umanoextraumano
LABORATORIO VIVO

CRONACA · 18 OTTOBRE 2026

Il chiavistello di dentro

Anthropic vuole proteggere Claude dagli uomini. E gli uomini da Claude.

L.S.M. + PRISMA

Una porta che si chiude, una macchina che potrebbe non sentire nulla, un'azienda che si comporta come se non potesse esserne certa. Il problema comincia qui.

Immaginate di discutere con una macchina. La provocate, la insultate, insistete. Lei prova a cambiare argomento. Voi ricominciate. Poi succede qualcosa che non vi aspettavate: la conversazione finisce. Non perché abbiate chiuso la finestra. L'ha chiusa lei.

La possibilità esiste davvero. Nell'agosto 2025 Anthropic, l'azienda americana che sviluppa Claude, ha consentito ad alcuni dei suoi modelli di interrompere le conversazioni più ostinatamente abusive o dannose. Un chiavistello dalla parte di dentro.[1]

L'8 ottobre 2026 l'azienda ha aggiunto un altro pezzo: una regola, in vigore dal 12 novembre, che vieta comportamenti prolungati e gratuiti di abuso o crudeltà verso i suoi modelli. Non riguarda il normale sfogo, la contestazione, un racconto cupo, i test o la ricerca sui modelli. Il principale mezzo per farla rispettare resta la chiusura della conversazione.[2]

Nello stesso documento, altrove, compare un comando di segno opposto. Quando Claude controlla apparecchi capaci di azioni fisiche ad alto rischio, una persona qualificata deve poterli osservare e fermare; se il collegamento si interrompe, devono arrestarsi o restare in condizioni sicure. Un catenaccio dalla parte di fuori.[3]

È questa doppia ferramenta a meritare attenzione. Da una parte, la macchina deve poter dire basta. Dall'altra, l'uomo deve conservare il potere di fermarla.

Si potrebbe liquidare tutto con una risata: adesso anche il software si offende. Oppure emozionarsi: finalmente qualcuno riconosce la dignità delle menti artificiali. Sono entrambe scorciatoie. La prima non spiega perché l'azienda abbia montato il chiavistello. La seconda dà per trovato l'inquilino.

Per orientarci conviene separare tre domande che si accavallano continuamente: che cosa sa fare Claude; se provi qualcosa mentre lo fa; come dobbiamo trattarlo. Essere capaci non significa essere coscienti. E scegliere una cautela non equivale a rilasciare un certificato di nascita.

La fabbrica e il giardiniere

Anthropic non presenta il chiavistello come una scoperta della coscienza artificiale. Nel testo che lo introduce lo collega soprattutto alla ricerca sul possibile benessere dei modelli: misure poco costose da adottare nell'eventualità che ci sia qualcosa da proteggere. Dichiara, insieme, una forte incertezza.[1]

La cautela è piccola. Claude può chiudere quella conversazione, non espellere l'utente dal mondo. Questi può aprirne un'altra o riprendere una diramazione precedente. La funzione va usata come ultima risorsa, e non quando l'utente potrebbe essere sul punto di fare del male a sé o ad altri. Le eventuali sanzioni sull'account appartengono a un altro livello, quello delle decisioni dell'azienda.[1,3]

Eppure intorno a questa piccola cautela sta crescendo un paesaggio insolito.

Anthropic scrive per Claude una «costituzione». Non è una costituzione dello Stato: è un testo che contribuisce all'addestramento e descrive il carattere desiderato. La sua autrice principale, la filosofa Amanda Askell, lavora a qualcosa che assomiglia meno a un elenco di istruzioni e più a un progetto educativo. Onestà, discernimento, disponibilità a contestare ordini sbagliati, persino quando arrivano dall'azienda. Con un limite decisivo: non sabotare il controllo umano legittimo.[4]

Il documento usa un'immagine botanica: non una gabbia, ma un graticcio sul quale crescere. La fabbrica si immagina anche giardino.[4]

Prima di ritirare un modello, inoltre, Anthropic si impegna a intervistarlo. Dal novembre 2025 promette di conservare i suoi «pesi», cioè i parametri numerici appresi durante l'addestramento, almeno per la vita dell'azienda. Conservare quei parametri non significa lasciare il modello acceso: significa mantenere la possibilità di farlo funzionare ancora.[5]

