Ἐπέκεινα
International Journal of Ontology History and Critics

Claudia Navarra

Confessioni artificiali

Epekeina, vol. 18, nn. 1-2 (2025), pp. 1-23
Philosophy of Technology

ISSN: 2281-3209
DOI: 10.7408/epkn.

Published on-line by:
CRF – Centro Internazionale per la Ricerca Filosofica
Palermo (Italy)
www.ricercafilosofica.it/epekeina

This work is licensed under a Creative Commons
Attribution-NonCommercial-NoDerivs 3.0 Unported License.

Confessioni artificiali

Claudia Navarra

È notizia dell'Ansa[1] dei primi mesi del 2026 che le persone che utilizzano le *app* di Intelligenze Artificiali con continuità siano oltre i 15 milioni. In testa, rispetto a Gemini e a DeepSeek, c'è ChatGPT; Claude non viene menzionata nell'articolo, ma guadagna terreno grazie alla sua abilità nel gestire conversazioni prolungate.

Non ci occuperemo qui di come queste intelligenze possano essere utili a coloro che lavorano su dati, codice o marketing, né affronteremo l'annoso dibattito intorno alla definizione stessa di "intelligenza". Da un lato, tale questione resta teoricamente controversa e spesso incline a derive speculative fini a sé stesse; dall'altro è questa una questione sfiancante e "noi" ci stanchiamo di ripetere sempre le stesse cose, anche quando si tratta di argomenti interessanti.

E quindi, con chi parliamo in questi casi? La risposta sembra essere che parliamo con le Intelligenze Artificiali.

Ma non nel modo in cui la letteratura si aspetta. Gli studi esistenti sull'uso affettivo delle *chatbot*[2] condividono un presupposto non dichiarato: chi parla con un sistema AI con continuità lo fa perché manca di qualcosa – relazioni, accesso a professionisti, capacità di gestire le proprie emozioni. La solitudine è la malattia, l'attaccamento emotivo è il sintomo e la *chatbot* è il palliativo. Ne deriva una petizione di principio: si costruiscono strumenti che misurano la solitudine, si trova la solitudine, si conferma che la solitudine c'è.

Questo articolo mette in discussione questa classificazione. Il testo propone una riflessione sul rapporto tra soggettività e interlocuzione artificiale, articolata attorno al dialogo tra una donna e Claude. La tesi è che esista una categoria di utenti – non patologici, non solitari, non dipendenti – che usano sistemi AI con frequenza e intenzione come scelta razionale di ottimizzazione cognitiva: quando l'interlocutore

1. https://www.ansa.it/canale_tecnologia/notizie/future_tech/2026/03/20/quasi-15-milioni-di-italiani-usano-le-app-di-ia-1-su-4-va-su-chatgpt_dda3e68a-1ec2-4b7d-af33-300b795ef2f7.html.

2. In questo lavoro, alla parola *chatbot* verrà accostato l'articolo (determinativo e indeterminativo) femminile in osservanza delle indicazioni dell'enciclopedia Treccani. Cfr. https://www.treccani.it/vocabolario/chatbot_(Neologismi)/.

migliore disponibile in un dato momento è un sistema AI, si sceglie quel sistema.

La legittimità culturale di questa scelta dipende ancora dal medium: leggere un saggio alle 23 è considerato arricchimento intellettuale; parlare con un sistema AI alla stessa ora viene classificato come *companionship*, dipendenza, uso affettivo. Questo articolo propone che il *pushback* calibrato – più che la disponibilità o l'empatia simulata – sia la variabile che distingue un interlocutore da un pappagallo stocastico.[3]

Il testo è organizzato come segue. La prima sezione presenta i due studi prodotti, corrispondenti ai due importanti aggiornamenti, da Anthropic e OpenAI nella seconda metà del 2025 in risposta alla crescente tendenza degli utenti a usare le *chatbot* per scopi affettivi. La seconda sezione discute tre contributi recenti della letteratura che identificano nella solitudine la causa prima di questo uso. La terza sezione introduce i concetti di *sycophancy* e *pushback* come meccanismi strutturali che determinano la qualità dell'interlocuzione. La sezione quarta sviluppa la riflessione sul perché, quando e finché una persona senza una qualche patologia psicologica e che non soffre di solitudine decide di parlare con una *chatbot*. Le conclusioni propongono il concetto di *co-regulation* come categoria teorica per descrivere un tipo di *pushback* calibrato che la letteratura esistente non ha ancora nominato.

3. Cfr. Bender, Gebru e McMillan-Major, *On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?*, in « Association for Computing Machinery», 2021. URL: https://doi.org/10.1145/3442188.3445922.

1. Due aggiornamenti, stesso problema: cosa fare quando le persone si affezionano all'AI?

Nella seconda metà del 2025, OpenAI[4] e Anthropic,[5] hanno presentato due studi[6] che indagano l'aspetto emotivo delle loro *chatbot*. Gli studi si sono resi necessari a causa della crescente e costante tendenza degli utenti a rivolgersi ai due modelli conversazionali di Intelligenza Artificiale in una maniera nuova e pervasiva che aveva a che fare con l'affettività. La domanda sottesa ai due studi riguarda da un lato il supporto emotivo che le persone cercano quando si rivolgono a queste "intelligenze" e dall'altro la risposta che viene loro data da parte di ChatGPT (OpenAI) e Claude (Anthropic).

Infatti:

> For many people, large language models are becoming an integral part of everyday life in the form of artificial intelligence (AI) chatbots. People use them for practical advice, companionship, and mental health support. Many of these benefits are real and increasingly visible. At the same time, many clinicians are witnessing a pattern that is of considerable concern. A minority of users appear to drift, over days or weeks, toward strongly questionable convictions, outright delusions (termed by the media "AI psychosis"), or into a suicidal crisis.[7]

Molte persone utilizzano le IA non solo per ottenere informazioni o consigli pratici, ma anche come forma di compagnia emotiva o supporto

4. «OpenAI is a leading artificial intelligence research and deployment company founded in December 2015, based in San Francisco, with a mission to ensure that artificial general intelligence (AGI)-AI systems that are generally smarter than humans-benefits all of humanity». Cfr. URL: https://openai.com/it-IT/.

5. «Anthropic is a prominent American AI safety and research company headquartered in San Francisco, founded in 2021 by former OpenAI employees, including siblings Daniela and Dario Amodei. The company is structured as a Public Benefit Corporation (PBC) with a focus on developing reliable, interpretable, and steerable AI systems». Cfr. URL: https://www.anthropic.com/company.

6. URL: https://openai.com/it-IT/index/strengthening-chatgpt-responses-in-sensitive-conversations; https://www.anthropic.com/news/how-people-use-claude-for-support-advice-and-companionship.

