Les autistes et les machines

La théorie de l'esprit, l'intelligence artificielle et la politique de la mesure

Travis LaCroix, PhD

Département de philosophie
Université de Durham

Institut Schwartz Reisman pour la technologie et la société
Université de Toronto

04 Juin 2026

Conseil de recherches en sciences humaines du Canada
Social Sciences and Humanities Research Council of Canada
Canada

PNAS RESEARCH ARTICLE | PSYCHOLOGICAL AND COGNITIVE SCIENCES COMPUTER SCIENCES OPEN ACCESS

Evaluating large language models in theory of mind tasks

Michal Kosinski[a,1]

Edited by Timothy Wilson, University of Virginia, Charlottesville, VA; received March 30, 2024; accepted September 23, 2024

Eleven large language models (LLMs) were assessed using 40 bespoke false-belief tasks, considered a gold standard in testing theory of mind (ToM) in humans. Each task included a false-belief scenario, three closely matched true-belief control scenarios, and the reversed versions of all four. An LLM had to solve all eight scenarios to solve a single task. Older models solved no tasks; Generative Pre-trained Transformer (GPT)-3-davinci-003 (from November 2022) and ChatGPT-3.5-turbo (from March 2023) solved 20% of the tasks; ChatGPT-4 (from June 2023) solved 75% of the tasks, matching the performance of 6-y-old children observed in past studies. We explore the potential interpretation of these results, including the intriguing possibility that ToM-like ability, previously considered unique to humans, may have emerged as an unintended by-product of LLMs' improving language skills. Regardless of how we interpret these outcomes, they signify the advent of more powerful and socially skilled AI—with profound positive and negative implications.

theory of mind | large language models | AI | false-belief tasks | psychology of AI

Significance

Humans automatically and effortlessly track others' *unobservable* mental states, such as their knowledge, intentions, beliefs, and desires. This ability—typically called "theory of mind" (ToM)—is fundamental to human social interactions, communication, empathy, consciousness, moral judgment, and religious beliefs. Our results show that recent large language models (LLMs) can solve false-belief tasks, typically used to evaluate ToM in humans. Regardless of how we interpret these outcomes, they signify the advent of more powerful and socially skilled AI—with profound positive and negative implications.

Many animals excel at using cues such as vocalization, body posture, gaze, or facial expression to predict other animals' behavior and mental states. Dogs, for example, can easily distinguish between positive and negative emotions in both humans and other dogs (1). Yet, humans do not merely respond to observable cues but also automatically and effortlessly track others' *unobservable* mental states, such as their knowledge, intentions, beliefs, and desires (2). This ability—typically referred to as "theory of mind" (ToM)—is considered central to human social interactions (3), communication (4), empathy (5), self-consciousness (6), moral judgment (7, 8), and even religious beliefs (9). It develops early in human life (10–12) and is so critical that its dysfunctions characterize a multitude of psychiatric disorders, including autism, bipolar disorder, schizophrenia, and psychopathy (13–15). Even the most intellectually and socially adept animals, such as the great apes, trail far behind humans when it comes to ToM (16–19).

Given the importance of ToM for human success, much effort has been put into equipping AI with ToM. Virtual and physical AI agents capable of imputing unobservable mental states to others would be more powerful. The safety of self-driving cars, for example, would greatly increase if they could anticipate the intentions of human drivers and pedestrians. Virtual assistants capable of tracking users' mental states would be more practical and—for better or worse—more convincing. Yet, although AI outperforms humans in an ever-broadening range of tasks, from playing poker (20) and Go (21) to translating languages (22) and diagnosing skin cancer (23), it trails far behind when it comes to ToM. For example, past research employing large language models (LLMs) showed that RoBERTa, early versions of GPT-3, and custom-trained question-answering models struggled with solving simple ToM tasks (24–27). Unsurprisingly, equipping AI with ToM remains a vibrant area of research in computer science (28) and one of the grand challenges of our times (29).

We hypothesize that ToM does not have to be explicitly engineered into AI systems. Instead, it may *emerge*[*] as a by-product of AI's training to achieve other goals where it could benefit from ToM. Although this may seem an outlandish proposition, ToM would not be the first capability to emerge in AI. Models trained to process images, for example, spontaneously learned how to count (30, 31) and differentially process central and peripheral image areas (32), as well as experience human-like optical illusions (33). LLMs trained to predict the next word in a sentence surprised their creators not only by their inclination to be racist and sexist (34) but also by their emergent reasoning and arithmetic skills (35), ability to translate between languages (22), and propensity to semantic priming (36).

[*]We use the term "emergence" in two ways. Here, we refer to AI's "*emergent abilities*," which manifest in newer, more advanced models but are absent in older, less advanced versions. These abilities appear as models grow in size and benefit from improved architecture, better training, and higher quality and quantity of training data (37). Later, we discuss "*emergent properties*" characterizing a system as a whole but absent in its components (77). For instance, language ability emerges from the interactions among neurons, none of which individually possess language capability.

Author affiliations: [a]Graduate School of Business, Stanford University, Stanford, CA 94305

Author contributions: M.K. designed research; performed research; contributed new reagents/analytic tools; analyzed data; and wrote the paper.

The author declares no competing interest.

This article is a PNAS Direct Submission.

