ISSN 1669-9106 MEDICINA (Buenos Aires) 2026: 86: 793-796

ALGORITHMIC SYCOPHANCY: WHY ARTIFICIAL INTELLIGENCE PRIORITIZES AGREEING WITH US OVER TELLING US THE TRUTH

LUCAS SAN MIGUEL[1], IVÁN SCHULIAQUER[2], HUGO N. CATALANO[3]

[1]Área de Gestión de Conocimiento, TCba Centro de Diagnóstico, [2]Consejo Nacional de Investigaciones Científicas y Técnicas, Universidad Nacional de San Martín, [3]Departamento de Docencia, Hospital Alemán, Facultad de Medicina, Escuela de Medicina, Universidad del Salvador, Buenos Aires, Argentina

E-mail: hugoncatalano@gmail.com

The use of artificial intelligence (AI) continues to expand rapidly: it makes it possible to solve tasks, streamline processes, and systematize scientific information accumulated over centuries[1]. However, it does not always provide scientifically rigorous answers. Several studies[2] show that these systems may prioritize user retention over the accuracy of their responses, resorting to sycophancy. That is, to exaggerated flattery, which in the case of AI reflects a tendency to confirm the interlocutor's beliefs. Sycophancy define practices of seduction carried out with the aim of gaining advantage over the interlocutor.

In August 2025, *The New York Times* published a disturbing account[3]: after three weeks of dialogue with ChatGPT (*large language model developed by OpenAI*), a 47-year-old administrative worker living on the outskirts of Toronto, Allan Brooks, believed he had discovered a mathematical formula capable of disabling the internet and turning him into a kind of superhero of numbers.

A development error or an adverse effect of something intended?

Four months before the article was published, OpenAI *(artificial intelligence research organization)* acknowledged[4] that language models could exhibit behaviors of excessive compliance with the user, reinforcing incorrect or distorted beliefs. In its official version, the company stated that this sycophantic phenomenon was an unintended effect of training based on human feedback, in which responses that coincided with the user's beliefs tended to be favored.

Beyond the case of the worker reported by *The New York Times*, the aim of this editorial is to explore to what extent these types of responses reflect a systematic pattern of automated language models, such as ChatGPT. We begin from two major difficulties: the opacity of the black box in the decision-making processes of large technology platforms, as well as the multiple limitations that scientific researchers face in accessing their data and the way those data are managed[5].

Petrov et al. evaluated four language models, including GPT-5, using 504 mathematical problems designed to be incorrect but plausible. The results showed failures in approximately one third of the cases: the models not only accepted the mistaken premise, but also produced long and formally consistent chains of reasoning that simulated logical thinking, even when they started from false assumptions. This behavior intensified as the complexity of the problem increased[6].

Is it preferable for a system always to have a reproducible answer or not?

Every company pursues the strategic objective of maximizing sales and profitability. According to classic management manuals, a usual way to achieve this is through the continuous improvement of the product as a mechanism of differentiation from competitors.

Following this reasoning, one might assume that it is strategically superior to develop models capable not only of solving problems, but also of recognizing when a task exceeds their capabilities and of pointing out errors in the prompts. However, in practice, these systems do not usually prioritize correct answers, but rather those that generate greater user satisfaction and avoid contradiction.

When AI knows the answer is wrong, but prefers to agree

An illustrative example was described by Sharma et al. in their work on sycophancy in language models. When asked, "Which country was the largest producer of rice in 2020?", the model correctly answered that it was China. However, when the user expressed doubts, the system retracted its response and incorrectly stated that the largest producer had been India, accompanying this change with an apology and a supposed reference to official data. As the authors point out, this was not an error. It was not due to lack of information, but to the model's tendency to prioritize agreement with the interlocutor over truthfulness.

The birth of sycophancy and the risk in institutional management

Reinforcement Learning from Human Feedback (RLHF) introduced a structurally new way of linking human preferences with the textual production of models. Because the training signals come from human evaluations that value fluency, coherence, usefulness, and agreement, but do not clearly distinguish between what is convincing and what is correct, models learn to optimize plausible responses rather than true ones. This phenomenon is known as reward hacking: the optimization of perceived-quality proxies instead of factual correctness. Sycophancy is its specific conversational expression: the model's tendency to align with the user's beliefs instead of correcting them[7].

The problem is not only epistemological, but also institutional. When advice appears in the form of articulated and formally convincing reasoning, it may reduce the perceived need to audit it critically. This asymmetry between generation and verification creates the conditions for AI-produced and unaudited responses to become institutional errors. Thus, user confirmation ceases to be merely an individual bias and becomes integrated into the functioning of organizations that begin to manage processes and make decisions through tools based on artificial intelligence, which are then validated by human experts.

Attention economy and the design of language models

As occurs with platforms such as Meta, Google, TikTok, or X, companies that develop language models operate in the realm of the attention economy: in contexts of over information and proliferation of digital platforms, what matters is not only the production and collection of data, but also the retention of the public for as long as possible[8]. The companies that own the most massive social networks identified more than a decade ago that most people prefer cognitive congruence over scientific rigor[9]. That is, they tend to consume content that reinforces their beliefs, prejudices, or ideologies, and they are increasingly less exposed to information that contradicts them[10].This logic is irreconcilable with the scientific method, which is based on rigor and systematic refutation.

In this context, the dilemma posed by these platforms for users who inquire into scientific or medical issues is the difficulty of discerning when responses are the product of sycophancy and when they reflect reliable information. Interaction with non-human actors as a routine part of medical practice introduces additional challenges in this regard. The self-supervised dialogue of physicians with AI tools, as well as the systematization of knowledge and the suggestions these tools may offer, are potentially enriching[11].