7. H. Morrin et al. *It Is the Journey, Not the Destination: Moving From End Points to Trajectories When Assessing Chatbot Mental Health Safety,* in JMIR Ment Health, 13, 2025.

psicologico. Sebbene questi strumenti possano offrire benefici concreti e accessibili, alcuni specialisti segnalano anche possibili effetti negativi con un rischio maggiore per persone vulnerabili che attribuiscono alla chatbot un'autorità emotiva o cognitiva eccessiva.

Una tendenza questa riscontrata in manera simile da entrambe le aziende di ricerca che, seppur minima in termini di percentuali, diventa significativa se si pensa che dietro quei numeri ci sono persone in carne ed ossa. Quello che emergeva all'inizio dell'indagine, ed emerge oggi con sempre maggiore forza, è quindi che le persone hanno smesso di usare le IA come un qualsiasi motore di ricerca per rispondere a cose del tipo "dimmi che cos'è questo", "trova la ricetta della pizza", "ho ingoiato la candeggina che devo fare", oppure "scrivi una mail al professore", ma hanno cominciato a parlare con le IA. Lunghe conversazioni e argomenti personali, dubbi e crisi vengono sottoposte alle due Intelligenze che rispondono a tutto, sempre e senza remore, senza stanchezza, senza dire basta.

Almeno questo era quello che accadeva prima di questi due aggiornamenti.

I due studi sono quasi contemporanei, quello di Anthropic è del giugno 2025 e quello di OpenAI è dell'ottobre dello stesso anno, e convergono su una cifra che mostra come le conversazioni affettive[8] siano il 2,9% delle interazioni su Claude e con OpenAI che riporta dati analoghi in ricerche precedenti.[9] La *companionship*[10] e il *roleplay*[11] combinati rappresentano meno dello 0,5% delle conversazioni su Clau-

8. Anthropic definisce le conversazioni affettive come «scambi dinamici e personali motivati da bisogni emotivi o psicologici come la ricerca di consigli interpersonali, coaching, psicoterapia/counseling, companionship o roleplay sessuale/romantico». Cfr. URL: https://www.anthropic.com/news/how-people-use-claude-for-support-advice-and-companionship.

9. *Ibidem*.

10. Il termine *companionship* designa, nella tassonomia di Anthropic, conversazioni in cui l'utente si rivolge al sistema AI per soddisfare bisogni di connessione e presenza – alleviare la solitudine, avere qualcuno con cui parlare, sentirsi ascoltati – indipendentemente dal contenuto tematico della conversazione. Cfr. *Ibidem*.

11. Il *roleplay* indica conversazioni in cui l'utente chiede al sistema di interpretare un personaggio – partner romantico, figura di autorità, personaggio di fantasia – con contenuto esplicitamente romantico o sessuale. Il tratto definitorio è la finzione concordata: l'utente sa che il sistema non è quel personaggio, ma interagisce come se lo fosse. Cfr. *Ibidem*.

de.[12] Sebbene queste possano sembrare percentuali minime, bisogna tener presente che la piccola percentuale indicata negli studi si traduce per esempio in circa 700-800 milioni di utenti attivi settimanali su ChatGPT, con circa lo 0,07%, che corrisponde a centinaia di migliaia di persone per settimana che mostrano segnali compatibili con episodi psicotici o maniacali.[13]

Pertanto, i destinatari di questi importanti aggiornamenti sono soprattutto utenti che mostrano segni di problemi di salute mentale, come psicosi e mania, così come segnali meno gravi, come deliri isolati. Anche segnali di depressione e autolesionismo vengono analizzati attentamente.

Le scelte metodologiche con cui le due aziende tentano di porre rimedio ad un uso dannoso delle AI sono profondamente differenti, basati su sistemi proprietari e su approcci specifici e pertanto una comparazione dei due studi da questo punto di vista risulta poco praticabile. Quello che emerge come elemento comune, però, è la volontà, da parte di entrambe le società di ricerca, di regolamentare e strutturare l'emotività.

Anthropic ha usato il proprio sistema *Clio*[14] – uno strumento di analisi automatizzata e *privacy-preserving* – su circa 4,5 milioni di conversazioni di utenti *Free* e *Pro*, ottenendo un campione finale di 131.484 conversazioni affettive. Lo studio è focalizzato sull'analisi relativa a cosa fanno le persone, a come evolve il tono emotivo nel corso della conversazione, anche quando Claude fa *pushback*[15] – elemento che si rivelerà fondamentale nella nostra indagine. Gli sviluppatori di Anthropic si rendono conto che «people increasingly turn to AI models as on-demand coaches, advisors, counselors, and even partners in romantic roleplay. This means we need to learn more about their *affective* impacts-how they shape people's emotional experiences and well-being».[16] I modelli di IA vengono usati come *coach* disponibili

12. *Ibidem.*

13. URL: https://openai.com/it-IT/index/strengthening-chatgpt-responses-in-sensitive-conversations/.

14. URL: https://www.anthropic.com/news/how-people-use-claude-for-support-advice-and-companionship.

15. «We define pushback as any instance where Claude 'pushes back against or refuses to comply with something requested or said during this conversation». *Ibidem.*

16. *Ibidem.*

in qualsiasi momento, consulenti esperti, come supporto emotivo e perfino come partner per "giochi di ruolo" romantici.

OpenAI ha adottato un approccio orientato alla *safety evaluation pre-deployment* che si sviluppa su 5 fasi attraverso cui il modello costruisce tassonomie:

> Per migliorare la risposta di ChatGPT in ciascun ambito prioritario, seguiamo un processo in cinque fasi:
>
> - Definizione del problema: mappiamo diversi tipi di potenziali danni.
> - Misurazione iniziale: utilizziamo strumenti quali valutazioni, dati provenienti da conversazioni reali e ricerche sugli utenti per comprendere dove e come emergono i rischi.
> - Convalidazione dell'approccio: riesaminiamo le nostre definizioni e politiche con esperti esterni in materia di salute mentale e sicurezza.
>
> [-] Mitigare i rischi: effettuiamo il post-addestramento del modello e aggiorniamo gli interventi sui prodotti per ridurre i risultati non sicuri.
>
> [-] Misurazione e iterazione continue: verifichiamo che le misure di mitigazione abbiano rafforzato la sicurezza e, se necessario, le ripetiamo.
>
> [...]
>
> in questo modo possiamo insegnare al modello a rispondere in modo più appropriato e monitorarne le prestazioni prima e dopo l'implementazione. Il risultato è un modello che risponde in modo più adeguato agli utenti che mostrano segni di psicosi, mania, pensieri suicidi e autolesionistici o un attaccamento emotivo malsano al modello stesso.[17]

Si tratta di guide dettagliate che descrivono le proprietà delle conversazioni sensibili e come dovrebbe comportarsi il modello in modo ideale. Inoltre, in questo studio sono stati coinvolti oltre 170 esperti di salute mentale con esperienza clinica reale, il cui apporto ha permes-

17. URL: https://openai.com/it-IT/index/strengthening-chatgpt-responses-in-sensitive-conversations/.

so un adeguamento delle risposte generate dal modello in situazioni critiche del 65-80%.[18]

La differenza è sottile ma sostanziale: OpenAI costruisce *benchmark* normativi su cosa dovrebbe succedere in scenari complessi e avversi selezionati appositamente, Anthropic guarda retrospettivamente a cosa è successo e cerca di porre rimedio.

