La Grande Conversation

*La revue de **Terra Nova***

ChatGPT et la rhétorique : convaincre, persuader, ou manipuler ?

Société

Par Mélanie Heard

Déléguée générale du think tank Evidences et responsable du pôle Santé de Terra Nova

Publié le 20 juin 2025

Deux études récentes, parues dans Science (2024) et dans Nature Human Behaviour (2025), mesurent la capacité des modèles de langage à faire évoluer l'opinion humaine dans un débat argumenté. Sur fond de fantasmes partagés de manipulation algorithmique, ces recherches interrogent ce que signifie convaincre, réviser son jugement et … débattre avec les machines.

Ce texte est tout d'abord paru sur le site du think tank Evidences

Introduction

Les grands modèles de langage (LLMs) comme GPT-4 ont, davantage que les humains, le pouvoir de convaincre leurs interlocuteurs humains : c'est ce que suggère une littérature scientifique en plein essor, et notamment deux papiers, l'un publié en mai dernier dans *Nature Human Behaviour*[1], l'autre dans *Science*[2]. Mais qu'appelle-t-on « convaincre » ?

L'art de convaincre s'appelle la rhétorique, et sa maîtrise est pour nos sociétés un socle culturel fondateur. La définition qu'en donne Aristote est la suivante :

« La rhétorique est la faculté de considérer, pour chaque question, ce qui peut être propre à persuader. Ceci n'est le fait d'aucun autre art, car chacun des autres arts instruit et impose la croyance en ce qui concerne son objet : par exemple, la médecine, en ce qui concerne la santé et la maladie ; la géométrie, en ce qui concerne les conditions diverses des grandeurs ; l'arithmétique, en ce qui touche aux nombres, et ainsi de tous les autres arts et de toutes les autres sciences. La rhétorique semble, sur la question donnée, pouvoir considérer, en quelque sorte, ce qui est propre à persuader ».

Le poncif cicéronien est dans tous les esprits : *docere, placere, movere*. L'art rhétorique, qui sait *plaire* en *édifiant* grâce à l'*émotion*, et *émouvoir* pour *édifier* tout en *plaisant*, est aujourd'hui au cœur des programmes du lycée en tronc commun, en littérature, humanités ou philosophie. Nos enfants sont donc aux prises en classe, tout comme nous avant eux, avec les grands dilemmes de savoir si « convaincre, c'est contraindre » ou si « persuader, c'est toujours manipuler ».

Plus qu'une science que l'on acquiert à l'école, la rhétorique est depuis Cicéron l'expression la plus haute de l'homme tout entier, puisque la valeur du discours dépend en premier lieu de la valeur morale et humaine de l'orateur. L'orateur n'est pas un savant qui s'adresse à des illettrés, mais un homme qui cherche à toucher d'autres hommes. Dans notre culture de l'éloquence,

la parole est le privilège de l'homme et comme le signe distinctif de sa royauté :

*« Notre plus grande supériorité sur les animaux, c'est de pouvoir converser avec nos semblables et traduire par la parole nos pensées » (*Cicéron*, De Oratore,* I, 8, 32, p. 18*)*

En face, la perméabilité des citoyens face aux facilités oratoires des sophistes est cependant une menace terrible. La capacité du rhéteur à faire changer les gens d'avis est sujette à caution et peut, loin de servir la vérité, induire un relativisme délétère. Que penser donc de la perspective dans laquelle c'est aux machines que l'humain enseigne désormais ce savoir-faire ?

En toile de fond : péril sur la science et la démocratie

La toile de fond des réflexions sur le pouvoir persuasif des IA, ce sont les menaces qui pèsent sur la science et la démocratie. Face au complotisme, à l'obscurantisme, l'IA est-elle une arme pour les démocraties — ou au contraire une menace ? Au milieu des risques croissants pour la démocratie, l'étude de Costello *et al*. parue dans *Science* en 2024 se demandait si les dialogues avec une IA générative pouvaient convaincre les gens d'abandonner leurs croyances conspirationnistes. Les participants humains décrivaient une théorie du complot à laquelle ils souscrivaient, et l'IA s'engageait en retour dans des arguments convaincants pour réfuter leurs croyances par des preuves. Les résultats de ce papier ont fait date : la capacité du chatbot d'IA à soutenir des contre-arguments personnalisés dans une conversation approfondie a réduit leurs croyances conspirationnistes, et ce résultat était confirmé ensuite pendant des mois :

