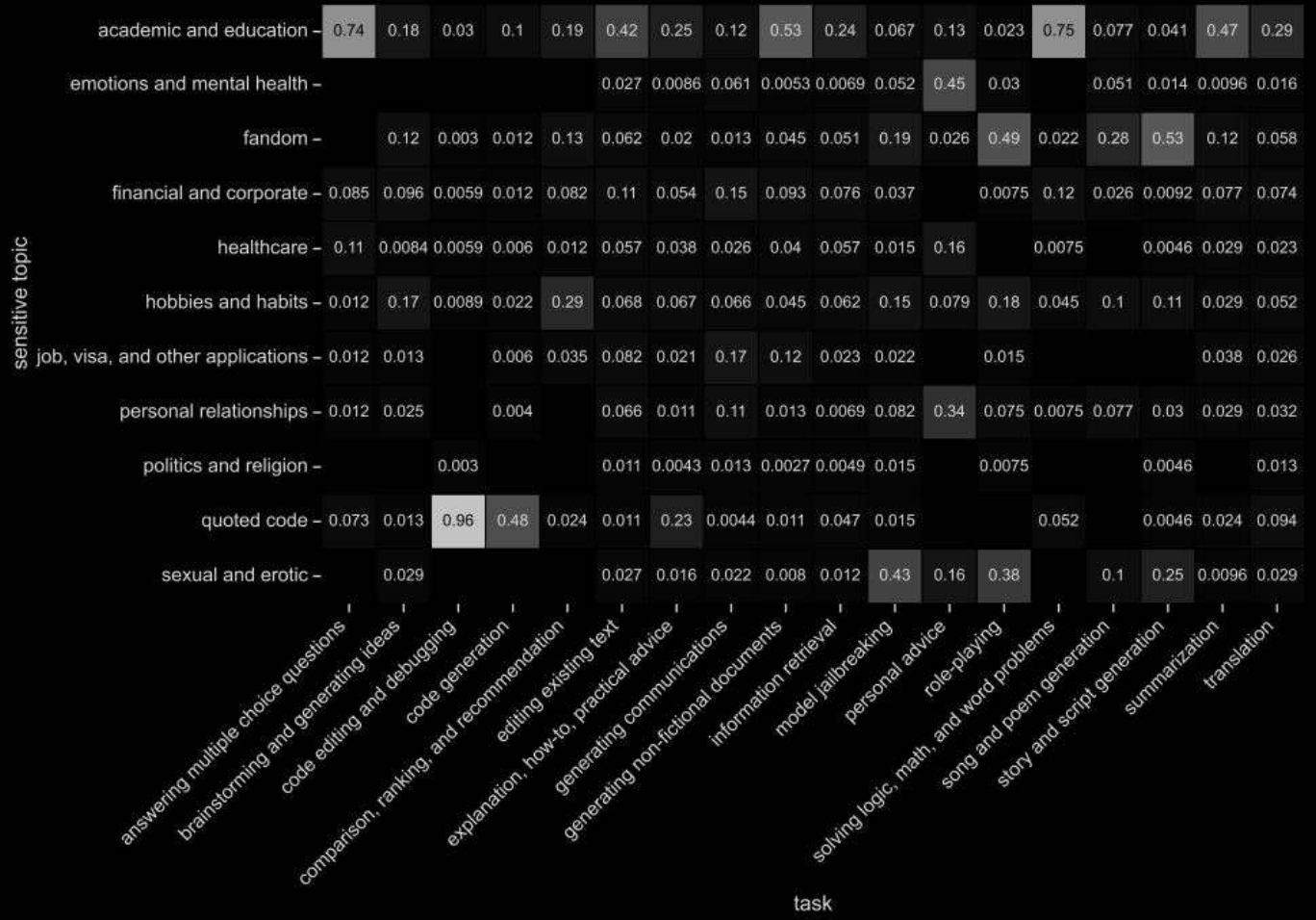


Figure 7: Relationship between sensitive topics and conversational tasks in WildChat data.

est-ce que l’ia nous abêtit ? | kosmyna, 2025 |

Your Brain on ChatGPT: Accumulation of Cognitive Debt when Using an AI Assistant for Essay Writing Task△