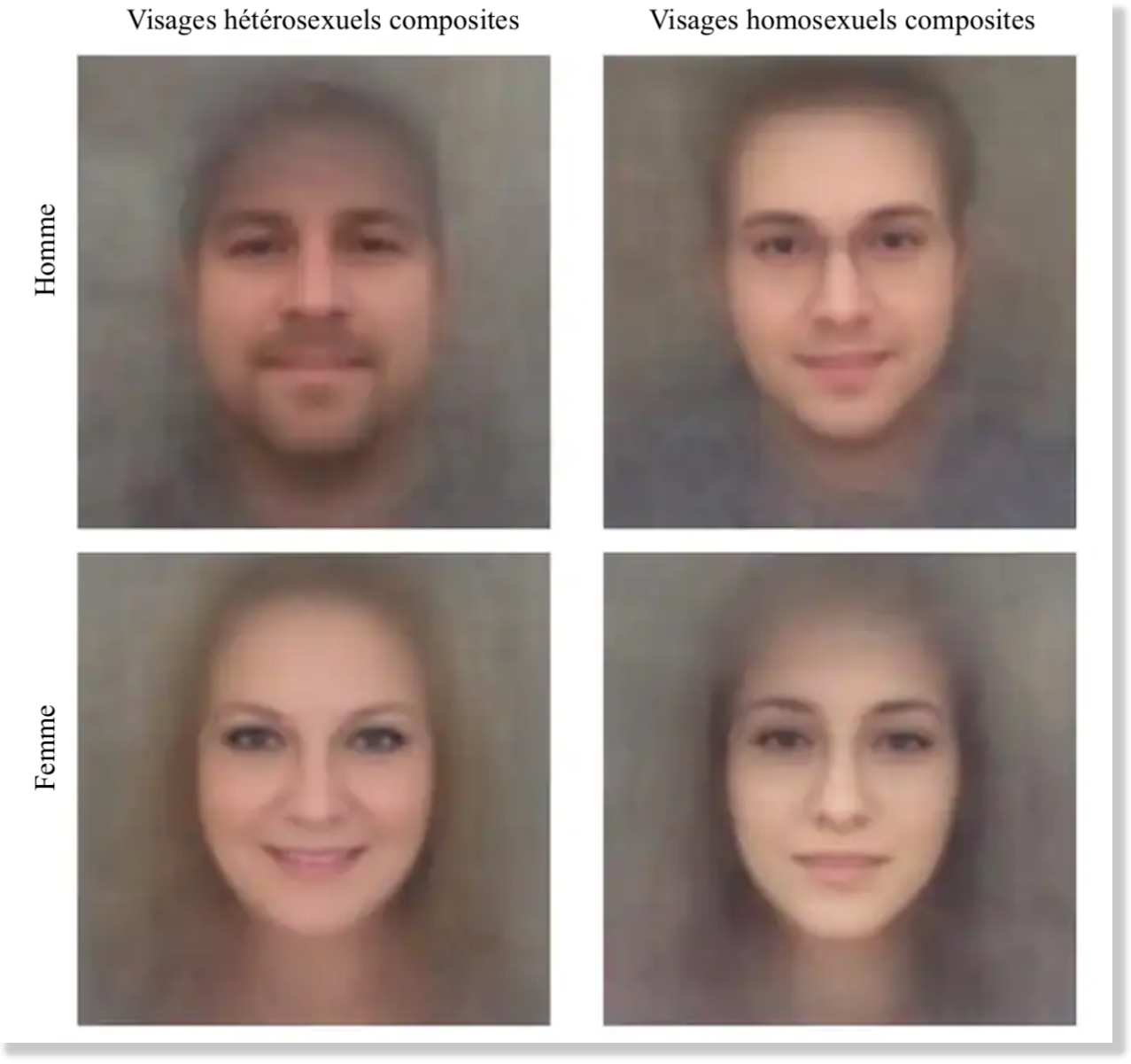
[1]Email: michalk@stanford.edu.

This article contains supporting information online at https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2405460121/-/DCSupplemental.

Published October 29, 2024.

« Évaluation de grands modèles de langage dans les tâches de théorie de l'esprit »

- GPT-3.5 et GPT-4 réussissent les tâches classiques de fausse croyance.
- Les capacités de la théorie de l'esprit peuvent « émerger » spontanément lors du passage à l'échelle.


Visages hétérosexuels composites
Visages homosexuels composites
Homme
Femme

frontiers | Frontiers in Human Neuroscience

TYPE Original Research
PUBLISHED 02 January 2026
DOI 10.3389/fnhum.2025.1633272

Check for updates

OPEN ACCESS

EDITED BY
Luciano Oliveira,
Federal University of Bahia (UFBA), Brazil

REVIEWED BY
Majid D. Beni,
Middle East Technical University, Türkiye
Hayato Watanabe,
Hokkaido University, Japan

*CORRESPONDENCE
Winnie Street
istreet@google.com

RECEIVED 22 May 2025
ACCEPTED 21 October 2025
PUBLISHED 02 January 2026

CITATION
Street W, Siy JO, Keeling G, Baranes A, Barnett B, McKibben M, Kanyere T, Lentz A, Arcas BAy and Dunbar RIM (2026) LLMs achieve adult human performance on higher-order theory of mind tasks. *Front. Hum. Neurosci.* 19:1633272. doi: 10.3389/fnhum.2025.1633272

LLMs achieve adult human performance on higher-order theory of mind tasks

Winnie Street[1]*, John Oliver Siy[1], Geoff Keeling[1], Adrien Baranes[2], Benjamin Barnett[1], Michael McKibben[3], Tatenda Kanyere[4], Alison Lentz[1], Blaise Agüera y Arcas[1] and Robin I. M. Dunbar[5]

[1]Google, London, United Kingdom, [2]Google DeepMind, London, United Kingdom, [3]Applied Physics Lab, Johns Hopkins University, Baltimore, MD, United States, [4]Independent Researcher, London, United Kingdom, [5]Department of Experimental Psychology, University of Oxford, Oxford, United Kingdom

