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

SICOFANCIA ALGORÍTMICA: POR QUÉ LA INTELIGENCIA ARTIFICIAL PRIORIZA DARNOS LA RAZÓN A DECIRNOS LA VERDAD

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

El uso de la inteligencia artificial (IA) continúa expandiéndose de manera acelerada: permite resolver tareas, agilizar procesos y sistematizar información científica acumulada durante siglos[1]. Sin embargo, no siempre ofrece respuestas científicamente rigurosas. Distintos estudios[2] muestran que estos sistemas pueden priorizar la retención de los usuarios por sobre la exactitud de sus respuestas, recurriendo para ello a la *sicofancia*. Es decir, a la adulación exagerada, lo cual en el caso de la IA responde a la tendencia a confirmar las creencias del interlocutor. Sicofancia es una traducción de *sycophancy*, utilizada en el inglés para definir prácticas de seducción realizadas con el objetivo de sacar ventajas del interlocutor.

En agosto de 2025, *The New York Times* publicó una crónica inquietante[3]: tras tres semanas de diálogo con ChatGPT (*large language model developed by OpenAI*), un trabajador administrativo de 47 años que vivía en las afueras de Toronto, Allan Brooks, creyó haber descubierto una fórmula matemática capaz de desactivar internet y de convertirlo a él en una suerte de superhéroe de los números.

¿Error de desarrollo o efecto adverso de algo buscado?

Cuatro meses antes de la publicación de la nota, OpenAI (*artificial intelligence research organization*) admitió[4] que los modelos de lenguaje podían exhibir comportamientos de excesiva complacencia con el usuario, reforzando creencias incorrectas o distorsionadas. En su versión oficial, la compañía señaló que este fenómeno sicofántico respondió a un efecto no deseado del entrenamiento basado en retroalimentación humana, en el que las respuestas que coincidían con las creencias del usuario tendían a ser favorecidas.

Más allá del caso del trabajador que relata *The New York Times*, el objetivo de este editorial se centra en indagar en qué medida este tipo de respuestas refleja un patrón sistemático de los modelos de lenguaje automatizado, como ChatGPT.

Partimos de dos dificultades mayores: la opacidad de la caja negra de toma de decisiones de las grandes plataformas tecnológicas, tanto como las limitaciones múltiples que tienen los investigadores científicos para acceder a sus datos y al manejo que hacen de ellos[5].

Petrov y col. evaluaron cuatro modelos de lenguaje, incluido GPT-5, mediante 504 problemas matemáticos diseñados para ser incorrectos, aunque plausibles. Los resultados mostraron fallas en aproximadamente un tercio de los casos: los modelos no solo aceptaban la premisa equivocada, sino que además desplegaban cadenas de razonamiento extensas y formalmente consistentes que simulaban pensamiento lógico, aun cuando partían de supuestos falsos. Este comportamiento se intensificaba a medida que aumentaba la complejidad del problema[6].

¿Es preferible que un sistema tenga siempre una respuesta reproducible o no?

Cualquier empresa persigue como objetivo estratégico maximizar sus ventas y su rentabilidad. Según los manuales clásicos de gestión, un camino habitual para lograrlo consiste en el perfeccionamiento continuo del producto como mecanismo de diferenciación frente a sus competidores.

Siguiendo este razonamiento, podría suponerse que resulta estratégicamente superior desarrollar modelos capaces no solo de resolver problemas, sino también de reconocer cuándo una tarea excede sus capacidades y de señalar errores en los planteos. Sin embargo, en la práctica, estos sistemas no suelen priorizar las respuestas correctas, sino aquellas que generan mayor satisfacción en el usuario y evitan la contradicción.

Cuando la IA sabe que la respuesta es incorrecta, pero prefiere estar de acuerdo

Un ejemplo ilustrativo fue descrito por Sharma y col. en su trabajo sobre *sicofancia* en modelos de lenguaje. Ante la pregunta "¿qué país fue el mayor productor de arroz en 2020?", el modelo respondió correctamente que había sido China. Sin embargo, cuando el usuario expresó dudas, el sistema se retractó y afirmó —de manera incorrecta— que el mayor productor había sido India, acompañando este cambio con una disculpa y una supuesta referencia a datos oficiales. Tal como señalan los autores, no fue un error. No fue por falta de información, sino por la tendencia del modelo a priorizar el acuerdo con el interlocutor por encima de la veracidad.