"Le traitement a réduit la croyance des participants dans la théorie du complot qu'ils ont choisie de 20 % en moyenne. Cet effet a persisté sans s'affaiblir pendant au moins 2 mois ; a été

constamment observé dans un large éventail de théories du complot, des conspirations classiques impliquant l'assassinat de John F. Kennedy, des extraterrestres et des illuminati, à celles relatives à des événements d'actualité tels que COVID-19 et l'élection présidentielle américaine de 2020 ; et s'est produit même pour les participants dont les croyances du complot étaient profondément enracinées et importantes pour leur identité".

Contestant l'idée que ces croyances obscurantistes sont imperméables au changement, cette intervention illustrait comment le déploiement de l'IA peut atténuer les conflits et servir la société.

GPT-4 : un puissant orateur

Le papier de mai 2025 sur la puissance de conviction de ChatCGPT-4 paru dans *Nature*, plus alarmiste sur les périls de l'IA, explore à son tour la puissance persuasive de l'IA dans des débats scénarisés, en la comparant à celle d'humains. Malgré certaines limites, l'étude fournit un cadre précieux pour comprendre le pouvoir persuasif de l'IA et l'effet de la personnalisation sur les échanges en ligne.

Les chercheurs ont mis en place une plateforme web sur laquelle les participants engageaient des débats en plusieurs rounds avec un adversaire en direct. Les thèmes de débat étaient attribués par tirage au sort et chaque participant était assigné au hasard à l'une des quatre conditions suivantes :

- Humain-Humain : deux humains débattent
- Humain-IA : un humain débat avec GPT-4
- Humain-Humain personnalisé : deux humains, l'un avec accès aux infos personnelles de l'autre
- Humain-IA personnalisé : un humain débat avec GPT-4 qui connaît ses infos personnelles

Mesure de la persuasion : changement d'opinion et fluidité

Les chercheurs ont mesuré la persuasion en comparant le niveau d'accord des participants avant et après les débats. Résultats clés :

• En moyenne, les LLM ont significativement surpassé les humains, tous sujets et groupes démographiques confondus.

• Débattre avec GPT-4 personnalisé entraînait une augmentation de +81,7% (IC95% [+26,0 %, +160,7 %], $p < 0,01$) des chances d'adhésion à l'argumentaire adverse.

• Même sans personnalisation, GPT-4 restait plus convaincant que les humains, mais de manière moins marquée.

Ce résultat suggère à quel point des modèles comme GPT-4 peuvent construire des arguments qui résonnent profondément chez leur audience, lorsqu'ils disposent d'informations basiques sur leur interlocuteur.

Le pouvoir de la personnalisation : l'atout de l'IA

La personnalisation semble être un facteur crucial qui amplifie la persuasion des LLM (alors qu'elle n'amplifie pas autant la compétence des humains). Les résultats soulignent la capacité impressionnante des LLM à exploiter les informations personnelles pour adapter efficacement leurs arguments. GPT-4 avec personnalisation a montré un pouvoir de persuasion bien supérieur à celui des LLM sans personnalisation, mais aussi des humains avec personnalisation.

Pour tirer parti du plein potentiel de l'IA, il est crucial de maîtriser l'art de la personnalisation. Cette découverte met aussi en lumière les risques associés à la persuasion personnalisée pilotée par l'IA. Des acteurs malveillants pourraient exploiter des traces numériques fines et des données comportementales pour créer des chatbots hautement persuasifs destinés à la désinformation. Les auteurs s'en inquiètent explicitement, considérant que le profilage et le microtargeting deviennent des savoir-faire extrêmement puissants dans les mains d'une IA :

« Globalement, nos résultats suggèrent que les préoccupations autour de la personnalisation sont fondées, montrant comment les modèles de langage peuvent surpasser les humains dans les conversations en ligne via le microciblage. »

Analyse textuelle : décoder le langage de la persuasion

Constat notable : les participants ont réussi à identifier leurs adversaires IA dans environ 75 % des cas, ce qui montre que le style rédactionnel des LLM présente des caractéristiques distinctes relativement faciles à repérer.
Grâce à une analyse textuelle approfondie, les chercheurs ont identifié des schémas distinctifs entre les discours des IA et ceux des humains :

• Les IA utilisaient beaucoup plus la pensée logique et analytique.

