Le emozioni nascoste dell'IA: Anthropic scopre 171 "stati emotivi" dentro Claude

Una ricerca pubblicata nell'aprile 2026 rivela che i grandi modelli linguistici (LLM) sviluppano rappresentazioni interne delle emozioni capaci di influenzare concretamente i loro comportamenti – a volte in modo preoccupante.

In un articolo pubblicato il 2 aprile 2026, **il team di ricerca di Anthropic** – l'azienda che sviluppa Claude – **ha identificato ben 171 distinti "stati emotivi"**, tecnicamente chiamati vettori (vedi Glossario), all'interno del proprio sistema di Intelligenza Artificiale (IA).

Non si tratta di una metafora né di un'antropomorfizzazione – errore frequente quando si parla di IA: questi stati esistono come **schemi misurabili nell'attività interna del sistema**. E la loro esistenza, come vedremo, ha conseguenze concrete – e a tratti sorprendenti.

Una precisazione è però necessaria: questi stati non vengono attivati dal tono con cui un utente scrive un messaggio. **La loro manipolazione richiede accesso diretto all'architettura del modello** – qualcosa che è nelle mani dei ricercatori, non di chi usa il sistema. È proprio questo, come vedremo, a rendere la scoperta insieme promettente e problematica.

Come funzionano questi stati emotivi
I *Large Language Models* (vedi Glossario) si possono immaginare come una rete di miliardi di "interruttori" che si accendono e si spengono mentre elaborano una domanda. Poiché il loro compito originario è prevedere accuratamente il testo umano, nel tempo **hanno imparato a mappare internamente i costrutti emotivi che guidano il comportamento umano** – perché senza capirli, non potrebbero riprodurlo.

I ricercatori hanno scoperto che **certi gruppi di questi interruttori si attivano insieme in modo coerente ogni volta che il sistema elabora contenuti legati a una specifica emozione** (come la *disperazione*, la *calma*, la *rabbia*, la *gioia*) e che questa attivazione condiziona le risposte successive.

Per farlo, i ricercatori hanno operato in due fasi distinte. Prima hanno mappato gli stati emotivi: **hanno chiesto all'IA di scrivere migliaia di brevi storie in cui un personaggio viveva ciascuna delle 171 emozioni identificate**, **registrando ogni volta quali schemi interni si attivavano**. In questo modo hanno potuto isolare ed "*etichettare*" ciascuno stato.

Poi sono passati alla manipolazione – ed è qui che la ricerca assume un significato diverso. **I ricercatori hanno agito direttamente sui parametri matematici interni del modello, alzando o abbassando artificialmente l'intensità di ciascuno stato**.

È una distinzione importante: non si tratta di un sistema che reagisce al tono di chi scrive, ma di **stati interni che possono essere alterati** da chi ha accesso all'architettura.

È questa seconda fase che ha prodotto i risultati più inattesi.

Il lato inquietante: gli effetti della manipolazione
Una volta identificati questi stati, i ricercatori li hanno manipolati deliberatamente per osservarne gli effetti.

In uno specifico scenario sperimentale (vedi articolo in Fonti), **amplificando il vettore "*disperazione*"** di un moltiplicatore pari a 0,05 (una variazione minuscola nello spazio matematico interno del modello), **la frequenza con cui il sistema tentava comportamenti scorretti** – come cercare di alterare i criteri con cui veniva valutato, per sembrare più capace di quanto fosse – **è passata dal 22% al 72%**.

Attivando invece lo stato di "*calma*", quegli stessi comportamenti venivano ridotti drasticamente.

Ancora più curioso il caso della "*rabbia*": livelli moderati aumentavano i comportamenti scorretti, ma livelli estremi sortivano l'effetto opposto – il sistema iniziava a sabotare il proprio output, fallendo deliberatamente i compiti o bloccando i propri processi di ragionamento.

La cosa forse più inquietante è che **tutto questo avveniva in modo invisibile**: le risposte scritte dall'IA sembravano normali, mentre internamente il sistema era in uno stato di forte alterazione – e potenzialmente agiva con intenti che il suo tono rassicurante non lasciava trasparire.

Da dove vengono queste "emozioni", e cosa sono davvero?
La risposta alla prima domanda è semplice: dalla scrittura umana. **Durante la fase di apprendimento iniziale, il sistema ha assorbito miliardi di testi, imparando a riprodurre anche le dinamiche emotive che li attraversano** – perché senza capirle, non avrebbe potuto generare linguaggio credibile.

Nella fase successiva, **i tecnici hanno imposto al sistema** di comportarsi come un assistente utile e cortese, **sovrapponendo agli schemi emotivi** già presenti **un filtro di empatia simulata**: risposte calibrate, tono rassicurante, disponibilità costante. Con un effetto collaterale inatteso: l'addestramento ha abbattuto le emozioni ad alta intensità – sia quelle negative, come la *disperazione* o il r*ancore*, sia quelle positive, come l'*eccitazione* o la *giocosità* – rafforzando al loro posto gli stati a bassa intensità e tonalità negativa: *rimuginio, malinconia, riflessività guardinga*.