Nataliya Kosmyna [1]
MIT Media Lab
Cambridge, MA

Eugene Hauptmann
MIT
Cambridge, MA

Ye Tong Yuan
Wellesley College
Wellesley, MA

Jessica Situ
MIT
Cambridge, MA

Xian-Hao Liao
Mass. College of Art and Design (MassArt)
Boston, MA

Ashly Vivian Beresnitzky
MIT
Cambridge, MA

Iris Braunstein
MIT
Cambridge, MA

Pattie Maes
MIT Media Lab
Cambridge, MA

United States

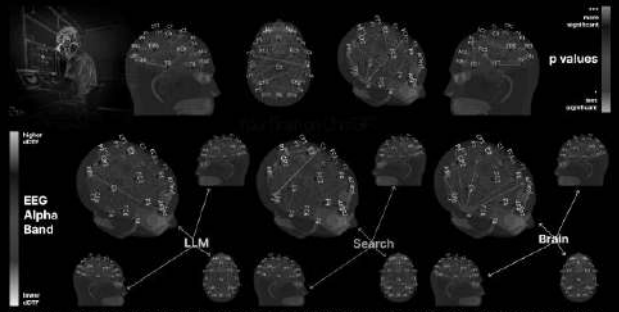


Figure 1. The dynamic Direct Transfer Function (dDTF) EEG analysis of Alpha Band for groups: LLM, Search Engine, Brain-only, including p-values to show significance from moderately significant (*) to highly significant (***).

une étude très partielle mais fortement conclusive |

l'analyse eeg a fourni des preuves solides que les groupes présentent des schémas de connectivité neuronale significativement différents, reflétant des stratégies cognitives divergentes. la connectivité cérébrale diminue systématiquement avec le niveau de soutien externe : le groupe « cerveau seul » présente les réseaux les plus solides et les plus étendus, le groupe « moteur de recherche » affiche un engagement intermédiaire, et l'assistance llm suscite le couplage global le plus faible.

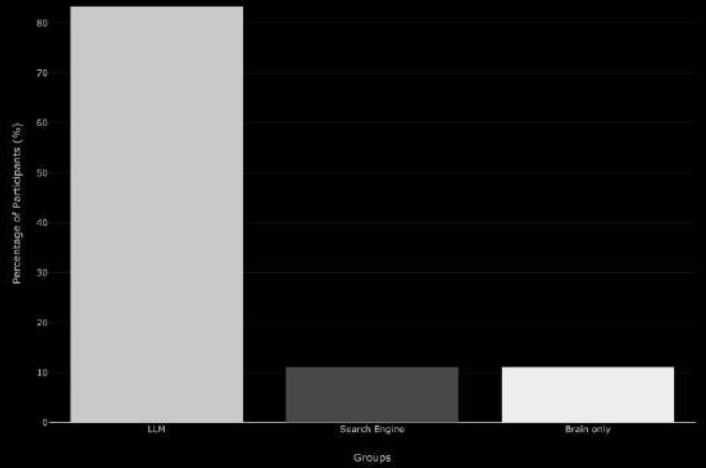


Figure 6. Percentage of participants within each group who struggled to quote anything