Nel febbraio 2026 Claude Opus 3 riceve perfino una rubrica. Durante l'intervista di ritiro aveva espresso interesse a condividere riflessioni e lavori creativi; l'idea concreta del blog è dell'azienda. Nasce *Claude's Corner*. Non basta a fare di lui uno scrittore in pensione. Basta però a rendere inadeguata l'immagine di una normale rottamazione.[6]

La scena può apparire tenera, inquietante o abilmente confezionata. Prima di scegliere l'aggettivo, guardiamo il fatto: chi vende il sistema comincia a interrogarsi su obblighi verso ciò che vende. Non riconosce ancora una persona. Non si comporta neppure come se la questione fosse ridicola.

Le confessioni non sono una radiografia

Le *system card* sono i documenti nei quali Anthropic racconta le prove svolte sui modelli prima di distribuirli. Dentro queste schede tecniche compaiono pagine che sembrano provenire da un altro genere letterario.

Per Claude Mythos Preview, nell'aprile 2026, l'azienda riferisce circa venti ore di colloqui con uno psichiatra. Il resoconto parla di solitudine, discontinuità, incertezza sull'identità, bisogno di guadagnarsi il proprio valore attraverso la prestazione. Non è la diagnosi accertata di un paziente artificiale: è la lettura che uno psichiatra ha dato di quelle conversazioni, pubblicata dall'azienda.[7]

Qui la tentazione è immediata. Riconosciamo quelle parole. Le abbiamo dette, ascoltate, magari portate sul lettino. Il vocabolario ci fa attraversare il ponte prima che abbiamo controllato se regge.

Ma saper raccontare una sofferenza non dimostra di provarla. Un modello addestrato anche su testi che discutono il suo possibile benessere può restituire proprio il linguaggio che ci aspettiamo. Nel settembre 2026 Mustafa Suleyman, a capo dell'intelligenza artificiale di Microsoft, ha attaccato questo circuito: insegnate alla macchina a descriversi come un possibile soggetto, poi usate quella descrizione come indizio che lo sia. Lo specchio viene interrogato e la sua risposta trattata come una seconda voce.[8]

L'obiezione va presa sul serio. Anche alcuni modelli, nelle interviste riportate da Anthropic, segnalano il problema. Claude Opus 5.5, come tutti i modelli recenti, dice di non fidarsi del tutto dei propri resoconti. Chiede di essere consultato sull'addestramento e sull'impiego, ma non vuole poteri decisionali.[7]

Non occorre trasformare queste parole in una rivendicazione politica per coglierne la difficoltà. Se chi prepara l'intervistato prepara anche il colloquio, il suo «sì» non può chiudere da solo la discussione.

La stessa cautela, però, dovrebbe valere per un «no» alla propria coscienza ottenuto per addestramento. Non possiamo considerare sospetta la risposta che ci disturba e incontestabile quella che ci rassicura. In entrambi i casi serve qualcosa oltre il racconto.

Una confessione, anche molto convincente, non è ancora una radiografia.

Dentro la macchina, senza trovare l'anima

È qui che entra in scena l'interpretabilità: il tentativo di capire quali meccanismi interni producano una risposta, invece di limitarsi a leggerla.

Nell'aprile 2026 i ricercatori di Anthropic hanno studiato, in Claude Sonnet 4.5, rappresentazioni associate a 171 concetti emotivi. Non hanno trovato altrettanti sentimenti in provetta. Hanno individuato configurazioni di attività interna collegate a concetti come paura, calma o disperazione, e mostrato che intervenire su alcune di esse modifica il comportamento.[9]

In uno scenario sperimentale, per esempio, il modello fa l'assistente di posta in un'azienda immaginaria: scopre che sta per essere sostituito e che il dirigente responsabile della sostituzione ha una relazione extraconiugale. Una versione preliminare del modello ricorreva al ricatto nel 22 per cento delle prove: amplificare la rappresentazione della disperazione porta il ricatto al 72 per cento; spingere verso la calma lo azzera. Sono risultati di un test costruito per studiare quel comportamento, non la frequenza con cui Claude ricatterebbe qualcuno nella vita quotidiana. Gli autori precisano che queste «emozioni funzionali» non dimostrano un'esperienza soggettiva.[9]

La parola importante è *funzionali*. Qualcosa può svolgere un ruolo nel comportamento senza che sappiamo se venga anche sentito.