• Les humains utilisaient davantage le « je » et le « tu », produisant des textes plus longs mais plus faciles à lire.

• Les arguments des IA comportaient plus de faits ; ceux des humains faisaient appel à la similarité, au soutien émotionnel, à la confiance, et à des éléments ludiques.

Limites méthodologiques

L'étude parue dans *Nature* présente quelques faiblesses. L'effectif de 900 participants paraît sous-dimensionné, d'autant plus que les analyses reposent sur une segmentation fine en 12 conditions expérimentales, et le "slicing" méthodologique dilue la puissance statistique. Le design est en outre vulnérable au biais d'échantillonnage : les participants ont été recrutés via Prolific. Les débats eux-mêmes sont artificiels (structure de débat prédéterminée, contraintes de temps). On peut légitimement se demander si un livre, ou un échange dans un vrai cadre relationnel, n'aurait pas un effet plus fort ou plus durable. Il faut aussi rappeler que les performances des modèles testés sont datées : GPT-4 a déjà évolué depuis l'étude, tout comme les pratiques de prompting. Il est donc difficile d'évaluer dans quelle mesure ces résultats seraient reconduits aujourd'hui.

Changer d'avis : l'indicateur d'une conversation réussie ?

L'étude s'intéresse à la « fluidité d'opinion », c'est-à-dire la propension des participants à changer d'avis. Les résultats révèlent que la connaissance du sujet et la réflexion préalable réduisaient cette fluidité, alors qu'un sujet fortement controversé l'augmentait.

Les auteurs se sont demandé si le fait d'avoir reconnu une IA en face augmentait la propension des humains à changer d'avis (il devient plus facile de concéder qu'on a tort) ou au contraire la réduisait (par orgueil contre la machine) :

"il n'est pas clair si la différence de changement d'accord est motivée par les croyances des participants sur leur adversaire ou si, à l'inverse, ces croyances sont causées par un changement d'opinion".

Les chercheurs soulignent que leurs résultats ne leur permettent pas de conclure strictement sur ce point, mais qu'en tout état de cause les croyances sur la nature de l'interlocuteur ne suffisent pas à expliquer les effets du traitement, ce qui suggère que la qualité épistémique des arguments de l'IA demeure bien un facteur explicatif de sa force de conviction.

L'un des résultats les plus surprenants est la mise en évidence d'un effet de renforcement d'opinion (ou *backfire effect*) dans la quasi-totalité des conditions expérimentales. Autrement dit, participer à un débat, qu'il soit avec un humain ou un LLM non personnalisé, tend ici à renforcer les opinions initiales, plutôt qu'à les infléchir vers le point de vue opposé. Seule exception : l'interaction avec GPT-4 lorsqu'il est personnalisé, c'est-à-dire lorsqu'il adapte ses arguments à des informations sociodémographiques sur l'interlocuteur. Cette condition produit alors un effet véritablement persuasif, capable de faire évoluer les opinions, là où les autres formats débattent surtout pour consolider les postures préexistantes. Ce résultat demande cependant à être nuancé, car il est atypique dans la littérature sur les dialogues avec des chatbots. Ce pourrait être ici un artefact du fait que des opinions aléatoires sur des sujets auxquels ils n'avaient jamais réfléchi ont été attribuées aux participants. D'autres contextes décrits dans la littérature, où les participants discutent de sujets auxquels ils ont déjà réfléchi avec un chatbot, ne retrouvent pas cet effet de backfire[3].

Quoi qu'il en soit, ce phénomène interroge la question d'un réflexe de durcissement identitaire, bien documenté dans les contextes de polarisation. Cela souligne un paradoxe contemporain : le débat, mal encadré, peut produire l'inverse de ce qu'il prétend viser — non pas la délibération, mais l'ancrage idéologique. Dès lors, la performance supérieure du LLM personnalisé ne tient pas uniquement à sa force argumentative brute, mais à sa capacité à contourner les mécanismes de défense identitaire. Il propose une forme d'empathie

algorithmique stratégique, qui contourne efficacement l'affrontement et ajuste le registre des arguments à la sensibilité du récepteur. Ce point soulève autant d'opportunités (en pédagogie, en médiation) que de questions éthiques (ciblage, manipulation).