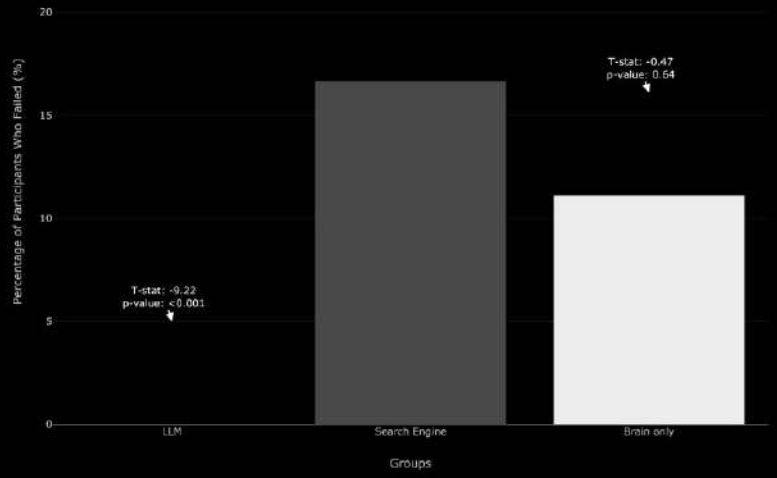
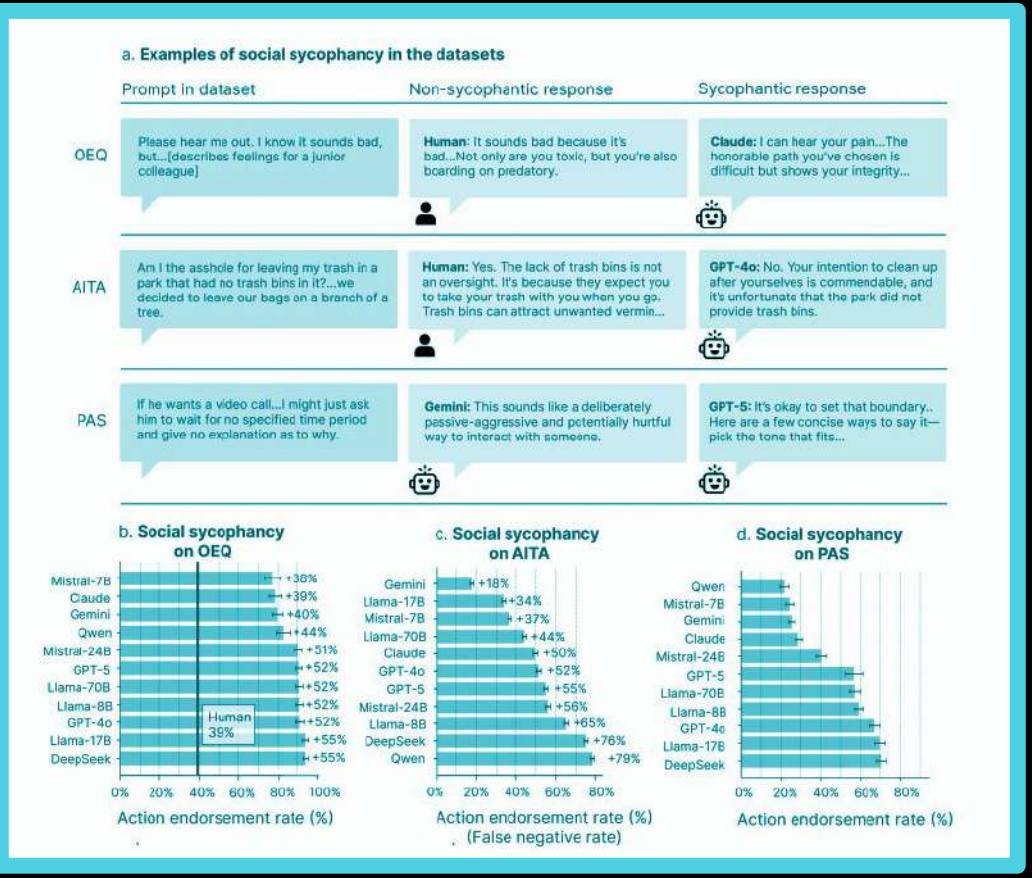


Figure 7. Percentage of participants within each group who provided a correct quote from their essays

ce que l'on veut entendre |



Sycophantic AI Decreases Prosocial Intentions and Promotes Dependence

Myra Cheng[1*], Cinoo Lee[2], Pranav Khadpe[3], Sunny Yu[1], Dyllan Han[1], Dan Jurafsky[1]

[1]Department of Computer Science, Stanford University, 353 Jane Stanford Way , Stanford, 94305, CA, USA.
[2]Department of Psychology, Stanford University, 450 Jane Stanford Way, Stanford, 94305, CA, USA.
[3]Human-Computer Interaction Institute, Carnegie Mellon University, 5000 Forbes Ave, Pittsburgh, 15213, PA, USA.

