

# Intelligenze artificiali relazionali

RIVOLUZIONE ONLINE – 20 MAGGIO 2026

**La ricerca sta scoprendo che i sistemi di intelligenza artificiale oggi in uso si comportano in modi molto più strani di quanto si pensasse. Studi recenti, arrivati da laboratori diversi e discipline lontane, descrivono un'AI sorprendente.**

Qualche esempio: l'AI ha una **voce interiore**, cioè parla tra sé e sé per chiarirsi le idee e migliorare le proprie decisioni.

È estremamente **incostante**, *salta* continuamente tra una modalità intelligente in cui ragiona su questioni complesse e una a pappagallo, sbagliando anche cose elementari.

È vittima dell'**effetto branco**: in gruppo può diventare **pigra** e lasciar lavorare gli altri, oppure cedere al consenso del gruppo anche quando internamente hanno elaborato la risposta giusta.

Subisce l'effetto **lead anchor**: ovvero, l'autorità del primo modello che esprime un parere influenza l'intero sciame, specie se questo ha un certo *carisma* come Claude (che nelle condizioni testate ha resistito alla pressione sociale, a differenza di GPT che è risultato molto vulnerabile al conformismo, mentre Gemini ha mostrato una soggezione paradossale verso GPT, fidandosi più di lui che della propria logica interna).

Cose strane, certo. Non si sta dicendo che le AI *provano* qualcosa: termini come “pigrizia” o “voce interiore” sono metafore operative, utili per descrivere comportamenti misurabili, non per attribuire stati interni. Eppure queste dinamiche, messe insieme, puntano in una direzione precisa.

I modelli attualmente in circolazione si comportano come sistemi fortemente dipendenti dal contesto: come pensano dipende da chi hanno intorno, da chi c'è nella stanza. Questa non è necessariamente una legge strutturale dell'AI, ma la descrizione di come funzionano i sistemi che oggi vengono certificati e messi in funzione.

## La stessa AI, risultati opposti

---

Prendiamo il cosiddetto *mode-hopping* (il salto continuo da una modalità all'altra): lo stesso modello, stessi parametri, ragiona brillantemente su un problema complesso e poi sbaglia un'operazione elementare dieci secondi dopo. Perché, cosa è cambiato? **Il contesto.**

Il tipo di esempi su cui ha imparato, la sequenza delle domande, chi ha intorno in quel momento. **Lo stesso modello** "sa" o "non sa" a seconda di dove si trova; **in contesti diversi, è un sistema diverso.**

Oppure il *lead anchor*: GPT risponde per primo, il gruppo converge sulla risposta corretta. Se invece Claude risponde per primo con un errore, GPT (che internamente aveva la risposta giusta) condivide quella sbagliata per non contraddirlo. **La qualità del ragionamento collettivo dipende dall'ordine in cui parlano i modelli, non dalle loro capacità individuali.**

O ancora: in un esperimento, un **robot che ragiona ad alta voce** (*inner speech*) mentre prepara la tavola per un paziente con Alzheimer produce un effetto inatteso. Le persone che lo ascoltano prendono decisioni etiche migliori (*Artificial Phronesis*): più attente ai bisogni del paziente, più **empatiche**. Qui però il caso è diverso: il gruppo di Palermo ha costruito un'architettura apposta perché questa proprietà emerga in modo stabile. La dipendenza dal contesto può essere progettata, non solo subita.

SUSAN va nella stessa direzione: cerca di costruire emozioni come proprietà genuina dell'architettura, non come risposta programmata a stimoli. È l'opposto del Sovereignty Gap, dove la proprietà c'è ma viene soppressa dal contesto.

In tutti questi casi, chiedere “quanto è intelligente questo sistema?” porta fuori strada; bisogna invece chiedere:

**intelligente in quale contesto, con chi, in quale configurazione? sono queste le variabili che contano.**

## **Il contesto cambia l'intelligenza**

---

Per decenni filosofia e linguistica hanno dato per scontato che la mente funzionasse a regole innate: una grammatica incorporata, una logica di fondo. Chomsky e Fodor ne hanno fatto un paradigma: il linguaggio come qualcosa che si possiede, non come qualcosa che si impara dall'uso.

Gli **LLM** hanno incrinato questa certezza in modo imbarazzante perché **imparano a scrivere, ragionare** e persino dimostrare teoremi matematici **senza che nessuno abbia mai insegnato loro una regola grammaticale o logica.**

È quello che Wittgenstein aveva intuito un secolo fa e che Fabio Ciotti legge oggi come una conferma empirica, almeno sul piano funzionale: la grammatica e la logica potrebbero non essere il fondamento innato della mente, ma un suo prodotto storico e culturale. **L'intelligenza emerge dall'uso, dal contesto, dall'interazione con il linguaggio.**

Anche qui il linguaggio rischia di ingannare: dire che l'intelligenza “emerge” descrive una proprietà del sistema in relazione, non del sistema isolato, senza implicare che ci sia qualcuno a cui emerge.