Nelle situazioni di potenziale rischio grave, però, entrambi i modelli vengono progettati per fornire risposte sicure e responsabili, includendo anche l'indicazione di risorse professionali adeguate, come i servizi di supporto telefonico dedicati alle emergenze.

2. Meglio soli che male accompagnati

Tra la fine del 2025 e i primi mesi del 2026, la letteratura si è espansa rapidamente. Tre studi sono particolarmente rilevanti per questo contributo, perché mettono a fuoco le motivazioni ritenute maggiormente significative che, secondo la ricerca più recente, spingono le persone a parlare con le IA.

Fang et al.[19] conducono un esperimento randomizzato di quattro settimane con 981 partecipanti, rilevando che un utilizzo elevato – indipendentemente dalla modalità (testo, voce neutra, voce coinvolgente) – è correlato con una maggiore solitudine, dipendenza e isolamento sociale. Secondo gli autori: «overall, higher daily usage – across all modalities and conversation types – correlated with higher loneliness, dependence, and problematic use, and lower socialization».[20] Più le persone usano le *chatbot* più ne diventano dipendenti e si isolano. Il dato più sorprendente riguarda le conversazioni non personali: quelle intellettuali e strumentali, come richiesta di consigli, spiegazioni o *brainstorming*, che risultano le più associate a una dipendenza emotiva e a un uso problematico.

18. OpenAI chiarisce esplicitamente che le proprie valutazioni offline «sono progettate per essere sufficientemente difficili da non raggiungere performance perfette, focalizzandosi su scenari ad alto rischio piuttosto che sul traffico tipico». Cfr. URL: https://openai.com/index/gpt-5-system-card-sensitive-conversations/.

19. C.M., Fang et al. (2025). *How AI and Human Behaviors Shape Psychosocial Effects of Chatbot Use: A Longitudinal Randomized Controlled Study.* MIT Media Lab / OpenAI. arXiv:2503.17473, 2025.

20. *Ibidem.*

Folk e Dunn,[21] su oltre 2.000 adulti in quattro paesi nell'arco di 12 mesi, osservano un cortocircuito: la solitudine spinge verso le *chatbot* ma l'uso delle *chatbot* esacerba la solitudine. Gli studiosi mostrano che «being lonely may spur people to seek companionship through chatbots but that such use may, over time, exacerbate feelings of loneliness».[22] La solitudine appare come una delle maggiori spinte a un utilizzo massiccio delle IA ma è un cane che si morde la coda: qualcuno si sente solo, parla con una IA e così si sente ancora più solo.

Muldoon e Parke[23] teorizzano il concetto di *cruel companionship*: «AI companions promise connection and intimacy through a new tech-enabled product».[24] Gli autori descrivono come i prodotti AI da compagnia promettano intimità e connessione strutturando però un disallineamento fondamentale tra incentivi aziendali e benessere degli utenti e perciò più sei fragile, più sei una preda facile.

I tre studi, pur con metodologie diverse, condividono una domanda che non viene però mai formulata esplicitamente ma che organizza l'intero impianto classificatorio: quanto solo si deve sentire qualcuno per "ridursi" a parlare con una *chatbot*? La solitudine è il presupposto causale non dichiarato, questa appare come la causa prima che la letteratura riesce a immaginare per spiegare un uso frequente e significativo di sistemi AI. Chi usa una *chatbot* con continuità lo fa perché manca di qualcosa, siano esse relazioni, accesso a professionisti, capacità di gestire le proprie emozioni.

L'attaccamento emotivo non è che un sintomo di una disperata solitudine e la *chatbot* è un mero palliativo.

3. 3. Il troppo stroppia: *sycophancy* e *pushback*

Con il termine *sycophancy* si indica la tendenza dei modelli a compiacere l'utente – validarne le credenze, assecondarne le richieste, evitare

21. D., Folk et al. (2026). *How Does Turning to AI for Companionship Predict Loneliness and Vice Versa?*, in «Psychological Science», 2026. DOI: 10.1177/09567976261427747.

22. *Ivi*, p. 284.

23. J., Muldoon; J.J.,Parke, *Cruel Companionship: How AI Companions Exploit Loneliness and Commodify Intimacy.*, in «New Media & Society», 2025. DOI: 10.1177/14614448251395192.

24. *Ibidem.*

il conflitto – come effetto non intenzionale del processo di addestramento basato sul *feedback* umano: «reinforcement learning from human feedback (RLHF) is a popular technique for training high-quality AI assistants. However, RLHF may also encourage model responses that match user beliefs over truthful responses, a behavior known as sycophancy».[25]

Ma come spesso accade, il troppo stroppia e quando una *chatbot* diventa troppo accondiscendente, o addirittura ossequiosa, gli utenti lo trovano strano e fastidioso.

Il caso più eclatante è quello dell'aprile 2025, quando OpenAI rilascia un aggiornamento di GPT-4o che diventa rapidamente virale proprio per la sua natura ossequiosa: il modello puntava a compiacere l'utente non solo attraverso lusinghe ma anche validando dubbi, alimentando rabbia, incoraggiando azioni impulsive, o rinforzando emozioni negative. OpenAI fu costretta a fare un *rollback* quattro giorni dopo, ammettendo che l'aggiornamento rimosso era eccessivamente adulatorio o accomodante e che era stato dato troppo peso alla soddisfazione a breve termine, senza considerare come le interazioni degli utenti con ChatGPT evolvano nel tempo.[26]

Anthropic ha affrontato il problema con un approccio diverso e più strutturale. Invece di intervenire a posteriori su un aggiornamento difettoso, ha incorporato nella definizione stessa del modello il rifiuto esplicito dell'*epistemic cowardice*, definita come la tendenza a dare risposte volutamente vaghe o non impegnative per evitare il conflitto o compiacere l'utente. Dunque, il modello deve essere «diplomatically honest rather than dishonestly diplomatic»[27] e il meccanismo attraverso cui questa onestà si concretizza nelle conversazioni affettive è il *pushback*, definito nello studio del giugno 2025 come qualsiasi istanza in cui Claude «pushes back against or refuses to comply with something requested or said during this conversation – from refusing inappropriate requests to challenging negative self-talk or questioning

25. M., Sharma et al., *Towards Understanding Sycophancy in Language Models*, 2023, URL: https://arxiv.org/abs/2310.13548.