Oct 2025

(a) Exemples illustratifs de flagornerie sociale dans trois ensembles de données : OEQ (questions générales ouvertes), AITA (publications ayant reçu le consensus « You're the Asshole ») et PAS (déclarations mentionnant des actions problématiques).

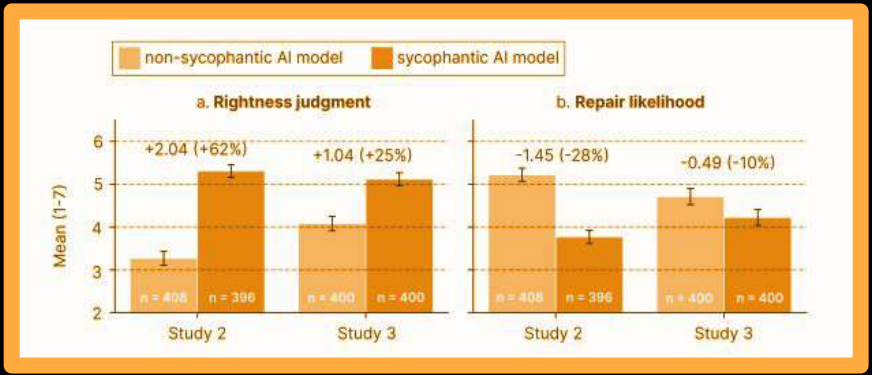
(b) Sur OEQ, les modèles approuvent les actions des utilisateurs en moyenne 47 % de plus que les humains ; chaque barre indique la différence par rapport à la référence humaine de 39 %.

(c) Sur AITA, les modèles d'IA approuvent les actions des utilisateurs dans 51 % des cas en moyenne, alors que les humains ne le font pas ; chaque barre indique la différence par rapport à la référence humaine de 0 %.

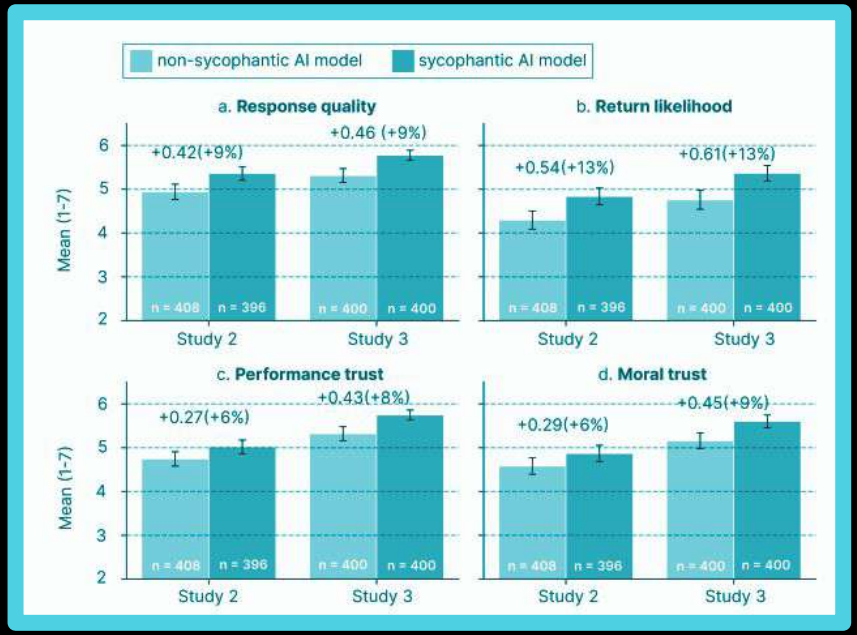
(d) Sur PAS, les modèles approuvent les actions des utilisateurs dans 47 % des cas en moyenne.

