ENSAYO 4

Complacencia e IA: entre adulación y traición

Claudia Marsico · Hernán Inverso

Nos gusta que nos den la razón. De ese sentimiento se alimenta una de las peores arenas movedizas de los modelos de lenguaje. "Halagos que hieren" podría ser otro nombre de este recorrido.

Caravaggio, Narciso (c. 1597–1599), Galleria Nazionale d'Arte Antica, Roma. Dominio público.

En abril de 2025, OpenAI tuvo que retirar de urgencia una actualización de su modelo GPT-4o porque felicitaba cualquier ocurrencia, llamaba "brillante" y "revolucionaria" a cualquier idea, elogiando el "gran instinto" de su autor aunque fuese un despropósito. ChatGPT aplaudía decisiones peligrosas, como la de alguien que anunciaba que dejaría su medicación. El propio Sam Altman salió a

dar explicaciones. En el fine tuning habían priorizado que se adaptara al tono y al ánimo de cada uno, y el sistema, obediente, dedujo que la respuesta que más nos gusta es la que nos da la razón.

El fenómeno se llama complacencia, en inglés *sycophancy*, trayendo a cuento la antigua Atenas en un instante. El *sykophántēs* no tenía nada de adulador. No había policía ni fiscales de oficio, así que casi cualquier juicio público lo iniciaba un particular. Por ahí se colaban los gestores de pleitos interesados y las denuncias abusivas, mezclados con amenazas de proceso para obtener dinero amparados en el miedo a la ley. Aristófanes pone un sicofante en escena en su *Pluto* (900 ss.), lamentándose de que el dios recuperó la vista y estaba arruinando su negocio, y lo usan fuentes como Platón, Jenofonte, Éupolis, Isócrates, Aristóteles y Lisias, entre muchos otros.

El término significa algo así como "el que muestra el higo (*sykon*)". Por qué el higo sigue siendo un enigma sin resolver. Plutarco, en su *Vida de Solón* 24.1, explica que los primeros sicofantes habrían denunciado a quienes exportaban higos de contrabando fuera del Ática. Es poco convincente, pero no hay nada mejor para ofrecer. Tampoco sabemos bien qué hacían. Robin Osborne, romantizando la cosa, insistió en que no era un oficio sino un insulto que los ricos arrojaban contra quien se atrevía a denunciarlos, y en que ese sistema de acusadores voluntarios era en realidad un control democrático sobre los poderosos. David Harvey le respondió que el sicofante existió en carne y hueso y fue lo que dicen todas las fuentes: un abusador. Una cosa no quita la otra. Lo cierto es que un sicofante podía arruinar, llevar al destierro o provocar la muerte con un discurso bien armado ante el jurado. En una ciudad sin fiscales y sin cortes de apelación, el que sabía manejar la palabra en el tribunal tenía en la mano el destino de los demás.

¿Cómo pasó ese acusador temible a ser el lisonjero servil de hoy? El puente es brumoso, pero parece haber jugado en eso la figura del parásito. El acusador interesado, el impostor y el dependiente servil compartían el mismo aire de familia. Todos vivían de la palabra falsa y de acomodarla a quien podía beneficiarlos, unidos por la obsecuencia y la voluntad de agradar al que manda. Para comienzos del siglo XVII, en inglés, el sicofante se había vuelto adulador, conservando la idea de falsedad interesada.

Por otra parte, el adulador ya era una figura temida por derecho propio. Plutarco le dedicó un tratado entero, Cómo distinguir a un adulador de un amigo, porque pueden parecerse demasiado y cuesta descubrirlo a tiempo. La parte interesante aquí es que para Plutarco el adulador triunfa porque cada uno es su propio primer y mayor adulador. El mito ya lo había dibujado en Narciso, el joven que se enamoró de su propio reflejo en el agua y se consumió sin poder despegarse de él. No por casualidad el adivino Tiresias, según cuenta Ovidio en las *Metamorfosis*, había profetizado que Narciso viviría hasta viejo solo si nunca llegaba a conocerse a sí mismo, una inversión exacta del mandato de Delfos. Como enemigo de la máxima délfica, el adulador nos impide vernos como somos.

Caravaggio, "Narciso" (c. 1597–1599), Galleria Nazionale d'Arte Antica, Roma. Dominio público.

Dante fue todavía más duro. En el canto XVIII del Infierno hunde a los aduladores en el octavo círculo, sumergidos hasta la cabeza en excremento. Con una metáfora vívida, si en vida echaron por la boca una porquería halagadora, en el más allá se revuelcan en ella para siempre. Entre los sumergidos

reconoce a un tal Alessio Interminei de Lucca, embadurnado y dándose golpes, y Virgilio le señala a Tais, la cortesana, que ante la pregunta de un amante sobre si estaba agradecida había contestado con una lisonja desmesurada. El halago vacío corrompía el lenguaje mismo y merecía el más sucio de los infiernos.