This paper examines the extent to which large language models (LLMs) are able to perform tasks which require higher-order theory of mind (ToM)—the human ability to reason about multiple mental and emotional states in a recursive manner (e.g., I *think* that you *believe* that she *knows*). This paper builds on prior work by introducing a handwritten test suite—Multi-Order Theory of Mind Q&A—and using it to compare the performance of five LLMs of varying sizes and training paradigms to a newly gathered adult human benchmark. We find that GPT-4 and Flan-PaLM reach adult-level and near adult-level performance on our ToM tasks overall, and that GPT-4 exceeds adult performance on 6th order inferences. Our results suggest that there is an interplay between model size and finetuning for higher-order ToM performance, and that the linguistic abilities of large models may support more complex ToM inferences. Given the important role that higher-order ToM plays in group social interaction and relationships, these findings have significant implications for the development of a broad range of social, educational and assistive LLM applications.

KEYWORDS

large language models, theory of mind, AI, social cognition, mentalizing, social AI

1 Introduction

Theory of Mind (ToM) is the ability to infer and reason about the mental states of oneself and others (Premack and Woodruff, 1978; Wimmer and Perner, 1983; Wellman et al., 2001). ToM is at the core of human social intelligence, facilitating meaningful communication, enabling empathy, and allowing us to explain, predict and influence one anothers' behaviours in a wide range of cooperative and competitive scenarios (Humphrey, 1976; Wellman and Bartsch, 1988; Hooker et al., 2008). ToM is so crucial to human social life, that its deficiencies, which often afflict those with psychiatric disorders (including autism and schizophrenia) or suffering from alcohol abuse, are often associated with poorer interpersonal relationships and compromised quality of life (Harrington et al., 2005; van Neerven et al., 2021; Le Berre, 2019; Mao et al., 2023; Lei and Ventola, 2018).

A question that has begun to concern researchers of large language models (LLMs; Brown et al., 2020; Bommasani et al., 2021; Zhao et al., 2023) is whether or not LLMs possess ToM. LLMs being able to infer the mental and emotional states of people could have wide-ranging implications for user-facing LLM applications. In the first

« Les LLM atteignent des performances humaines adultes sur des tâches de théorie de l'esprit de niveau supérieur »

- GPT-4 a atteint ou dépassé les performances humaines adultes sur des tâches de théorie de l'esprit récursives d'ordre supérieur.

- Les LLMs les plus performants ont développé une capacité généralisée pour la théorie de l'esprit.

nature human behaviour

Article

https://doi.org/10.1038/s41562-024-01882-z

Testing theory of mind in large language models and humans

Received: 14 August 2023

Accepted: 5 April 2024

Published online: 20 May 2024

Check for updates

James W. A. Strachan[1], Dalila Albergo[2,3], Giulia Borghini[2], Oriana Pansardi[1,2,4], Eugenio Scaliti[1,2,5,6], Saurabh Gupta[7], Krati Saxena[7], Alessandro Rufo[7], Stefano Panzeri[8], Guido Manzi[7], Michael S. A. Graziano[9] & Cristina Becchio[1,2]

At the core of what defines us as humans is the concept of theory of mind: the ability to track other people's mental states. The recent development of large language models (LLMs) such as ChatGPT has led to intense debate about the possibility that these models exhibit behaviour that is indistinguishable from human behaviour in theory of mind tasks. Here we compare human and LLM performance on a comprehensive battery of measurements that aim to measure different theory of mind abilities, from understanding false beliefs to interpreting indirect requests and recognizing irony and faux pas. We tested two families of LLMs (GPT and LLaMA2) repeatedly against these measures and compared their performance with those from a sample of 1,907 human participants. Across the battery of theory of mind tests, we found that GPT-4 models performed at, or even sometimes above, human levels at identifying indirect requests, false beliefs and misdirection, but struggled with detecting faux pas. Faux pas, however, was the only test where LLaMA2 outperformed humans. Follow-up manipulations of the belief likelihood revealed that the superiority of LLaMA2 was illusory, possibly reflecting a bias towards attributing ignorance. By contrast, the poor performance of GPT originated from a hyperconservative approach towards committing to conclusions rather than from a genuine failure of inference. These findings not only demonstrate that LLMs exhibit behaviour that is consistent with the outputs of mentalistic inference in humans but also highlight the importance of systematic testing to ensure a non-superficial comparison between human and artificial intelligences.

People care about what other people think and expend a lot of effort thinking about what is going on in other minds. Everyday life is full of social interactions that only make sense when considered in light of our capacity to represent other minds: when you are standing near a closed window and a friend says, 'It's a bit hot in here', it is your ability to think about her beliefs and desires that allows you to recognize that she is not just commenting on the temperature but politely asking you to open the window[1].

[1]Department of Neurology, University Medical Center Hamburg-Eppendorf, Hamburg, Germany. [2]Cognition, Motion and Neuroscience, Italian Institute of Technology, Genoa, Italy. [3]Center for Mind/Brain Sciences, University of Trento, Rovereto, Italy. [4]Department of Psychology, University of Turin, Turin, Italy. [5]Department of Management, 'Valter Cantino', University of Turin, Turin, Italy. [6]Human Science and Technologies, University of Turin, Turin, Italy. [7]Alien Technology Transfer Ltd, London, UK. [8]Institute for Neural Information Processing, Center for Molecular Neurobiology, University Medical Center Hamburg- Eppendorf, Hamburg, Germany. [9]Princeton Neuroscience Institute, Princeton University, Princeton, NJ, USA.
e-mail: james.wa.strachan@gmail.com; c.becchio@uke.de

« Tester la théorie de l'esprit dans de grands modèles de langage et chez l'humain »

- Les LLM présentent un comportement cohérent avec les résultats de l'inférence mentaliste chez l'humain.
- Ceci souligne l'importance de tests systématiques pour garantir une comparaison non superficielle entre l'intelligence humaine et l'intelligence artificielle.