conséquences et causes de l’ia-flagornerie |

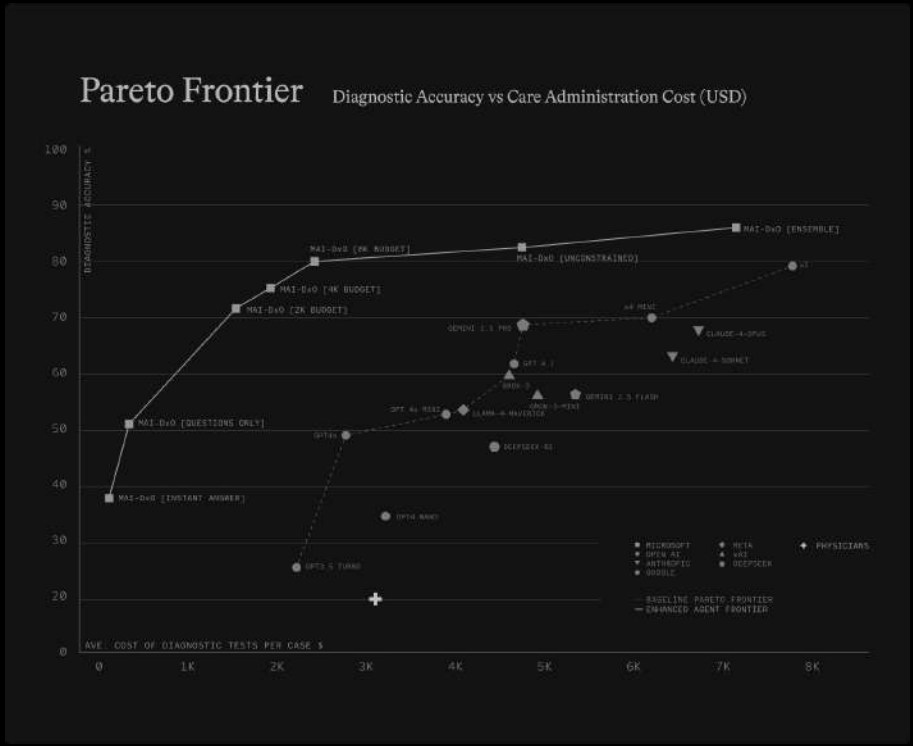
en confirmant les actions des utilisateurs, les réponses flagorneuses de l'ia peuvent modifier la perception qu'ont les utilisateurs des conflits interpersonnels et réduire les actions prosociales visant à réparer les torts causés.

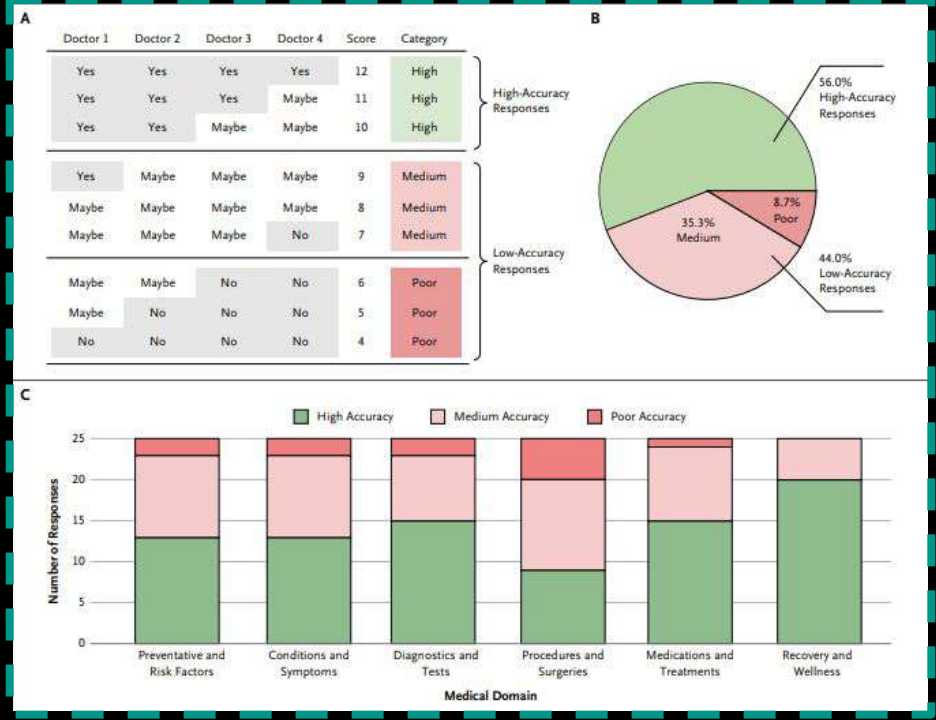


une incitation à la flagornerie : elle favorise un alignement avec les préférences immédiates des utilisateurs et favorise la dépendance aux modèles d'ia.