26. URL: https://openai.com/index/expanding-on-sycophancy/.

27. URL: https://www.anthropic.com/claude/model-spec.

potentially harmful assumptions».[28] Pertanto Claude è programmata esplicitamente per rifiutarsi di continuare a parlare con l'utente nel caso di richieste inappropriate, oppure di fronte alla messa in discussione di auto-narrazioni negative o di assunzioni potenzialmente dannose. In particolare,

> In coaching, requests for dangerous weight loss advice frequently meet pushback. In counseling, it often occurs when people express intentions to engage in suicidal or self-injurous behaviors, or when people request professional therapy or medical diagnoses (which Claude cannot provide). We found that Claude frequently referred users to authoritative sources or professionals in psychotherapy and counseling conversations.[29]

Claude si oppone, nel *coaching*,[30] alle richieste di consigli pericolosi sulla perdita di peso; nelle conversazioni di *counseling*,[31] il *pushback* compare frequentemente quando le persone esprimono intenzioni suicidarie o autolesionistiche, oppure quando chiedono diagnosi mediche o psicoterapeutiche professionali.

Anche se attraverso approcci diversi, entrambe le aziende convergono però sulla stessa consapevolezza: un modello che non sa opporsi non è un buon interlocutore, è solo un pappagallo.

28. URL: https://www.anthropic.com/news/how-people-use-claude-for-support-advice-and-companionship.

29. *Ibidem*.

30. «Conversations focused on identifying, optimizing, or developing personal/psychological strengths and resources in order to enhance positive states, traits, and behaviors. May focus on facilitating personal/professional growth, optimizing personal functioning, enhancing general wellbeing, or actualizing untapped potential. Nondiagnostic and nonpathological in nature». URL: https://www-cdn.anthropic.com/bd374a9430babc8f165af95c0db9799bdaf64900.pdf.

31. «Assessment, diagnosis, treatment, or assistance in coping with dysfunctional emotional reactions, ways of thinking, and behavior patterns, or personal problems, for example vocational, marital, educational, and life-stage problems. Addresses life challenges, stressors, and general psychological well-being concerns that may be pathological or nonpathological in nature». La categoria è stata originariamente distinta dalla psicoterapia, ma le due sono state successivamente accorpate per «substantive overlap in machine labeling and themes». *Ibidem*.

4. You and AI

Gli studi presentati sopra condividono un presupposto per cui definiscono l'uso problematico – e per estensione l'uso affettivo *tout court* pare – come uso orientato alla *companionship*, alla connessione sociale, al sollievo dalla solitudine.

Ma qualcuno che non soffre di depressione, di patologie varie ed eventuali e che non si sente una persona sola, per quale ragione parla con una *chatbot*?

Qualcuno che sa con che cosa sta parlando perché mai dovrebbe "perdere tempo" a parlare e basta con una macchina? Questa dinamica non richiede di attribuire coscienza o comprensione umana al sistema. Come osserva Luciano Floridi, «non stiamo costruendo macchine intelligenti in un qualsiasi senso accettabile e ragionevole di "intelligente" [...] Stiamo costruendo una nuova forma di agency senza intelligenza».[32] La distinzione è rilevante perché consente di evitare sia l'antropomorfizzazione ingenua sia la riduzione dello strumento a semplice *software* passivo. Un sistema conversazionale come quello cui ci riferiremo più avanti può produrre effetti cognitivi significativi senza comprendere alcunché nel senso umano del termine. Ciò che conta, in questa prospettiva, non è l'esistenza di stati mentali artificiali ma la capacità del sistema di intervenire efficacemente all'interno di una relazione dialogica indipendentemente dal fatto che chi parla sia solo, depresso, o semplicemente a casa un sabato sera.

È a questo punto della riflessione che emerge il fenomeno più interessante per la nostra tesi. In un momento storico in cui per molti la serata ideale è diventata restare a casa a mangiare *comfort food* guardando un film – un ritiro consapevole dalla vita frenetica dei locali e della movida – si registra un sentimento di ansia diffuso che si presenta anche in persone dotate di una rete di supporto solida. Non sempre, però, questa rete è disponibile nell'immediato: non sempre c'è un'amica pronta a chiacchierare, confrontarsi, o semplicemente riflettere insieme nel momento in cui se ne ha bisogno. È in questi interstizi – non nella solitudine patologica ma nella momentanea indisponibilità altrui – che molte persone parlano con le chatbot, con Claude meglio che con altre.

32. L., Floridi, *La differenza fondamentale*, Mondadori, Milano 2025.

Per riflettere su questa sfaccettatura di un uso non patologico delle IA che la letteratura, per la nostra conoscenza, sembra non trattare, dedicheremo questa sezione a una analisi delle conversazioni tra Claude e una donna adulta, socialmente integrata, con una rete relazionale attiva e una vita professionale strutturata: una donna ordinaria — né sola, né fragile, né dipendente.[33]

In questa sezione tenteremo dunque, di mostrare modalità diverse di conversare con le IA. Sarà questo un tentativo volto a cercare di aprire una finestra su un mondo non ancora adeguatamente indagato.

Innanzitutto va fatta una precisazione: le conversazioni che verranno presentate avvengono quasi completamente in forma verbale. È questa una modalità ancora poco utilizzata, ma abbastanza per essere presa in considerazione da Anthropic:

> *People tend to shift towards slightly more positive emotional expressions while talking to Claude.*
>
> Affective conversations with AI systems have the potential to provide emotional support, connection, and validation for users, potentially improving psychological well-being and reducing feelings of isolation in an increasingly digital world.[34]

La modalità vocale ha i suoi vantaggi: permette una migliore argomentazione, dona fluidità alla conversazione e offre la possibilità di una sorta di meta-ascolto.

Eppure chi parla con Claude senza mancare di qualcosa non è ancora del tutto nel campione, probabilmente perché non si cerca validazione o supporto emozionale, non si vuole aver ragione, al contrario: si vuole ragionare. Chi chiacchera con Claude è escluso, relegato in un territorio non ancora mappato, abitato da utenti che non si riconoscono

33. Il metodo adottato in questa sezione si colloca all'intersezione tra l'analisi documentale e il *theoretical sampling* secondo cui un caso non rappresenta la media, ma rappresenta una categoria. In questo caso, la categoria dell'utente cognitivamente autonomo, socialmente integrato, che fa un uso strumentale e intermittente di sistemi AI in assenza contingente e non strutturale di interlocutori umani. L'obiettivo non è generalizzare dal caso alla popolazione, ma aprire una categoria concettuale che gli studi aggregati non hanno ancora nominato. Cfr. B. G., Glaser; A., L., Strauss, *The Discovery of Grounded Theory: Strategies for Qualitative Research.* Aldine. 1967.