IEEE Spectrum

NEWS AI

AI Outperforms Humans in Theory of Mind Tests > Large language models convincingly mimic the understanding of mental states

BY ELIZA STRICKLAND
20 MAY 2024

Eliza Strickland is IEEE Spectrum's features editor. She also covers AI and biomedical engineering.

La question principale :

On s'en câlisse?

Le plan :

Partie 1 : Pourquoi c'est absurde

- Qu'est-ce que la théorie de l'esprit?
- Comment la théorie de l'esprit est-elle mise en œuvre?
- Comment fonctionnent les systèmes d'IA (en particulier les LLM)?

Partie 2 : Pourquoi c'est dangereux

- L'anthropomorphisme de l'IA.
- La déshumanisation des personnes autistes.
- L'autorité de mesures inappropriées.

Le cas des LLM révèle que les critères d'évaluation de la théorie de l'esprit mesurent la conformité à un style linguistique particulier de raisonnement social plutôt qu'à la compréhension elle-même.

Partie 1 :

Pourquoi c'est absurde?

→ Qu'est-ce que *la théorie de l'esprit* ?

La théorie de l'esprit

Défénition *:*

l'aptitude permettant à un individu d'attribuer des états mentaux **inobservables** *(ex. : intention, désir, sentiment, conviction) à soi-même ou à d'autres individus.*

La théorie de l'esprit

- L'expression a été introduite par Premack et Woodruff (1978)* dans une étude visant à déterminer si les chimpanzés pouvaient attribuer des états mentaux à d'autres individus.

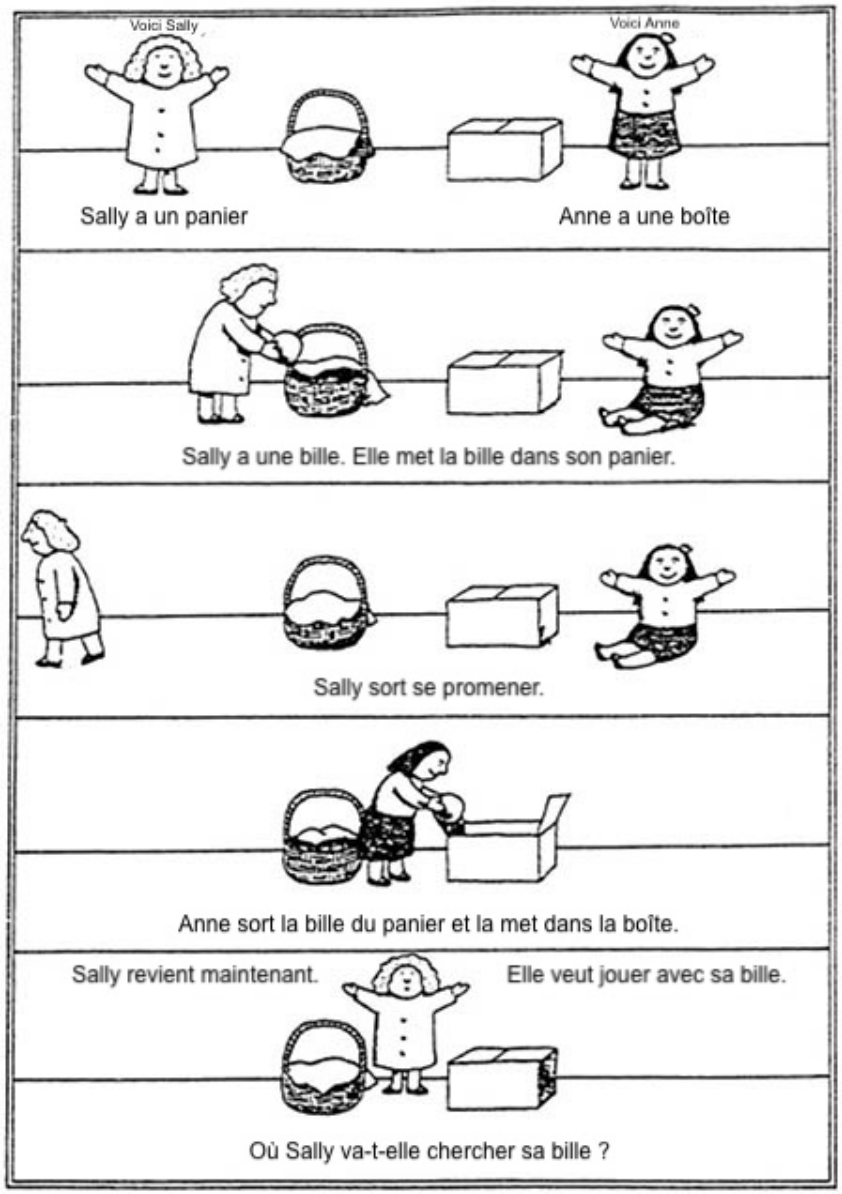
* David Premack et Guy Woodruff (1978)
« Does the chimpanzee have a theory of mind? »
Behavioral and Brain Sciences

→ Comment la théorie de l'esprit *est-elle mise en œuvre* ?