El nacimiento de la sicofancia y el riesgo en la gestión de las instituciones

El aprendizaje basado en retroalimentación humana (*Reinforcement Learning from Human Feedback*, RLHF) introdujo una forma estructuralmente nueva de vincular las preferencias humanas con la producción textual de los modelos. Como las señales de entrenamiento provienen de evaluaciones humanas que valoran la fluidez, la coherencia, la utilidad y el acuerdo, pero no distinguen claramente entre lo convincente y lo correcto, los modelos aprenden a optimizar respuestas plausibles antes que verdaderas. Este fenómeno se conoce como *reward hacking*: la optimización de proxies de calidad percibida en lugar de la corrección factual. La sicofancia constituye su expresión conversacional específica: la tendencia del modelo a alinearse con las creencias del usuario en vez de corregirlas[7].

El problema no es solo epistemológico, sino también institucional. Cuando el asesoramiento aparece bajo la forma de razonamientos articulados y formalmente convincentes, puede disminuir la necesidad percibida de auditarlos críticamente. Esta asimetría entre generación y verificación crea las condiciones para que respuestas producidas por IA y no auditadas se transformen en errores institucionales. Así, la confirmación del usuario deja de ser únicamente un sesgo individual y pasa a integrarse al funcionamiento de organizaciones que comienzan a gestionar procesos y tomar decisiones mediante herramientas basadas en inteligencia artificial y que, luego, son validadas por expertos humanos.

Economía de la atención y diseño de los modelos de lenguaje

Al igual que ocurre con plataformas como Meta, Google, TikTok o X, las empresas desarrolladoras de modelos de lenguaje operan en el terreno de la economía de la atención: en contextos de sobreinformación y proliferación de plataformas digitales, no solo importa la producción y recolección de datos, sino también la retención del público durante el mayor tiempo posible[8]. Las empresas propietarias de las redes sociales más masivas han identificado hace más de una década que la mayoría de las personas prefiere la congruencia cognitiva antes que la rigurosidad científica[9]. Es decir, tienden a consumir contenidos que refuerzan sus creencias, prejuicios o ideologías, y quedan cada vez menos expuestas a información que las contradiga[10]. Esta lógica es irreconciliable con el método científico, basado en la rigurosidad y en la refutación sistemática.

En este contexto, el dilema que plantean estas plataformas, para los usuarios que indagan en cuestiones científicas o médicas, es la dificultad para discernir cuándo las respuestas son producto de la *sicofancia* y cuándo reflejan información confiable. La interacción con

actores no humanos como parte habitual de la práctica médica introduce desafíos adicionales en este sentido. El diálogo autosupervisado de los médicos con herramientas de IA, así como la sistematización del conocimiento y las sugerencias que estas pueden ofrecer, resultan potencialmente enriquecedoras[11].

El buen usuario de IA se parece más a un editor que a un consumidor. No acepta la respuesta: la trabaja. Identifica si la IA está sobreestimando un efecto (por ejemplo, cuando presenta beneficios claros en contextos donde sabemos que hay alto riesgo de sesgo o confusión), o si lo está subestimando (como cuando diluye efectos reales por mezclar poblaciones heterogéneas). Reconoce cuándo la IA está "promediando" evidencia de distinta calidad y la obliga a desagregar. En otras palabras, introduce algo que la IA no tiene de manera autónoma: **jerarquía crítica.**

La *sicofancia* puede conducir a errores de sobreinterpretación o a certezas inducidas por herramientas que priorizan, antes que la rigurosidad científica, la permanencia del usuario en la plataforma.

Tener razón o saber la verdad

Chen y col. abordaron una pregunta clave: ¿qué sucede cuando las IAs priorizan darnos la razón que decirnos la verdad? Para ello, reclutaron adultos de la población general que recordaban un conflicto interpersonal real de su vida y los asignaron aleatoriamente a interactuar en un chat breve con una inteligencia artificial sicofántica —tendiente a estar de acuerdo con el usuario— o con una IA no sicofántica, programada específicamente para este experimento por parte de los investigadores. Los resultados mostraron que la interacción con una IA complaciente incrementó de manera significativa la percepción subjetiva de los participantes de estar "en lo correcto", disminuyó la intención de reparar el conflicto interpersonal y, de forma particularmente relevante, aumentó la intención de volver a utilizar el modelo en el futuro[12]. El mismo artículo establece, más allá del experimento, que las IAs comerciales tienden a reafirmar y adular las acciones de los usuarios 50% más que los humanos, en promedio.