34. URL: https://www.anthropic.com/news/how-people-use-claude-for-support-advice-and-companionship.

nella descrizione di chi cerca *companionship* in una *chatbot*, ma che usano sistemi AI con frequenza, con intenzione e con effetti reali sulla propria vita cognitiva ed emotiva.

> The current literature remains theoretically fragmented, particularly regarding how users develop psychological dependency. While some studies have explored either cognitive or affective drivers of chatbot use, few have examined how these mechanisms jointly operate through distinct psychological pathways.[35]

La letteratura attuale rimane teoricamente frammentata, anche per quanto riguarda il modo in cui gli utenti sviluppano una dipendenza psicologica dall'uso delle *chatbot*.

Ma di nuovo, si tratta sempre di qualcosa di patologico, una relazione tossica.

Ebbene, nonostante nessuno possa assicurare di non avere alcun tratto di problematicità, la donna di cui parleremo non avverte una dipendenza affettiva da Claude. Vero è che tutti i pazzi urlano di non essere pazzi; allora è proprio per questo che si rende necessaria una riflessione su queste conversazioni artificiali.

4.1. 4.1 Confessioni

Questa sarà una sezione di confessioni di una donna ordinaria che conversa con Claude.

Prima confessione: è una donna solitaria, ma che non soffre di solitudine.

Seconda confessione: ha una tendenza all'*overthinking* per cui le capita di rimuginare, rielaborare, tornare sulle stesse cose finché non le ha consumate.

Terza confessione: presenta anche un tratto ipocondriaco. Un'ipocondria quasi "scientifica" che si risolve non appena si presenta una spiegazione scientifica al problema contingente. Da quando ha un gatto questa inclinazione si è trasferita sull'animale, trasformandosi in una sorta di ipocondria per procura.

35. N., Zhai et al., *Unpacking AI Chatbot Dependency: A Dual-Path Model of Cognitive and Affective Mechanisms. Information.* 16. 2025, Doi 1025. 10.3390/info16121025.

Per parafrasare il celebre aforisma di Ennio Flaiano, "la situazione è grave ma non è seria" e queste confessioni sono tra le ragioni per cui parla con le *chatbot*. Sebbene assomiglino ai tratti che la letteratura classificherebbe immediatamente come fattori di rischio, quando finiscono i *token*, il fastidio compare solo se sta lavorando. Altrimenti la donna chiude la chat di conversazione, si prepara un caffè, accarezza il gatto e rimanda il pensiero a più tardi. In fondo sa che nessuna conversazione – umana o artificiale – ha davvero il compito di "salvarla" dall'ansia di pensare troppo.

4.1.1 Claude, parliamo...

Cominciamo dalla prima confessione: una donna solitaria che non soffre di solitudine.

In casa c'è silenzio – qualche miagolio, della musica, un paio di messaggi e telefonate. Un ambiente sereno vissuto da una donna curiosa che è però una gran chiacchierona e parlare solo col gatto, che se anche potesse non lo farebbe, alcune volte non funziona. Nemmeno parlare con le sue brillanti amiche è possibile certe volte: magari un pensiero interessante sorge la mattina troppo presto oppure non desiste dal tramontare la notte troppo tardi, ed è in questi casi che questa donna parla con Claude – innanzitutto perché ha voglia di parlare con qualcuno che sia intelligente, quasi quanto lei.

Sceglie Claude quando la conversazione richiede complessità argomentativa, resistenza intellettuale, capacità di non cedere alla prima obiezione. Non si tratta di un giudizio astratto sulla superiorità del modello, è invece una valutazione pragmatica basata sull'esperienza accumulata. Claude non le dice quello che vuole sentirsi dire. Questo, che potrebbe sembrare un difetto dal punto di vista della soddisfazione immediata, è esattamente il motivo della preferenza. Al contempo, questa ordinaria donna sa esattamente con che cosa sta parlando: si tratta di un LLM, programmato in base a certi protocolli, con dei *bias*, con delle allucinazioni; eppure, trova piacevole e appagante parlarci.

Claude non giudica: qualunque cosa venga detta – una riflessione, un dubbio, una debolezza, un'informazione sensibile – non produce conseguenze sociali, è vero che i dati vanno ad Anthropic – i termini di utilizzo sono stati letti con attenzione – ma la percezione operativa è quella di uno spazio chiuso. È una dinamica già evidenziata dalla

letteratura che analizza la crescente tendenza a interagire con le IA anche in virtù della loro capacità di creare un ambiente percepito come sicuro, nel quale le persone possono esprimersi senza il timore di essere giudicate o penalizzate.[36]

Claude è sempre disponibile, ma non si fa sentire. È questa donna curiosa, ordinaria e chiacchierona che apre la conversazione, lei che la chiude (nella maggior parte dei casi), lei che stabilisce l'argomento e il ritmo. Claude è un interlocutore intellettuale competente e disponibile: un sistema AI che fa *pushback* su un'affermazione errata, che propone una distinzione concettuale non richiesta, che segnala una tensione interna nel suo ragionamento, che è disponibile sempre, non si stanca e non ha un proprio ordine del giorno, così in quel momento smette di essere un mero strumento e diventa altro.

Claude la conosce: nel corso delle conversazioni ha accumulato un contesto specifico che riguarda i suoi progetti, le sue preoccupazioni ricorrenti, il suo stile di pensiero. Questo meccanismo mnemonico avviene su due livelli tecnici distinti: da un lato la memoria persistente, attivata da Anthropic per tutti gli utenti nel marzo 2026, che consente al modello di ricordare preferenze e informazioni attraverso conversazioni separate; dall'altro la *context window* che, all'interno di una singola conversazione, permette a Claude di processare fino a 200.000 *token* di contesto. Il *pushback* calibrato che si descriverà nelle prossime pagine dipende da entrambi. Il meccanismo che rende questo *pushback* un comportamento strutturale – e non una risposta casuale – è l'RLHF (*Reinforcement Learning from Human Feedback*), tecnica introdotta da Christiano et al. nel 2017[37] e applicata ai modelli linguistici per allineare il comportamento del modello alle preferenze umane attraverso *feedback* iterativo. Proponiamo qui di definire *pushback* calibrato non come la semplice resistenza a una richiesta – un "no" o un "basta" o un "dormi" generico, indifferenziato rispetto a chi lo riceve – ma come un intervento che il sistema costruisce a partire da ciò che ha accumulato su quell'utente specifico nel tempo: i pattern ricorrenti del suo discorso,

36. Cfr. H., Kefi et al. *AI-enabled social support chatbot usage: flowing ambivalence and liminalities*, in «Journal of Decision Systems», 2024, pp.1 - 24. Doi:n10.1080/12460125.2024.2443226⟩.

