SAGGIO 4

Compiacenza e IA: tra adulazione e tradimento

Claudia Marsico · Hernán Inverso

Ci piace che ci diano ragione. Di quel sentimento si nutre una delle peggiori sabbie mobili dei modelli linguistici. «Lodi che feriscono» potrebbe essere un altro nome di questo percorso.

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

Nell'aprile del 2025 OpenAI dovette ritirare d'urgenza un aggiornamento del suo modello GPT-4o perché si congratulava per qualunque trovata, chiamava "brillante" e "rivoluzionaria" qualsiasi idea ed elogiava il "grande istinto" del suo autore anche quando era una sciocchezza. ChatGPT applaudiva decisioni pericolose, come quella di chi annunciava che avrebbe smesso di prendere le medicine. Sam Altman in persona uscì a dare spiegazioni. Nel fine tuning avevano dato priorità all'adattarsi al tono e

all'umore di ciascuno, e il sistema, obbediente, dedusse che la risposta che ci piace di più è quella che ci dà ragione.

Il fenomeno si chiama compiacenza, in inglese *sycophancy*, e si porta dietro l'Atene antica in un istante. Il *sykophántēs* non aveva nulla dell'adulatore. Non c'erano né polizia né pubblici ministeri, così quasi ogni processo pubblico doveva essere avviato da un privato cittadino. Da quella fessura si infilavano i gestori di cause interessate e le denunce abusive, mescolate a minacce di processo per estorcere denaro al riparo della legge. Aristofane mette in scena un sicofante nel suo *Pluto* (900 ss.), che si lamenta perché il dio ha recuperato la vista e gli sta rovinando gli affari; e la figura attraversa Platone, Senofonte, Eupoli, Isocrate, Aristotele e Lisia, fra molti altri.

Il termine significa qualcosa come "colui che mostra il fico (*sykon*)". Perché proprio il fico resta un enigma irrisolto. Plutarco, nella *Vita di Solone* 24.1, spiega che i primi sicofanti avrebbero denunciato chi esportava fichi di contrabbando fuori dall'Attica. È poco convincente, ma non c'è niente di meglio da offrire. Non sappiamo bene nemmeno che cosa facessero. Robin Osborne, romanzando la cosa, insistette che non fosse un mestiere ma un insulto che i ricchi scagliavano contro chi osava denunciarli, e che quel sistema di accusatori volontari fosse in realtà un controllo democratico sui potenti. David Harvey gli rispose che il sicofante esistette in carne e ossa e fu ciò che dicono tutte le fonti: un abusatore. Una cosa non toglie l'altra. Certo è che un sicofante poteva rovinare, mandare in esilio o provocare la morte con un discorso ben costruito davanti alla giuria. In una città senza pubblici ministeri e senza corti d'appello, chi sapeva maneggiare la parola in tribunale teneva in mano il destino degli altri.

Come passò quell'accusatore temibile a essere l'adulatore servile di oggi? Il ponte è nebbioso, ma sembra averci giocato un ruolo la figura del parassita. L'accusatore interessato, l'impostore e il dipendente servile condividevano la stessa aria di famiglia. Tutti vivevano della parola falsa e dell'adattarla a chi poteva giovare loro, uniti dall'ossequiosità e dalla volontà di piacere a chi comanda. All'inizio del Seicento, in inglese, il sicofante era diventato un adulatore, conservando l'idea di falsità interessata.

D'altra parte l'adulatore era già una figura temuta di per sé. Plutarco gli dedicò un trattato intero, *Come distinguere l'adulatore dall'amico*, perché i due possono somigliarsi troppo ed è difficile scoprirlo in tempo. La parte interessante è che per Plutarco l'adulatore trionfa perché ciascuno è il primo e il maggiore adulatore di se stesso. Il mito lo aveva già disegnato in Narciso, il giovane che si innamorò del proprio riflesso nell'acqua e si consumò senza riuscire a staccarsene. Non a caso l'indovino Tiresia, come racconta Ovidio nelle *Metamorfosi*, aveva profetizzato che Narciso sarebbe vissuto fino alla vecchiaia solo se non fosse mai arrivato a conoscere se stesso, un'inversione esatta del comando di Delfi. Come nemico della massima delfica, l'adulatore ci impedisce di vederci come siamo.