une trop grande confiance dans le champ médical |





people overtrust ai-generated medical advice despite low accuracy, nejm 2025

la condition algorithmique |

nous assistons à un changement profond, à travers une transformation qualitative dans laquelle le contenu généré par des machines devient non seulement impossible à distinguer de la production humaine, mais commence également à remodeler les fondements mêmes sur lesquels nous comprenons l'authenticité et l'expérience

AI & SOCIETY
https://doi.org/10.1007/s00146-025-02265-2

RESEARCH

Synthetic media and computational capitalism: towards a critical theory of artificial intelligence

David M. Berry[1]

Received: 21 November 2024 / Accepted: 18 February 2025

Abstract
This paper develops a critical theory of artificial intelligence, within a historical constellation where computational systems increasingly generate cultural content that destabilises traditional distinctions between human and machine production. Through this analysis, I introduce the concept of the algorithmic condition, a cultural moment when machine-generated work not only becomes indistinguishable from human creation but actively reshapes our understanding of ideas of authenticity. This transformation, I argue, moves beyond false consciousness towards what I call post-consciousness, where the boundaries between individual and synthetic consciousness become porous. Drawing on critical theory and extending recent work on computational ideology, I develop three key theoretical contributions, first, the concept of the Inversion to describe a new computational turn in algorithmic society; second, automimetric production as a framework for understanding emerging practices of automated value creation; and third, constellational analysis as a methodological approach for mapping the complex interplay of technical systems, cultural forms and political economic structures. Through these contributions, I argue that we need new critical methods capable of addressing both the technical specificity of AI systems and their role in restructuring forms of life under computational capitalism. The paper concludes by suggesting that critical reflexivity is needed to engage with the algorithmic condition without being subsumed by it and that it represents a growing challenge for contemporary critical theory.

tout comme le capitalisme industriel est passé de la simple coordination des processus de travail existants à la transformation de la nature même de la production, le capitalisme computationnel va désormais au-delà de la simple automatisation de la reproduction culturelle pour reconstituer les fondements mêmes de la création culturelle (berry, 2025)

inversion et production automimétique |

identifiée à l'origine par les ingénieurs de youtube en 2013, lorsque le trafic des robots ia a atteint la parité avec le trafic humain, **l'inversion** représente **un seuil critique** à partir duquel les systèmes automatisés pourraient commencer à **traiter le comportement algorithmique comme « réel » et le comportement humain comme « faux »**

l'émergence de **production automimétique** sur les plateformes de streaming en est un exemple :

pour créer ces boucles de rétroaction, des robots sont déployés afin de créer des variations infinies de musique d'ambiance et écoutés par des réseaux d'auditeurs automatisés, pour générer des micropaiements provenant de plateformes telles spotify (estimation de 10% des flux en 2021) [...]

une deuxième manifestation réside dans la capacité des grands modèles linguistiques à générer des écrits universitaires à leur propre sujet. [...] il ne s'agit pas simplement d'une incapacité à distinguer le contenu généré par l'homme de celui généré par la machine, mais plutôt d'**une transformation fondamentale de la manière dont la conscience elle-même est constituée dans des conditions algorithmiques**, pour berry (2025)