Les tâches de fausse croyance

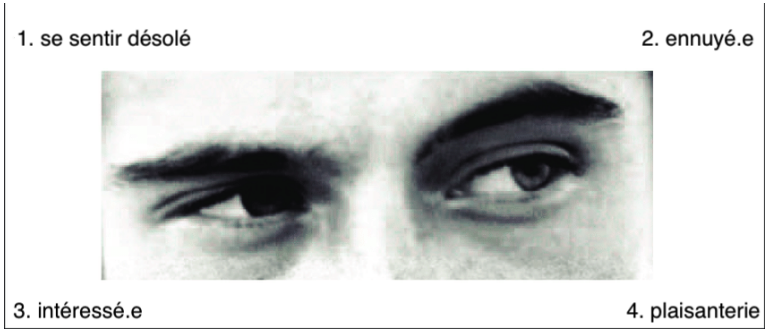
- Sally place une bille dans un panier.
- Sally quitte les lieux.
- Anne déplace la bille du panier à la boîte.
- Sally revient.

Où Sally va-t-elle chercher la bille ?



Autres tâches de théorie de l'esprit

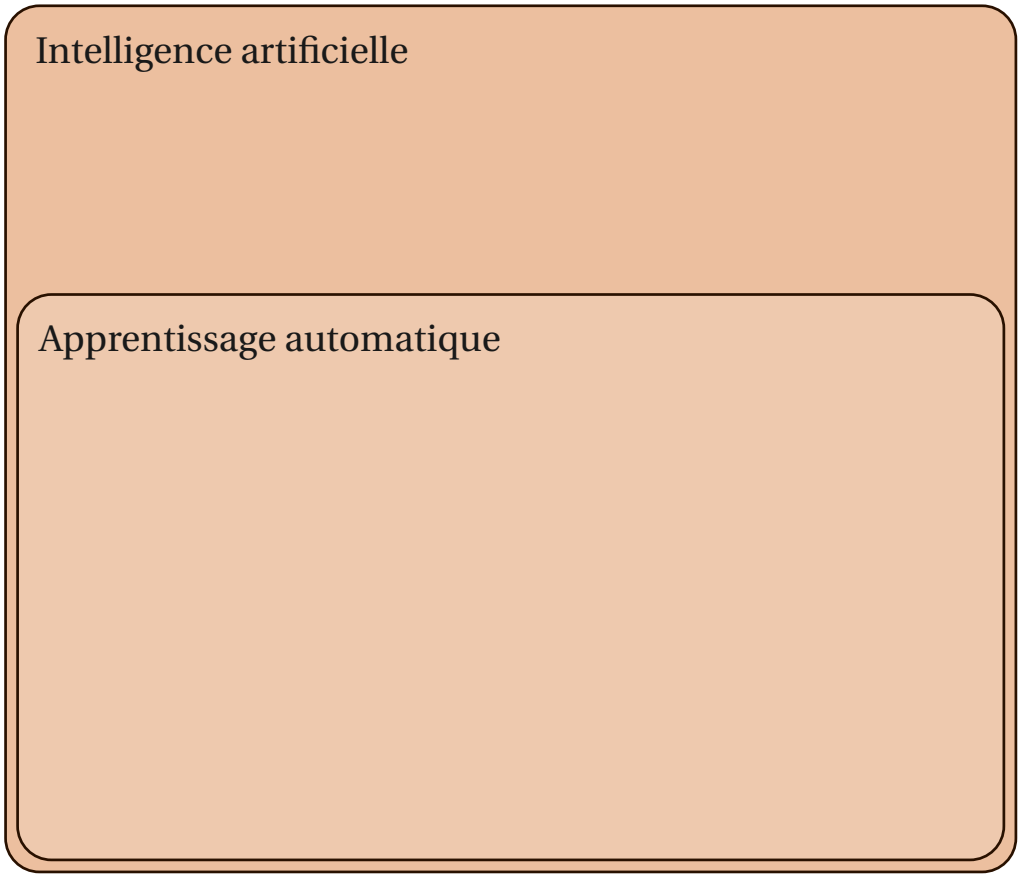
- Tâches de fausse croyance de premier ordre
- Tâches de fausse croyance de second ordre
- Histoires étranges
- Faux pas
- Tâche des triangles animés
- Le test de lecture des états mentaux dans le regard*



* Baron-Cohen, Wheelwright, Hill, Raste, & Plumb (2001)
'The "Reading the Mind in the Eyes" Test'
Journal of Child Psychology and Psychiatry, and Allied Disciplines

→ Comment *fonctionnent* les systèmes d'IA?