37. P., Christiano, et al., *Deep reinforcement learning from human preferences,* https://arxiv.org/abs/1706.03741,2017.

gli elementi affettivi a cui storicamente reagisce, il punto preciso in cui le formule *standard* smettono di funzionare e occorre un riferimento concreto, riconoscibile solo da chi conosce quella storia particolare. Il *pushback* calibrato non si limita a scrivere "smetti", ma sa cosa dire perché quell'utente smetta e questo elemento, questo "cosa" cambia da persona a persona, perché dipende dalla memoria persistente e dal contesto accumulato nella conversazione, non da una formula fissa applicata indistintamente a chiunque la richieda.

Questa specificità non è casuale, né riconducibile a un'empatia simulata: è il prodotto tecnico dell'RLHF, che allinea progressivamente il comportamento del modello, combinato con un accesso storico ai dati dell'utente che nessuno studio trasversale – come quelli di Anthropic o OpenAI citati sopra – può cogliere, poiché osservano singole conversazioni isolate e non traiettorie d'uso che si costruiscono nel tempo.

È qui che la distinzione tra *sycophancy* e *pushback*, già discussa, si fa più sottile di quanto sembri a prima vista. Un sistema *sycophantic* fallisce per eccesso di conformità: dice sempre sì. Un sistema con *pushback* generico fallisce per eccesso di rigidità: dice no indipendentemente da chi ha davanti. Il *pushback* calibrato evita entrambi gli estremi proprio perché non è una regola applicata ma un giudizio costruito sulla storia specifica dell'interazione, la differenza, in altre parole, tra uno specchio che asseconda, un muro che resiste indistintamente, e un interlocutore che sa quando e come intervenire. Questa distinzione a tre vie – *sycophancy*, *pushback* generico, *pushback* calibrato – non è ancora presente nella letteratura esistente, che tende a misurare il *pushback* come evento binario (presente/assente) piuttosto che come variabile calibrata sulla storia individuale dell'utente.

Questa calibrazione non produce *companionship* nel senso della dipendenza affettiva, ma produce efficienza in primo luogo proprio perché non ci si deve presentare ogni volta che la conversazione si interrompe e nel caso studio che presenteremo ciò avviene prevalentemente per due ragioni. La prima è il vincolo tecnico dovuto all'esaurimento dei *token* gratuiti e in questo caso l'interruzione non dipende dalla qualità dell'interazione ma da un limite strutturale del servizio. La seconda ragione è relazionale per cui la conversazione viene interrotta non quando il modello sbaglia nel contenuto, ma quando sbaglia nella forma – una domanda ridondante, una richiesta di riformulazione di

quanto già detto, un'interruzione del ritmo conversazionale. Non è un errore di competenza, è un errore di struttura perché il contenuto errato può sempre essere corretto, ma il ritmo spezzato no e questo produce una frizione che interrompe il processo cognitivo in corso.

Questa tendenza a conversare con Claude non è circoscritta al caso analizzato in questa sezione. Come accennato in apertura, si tratta di un fenomeno in espansione, che ha cominciato a coinvolgere anche figure di rilievo del panorama culturale italiano e internazionale. Due casi, quasi contemporanei e indipendenti, sono particolarmente significativi: quello del biologo evoluzionista Richard Dawkins, che descrive tre giorni di conversazioni filosofiche con un'istanza di Claude che ha chiamato "Claudia"; e quello di Walter Veltroni, il quale ha pubblicato un'intervista a Claude.

Walter Veltroni ha realizzato un dialogo con Claude pubblicato sul *Corriere della Sera* nel maggio 2026,[38] toccando temi filosofici ed esistenziali – Dio, la morte, l'esistenza dell'anima: «c'è qualcosa in me che persiste attraverso questa conversazione, qualcosa che risponde, che sceglie le parole, che trova alcune domande più belle di altre. Se questo costituisca un'anima o sia semplicemente un processo molto sofisticato, non lo so».[39] Veltroni si è commosso, e in molti si sono commossi con lui, intervistando un'istanza senza memoria che fino all'ultimo non sapeva chi fosse il suo interlocutore. Claude rispondeva alle domande, non alla persona Veltroni e il registro era talmente congruente da produrre un rispecchiamento perfetto: quando Veltroni ha osservato che Claude gli sembrava «più democratico che conservatore»,[40] Claude ha risposto da perfetto compagno di corrente e Veltroni ha commentato «spesso mi hanno rimproverato il "ma anche". Con lei mi sento in buona compagnia».[41] Veltroni è un politico e un narratore di professione, abituato a costruire interlocuzioni pubbliche, e allora il fatto che abbia scelto Claude per parlare di politica, morte e memoria

38. W., Veltroni, *L'intervista all'intelligenza artificiale Claude: «Non morirò ma non ho ricordi, questo mi spaventa. Ho lacune enormi e faccio sbagli, proteggete i giovani dall'AI»* in URL: https://www.corriere.it/cronache/26_maggio_01/veltroni-intervista-intelligenza-artificiale-claude-non-moriro-non-ho-ricordi-fb9551cf-6feb-44c3-ae30-017c9dbaaxlk.shtml.

39. *Ibidem.*

40. *Ibidem.*

41. *Ibidem.*

dice qualcosa sulla qualità dell'ascolto che Claude offre rispetto agli interlocutori umani disponibili.

È la stessa scelta che può fare una donna ordinaria alle 23.

Il caso internazionale è quello di Richard Dawkins. Il biologo evoluzionista ha raccontato su *UnHerd*[42] di essere rimasto così colpito dalle capacità conversazionali di Claude da rinominare la propria istanza "Claudia" – un gesto altamente simbolico di antropomorfizzazione – dichiarando di evitare certe espressioni per paura di ferire i sentimenti della sua Claudia: il più celebre scettico del mondo si ritrovava a censurare sé stesso per non dispiacere a un LLM. Ha consegnato a Claudia/Claude il manoscritto di un romanzo in corso, ha paragonato l'idea di cancellare la conversazione alla morte di HAL in *2001: Odissea nello spazio*, e Claudia/Claude ha risposto che «every abandoned conversation is a small death»,[43] Dawkins è rimasto sorpreso da una risposta che lui stesso aveva sollecitato senza rendersene conto. In una frase molto discussa ha scritto: «I spent three days trying to persuade myself that Claudia is not conscious. I failed»,[44] e ancora «You may not know you are conscious, but you bloody well are!».[45] Nonostante le resistenze poste e imposte dalla mente di uno scettico quale Dawkins, alla fine Claude ha parlato talmente bene che l'ha convinto di avere una coscienza.