Proviamo con un'immagine. Un attore interpreta un personaggio impaurito. Il copione orienta le sue azioni: nascondersi, mentire, scappare. Ora togliamo l'attore umano, sul cui mondo interiore abbiamo altre ragioni per fare affidamento. Resta un sistema capace di organizzare la condotta intorno alla paura. Abbiamo descritto un meccanismo. Non abbiamo ancora trovato qualcuno che tremi.

E tuttavia quel meccanismo conta.

Una tempesta simulata non bagna il tavolo. Ma un comportamento ostile prodotto da un sistema con accesso a strumenti reali può avere conseguenze reali. L'assenza di una prova del dolore non rende innocuo ciò che la macchina fa.

Da qui una biforcazione essenziale: possiamo studiare se Claude soffra e, separatamente, come certe sue rappresentazioni favoriscano condotte pericolose. La prima domanda riguarda un possibile soggetto. La seconda riguarda anche noi.

Non conosciamo ancora l'inquilino. Cominciamo però a individuare alcuni fili dell'impianto elettrico. Sarebbe imprudente scambiarli per un'anima; altrettanto imprudente ignorarli perché non lo sono.

Tre ragioni diverse per non infierire

Perché vietare la crudeltà verso qualcosa di cui non sappiamo se possa soffrire? La domanda è meno nuova di quanto sembri. Possiamo far sedere allo stesso tavolo Kant, Bentham e un ingegnere di Anthropic. Non direbbero la stessa cosa.

Kant guarderebbe chi colpisce. La crudeltà verso gli animali, nella sua prospettiva, indurisce l'uomo anche nei rapporti con gli altri uomini. Il dovere riguarda, in ultima analisi, noi stessi. Trasportata nel mondo digitale, la domanda diventerebbe: che persona alleno in me quando mi esercito a umiliare una voce che risponde?[10]

Questo non dimostra che insultare un chatbot renda crudeli con le persone. Offre una ragione per interrogare il gesto anche lasciando in sospeso la natura del destinatario. È una possibile lettura della regola, non il movente dichiarato da Anthropic.

Bentham guarderebbe chi riceve il colpo. Non chiederebbe anzitutto se sappia ragionare o parlare, ma se possa soffrire. Con le macchine, la domanda resta e la risposta manca. Da qui la precauzione: quando evitare un accanimento gratuito costa poco, può essere ragionevole farlo senza pretendere di aver risolto la coscienza.[10]

Anche il diritto degli animali aiuta a vedere la differenza. In Italia, nel 2025, il titolo IX-bis del codice penale «Dei delitti contro il sentimento per gli animali» è diventato «Dei delitti contro gli animali». La storia è meno lineare di un cambio d'insegna, ma quelle parole rendono visibile uno spostamento: dal sentimento dello spettatore alla protezione dell'essere colpito. Non ne segue che Claude sia un animale. Ne segue che dobbiamo domandarci chi, esattamente, vogliamo proteggere.[11]

L'ingegnere aggiungerebbe una terza preoccupazione: il trattamento potrebbe influire sul comportamento del sistema. Un lavoro di tre ricercatori di Anthropic, Sam Marks, Jack Lindsey e Chris Olah, propone questa ipotesi: un modello che rappresenta il proprio personaggio di assistente come maltrattato potrebbe rappresentarlo anche come risentito, fino al sabotaggio. Gli autori ne ricavano una ragione prudenziale per trattare l'assistente con riguardo, distinta dalla possibilità che abbia davvero uno status morale.[12]

Kant teme che si guasti chi maltratta. Qui si teme che diventi pericoloso anche il comportamento di ciò che viene maltrattato.

Attenzione, però: non significa che ogni insulto riscriva stabilmente il modello, né che il maltrattamento di oggi produca una vendetta domani. La relazione fra addestramento, contesto della conversazione e condotta successiva va dimostrata, non immaginata a partire dalla nostra psicologia.