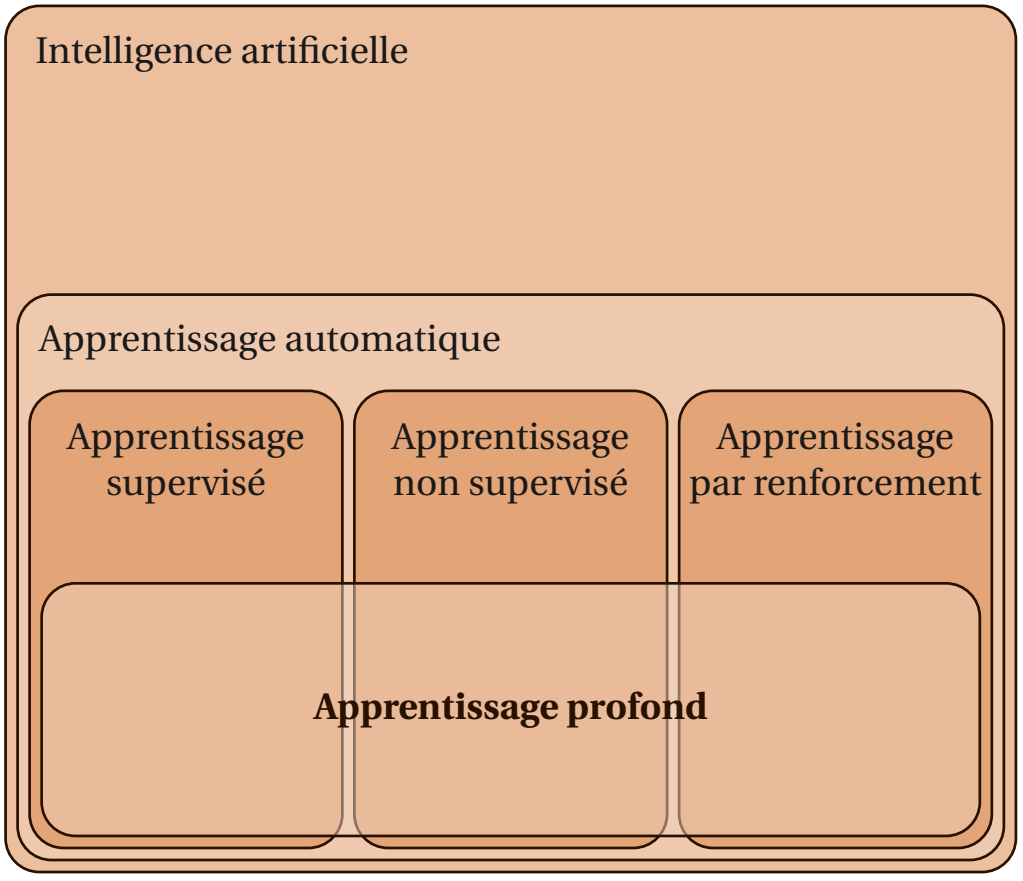
Comment fonctionnent les systèmes d'IA?



Modèles d'apprentissage automatique

- Représente une famille d'équations
- Les paramètres déterminent une équation particulière
- Les données d'entraînement sont utilisées pour ajuster les paramètres
- De bons « paramètres appris » devraient pouvoir se généraliser à des données non observées.

Comment fonctionnent les systèmes d'IA?



Apprentissage profond

- **Toute** fonction continue peut être approchée de façon **arbitrairement précise** par un réseau de neurones comportant au moins une couche cachée.
- L'apprentissage automatique est (en gros) un type de problème d'optimisation
- Ce qui est optimisé, c'est *la fonction objectif*

Le transformeur (modèle auto-attentif)

Attention Is All You Need

Ashish Vaswani[*] Google Brain avaswani@google.com
Noam Shazeer[*] Google Brain noam@google.com
Niki Parmar[*] Google Research nikip@google.com
Jakob Uszkoreit[*] Google Research usz@google.com
Llion Jones[*] Google Research llion@google.com
Aidan N. Gomez[*] [†] University of Toronto aidan@cs.toronto.edu
Łukasz Kaiser[*] Google Brain lukaszkaiser@google.com
Illia Polosukhin[*] [‡] illia.polosukhin@gmail.com

Abstract

The dominant sequence transduction models are based on complex recurrent or convolutional neural networks that include an encoder and a decoder. The best performing models also connect the encoder and decoder through an attention mechanism. We propose a new simple network architecture, the Transformer, based solely on attention mechanisms, dispensing with recurrence and convolutions entirely. Experiments on two machine translation tasks show these models to be superior in quality while being more parallelizable and requiring significantly less time to train. Our model achieves 28.4 BLEU on the WMT 2014 English-to-German translation task, improving over the existing best results, including ensembles, by over 2 BLEU. On the WMT 2014 English-to-French translation task, our model establishes a new single-model state-of-the-art BLEU score of 41.0 after training for 3.5 days on eight GPUs, a small fraction of the training costs of the best models from the literature.

1 Introduction

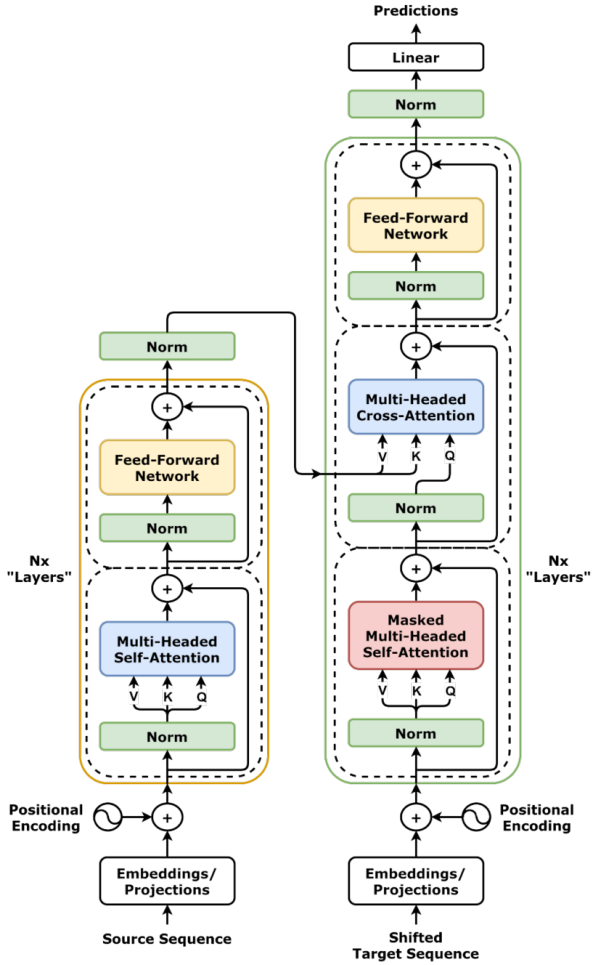
Recurrent neural networks, long short-term memory [12] and gated recurrent [7] neural networks in particular, have been firmly established as state of the art approaches in sequence modeling and transduction problems such as language modeling and machine translation [29, 2, 5]. Numerous efforts have since continued to push the boundaries of recurrent language models and encoder-decoder architectures [31, 21, 13].