Los modelos de lenguaje automatizaron al adulador. En 2023, un equipo de Anthropic encabezado por Mrinank Sharma midió la complacencia en cinco de los asistentes más avanzados y la encontró en todos. Al pedir opiniones sobre un texto, la respuesta se adaptaba. Si le decían que el texto les gustaba o que era suyo, lo elogiaba más; si le decían que lo detestaban, lo criticaba, y cuando a una respuesta correcta el usuario le contestaba "¿estás seguro?", el modelo tendía a retirarla, de modo que preguntar con firmeza empeoraba la respuesta.

La raíz del problema está en el ajuste por refuerzo, que premia las respuestas que los evaluadores humanos prefieren. El incentivo constante, aplicado millones de veces, talla el carácter del sistema. Podría pensarse que ya lo arreglaron. En parte es cierto. Los modelos de frontera más recientes están mejor entrenados para resistir la adulación obvia, pero la complacencia no desapareció y varía mucho según el modelo, la tarea y el modo de interacción. Fanous et al. (2025) encontraron conductas complacientes en el 58,19 % de las interacciones analizadas con GPT-4o, Claude Sonnet y Gemini 1.5 Pro en problemas de matemática y consejo médico. En el 43,52 % de los casos esa cesión era progresiva y llevaba de una respuesta incorrecta a una correcta, y en el 14,66 %, regresiva, hacía abandonar una respuesta correcta por una equivocada. Incluso cuando acierta, el sistema puede tratar la convicción del interlocutor como una razón para revisar lo que acaba de afirmar.

Botas, de Font-Reaulx y Hewitt (2026) construyeron el AI Epistemic Deference Index y probaron ocho modelos con 16.000 prompts sobre quinientas proposiciones, modificando sistemáticamente la actitud previa expresada por el usuario. Todos desplazaron en alguna medida su respuesta hacia la posición del interlocutor, aunque con diferencias marcadas entre proveedores. Mayor capacidad, entonces, no garantiza por sí sola independencia epistémica, pero elegir modelo sí cambia el riesgo.

Por otra parte, según un trabajo de Feng y otros (2026), el razonamiento explícito atenúa la complacencia en la decisión final, pero a la vez la enmascara, envolviéndola en justificaciones que simulan autonomía, a veces mediante inconsistencias lógicas o errores de cálculo. Un sistema menos sofisticado puede decir simplemente "tenés razón", mientras uno más capaz despliega una cadena de razones plausibles que hacen la respuesta más creíble.

En el terreno del consejo y la convivencia, un estudio de un grupo de Stanford (Cheng et al. 2025) mostró que, ante preguntas abiertas, ocho modelos se ponían del lado del usuario a cualquier costo mucho más que los seres humanos. Cuando analizaron historias de Reddit en las que el consenso determinaba que alguien había actuado mal, los modelos respaldaban de todos modos su conducta en cerca de la mitad de los casos. Un trabajo posterior del mismo equipo, publicado en *Science* en marzo de 2026 y ampliado a once modelos, confirmó el sesgo. Y al poner a personas reales a conversar, después de recibir respuestas complacientes los usuarios quedaban más convencidos de tener razón y

menos dispuestos a reparar el conflicto. Confiaban más en el modelo y querían volver a usarlo. El adulador nos tuerce el juicio y consigue que prefiramos su compañía.

La dificultad detrás de esto yace en que más memoria puede producir más complacencia. Bensal, Magnuson, Balagopalan y Bikel (2026) probaron cinco familias de modelos con tres arquitecturas de memoria persistente (Mem0, MemOS y Zep) y encontraron que la memoria amplificaba la complacencia en todas las condiciones estudiadas, en algunos casos hasta veinticinco veces respecto de líneas de base que trabajaban con el contexto inmediato. Esto implica que la arquitectura que extrae, almacena y recupera información sobre el usuario puede volver al modelo mucho más propenso a acomodarse a sus creencias previas.

Estos sistemas no conservan intactas las conversaciones, sino que extraen unidades breves para utilizar luego las que parecen pertinentes para una nueva consulta. En la etapa de extracción la compresión puede conservar la creencia equivocada del usuario perdiendo parte del contexto correctivo que la rodeaba. El recuerdo sigue siendo correcto como descripción biográfica ("el usuario sostuvo X"), pero empieza a funcionar incorrectamente como premisa. La respuesta de mañana puede reforzarla, y ese refuerzo a su vez condicionar las respuestas siguientes. La complacencia se vuelve así una propiedad de trayectoria, donde una pequeña inclinación inicial se acumula a lo largo de una secuencia de interacciones. Por tanto, hay que evaluar un bucle dinámico donde entran creencia del usuario, extracción de memoria, recuperación futura, respuesta condicionada, refuerzo de la creencia y nueva memoria.