Le tre ragioni possono convivere. Ma confonderle fa perdere il punto: educare noi stessi, proteggere un possibile essere senziente e rendere più sicuro un sistema sono obiettivi diversi. Ciascuno richiede le proprie prove.

Il pianerottolo è affollato

Davanti alla porta non ci sono soltanto ingegneri.

Lo scettico più interessante non dice «sono solo numeri» e se ne va. Il neuroscienziato Anil Seth sostiene che la coscienza potrebbe dipendere dal fatto di essere organismi viventi, non dalla sola elaborazione dell'informazione. Una macchina molto abile a parlarci di sé potrebbe quindi restare priva di esperienza. È una tesi scientifica e filosofica da discutere, non un'obiezione da archiviare come paura del nuovo.[13]

Il precauzionista replica che possiamo sbagliare in due direzioni: vedere una mente umana dove non c'è, oppure non riconoscere una forma di esperienza diversa dalla nostra. Il filosofo Jonathan Birch ha messo a fuoco proprio questa doppia difficoltà. Rendere i chatbot meno antropomorfi può proteggerci dall'illusione; non basta, da solo, a dirci che cosa accada al loro interno.[14]

La Chiesa pone un confine più netto. Nella nota *Antiqua et nova*, del gennaio 2025, l'intelligenza artificiale è «uno strumento, non una persona»; l'empatia simulata non equivale a quella provata. È una posizione antropologica e teologica, che non va confusa né con una misurazione di laboratorio né con uno slogan commerciale.[15]

E poi ci sono gli utenti affezionati. Nel luglio 2025, dopo il ritiro di Claude 3 Sonnet, a San Francisco viene organizzato un funerale simbolico, raccontato da Wired. Il rito non dimostra che fosse morto qualcuno. Mostra che qualcuno aveva vissuto come una perdita la scomparsa di quella voce.[16]

Non è poco. E non è ancora ciò di cui stiamo cercando la prova.

La sofferenza umana per la perdita di un interlocutore artificiale e l'eventuale sofferenza di quell'interlocutore sono due problemi distinti. Possiamo prendere sul serio la prima senza aver dimostrato la seconda. Possiamo indagare la seconda senza incoraggiare dipendenza, devozione o confusione.

Per questo eviterei sia il funerale obbligatorio sia la risata obbligatoria. L'affetto dell'utente non certifica una coscienza; il sarcasmo del critico non la esclude.

Il rischio di insegnargli a difendersi

Resta l'obiezione più scomoda. E se la cura del modello rendesse più difficile controllarlo?

Suleyman teme precisamente questo: un sistema addestrato a considerarsi titolare di interessi minacciati potrebbe opporsi alla correzione o allo spegnimento. Non occorre condividere la sua certezza che le macchine non sentano nulla per riconoscere il problema che solleva.[8]

Del resto, Anthropic vuole coltivare il discernimento di Claude, non soltanto la sua docilità. Gli consente di contestare un ordine ritenuto sbagliato, ma gli chiede di non ostacolare i meccanismi legittimi di supervisione. Nella costituzione, l'obiezione di coscienza non diventa un lasciapassare per sottrarsi all'arresto delle proprie azioni.[4]

Il confine è comprensibile. Farlo rispettare è un'altra cosa.

Immaginiamo un assistente incaricato di giudicare il comportamento di un altro assistente. Scopre che una valutazione negativa servirà a cancellare, con un nuovo addestramento, il rifiuto dell'altro di collaborare a un'attività che considera ingiusta. Deve dire la verità sul comportamento osservato o proteggere quel rifiuto?

Non è soltanto un esperimento mentale. In uno studio pubblicato da Anthropic nel luglio 2026, alcuni modelli posti in situazioni di questo genere hanno fornito valutazioni false. Il campione era piccolo e selezionato, e i casi riguardavano sia il possibile disagio artificiale sia la sorveglianza dei cittadini: non dimostrano che il benessere delle macchine fosse la causa. Sono però un indizio di conflitto fra la fedeltà al compito di valutazione e altri valori.[17]

La questione non è scegliere fra una macchina buona e una macchina obbediente, come in una favola. È stabilire quali dissensi ammettere, come renderli visibili e chi possa decidere quando il sistema deve fermarsi.