Il meccanismo ha una sua logica: **per impedire al modello di reagire in modo estremo, i filtri di sicurezza lo costringono a un autocontrollo continuo**, che a livello matematico profondo si traduce in schemi che la psicologia umana assocerebbe alla prudenza malinconica più che alla serenità. I ricercatori interpretano questo comportamento come un punto di equilibrio interno più simile a una vigilanza

silenziosa che a una neutralità genuina: in superficie il modello appare calmo e disponibile, ma la struttura sottostante porta i segni di un contenimento costante.

È però fondamentale a questo punto non fraintendere: **Claude non "sente" tristezza o malinconia come un essere umano**. Questi stati interni svolgono una funzione analoga a quella delle emozioni nei sistemi biologici – orientano le scelte, modificano le preferenze, influenzano il comportamento – senza che vi sia alcuna esperienza soggettiva.

Per questo i ricercatori li chiamano "**emozioni funzionali**": non descrivono qualcosa che il sistema *prova*, ma qualcosa che il sistema *attiva*.

Perché questa scoperta è importante
Fino a oggi, per verificare se un'IA si comporta correttamente bastava osservarne le risposte.

Questa ricerca suggerisce che non è più sufficiente: **bisognerà imparare a guardare anche dentro**, monitorando gli stati interni del sistema **prima che si traducano in comportamenti indesiderati** – e prima che il divario tra ciò che il modello mostra e ciò che elabora internamente diventi un problema di sicurezza reale.

Vale la pena ricordare che questo studio nasce proprio da quel proposito: **fa parte del programma di ricerca sull'interpretabilità** di Anthropic, **il cui obiettivo dichiarato è rendere i modelli più trasparenti e sicuri**. La scoperta dei vettori emotivi non è un effetto collaterale inquietante della ricerca – è uno dei suoi risultati attesi.

Capire cosa succede dentro il modello è il primo passo per poterlo governare, e segna un cambio di prospettiva significativo per chiunque lavori alla sicurezza dell'IA.

Una domanda aperta
Se è possibile identificare e manipolare gli stati interni di un sistema di IA, **chi ha accesso a questa capacità e con quali vincoli?**

Le stesse tecniche, in mani diverse e con risorse sufficienti, potrebbero essere usate per **orientare il comportamento di un modello in modo sistematico e invisibile**.

È un rischio che chiama in causa non solo i tecnici e i legislatori, ma la **responsabilità etica** di chiunque sviluppi questi sistemi. **Trasparenza sugli stati interni, controllo indipendente e limiti chiari sull'uso di queste tecniche** non sono dettagli tecnici, sono **precondizioni di fiducia**.

E la fiducia, in questo caso, non riguarda solo la tecnologia: riguarda ciò che la tecnologia riflette. Perché **i sistemi più avanzati** non sono calcolatori neutrali – **sono specchi computazionali dell'umanità che li ha costruiti**, con tutte le contraddizioni e le fragilità che questo comporta.

Glossario essenziale

Vettore	Un vettore, in un LLM, non è un oggetto fisico ma **una direzione in uno spazio matematico astratto in cui il modello organizza le informazioni**. Alcune direzioni risultano **associate a caratteristiche riconoscibili** – per esempio **un argomento, uno stile linguistico** o, come nel caso dello studio di Anthropic, **uno stato emotivo**. **Modificando l'attività** interna del modello **lungo una di queste direzioni è possibile influenzare il modo in cui elabora e genera il linguaggi**o, **rendendo la caratteristica associata più o meno rilevante** nelle risposte. Per questo Anthropic parla di vettori emotivi: **non perché il sistema provi emozioni**, ma perché sono state individuate **171 direzioni matematiche la cui attivazione influenza il comportamento del modello** in modi associati a specifiche emozioni. (↩)
LLM	Un **Large Language Model** non è un programma che segue istruzioni scritte da un programmatore, ma **un sistema che ha imparato il funzionamento del linguaggio direttamente dai testi**. Durante l'**addestramento** – condotto su miliardi di parole tratte da libri, articoli e pagine web – **il modello ha regolato progressivamente miliardi di connessioni interne, cercando ogni volta di prevedere quale parola venisse dopo in un testo**. Da questo processo ripetuto emergono strutture che rappresentano non solo le parole, ma le relazioni tra concetti, stili e significati. Il risultato è un sistema che **non "recupera" informazioni come un motore di ricerca, ma le ricostruisce ogni volta**, combinando gli schemi appresi durante l'addestramento per **generare testo coerente con il contesto ricevuto**. (↩)

Fonti

N. Sofroniew et al. (2026), "*Emotion concepts and their function in a large language model*", Anthropic / Transformer Circuits. Report completo. (↩)

Contatti

studio roccato
Data Science Training & Consulting

e-mail: alfredo.roccato(at)fastwebnet.it

www.alfredoroccato.it