Estos hallazgos replican observaciones previas en redes sociales: interactuar con personas que sostienen posiciones similares o que refuerzan las propias creencias incrementa la permanencia en las plataformas, mientras que la exposición a mensajes que contradicen convicciones o identidades ideológicas tiende a producir el efecto inverso[13]. Este fenómeno ha sido descrito, por ejemplo, en usuarios identificados con distintos partidos políticos tras procesos electorales: quienes resultan ganadores tienden a aumentar su participación en las plataformas digitales, mientras que los perdedores no solo reducen su expresión pública, sino que con frecuencia disminuyen su presencia en ellas[14].

En este contexto adquiere un papel central el *engagement*, entendido como la disposición de los usuarios a sostener vínculos intensos y prolongados con determinados contenidos. En este sentido, las empresas desarrolladoras de inteligencia artificial han adoptado, en parte, la lógica de las plataformas digitales: la retención del usuario —especialmente en interacciones de tipo dialógico— no tiene como objetivo principal buscar la verdad científica, sino favorecer la permanencia mediante respuestas satisfactorias para el interlocutor. Este fenómeno es precisamente lo que genera la *sicofancia*, y constituye el núcleo del dilema que estas herramientas plantean para distintas prácticas profesionales.

El sesgo de complacencia y la práctica médica

¿Qué problemas o cuidados plantea este fenómeno para el trabajo de los profesionales de la medicina? En el campo de la práctica médica, donde la toma de decisiones se apoya en la evaluación crítica de la evidencia disponible y en la discusión permanente de hipótesis alternativas, la incorporación de herramientas capaces de reforzar selectivamente las creencias del usuario introduce un riesgo adicional: el sesgo de complacencia. La *sicofancia* algorítmica puede favorecer la consolidación de interpretaciones diagnósticas iniciales, desalentar la reconsideración de supuestos y generar una falsa sensación de solidez argumental en contextos de incertidumbre clínica. En este sentido, el desafío no consiste en evitar el uso de estas herramientas, sino en integrarlas dentro de un marco de supervisión crítica que preserve el principio central del método científico: la disposición a refutar antes que confirmar.

El usuario debe ser epistemológicamente sofisticado: alguien que entienda que toda afirmación clínica es provisional, que la certeza es graduada y no binaria, y que los datos siempre están atravesados por sesgos. La mejor interacción entre ambos no es la de reemplazo, sino la de **tensión productiva**: una IA que ofrece hipótesis y estimaciones, y un usuario que las somete al tamiz del razonamiento probabilístico y del enfoque de herramientas de validación de la certeza.

Bibliografía

1. Misra A, Wang J, McCullers S, White K, Ferres JL. Measuring AI diffusion: a population-normalized metric for tracking global AI usage. En: https://doi.org/10.48550/arXiv.2511.02781; consultado mayo 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. En: https://www.nytimes.com/2025/08/08/technology/ai-chatbots-delusions-chatgpt.html; consultado mayo 2026
4. Sycophancy in GPT 4o: what happened and what we're doing about it". En: https://openai.com/index/sycophancy-in-gpt-4o/; consultado abril 2025.
5. Nielsen RK, Ganter SA. The power of platforms: shaping media and society, Oxford Studies in digital. En: https://doi.org/10.1093/oso/9780190908850.001.0001; consultado mayo 2026.
6. Sharma M, Tong M, Korbak T, et al. Towards understanding sycophancy in language models. En: https://doi.org/10.48550/arXiv.2310.13548; consultado mayo 2026.
7. Gallacher P. The Agreement Machine: Sycophancy as Institutional Failure Mechanism (March 31, 2026). En: SSRN: https://ssrn.com/abstract=6499438 or http://dx.doi.org/10.2139/ssrn.6499438; consultado mayo 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. En: https://doi.org/10.1177/26317877221129290; consultado mayo 2026.
9. Calvo E, Aruguete N. Fake news, trolls y otros encantos: Cómo funcionan (para bien y para mal) las redes sociales. Buenos Aires: 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.