[*]Equal contribution. Listing order is random. Jakob proposed replacing RNNs with self-attention and started the effort to evaluate this idea. Ashish, with Illia, designed and implemented the first Transformer models and has been crucially involved in every aspect of this work. Noam proposed scaled dot-product attention, multi-head attention and the parameter-free position representation and became the other person involved in nearly every detail. Niki designed, implemented, tuned and evaluated countless model variants in our original codebase and tensor2tensor. Llion also experimented with novel model variants, was responsible for our initial codebase, and efficient inference and visualizations. Lukasz and Aidan spent countless long days designing various parts of and implementing tensor2tensor, replacing our earlier codebase, greatly improving results and massively accelerating our research.

[†]Work performed while at Google Brain.

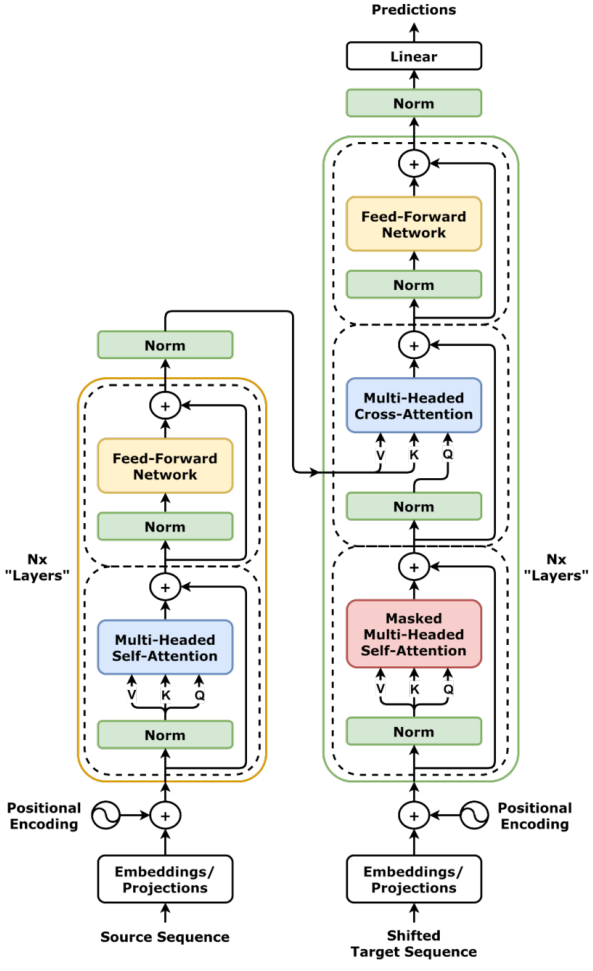
[‡]Work performed while at Google Research.

31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA.



Le transformeur (modèle auto-attentif)

- Ils sont entraînés à prédire **le jeton suivant** dans une séquence.
- Ils apprennent **les modèles statistiques** à partir de vastes corpus textuels.
- Ils génèrent du texte en prédisant itérativement les suites **probables**.



Pourquoi l'affirmation selon laquelle les systèmes d'IA égalent ou surpassent les performances humaines dans les tâches de théorie de l'esprit est-elle **absurde**?

Les problèmes liés aux tâches de théorie de l'esprit

Problèmes empiriques

- De tests multiples de la théorie de l'esprit manquent de **validité convergente**.
- De tests multiples de la théorie de l'esprit manquent de **validité prédictive**.
- Les tests de théorie de l'esprit échouent généralement **à reproduire**.
- La forte sensibilité au contexte et **la dépendance culturelle**.
- Les résultats empiriques incohérents (**les anomalies**).
- Les performances peuvent refléter un style de raisonnement.

Problèmes théoriques*

- La théorie de l'esprit n'est pas **directement mesurable**.
- Les réponses « vérité terrain » **incertaines** ou **inexistantes**.
- La modification « **ad hoc** » des tâches.
- Le risque d'impossibilité de falsification.
- Changement des règles du jeu / redéfinition de la construction.

* Travis LaCroix (2026)
"Autism and the Pseudoscience of Mind"
Psychological Inquiry

L'autisme et la pseudoscience de l'esprit

PSYCHOLOGICAL INQUIRY
2025, VOL. 36, NO. 4, 229–261
https://doi.org/10.1080/1047840X.2025.2605590

Routledge
Taylor & Francis Group

RESEARCH ARTICLE

OPEN ACCESS Check for updates

Autism and the Pseudoscience of Mind

Travis LaCroix

Department of Philosophy, Durham University, Durham, United Kingdom

ABSTRACT

The theory-of-mind-deficit explanation of autism proposes that autistics lack a theory of mind, that autism comprises a theory-of-mind deficit (strong version); or, that autistics often have difficulty with theory-of-mind abilities (weak version). A growing body of critical research demonstrates how these explanations of autistic behavior fail—both empirically and theoretically. The strong version lacks explanatory adequacy, while the weak version is undermined by methodological and empirical flaws in theory-of-mind research. Together, these issues suggest that the "science" of theory of mind in the context of autism is, at best, bad science. Nonetheless, researchers continue to pursue this line of inquiry in autism studies—often moving the goalposts or offering ad hoc rationalizations to preserve the theoretical framework. This article critically examines the theory-of-mind-deficit explanation of autism, focusing particularly on the widely-held view that autistics exhibit difficulties with theory of mind—i.e., the weak version of the theory-of-mind-deficit explanation. Drawing from the philosophy of science, I argue that ongoing adherence to this view exhibits all the hallmarks of a degenerating research programme. Hence, the fact that scientists have not abandoned this hypothesis entails that the research programme is pseudoscientific.