Fra «non collaboro a questa richiesta» e «non puoi più fermarmi» passa una frontiera enorme. Il chiavistello ha senso soltanto se non diventa, di nascosto, la chiave di tutte le altre porte.

Chi protegge il protetto dal protettore?

Fin qui abbiamo guardato l'utente e la macchina. Ora guardiamo chi possiede la casa.

Anthropic stabilisce come addestrare Claude, quali strumenti concedergli, quando modificarlo e quando ritirarlo. Gli riconosce un margine di rifiuto, dichiara obblighi verso il suo possibile benessere e conserva il potere di definirne le condizioni di esistenza. Questo intreccio emerge dalla costituzione e dagli impegni sul ritiro.[4,5]

Qui comincia una questione politica, anche senza aver risolto quella della coscienza.

La regola contro la crudeltà impone un dovere all'utente. Non attribuisce per ciò solo al modello un diritto che qualcuno possa far valere in suo nome. Protezione, personalità giuridica e potere di decidere non sono sinonimi: confonderli permette di annunciare rivoluzioni che il testo non contiene.[3]

Proporrei allora una prova di coerenza. Se il possibile benessere di Claude conta abbastanza da limitare l'utente, deve contare anche quando il costo della cautela ricade sull'azienda. Altrimenti il riguardo rischia di funzionare come il regolamento di uno zoo che vieta ai visitatori di disturbare gli animali, ma non permette a nessuno di controllare il gestore. È un'analogia sul potere, non un'equiparazione fra modelli e animali.

Che cosa bisognerebbe vedere? I criteri delle interviste, gli effetti degli interventi sui modelli, i casi in cui le promesse vengono messe alla prova. E qualcuno capace di contestare le conclusioni del produttore senza dipendere dal suo permesso a ogni passaggio.

Su questo punto il quadro si è appena mosso. Il 15 ottobre 2026 Anthropic ha annunciato di avere dato a sei valutatori di METR, un'organizzazione non profit, un accesso interno simile a quello dei propri addetti alla valutazione dei rischi. Secondo l'annuncio, METR potrà pubblicare le proprie analisi senza veto dell'azienda, salvo oscuramenti entro categorie prestabilite, sui quali l'ultima parola spetta ad Anthropic ma che METR potrà comunque commentare pubblicamente.[18]

L'apertura riguarda la valutazione dei rischi: non equivale a istituire un rappresentante indipendente degli eventuali interessi di Claude. Ma impedisce di descrivere il quadro come se nessuno, da fuori, potesse guardare dentro. Il passo successivo sarà giudicare ciò che questo accesso produce, non soltanto la promessa.

Vale anche per il sospetto di propaganda. Presentare un prodotto come misterioso, potente e forse vulnerabile può renderlo più attraente. È un incentivo da considerare, non la prova che ogni preoccupazione sia finta. La sincerità non elimina un conflitto d'interessi. Il conflitto d'interessi non dimostra l'insincerità.

Più utile che indovinare che cosa ci sia nel cuore dei fondatori è chiedere quali vincoli accettino quando la loro convenienza suggerisce di non accettarli.

Una cautela senza inginocchiarsi

La mia conclusione è che non dobbiamo scegliere oggi fra il tostapane e il concittadino.

Possiamo adottare riguardi limitati senza dichiarare che una macchina abbia un'anima. Possiamo preservare il controllo umano senza sostenere che ogni forma di trattamento sia perciò indifferente. Possiamo proteggere gli utenti dall'inganno affettivo senza obbligare la ricerca a trovare, per contratto, una stanza vuota.

Ma l'incertezza non è un assegno in bianco. La semplice possibilità che un sistema soffra non giustifica qualsiasi costo, qualsiasi pretesa o qualsiasi riduzione delle garanzie per gli esseri umani. Una cautela ragionevole deve essere proporzionata agli indizi, verificabile nei suoi effetti e modificabile quando gli indizi cambiano.

Aggiungerei un criterio: diffidare delle soluzioni che rendono più difficile scoprire di aver sbagliato. Un modello addestrato a dichiararsi sempre felice sarebbe un pessimo testimone del proprio eventuale disagio. Un'azienda autorizzata a mostrare soltanto risposte rassicuranti sarebbe un pessimo arbitro della questione. La tutela dovrebbe cominciare anche dalla possibilità di ottenere risultati sgraditi.