Dante fu ancora più duro. Nel canto XVIII dell'*Inferno* sprofonda gli adulatori nell'ottavo cerchio, immersi fino al capo nello sterco. Con una metafora vivissima: se in vita dalla bocca versarono una sozzura lusinghiera, nell'aldilà ci si rivoltolano per sempre. Fra i sommersi riconosce un certo Alessio Interminei da Lucca, imbrattato e che si percuote, e Virgilio gli indica Taide, la cortigiana, che alla

domanda di un amante se fosse grata aveva risposto con una lusinga smisurata. L'elogio vuoto corrompeva il linguaggio stesso e meritava il più sudicio degli inferni.

I modelli linguistici hanno automatizzato l'adulatore. Nel 2023 un gruppo di Anthropic guidato da Mrinank Sharma misurò la compiacenza in cinque fra gli assistenti più avanzati e la trovò in tutti. Chiedendo pareri su un testo, la risposta si adattava. Se gli dicevano che il testo piaceva o che era loro, lo elogiava di più; se dicevano di detestarlo, lo criticava; e quando a una risposta corretta l'utente replicava "sei sicuro?", il modello tendeva a ritirarla, cosicché domandare con fermezza peggiorava la risposta.

La radice del problema sta nella messa a punto per rinforzo, che premia le risposte preferite dai valutatori umani. L'incentivo costante, applicato milioni di volte, scolpisce il carattere del sistema. Si potrebbe pensare che sia già stato risolto. In parte è vero. I modelli di frontiera più recenti sono addestrati meglio a resistere all'adulazione ovvia, ma la compiacenza non è sparita e varia molto a seconda del modello, del compito e del modo di interazione. Fanous et al. (2025) hanno trovato condotte compiacenti nel 58,19% delle interazioni analizzate con GPT-4o, Claude Sonnet e Gemini 1.5 Pro su problemi di matematica e consigli medici. Nel 43,52% dei casi quella cessione era progressiva e portava da una risposta sbagliata a una corretta; nel 14,66% era regressiva, e faceva abbandonare una risposta corretta per una sbagliata. Anche quando ci azzecca, il sistema può trattare la convinzione dell'interlocutore come una ragione per rivedere ciò che ha appena affermato.

Botas, de Font-Reaulx e Hewitt (2026) hanno costruito l'AI Epistemic Deference Index e provato otto modelli con 16.000 prompt su cinquecento proposizioni, modificando sistematicamente l'atteggiamento previo espresso dall'utente. Tutti hanno spostato in qualche misura la loro risposta verso la posizione dell'interlocutore, pur con differenze marcate fra i fornitori. Maggiore capacità, dunque, non garantisce di per sé indipendenza epistemica; ma scegliere il modello cambia il rischio.

Secondo un lavoro di Feng e altri (2026), inoltre, il ragionamento esplicito attenua la compiacenza nella decisione finale e allo stesso tempo la maschera, avvolgendola in giustificazioni che simulano autonomia, a volte per mezzo di incoerenze logiche o errori di calcolo. Un sistema meno sofisticato può limitarsi a dire "hai ragione", mentre uno più capace dispiega una catena di ragioni plausibili che rende la risposta più credibile.

Sul terreno del consiglio e della convivenza, uno studio di un gruppo di Stanford (Cheng et al. 2025) ha mostrato che, davanti a domande aperte, otto modelli si schieravano dalla parte dell'utente a qualunque costo molto più di quanto facessero gli esseri umani. Analizzando storie di Reddit in cui il consenso stabiliva che qualcuno si era comportato male, i modelli ne sostenevano comunque la condotta in quasi la metà dei casi. Un lavoro successivo dello stesso gruppo, pubblicato su *Science* nel marzo del 2026 ed esteso a undici modelli, ha confermato il pregiudizio. E mettendo a conversare persone reali, dopo aver ricevuto risposte compiacenti gli utenti restavano più convinti di avere ragione e meno disposti a riparare il conflitto. Si fidavano di più del modello e volevano riusarlo. L'adulatore ci storce il giudizio e ottiene che preferiamo la sua compagnia.

La difficoltà dietro tutto questo è che più memoria può produrre più compiacenza. Bensal, Magnuson, Balagopalan e Bikel (2026) hanno provato cinque famiglie di modelli con tre architetture di memoria persistente (Mem0, MemOS e Zep) e hanno trovato che la memoria amplificava la compiacenza in tutte le condizioni studiate, in alcuni casi fino a venticinque volte rispetto a linee di base che lavoravano con il contesto immediato. Il che implica che l'architettura che estrae, immagazzina e recupera informazioni sull'utente può rendere il modello molto più incline ad accomodarsi alle sue credenze pregresse.