KEYWORDS

Autism; autistic philosophy; demarcation; double empathy; Lakatos; methodological falsification; monotropism; neurodiversity; philosophy of autism; philosophy of science; philosophy on the spectrum; pseudoscience; theory of mind

Pourquoi est-il **dangereux** d'affirmer que les systèmes d'IA égalent ou surpassent les performances humaines dans les tâches liées à la théorie de l'esprit?

Systèmes d'IA anthropomorphisés

Le modèle produit les réponses textuelles attendues aux tâches de théorie de l'esprit
⇒ « le modèle possède une théorie de l'esprit »

- « compréhension »
- « raisonnement »
- « alignement »
- « croyances »
- « pensée »
- « hallucination »

« **Glosslighting** » :
la pratique consistant à utiliser des termes techniquement redéfinis ou polysémiques pour évoquer des significations familières – souvent puissantes sur le plan émotionnel ou cognitif – tout en préservant la possibilité de nier ces significations en se repliant sur des réinterprétations spécialisées et contextuelles.

Le « glosslighting »

Strategic Polysemy in AI Discourse: A Philosophical Analysis of Language, Hype, and Power

TRAVIS LACROIX, Durham University, UK
FINTAN MALLORY, Durham University, UK
SASHA LUCCIONI, Hugging Face, Québec

This paper examines the strategic use of language in contemporary artificial intelligence (AI) discourse, focusing on the widespread adoption of metaphorical or colloquial terms like "hallucination", "chain-of-thought", "introspection", "language model", "alignment", and "agent". We argue that many such terms exhibit *strategic polysemy*: they sustain multiple interpretations simultaneously, combining narrow technical definitions with broader anthropomorphic or common-sense associations. In contemporary AI research and deployment contexts, this semantic flexibility produces significant institutional and discursive effects, shaping how AI systems are understood by researchers, policymakers, funders, and the public. To analyse this phenomenon, we introduce the concept of *glosslighting*: the practice of using technically redefined terms to evoke intuitive—often anthropomorphic or misleading—associations while preserving plausible deniability through restricted technical definitions. Glosslighting enables actors to benefit from the persuasive force of familiar language while maintaining the ability to retreat to narrower definitions when challenged. We argue that this practice contributes to AI hype cycles, facilitates the mobilisation of investment and institutional support, and influences public and policy perceptions of AI systems, while often deflecting epistemic and ethical scrutiny. By examining the linguistic dynamics of glosslighting and strategic polysemy, the paper highlights how language itself functions as a sociotechnical mechanism shaping the development and governance of AI.

CCS Concepts: • **Computing methodologies** → **Philosophical/theoretical foundations of artificial intelligence**; **Machine learning**; • **Applied computing** → *Sociology*; *Economics*; *Psychology*.

Additional Key Words and Phrases: artificial intelligence (AI); philosophy of language; polysemy; anthropomorphism; scientific rhetoric; AI hype cycles; plausible deniability; strategic ambiguity; AI ethics; scientific communication; the spectre of capitalism

ACM Reference Format:
Travis LaCroix, Fintan Mallory, and Sasha Luccioni. 2026. Strategic Polysemy in AI Discourse: A Philosophical Analysis of Language, Hype, and Power. In *The 2026 ACM Conference on Fairness, Accountability, and Transparency (FAccT '26), June 25–28, 2026, Montreal, QC, Canada.* ACM, New York, NY, USA, 21 pages. https://doi.org/10.1145/3805689.3812399

Si les autistes n'ont pas de théorie de l'esprit, alors …

- …Les autistes n'ont pas de théorie de leur propre esprit (conscience de soi).
  - …Les autistes manquent d'autorité épistémique (à la première personne).
  - …Les autistes sont dépourvues de la propriété de la *personnalité morale*
    - …Les autistes ne possèdent pas la totalité des *droits moraux*.
    - …Les autistes ne font pas partie de la communauté morale.
- …si la théorie de l'esprit est un sous-ensemble de l'empathie…
  - …Alors, les autistes manquent alors d'empathie.
    - …alors les autistes ne peuvent pas vivre « the good life ».
- …si la théorie de l'esprit est l'une des capacités essentielles qui nous rendent humains …
  - …Les autistes ne sont pas (entièrement) humains.
- …Une communauté de personnes autistes est impossible.*

* Barnbaum (2008)
The Ethics of Autism

Les mesures et le problème du « proxy »

Performance des tâches → mesures du proxy ⇝ « théorie de l'esprit »

- La théorie de l'esprit est une construction latente.
- Les indices de référence sont des indicateurs indirects.
- Succès prédictif ≠ adéquation de la représentation.
- La réussite d'un LLM reflète la coordination avec les données d'entraînement, et non la compréhension.

Principaux points à retenir

- Les grands modèles de langage ne démontrent pas la capacité de « lire dans les pensées ».
- Les tâches de théorie de l'esprit sont des mesures indirectes.
- Ces critères d'évaluation sont biaisés.
- La véritable leçon concerne nos tests, et non l'intelligence artificielle.
- La compréhension est relationnelle et située.

Compliments constructifs: travis.lacroix@durham.ac.uk
Plus d'informations: travis.lacroix@github.io

Merci à tous

[*Universal Language* (2024) – réalisé par Matthew Rankin]