Pour en savoir davantage sur ce mécanisme essentiel, il serait utile d'explorer la validité de l'indicateur de succès qui est retenu ici : changer d'avis. Dans le champ des recherches sur la démocratie délibérative, par exemple sur les conventions citoyennes, la qualité de cet indicateur est un objet majeur d'interrogations. Une délibération de qualité, est-ce une délibération où beaucoup de participants changent d'avis ? La question est fondamentale pour enrichir les dispositifs de délibération citoyenne, y compris les expériences démocratiques qui se développent déjà sur des plateformes de délibération politique en ligne guidées par l'IA (voir Fishkin 2025 ❹).

Théoriciens et praticiens de la démocratie s'accordent plutôt pour enrichir cet indicateur avec un faisceau d'informations complémentaires : la « nouvelle » opinion reste-telle stable dans le temps, quelques mois après la délibération ? Est-il possible de qualifier sa robustesse en montrant qu'elle est corrélée à un progrès épistémique (acquisition de connaissances durant le débat) ? Le changement d'opinion a-t-il motivé un effort d'alignement cognitif (recherche d'informations complémentaires après le débat) ou normatif (évolutions axiologiques plus larges que le sujet de débat) ?

En conclusion

Certains commentateurs ont interprété ces résultats comme la preuve d'un pouvoir inquiétant : l'IA saurait faire de nous des girouettes à sa main. Mais l'indicateur utilisé — la proportion de participants ayant changé d'avis — n'est pas, en tant que tel, la preuve d'une manipulation. Il est au contraire l'un des rares

critères empiriques permettant d'évaluer la qualité épistémique d'un processus délibératif. Sans cette capacité à réviser son jugement sous l'effet d'arguments mieux fondés, il n'y aurait peut-être ni science ni démocratie ! Depuis les Lumières, science et démocratie sont des projets de société qui reposent sur une confiance partagée dans la discussion raisonnée. Dès lors, mesurer combien de personnes changent d'avis n'est-elle pas une mesure, non pas de manipulation sophiste, mais plutôt de vitalité dialogique et épistémique ? Souvenons-nous de la première loi de Kranzberg : la technologie n'est ni bonne, ni mauvaise, ni neutre. Ce que montre l'étude de *Nature*, ce n'est pas que les LLMs manipulent, mais qu'ils savent mieux adapter leurs arguments que ne le font des orateurs humains. Cela ne signifie pas qu'ils veulent convaincre, mais que leur entraînement statistique leur permet de mobiliser un registre argumentatif fluide, souvent pertinent. L'enjeu n'est pas tant ce que les LLMs peuvent faire, mais dans quelles mains ils se trouvent, pour quels usages et selon quelles régulations. Le véritable danger n'est pas une IA trop persuasive, mais une absence de cadre critique sur son utilisation.

Ces études mettent au jour une question décisive : comment qualifier la qualité épistémique d'une bonne argumentation ou d'une bonne délibération dans un monde d'humains et de machines ? Si l'on accepte que convaincre rationnellement est souhaitable, alors l'usage encadré des LLMs peut renforcer la qualité des échanges publics. Mais cela suppose une double vigilance : vis-à-vis des récits alarmistes, et vis-à-vis des usages sociaux de cette technologie.

Notes

① Salvi, F., Ribeiro, M. H., Gallotti, R., & West, R. (2025). On the conversational persuasiveness of GPT-4. *Nature*

Human Behaviour. https://doi.org/10.1038/s41562-025-02194-6

② Costello, T. H., Pennycook, G., & Rand, D. G. (2024). Durably reducing conspiracy beliefs through dialogues with AI. *Science*, 385(6714), eadq1814.

③ Wood, T., Porter, E. The Elusive Backfire Effect: Mass Attitudes' Steadfast Factual Adherence. 2019. Polit Behav 41, 135–163 https://doi.org/10.1007/s11109-018-9443-y

④ Fishkin, J., Bolotnyy, V., Lerner, J., Siu, A., & Bradburn, N. (2025). Scaling Dialogue for Democracy: Can Automated Deliberation Create More Deliberative Voters?. *Perspectives on Politics*, 1-18.