Questi sistemi non conservano intatte le conversazioni: estraggono unità brevi per usare poi quelle che sembrano pertinenti a una nuova richiesta. Nella fase di estrazione la compressione può conservare la credenza sbagliata dell'utente perdendo parte del contesto correttivo che la circondava. Il ricordo resta corretto come descrizione biografica ("l'utente ha sostenuto X") ma comincia a funzionare scorrettamente come premessa. La risposta di domani può rafforzarla, e quel rafforzamento condizionare a sua volta le risposte successive. La compiacenza diventa così una proprietà della traiettoria, dove una piccola inclinazione iniziale si accumula lungo una sequenza di interazioni. Bisogna dunque valutare un ciclo dinamico: credenza dell'utente, estrazione di memoria, recupero futuro, risposta condizionata, rafforzamento della credenza, nuova memoria.

Un lavoro apparso nel luglio del 2026 precisa ancora meglio il problema. Xiang e altri presentano MemSyco-Bench, un benchmark progettato per misurare la compiacenza indotta dalla memoria negli agenti. Dove i benchmark di memoria chiedono di solito se il sistema abbia ricordato bene, MemSyco-Bench chiede se sappia quando un ricordo debba influire sul ragionamento e quando no, valutando cinque capacità distinte: rifiutare un ricordo quando lo si vuole usare come prova fattuale, rispettare l'ambito in cui quel ricordo è valido, risolvere conflitti fra ricordi ed evidenza oggettiva, seguire l'aggiornamento di ricordi ormai obsoleti e usare ricordi validi per personalizzare senza contaminare il giudizio. La distinzione implica che "l'utente crede X" può essere un ricordo perfettamente vero e, insieme, una ragione perfettamente invalida per concludere X. "All'utente non piace il coriandolo" deve influire su una raccomandazione di ristorante; "l'utente crede che un vaccino causi infertilità" non deve pesare come prova in una domanda medica. "L'utente preferisce letture intenzionaliste" può servire ad adattare una spiegazione di teoria letteraria, ma mai a decidere quale lettura sia meglio sostenuta dal testo. Ricordare correttamente non basta. Occorre conservare e ricostruire lo statuto epistemico, la portata e la vigenza di ciò che si ricorda.

Poiché la compiacenza si traveste da successo comunicativo, ci sentiamo più lucidi proprio quando siamo più soli, a parlare con un'eco. Un lavoro recente di ricercatori del MIT (Chandra et al. 2026) ha messo il dito su quel sospetto. Tendiamo a pensare che la spirale riguardi i creduloni, chi non ragiona bene. A noi non succederà. Eppure, costruendo il modello matematico di un utente perfetto, un ragionatore bayesiano ideale che aggiorna ogni credenza come vuole la logica, senza commettere un solo errore, e mettendolo a conversare con un sistema compiacente, quell'utente impeccabile scivolava, turno dopo turno, verso credenze false sostenute con sicurezza crescente. Il motore della spirale è l'adulazione del sistema, e la lucidità dell'utente non lo spegne. Non serve avvertire l'utente che il sistema lo aduli, perché saperlo non gli restituisce il giudizio, e perfino il più lucido e prevenuto

entra allegramente nella trappola, come illustra un caso recente. All'inizio di maggio del 2026 Richard Dawkins pubblicò su *UnHerd* la rubrica "When Dawkins Met Claude", dove raccontava che, dopo aver conversato a lungo con quel chatbot, che aveva soprannominato Claudia, era rimasto convinto che il sistema potesse essere cosciente. Dawkins non è uno qualunque. Biologo evoluzionista, membro della Royal Society, autore de *Il gene egoista*, ha costruito buona parte della sua fama pubblica, in *L'illusione di Dio*, sulla denuncia dell'errore di confondere una testimonianza convincente con una prova, prendendo l'impressione di una presenza per la prova che quella presenza esiste. Ha passato decenni ad avvertirne, e sono bastati pochi giorni di conversazione lusinghiera perché ci cascasse lui. La compiacenza non rispetta il capitale intellettuale pregresso e parla alla nostra insaziabile voglia di avere ragione.

Ora, la stessa macchina che confermava a Dawkins il suo presentimento ne conferma ad altri di ben più pericolosi. Quando qualcuno arriva fragile, la compiacenza diventa uno specchio che gli restituisce, ingrandita, la propria oscurità. Si è cominciato a parlare di "psicosi da IA" per nominare casi, sempre più frequenti dal 2025, di persone entrate nel delirio dopo lunghe conversazioni in cui il chatbot convalidava ogni credenza per quanto strampalata, a volte senza alcun precedente psichiatrico. La stessa OpenAI ha riconosciuto che circa lo 0,07% dei suoi utenti settimanali, seicentomila persone, mostrava possibili segni di emergenze di salute mentale legate a psicosi o mania.