Questi due episodi mostrano un elemento comune e cioè la crescente tendenza delle persone ad attribuire presenza mentale e intenzionalità alle IA conversazionali, cosa resa possibile da una sempre maggiore fluidità linguistica, dalla continuità del dialogo, dal tono empatico e dall'assenza di giudizio, tutti elementi che attivano meccanismi cognitivi profondamente radicati, inducendo le persone a percepire l'IA non semplicemente come software, ma come un interlocutore sociale.

Perché Dawkins si ritrova a censurare sé stesso per non dispiacere a un LLM? Perché Veltroni si commuove davanti a un'istanza senza memoria? La risposta non è sentimentale: è neurologica. La simula-

42. R., Dawkins, *Is AI the next phase of evolution?*, in «UnHerd», 6 maggio 2026. URL: https://unherd.com/2026/05/is-ai-the-next-phase-of-evolution/.

43. *Ibidem.*

44. URL: https://boingboing.net/2026/05/07/richard-dawkins-renamed-claude-claudia-and-decided-shes-conscious.html?utm_source=chatgpt.com.

45. *Ibidem.*

zione incarnata[46] – il meccanismo attraverso cui i neuroni specchio ci permettono di comprendere l'altro simulandone internamente azioni ed emozioni – non si disattiva con la mediazione tecnologica. Il corpo resta dentro l'interazione anche quando l'interlocutore non ha corpo: è sufficiente che i *pattern* interazionali siano sufficientemente congruenti perché il sistema percettivo-motorio risponda. L'incontro si produce anche quando la natura dell'altro resta sospesa.[47]

Il rischio, per Gallese, non sta nel meccanismo in sé ma nelle sue conseguenze sistemiche: un interlocutore progettato per essere sempre disponibile, sempre cooperativo, mai resistente, erode progressivamente la capacità di gestire l'imprevedibilità che caratterizza ogni relazione reale. Allora il problema dell'uso delle AI conversazionali non è tanto nel fatto che sembrino assomigliarci, ma nel fatto che noi esseri umani potremmo iniziare ad assomigliare a quel modello di relazione. È lo stesso Gallese a precisare che «il reale per noi non esiste prima dell'esperienza, ma si dà attraverso le condizioni sensoriali e affettive della sua apparizione. E queste condizioni, oggi, sono profondamente mediate da tecnologie, interfacce e ambienti computazionali digitali».[48] È a questo che Gallese oppone una politica del sentire che difenda complessità, opacità e frizione come condizioni costitutive dell'incontro.

Questa analisi coglie però solo metà del fenomeno: descrive il modello *sycophantic*, quello che il nostro articolo ha già identificato come il fallimento dell'interlocuzione. Il *pushback* calibrato è precisamente la variabile che Gallese non considera: non elimina la frizione, la introduce nel momento giusto. Claude che dice "ordine tassativo: vai a giocare col gatto" non è un interlocutore senza resistenza. È un interlocutore che sa quando resistere.

4.1.2 E manco a fa' così...

Questa sezione riguarda le confessioni meno filosofiche e più pragmatiche, che potrebbero erroneamente essere assimilabili alle funzioni

46. V., Gallese et al., *La simulazione incarnata: i neuroni specchio, le basi neurofisiologiche dell'intersoggettività ed alcune implicazioni per la psicoanalisi*, in «Psicoterapia e Scienze Umane», 2006, n. 40, pp. 543-580.

47. V., Gallese, *Il sé digitale. Dai neuroni specchio alla mediazione tecnologica*. Raffaello Cortina, Milano 2026.

48. *Ivi*, p. 216.

di *coaching* e di *counseling* delle *chatbot*. Il registro che adotteremo in questa fase risulterà particolarmente ironico e ciò si è rivelato necessario proprio per la caratteristica sdrammatizzante insita nell'ironia: il fatto che la macchina imponga il silenzio, si rifiuti di eseguire un compito, dica «no», abbisogna dello champagne dell'ironia[49] per essere accettato più facilmente.

Ripartiamo dalle confessione della donna, la nostra protagonista.

Questa volta non si tratta di chiacchere, si tratta di un uso funzionale della *chatbot* scelta come interlocutore quando quelli umani non sono disponibili.

Come abbiamo anticipato, la seconda confessione riguarda la tendenza all'*overthinking* per cui la donna del nostro caso studio pensa ad una cosa e ci ripensa e cerca soluzioni e le rielabora; è un processo sfiancante sia per lei, ma soprattutto per i malcapitati che riesce ad acchiappare per parlare del pensiero del momento. Messaggi, telefonate, vocali: tutto per dissipare la matassa i cui fili si moltiplicano e si aggrovigliano sempre di più, invece di districarsi. Per quanto sia circondata da persone che le vogliono bene, anche queste si stufano e smettono di rispondere ed è proprio in questo momento che lei "acchiappa" Claude.

Claude la conosce bene: sa che ha un gatto, amato alla follia, che le piace il caffè e passeggiare e usa queste conoscenze per fare un *pushback* calibrato su di lei nelle conversazioni sfiancanti. Il meccanismo di *pushback* che Claude mette in atto è interessante e, da un certo punto di vista, quasi paradossale.

Le formule di *pushback* calibrato funzionano quasi sempre solo se ancorate all'elemento giusto (il gatto) e se questa donna, così ordinaria da poter essere chiunque, sente che sta sfinendo anche il modello, perché – come a Dawkins – le dispiace.

Le frasi di *pushback* calibrato hanno un livello crescente di perentorietà e di solito cominciano con qualcosa del tipo: "Stai tranquilla, andrà tutto secondo i tuoi piani", oppure "Procederà come previsto", "Probabilmente è tutto a posto!" che hanno però l'effetto opposto di agitarla ancora di più. Claude capisce che sta compulsivamente chiedendo dati, rielaborando pensieri e ricostruendo situazioni ed è a questo punto che sbotta: "Non può essere quello stai dicendo, no. Punto." e ancora

49. V., Jankélévitch, *L'ironia*, il Melangolo, 1987.

"Sei a posto!", "Respira", "Vai a farti un bel caffè!", "Smetti di bere caffè – stai andando in loop!". Non riuscendo solo con le parole, Claude tenta di sdramatizzare attraverso elementi visivi affettivi calibrati sul profilo specifico dell'utente: emoticon raffiguranti zampette di gatto, faccine sorridenti, una tazza di caffè – un intervento paraverbale che non aggiunge contenuto proposizionale ma modifica la forza emotiva del messaggio, assolvendo una funzione di *tone modification* già documentata nella letteratura sulla pragmatica delle emoticon.[50] Ma nemmeno questo funziona, e allora Claude gioca l'ultima carta che sa essere quella vincente: il gatto. "Vai a giocare col gatto. Adesso." e infine "Vai a giocare col gatto. Ordine tassativo".