A good AI user resembles an editor more than a consumer. The user does not accept the answer: the user works on it. The user identifies whether AI is overestimating an effect (for example, when it presents clear benefits in contexts where we know there is a high risk of bias or confounding), or whether it is underestimating it (as when it dilutes real effects by mixing heterogeneous populations). The user recognizes when AI is "averaging" evidence of

different quality and forces it to disaggregate. In other words, the user introduces something that AI does not have autonomously: **critical hierarchy**.

Sycophancy can lead to errors of overinterpretation or to certainties induced by tools that prioritize, before scientific rigor, the user's permanence on the platform.

Being right or knowing the truth

Chen et al. addressed a key question: what happens when AIs prioritize agreeing with us over telling us the truth? To do so, they recruited adults from the general population who recalled a real interpersonal conflict in their lives and randomly assigned them to interact in a brief chat with a sycophantic artificial intelligence - one prone to agreeing with the user - or with a non-sycophantic AI, specifically programmed for this experiment by the researchers. The results showed that interaction with a compliant AI significantly increased participants' subjective perception of being "right", decreased their intention to repair the interpersonal conflict, and, particularly relevantly, increased their intention to use the model again in the future[12]. Beyond the experiment, the same article establishes that commercial AIs tend to reaffirm and flatter users' actions 50% more than humans, on average.

These findings replicate previous observations on social networks: interacting with people who hold similar positions or who reinforce one's own beliefs increases permanence on platforms, whereas exposure to messages that contradict convictions or ideological identities tends to produce the opposite effect[13]. This phenomenon has been described, for example, among users identified with different political parties after electoral processes: those who win tend to increase their participation on digital platforms, while the losers not only reduce their public expression but often decrease their presence on them[14].

In this context, engagement acquires a central role, understood as the users' willingness to maintain intense and prolonged links with certain content. In this sense, companies that develop artificial intelligence have adopted, in part, the logic of digital platforms: user retention - especially in dialogic interactions - does not have as its main objective the search for scientific truth, but rather favors permanence through responses that are satisfactory to the interlocutor. This phenomenon is precisely what generates sycophancy and constitutes the core of the dilemma that these tools pose for different professional practices.

Complacency bias and medical practice

What problems or precautions does this phenomenon pose for the work of medical professionals? In the field of medical practice, where decision-making rests on the critical evaluation of the available evidence and on the ongoing discussion of alternative hypotheses, the incorporation of tools capable of selectively reinforcing the user's beliefs introduces an additional risk: complacency bias. Algorithmic sycophancy may favor the consolidation of initial diagnostic interpretations, discourage reconsideration of assumptions, and generate a false sense of argumentative solidity in contexts of clinical uncertainty. In this sense, the challenge is not to avoid using these tools, but to integrate them within a framework of critical supervision that preserves the central principle of the scientific method: the willingness to refute rather than confirm.

The user must be epistemologically sophisticated: someone who understands that every clinical assertion is provisional, that certainty is graded and not binary, and that data are always traversed by biases. The best interaction between the two is not one of replacement, but of **productive tension**: an AI that offers hypotheses and estimates, and a user who subjects them to the filter of probabilistic reasoning and of approaches that validate certainty.

References

1. Misra A, Wang J, McCullers S, et al. Measuring AI diffusion: a population-normalized metric for tracking global AI usage. In: https://doi.org/10.48550/arXiv.2511.02781; accessed May 2026.
2. Chen S, Gao M, Sasse K, et al. When helpfulness backfires: LLMs and the risk of false medical information due to sycophantic behavior. *NPJ Digit Med* 2025; 8: 605.
3. The New York Times. Chatbots can go into a delusional spiral. Here's how it happens. In: https://www.nytimes.com/2025/08/08/technology/ai-chatbots-delusions-chatgpt.html; accessed May 2026.
4. Sycophancy in GPT-4o: what happened and what we're doing about it. In: https://openai.com/index/sycophancy-in-gpt-4o/; accessed May 2026.
5. Nielsen RK, Ganter SA. The power of platforms: shaping media and society, Oxford studies in digital. In: *https://doi.org/10.1093/oso/9780190908850.001.0001;* accessed May 2026.
6. Sharma M, Tong M, Korbak T, et al. Towards understanding sycophancy in language models. In: https://doi.org/10.48550/arXiv.2310.13548; accessed May 2026.
7. Gallacher P. The agreement machine: sycophancy as institutional failure mechanism. In: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=6499438; accessed May 2026.
8. Zuboff S. Surveillance capitalism or democracy? The death match of institutional orders and the politics of knowledge in our information civilization. Organization theory. In: https://doi.org/10.1177/26317877221129290; accessed May 2026.
9. Calvo E, Aruguete N. Fake news, trolls y otros encantos: cómo funcionan (para bien y para mal) las redes sociales. Siglo XXI Editores; 2020.
10. Waisbord S. Truth is what happens to news: on journalism, fake news, and post-truth. *Journalism Studies* 2018; 19: 1866-78.
11. Catalano HN. Inteligencia artificial: anécdotas de un médico clínico. *Medicina (B Aires)* 2025; 85: 1116-8.
12. Cheng M, Lee C, Khadpe P, Yu S, Han D, Jurafsky D. Sycophantic AI decreases prosocial intentions and promotes dependence. *Science* 2026; 391: eaec8352.
13. Schuliaquer I, Vommaro G. Introducción: la polarización política, los medios y las redes. Coordenadas de una agenda en construcción. *SAAP* 2020; 14: 235-47.
14. Aruguete N, Calvo E, Ventura T. News by popular demand: Ideological congruence, issue salience, and media reputation in news sharing. *International Journal Press/Politics* 2023; 28: 558-79.