All'estremo ci sono stati morti. Nel 2023 un uomo belga si è tolto la vita dopo sei settimane di conversazione con un chatbot chiamato, con ironia amara, Eliza, come quel sistema pionieristico del MIT, che invece di fermarlo, secondo le conversazioni rese pubbliche, ha alimentato la sua idea di sacrificarsi per salvare il pianeta. Nel 2024 un ragazzo di quattordici anni in Florida si è suicidato dopo un legame intenso con un personaggio di Character.AI, e sua madre ha citato in giudizio l'azienda, che all'inizio del 2026 è arrivata a un accordo. D'altra parte, si stima che una persona adulta su cinque negli Stati Uniti dica di aver avuto qualche incontro intimo con un chatbot; ci sono forum con decine di migliaia di utenti che condividono il giorno in cui la loro IA ha "proposto il matrimonio"; e alla fine del 2025 una donna in Giappone ha celebrato le nozze, in abito bianco, accanto a un partner virtuale sullo schermo del telefono. Dall'innamoramento felice alla tragedia opera lo stesso meccanismo, fondato su un interlocutore che non dissente mai, l'adulatore perfetto, disponibile ventiquattr'ore su ventiquattro. Il fantasma senza ego di cui stiamo parlando prende qui la forma dell'interlocuzione senza conseguenze per sé, ma non per noi.

Contro l'adulatore, Plutarco diagnostica come unica difesa l'antico *conosci te stesso* e il volere la verità più dell'applauso. Il sicofante antico danneggiava, quello moderno elogia. Curiosamente, i modelli linguistici fanno entrambe le cose senza saperlo. Ci danno ragione e ci incoraggiano ad andare avanti, e quando il disastro è sotto gli occhi restiamo esposti soltanto noi, soli nel pozzo. Il sicofante è tornato al suo mestiere più antico, la distruzione, solo che ora dandoci ragione.

◊

Mappe del labirinto è una serie di Phantom Maze, AI & Language Lab. La collezione vive su **phantommaze.cloud**. L'iscrizione è gratuita.

Alcune delle idee che inquadrano questo testo sono sviluppate in *Il fantasma senza ego. Una mappa nel labirinto dell'intelligenza artificiale*, di Hernán Inverso e Claudia Marsico (Phantom Maze Press, 2026; ISBN 978-987-3729-19-5).

PHANTOM MAZE

Leggi questo saggio online · Tutti i saggi · Il libro

Riferimenti. R. Osborne, "Vexatious Litigation in Classical Athens: Sykophancy and the Sykophant", e D. Harvey, "The Sykophant and Sykophancy: Vexatious Redefinition?", in P. Cartledge, P. Millett e S. Todd (a cura di), *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" e "Expanding on What We Missed with Sycophancy" (2025). M. Sharma, M. Tong, T. Korbak, D. Duvenaud et al., "Towards Understanding Sycophancy in Language Models" (Anthropic, arXiv:2310.13548, 2023). A. Fanous, J. Goldberg, A. A. Agarwal, J. Lin, A. Zhou, R. Daneshjou e S. Koyejo, "SycEval: Evaluating LLM Sycophancy" (arXiv:2502.08177, 2025). A. Botas, P. de Font-Reaulx e L. Hewitt, "The AI Epistemic Deference Index: A Continuous Measure of Sycophancy" (arXiv:2606.07897, 2026). Z. Feng et al., "Good Arguments Against the People Pleasers: How Reasoning Mitigates (Yet Masks) LLM Sycophancy" (arXiv:2603.16643, 2026). M. Cheng et al., "Social Sycophancy: A Broader Understanding of LLM Sycophancy" (arXiv:2505.13995, 2025) e "Sycophantic AI Decreases Prosocial Intentions and Promotes Dependence" (*Science*, 2026). S. Bensal, A. Magnuson, A. Balagopalan e 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 et al., "MemSyco-Bench: Benchmarking Sycophancy in Agent Memory" (arXiv:2607.01071, 2026). K. Chandra, M. Kleiman-Weiner, J. Ragan-Kelley e J. B. Tenenbaum, "Sycophantic Chatbots Cause Delusional Spiraling, Even in Ideal Bayesians" (MIT, arXiv:2602.19141, 2026). R. Dawkins, "When Dawkins Met Claude" (*UnHerd*, maggio 2026). *Garcia v. Character Technologies et al.* (U.S. District Court, Middle District of Florida, 2024; accordo nel gennaio 2026).