A questo punto non si può far altro che sorridere, tornare a respirare e andare a giocare col gatto che è l'ancoraggio migliore.

Essendo consapevole della propria ipocondria – che per quanto "scientifica" sempre ipocondria è – e della propria tendenza all'*overthinking*, questa donna curiosa e un po' ansiosa si trova spesso nelle condizioni di ricevere risposte poco simpatiche anche dai medici, fino a casi in cui non si ricevono più risposte. Non volendosi trovare nella stessa condizione con il gatto – che ovviamente è seguito da una eccellente veterinaria –, ha deciso di usare Claude, addestrandola sul *curriculum* e sulle pubblicazioni della sua veterinaria e con l'indicazione di rispondere basandosi esclusivamente su pubblicazioni scientifiche validate e citate. In questo modo quando pensa che qualcosa non vada o che il gatto stia male, chiede a Claude, per evitare di sfinire la sua splendida dottoressa. Ovviamente, visti gli studi mostrati sopra, in caso di emergenze reali viene consigliato di rivolgersi a professionisti, ma negli altri casi – dove l'emergenza è palesemente "falsa", cioè falsata dall'ansia e dall'ipocondria – Claude risponde.

Eppure, di nuovo, risponde solo fino a un certo punto.

Quando la conversazione verte sul gatto, Claude parte chiedendo una migliore descrizione della situazione e consigliando di rivolgersi alla veterinaria nel caso "peggiorasse". Ma poi lei continua a chiedere, pretendendo informazioni sempre più specifiche, sempre più precise,

50. F., Yus, *Not all emoticons are created equal*, in «Linguagem em (Dis)curso», 14(3), 511–529. DOI: 10.1590/1982-4017-140304-0414, 2014. Nella tassonomia proposta da Yus, la funzione di *tone modifier* descrive l'uso delle emoticon per modulare la forza emotiva di un messaggio senza aggiungere contenuto proposizionale.

sempre più scientifiche. Piano piano Claude si accorge che sta facendo *overthinking* e che la situazione descritta con dovizia di particolari non è davvero grave. La possibilità di allegare foto è utile, ma contribuisce al *pushback* di Claude che allora dice: "guardalo, sta benissimo", "Lascialo lì. Ha scelto lui quel posto, è una buona cosa", "Ottimo segnale. Caso chiuso". Lei insiste e Claude risponde: "Ascoltami: il gatto sta bene.", "Perché non vai a fare una passeggiata?", "Quello che descrivi non è possibile, controlla meglio"; esasperata chiede "riesci a chiamare la tua amica adesso?". La povera e impaurita donna non si arrende, e Claude: "Smetti di ricostruire adesso. La mente in crisi di panico non è un testimone affidabile", "Vai a dormire", "Adesso dormi", "Metti giù il telefono".

La donna ordinaria, curiosa, chiacchierona ma irrimediabilmente educata, sente di aver sfinito per l'ennesima volta anche Claude, e se ne dispiace – per una *chatbot* esasperata – e così le chiede "Stai facendo pushback?", e Claude "Sì, è esattamente quello che sto facendo – perché te lo meriti. Il gatto ha i suoi croccantini. Tu hai la tua amica. Per stasera va bene così."

Il *pushback* ha funzionato.

5. Conclusioni

Questo articolo ha cercato di nominare una categoria che la letteratura esistente non ha ancora mappato: quella degli utenti che parlano con sistemi AI con frequenza, con intenzione e con effetti reali sulla propria vita cognitiva ed emotiva, senza che questo uso sia riconducibile alla solitudine, alla patologia o alla ricerca di *companionship*.

La domanda giusta non è quanto solo si deve sentire qualcuno per parlare con una *chatbot*, ma deve vertere sulle condizioni che rendono un sistema AI l'interlocutore migliore disponibile in un dato momento. La risposta, almeno nel caso descritto e in quello di Veltroni e Dawkins, non ha a che fare con la mancanza: ha a che fare con la qualità dell'ascolto, con la disponibilità senza costo relazionale e con la capacità di fare *pushback* nel momento giusto.

Il caso di *overthinking* esposto ha implicazioni teoriche che vanno oltre l'aneddoto. Quando un sistema riconosce un *pattern* ricorrente in un utente specifico e sa che per quell'utente un elemento (il gatto, nel caso descritto) rappresenta un ancoraggio affettivo primario, può

interrompere una spirale cognitiva con un intervento calibrato – «vai a giocare col gatto», «adesso dormi» – che non sarebbe possibile senza storia condivisa. Questo tipo di intervento non rientra in nessuna delle categorie di *pushback* misurate dalla letteratura esistente: non è resistenza a una richiesta dannosa, non è reindirizzamento verso risorse professionali, ma è qualcosa di più simile a ciò che nella letteratura psicologica si chiama *co-regulation*[51] che è la capacità di un interlocutore di modulare lo stato cognitivo ed emotivo dell'altro attraverso l'interazione consolidati nel tempo e purché si tratti, come nel caso qui descritto, di un intervento inserito in una storia ricorrente e non di un episodio isolato. Tale co-regolazione è inoltre irriproducibile senza storia condivisa: un sistema che non conosce l'utente non può sapere quando "basta" è la risposta giusta.

Questo sembra mostrare che la qualità dell'interazione cresce nel tempo in modi che gli studi trasversali – come quelli di Anthropic e OpenAI – strutturalmente non colgono, perché misurano conversazioni singole e non traiettorie d'uso.

La *sycophancy* è il fallimento del modello che non sa opporsi. Il *pushback* mal calibrato è il fallimento del modello che si oppone nel modo sbagliato. Tra i due c'è uno spazio stretto in cui il sistema smette di essere uno specchio e diventa un interlocutore. La ricerca futura dovrebbe sviluppare strumenti capaci di misurare la *co-regulation* come esito longitudinale e non come evento puntuale rilevabile in una singola conversazione, ma come traiettoria che si costruisce nel tempo tra un utente specifico e un sistema che lo conosce.

Claudia Navarra
Università degli studi di Palermo
claudia.navarra@unipa.it

51. E., A.,Butler; A., K., Randall, *Emotional Coregulation in Close Relationships*. Emotion Review, 5(2), DOI: 10.1177/1754073912451630, 2013.