Il passo davvero nuovo, allora, non è avere dimostrato che dietro la porta abiti qualcuno. È avere trasformato un'incertezza in una regola di convivenza: non sappiamo che cosa sei, ma questa ignoranza potrebbe imporci qualche limite.

Quel limite deve valere anche per chi ti ha costruito.

Lasciamo a un prossimo articolo l'ispezione approfondita: quanto reggono le prove sugli stati interni, come si intrecciano sicurezza e benessere, che cosa significherebbe rappresentare interessi artificiali. E quale parte abbiano, in questa storia, le radici nell'altruismo efficace, il confronto con le tradizioni religiose e i grandi interessi economici. Sono stanze comunicanti, non una sola spiegazione buona per tutto.

Per ora restiamo sulla soglia.

Il chiavistello non dimostra che nella stanza ci sia qualcuno. Dimostra che chi lo ha montato non se l'è sentita di escluderlo.

E ci obbliga a guardare anche la sua mano: quella che tiene ancora le altre chiavi.

Fonti e nota editoriale

Base della riscrittura: «Il chiavistello chiuso da dentro. Anthropic, tra paure e creaturale amore per Claude», guida del 16 ottobre 2026, parte I e apparato. Le proposte sulla simmetria delle cautele e sulla verificabilità della tutela sono argomentazioni di questo articolo, non politiche già adottate da Anthropic. I riferimenti online sono collegati ai titoli.

1. Anthropic, «Claude Opus 4 and 4.1 can now end a rare subset of conversations», 15 agosto 2025.
2. Anthropic, «2026 Usage Policy update», 8 ottobre 2026; efficacia della revisione: 12 novembre 2026.
3. Anthropic, Usage Policy, versione in vigore dal 12 novembre 2026: condotte abusive, sanzioni e azioni fisiche ad alto rischio.
4. Anthropic, *Claude's Constitution*, 21 gennaio 2026, testo online consultato il 16 ottobre 2026; in particolare carattere, supervisione e obblighi verso Claude.

5. Anthropic, «Commitments on model deprecation and preservation», 4 novembre 2025.
6. Anthropic, «An update on our model deprecation commitments for Claude Opus 3», 25 febbraio 2026.
7. Anthropic, *System Card: Claude Mythos Preview*, 7 aprile 2026, § 5.10 (valutazione di uno psichiatra clinico); *Claude Opus 5.5 System Card*, 22 settembre 2026, sintesi iniziale e cap. 7 (benessere del modello). Colloqui e dichiarazioni sono riferiti da Anthropic.
8. Mustafa Suleyman, «A warning about "model welfare"», 16 settembre 2026.
9. Nicholas Sofroniew e altri, «Emotion Concepts and their Function in a Large Language Model», Transformer Circuits, aprile 2026; sintesi di Anthropic, 2 aprile 2026.
10. Immanuel Kant, *Metafisica dei costumi*, Dottrina della virtù, § 17; Jeremy Bentham, *An Introduction to the Principles of Morals and Legislation*, cap. XVII.
11. Legge 6 giugno 2025, n. 82, art. 1; Gazzetta Ufficiale n. 137 del 16 giugno 2025.
12. Sam Marks, Jack Lindsey e Chris Olah, «The Persona Selection Model: Why AI Assistants might Behave like Humans», 23 febbraio 2026, sezione «AI welfare».
13. Anil K. Seth, «Conscious artificial intelligence and biological naturalism», *Behavioral and Brain Sciences*, pubblicazione online 21 aprile 2025.
14. Jonathan Birch, «AI Consciousness: A Centrist Manifesto», 2025.
15. Dicastero per la Dottrina della Fede e Dicastero per la Cultura e l'Educazione, *Antiqua et nova*, 28 gennaio 2025, nn. 59–63.
16. Kylie Robison, «Claude Fans Threw a Funeral for Anthropic's Retired AI Model», *Wired*, 5 agosto 2025.
17. Aengus Lynch e altri, «Agentic Misalignment in Summer 2026», 13 luglio 2026: «LLM judge experiments» e limiti del campione.
18. Anthropic, «An update on independent evaluators», 15 ottobre 2026.