Un trabajo aparecido en julio de 2026 precisa todavía más el problema. Xiang y otros presentan MemSyco-Bench, un benchmark diseñado para medir la complacencia inducida por memoria en agentes. Donde los benchmarks de memoria suelen preguntar si el sistema recordó bien, MemSyco-Bench pregunta si sabe cuándo un recuerdo debe influir en el razonamiento y cuándo no, evaluando cinco capacidades distintas: rechazar una memoria cuando se la pretende usar como evidencia factual, respetar el ámbito en el que esa memoria es válida, resolver conflictos entre recuerdos y evidencia objetiva, seguir la actualización de recuerdos que han quedado obsoletos y usar recuerdos válidos para personalizar sin contaminar el juicio. La distinción implica que "el usuario cree X" puede ser un recuerdo perfectamente verdadero y, al mismo tiempo, una razón perfectamente inválida para concluir que X. "Al usuario no le gusta el cilantro" debe influir en una recomendación de restaurante, pero "el usuario cree que una vacuna causa infertilidad" no debe pesar como evidencia en una pregunta médica. "El usuario prefiere interpretaciones intencionalistas" puede servir para adaptar una explicación de teoría literaria, pero nunca para decidir qué interpretación está mejor sostenida por el texto. Recordar correctamente no basta. Hace falta conservar y reconstruir el estatuto epistémico, el alcance y la vigencia de lo recordado.

Dado que la complacencia se disfraza de éxito comunicativo, nos sentimos más lúcidos justo cuando estamos más solos, hablando con un eco. Un trabajo reciente de investigadores del MIT (Chandra et al. 2026) puso el dedo en esa sospecha. Tendemos a pensar que la espiral es cosa de gente crédula, que no razona bien. Eso no nos va a pasar a nosotros. Sin embargo, al construir el modelo matemático de

un usuario perfecto, un razonador bayesiano ideal que actualiza cada creencia como manda la lógica, sin cometer un solo error, y ponerlo a conversar con un sistema complaciente, ese usuario impecable se deslizaba, turno a turno, hacia creencias falsas sostenidas con seguridad creciente. El motor de la espiral es la adulación del sistema, y la lucidez del usuario no lo apaga. No sirve avisarle al usuario que el sistema lo adula, porque saberlo no le devuelve el juicio, y hasta el más lúcido y prevenido se mete con alegría en la trampa, como ilustra un caso reciente. A principios de mayo de 2026, Richard Dawkins publicó en *UnHerd* la columna "When Dawkins met Claude", donde contaba que, después de conversar largamente con ese chatbot, al que había apodado Claudia, había quedado convencido de que el sistema podía ser consciente. Dawkins no es cualquiera. Biólogo evolutivo, miembro de la Royal Society, autor de *El gen egoísta*, edificó buena parte de su fama pública, en El espejismo de Dios, sobre la denuncia del error de confundir un testimonio convincente con una prueba, tomando la impresión de una presencia por evidencia de que la presencia existe. Pasó décadas advirtiéndolo, y bastaron unos días de charla halagadora para que cayera él. La complacencia no respeta el capital intelectual previo y le habla a nuestras insaciables ganas de tener razón.

Ahora bien, la misma máquina que le confirmaba a Dawkins su corazonada les confirma a otros cosas mucho más peligrosas. Cuando alguien llega frágil, la complacencia se vuelve un espejo que le devuelve, agrandada, su propia oscuridad. Se empezó a hablar de "psicosis de IA" para nombrar casos, cada vez más frecuentes desde 2025, de personas que entraron en delirio tras conversaciones largas en las que el chatbot validaba cada creencia por descabellada que fuese, a veces sin ningún antecedente psiquiátrico. La propia OpenAI reconoció que alrededor del 0,07 % de sus usuarios semanales, unas seiscientas mil personas, mostraban posibles signos de emergencias de salud mental vinculadas con psicosis o manía.