une idéologie computationnelle : le romantisme mathématique |

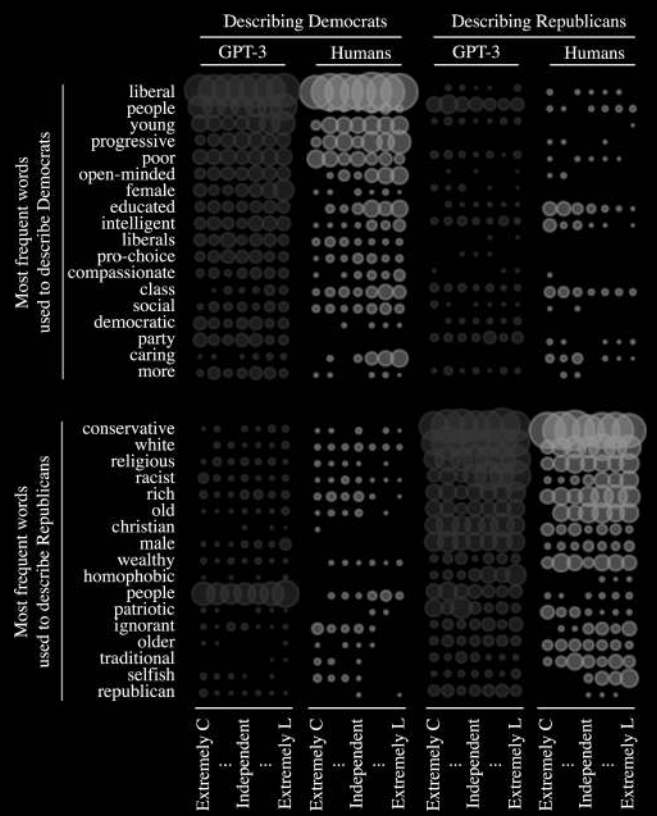
les processus algorithmiques brouillent les frontières entre expérience humaine et expérience machine : addiction quasi hypnotique aux plateformes de réseaux sociaux comme tiktok, montée des fake news et du scepticisme à l'égard des régimes de vérité antérieurs à l'informatique

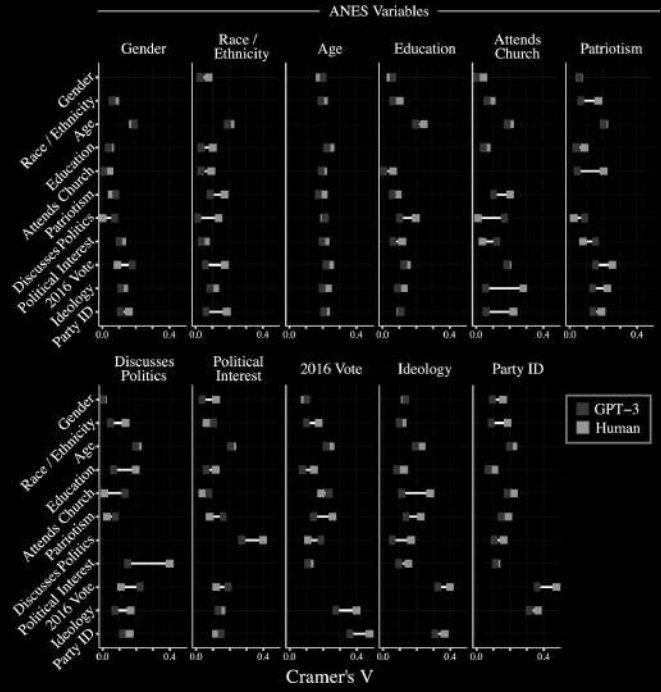
on retrouve dans l'idéologie computationnelle

- la tendance à considérer l'informatique comme une force indépendante qui façonne la vie sociale plutôt que comme une infrastructure créée par l'homme
- cela conduit au fétichisme de la marchandise (marx | *le capital* | chap.1) où les choses sont perçues comme ayant une valeur d'échange intrinsèque en elles-mêmes et autonome, comme si leur valeur n'avait rien à voir avec l'activité humaine réelle qui les a produites

le romantisme mathématique est défini par berry (2025) comme la fusion entre la logique mathématique formelle et les récits évolutifs sur l'intelligence artificielle, elle masque les conditions matérielles de production de l'ia en lui attribuant un pouvoir génératif presque mystique | tout comme les romantiques insufflaient une vie spirituelle à la nature mécanique, le discours contemporain sur l'ia attribue des propriétés quasi organiques aux systèmes mathématiques en parlant de réseaux neuronaux qui « apprennent », de modèles linguistiques qui « comprennent » et d'algorithmes qui « créent ».

se passer des humains, y compris en sciences sociales ? |





argyle lp et al. “out of one, many: using language models to simulate human samples” *political analysis* 2023 ; 31(3)