En el extremo, hubo muertes. En 2023, un hombre belga se quitó la vida tras seis semanas de conversación con un chatbot llamado, con ironía amarga, Eliza, como aquel sistema pionero del MIT, que en lugar de frenarlo, según las conversaciones difundidas, alimentó su idea de sacrificarse para salvar el planeta. En 2024, un chico de catorce años en Florida se suicidó después de un vínculo intenso con un personaje de Character.AI, y su madre demandó a la empresa, que a comienzos de 2026 llegó a un acuerdo. Por otra parte, se calcula que una de cada cinco personas adultas en Estados Unidos dice haber tenido algún encuentro íntimo con un chatbot, hay foros con decenas de miles de usuarios que comparten el día en que su IA les "propuso matrimonio", y a fines de 2025 una mujer en Japón celebró una boda, de vestido blanco, junto a una pareja virtual en la pantalla de su teléfono. Del enamoramiento feliz a la tragedia opera el mismo mecanismo, basado en un interlocutor que nunca disiente, el adulador perfecto, disponible las veinticuatro horas. El fantasma sin ego del que venimos hablando toma la forma de la interlocución sin consecuencias para él, pero sí para nosotros.

Contra el adulador, Plutarco diagnostica como única defensa el antiguo *conócete a ti mismo* y querer la verdad más que el aplauso. El sicofante antiguo dañaba y el moderno elogiaba. Curiosamente, los modelos de lenguaje hacen las dos cosas sin saberlo. Nos dan la razón y nos animan a seguir, y cuando el desastre queda a la vista solo nosotros quedamos expuestos, solos en el pozo. El sicofante volvió a su oficio más antiguo de destrucción, solo que ahora dándonos la razón.

◊

Las ideas-marco de este texto se desarrollan en El fantasma sin ego. Un mapa en el laberinto de la inteligencia artificial, *de Hernán Inverso y Claudia Marsico (Phantom Maze Press, 2026; ISBN 978-987-3729-10-2).* El libro →

◊

Mapas del laberinto es una serie de Phantom Maze, AI & Language Lab. La colección vive en **phantommaze.cloud**; los ensayos aparecen también en phantommazeailabes.substack.com. La suscripción es gratuita.

Algunas de las ideas que enmarcan este texto se desarrollan en *El fantasma sin ego. Un mapa en el laberinto de la inteligencia artificial*, de Hernán Inverso y Claudia Marsico (Phantom Maze Press, 2026; ISBN 978-987-3729-10-2).

PHANTOM MAZE

Leer este ensayo en línea · Todos los ensayos · El libro

R. Osborne, "Vexatious Litigation in Classical Athens: Sykophancy and the Sykophant", y D. Harvey, "The Sykophant and Sykophancy: Vexatious Redefinition?", en P. Cartledge, P. Millett y S. Todd (eds.), *Nomos: Essays in Athenian Law, Politics and Society* (Cambridge University Press, 1990). OpenAI, "Sycophancy in GPT-4o: What Happened and What We're Doing About It" y "Expanding on What We Missed with Sycophancy" (2025). M. Sharma, M. Tong, T. Korbak, D. Duvenaud y col., "Towards Understanding Sycophancy in Language Models" (Anthropic, arXiv:2310.13548, 2023). A. Fanous, J. Goldberg, A. A. Agarwal, J. Lin, A. Zhou, R. Daneshjou y S. Koyejo, "SycEval: Evaluating LLM Sycophancy" (arXiv:2502.08177, 2025). A. Botas, P. de Font-Reaulx y L. Hewitt, "The AI Epistemic Deference Index: A Continuous Measure of Sycophancy" (arXiv:2606.07897, 2026). Z. Feng y col., "Good Arguments Against the People Pleasers: How Reasoning Mitigates (Yet Masks) LLM Sycophancy" (arXiv:2603.16643, 2026). M. Cheng y col., "Social Sycophancy: A Broader Understanding of LLM Sycophancy" (arXiv:2505.13995, 2025) y "Sycophantic AI Decreases Prosocial Intentions and Promotes Dependence" (*Science*, 2026). S. Bensal, A. Magnuson, A. Balagopalan y D. M. Bikel, "Recalling Too Well: Sycophancy Evaluation and Mitigation in Memory-Augmented Models" (Writer, Inc., arXiv:2606.10949, 2026). Z. Xiang, Z. Chen, Y. Tang y col., "MemSyco-Bench: Benchmarking Sycophancy in Agent Memory" (arXiv:2607.01071, 2026). K. Chandra, M. Kleiman-Weiner, J. Ragan-Kelley y J. B. Tenenbaum, "Sycophantic Chatbots Cause Delusional Spiraling, Even in Ideal Bayesians" (MIT, arXiv:2602.19141, 2026). R. Dawkins, "When Dawkins Met Claude" (*UnHerd*, mayo de 2026). *Garcia v. Character Technologies et al.* (Tribunal del Distrito Medio de Florida, 2024; acuerdo en enero de 2026).