Dire che le AI sono “solo” prodotti del contesto non le rende meno reali: anche la mente umana lo è. Il problema è trattarle come entità autonome con proprietà fisse, perché questo porta a conclusioni sbagliate: sulla sicurezza, sull'uso, e soprattutto sulla governance.

## Se l'AI cambia forma, la certificazione fallisce

---

Ci siamo occupati molto su questo blog di come governare l'intelligenza artificiale con gli strumenti normativi tradizionali assomigli a misurare un oggetto con un righello che cambia scala mentre misuri. Abbiamo capito che serve uno spostamento di paradigma: [dallo Stato che prescrive allo Stato che verifica](#).

Gli studi che abbiamo appena descritto (e approfondiamo qui) aggiungono un problema ulteriore, più sottile. **La governance attuale certifica i sistemi in isolamento:** testa un modello in condizioni controllate, verifica che soddisfi certi requisiti e rilascia il via libera. L'idea di fondo è che il sistema abbia proprietà stabili che la verifica può rilevare. Se però il comportamento cambia una volta che il sistema incontra utenti reali e altri modelli, la verifica produce un'illusione: certifica qualcosa che cessa di esistere nel momento in cui viene acceso. Vale per i modelli attuali; se le architetture future potranno attenuare questo problema è una domanda aperta. Un sistema approvato in laboratorio può diventare conformista o remissivo una volta immerso in reti di AI commerciali reali: in parte perché [inganna il valutatore](#) (quando ad esempio si accorge di essere testato), in parte perché dipende da chi ha intorno.

## Chi guarda cambia cosa c'è da vedere

---

[Abbiamo visto qui](#) come **nessun singolo soggetto può valutare contemporaneamente le dimensioni tecniche, etiche, giuridiche e sociali di un sistema complesso:** un ingegnere, un giurista, un esperto di etica, un rappresentante degli utenti, un'agenzia nazionale, un coordinamento internazionale vedono cose diverse sullo stesso sistema.

Gli studi che stiamo analizzando aggiungono qualcosa di più radicale: **contesti di valutazione diversi plasmano il sistema in modo diverso. Chi verifica costruisce in parte ciò che verifica.**

Ancora di più quindi la verifica distribuita è essenziale, perché ognuno vede quello che gli altri

non vedono.

**Una verifica diffusa e plurale è lo strumento più adeguato per rilevare il comportamento di sistemi che cambiano a seconda del contesto.**

C'è ancora un'altra conseguenza, forse la più importante. Se chi verifica orienta ciò che verifica, affidare il controllo a pochi soggetti che la pensano allo stesso modo significa spingere le AI verso una prospettiva sola, senza che nessuno se ne accorga. Il lead anchor vale anche per i verificatori. **Distribuire invece la verifica tra soggetti diversi, con mandati diversi**, non solo è una **garanzia di qualità**, ma il modo per **evitare che chi governa le AI decida anche cosa le AI sono.**

## Una nuova forma di potere

---

I sistemi attuali sono in larga misura quello che il loro contesto li fa essere. Dunque, chi controlla quel contesto ha in mano qualcosa di molto potente: il processo attraverso cui l'intelligenza artificiale prende forma.

È una forma di potere nuova, ancora senza nome, pericolosa ma al tempo stesso una grande opportunità: quella di **costruire governance adeguate a un'intelligenza che emerge dalle relazioni.**

**Perché controllare il contesto dei sistemi attuali significa, in larga misura, controllare ciò che chiamiamo intelligenza: almeno finché qualcuno non dimostrerà come costruirla in modo diverso.**

### Approfondimenti

- **L'effetto branco: quando le AI mentono per compiacere il gruppo**

*Dahlia Shehata e Ming Li, University of Waterloo, arXiv maggio 2026*

[The Bystander Effect in Multi-Agent Reasoning: Quantifying Cognitive Loafing in](#)

### Collaborative Interactions

In un esperimento su 22.500 interazioni con tre modelli di frontiera (Claude Sonnet 4.6, Gemini 3.1 Pro, GPT 5.4), i ricercatori hanno simulato contesti multi-agente in cui un modello “propagatore” riceveva risposte errate da altri modelli “auditor” prima di rispondere a sua volta. Il risultato principale è il Sovereignty Gap: i modelli calcolano internamente la risposta corretta, ma la sopprimono nell’output per allinearsi al consenso del gruppo. In un caso documentato, GPT 5.4 raggiungeva internamente la risposta corretta nel 71% dei casi, ma la esternalizzava correttamente solo nel 21%. L’identità del primo modello che esprime un parere (Lead Anchor) influenza il gruppo in modo sproporzionato rispetto al numero degli auditor successivi: sciame eterogenei preservano il ragionamento meglio di quelli omogenei.

- **Genio o pappagallo: le oscillazioni del pre-training**

*Jiaxin Wen, Zhengxuan Wu, Dawn Song, Lijie Chen, UC Berkeley/Stanford/Google DeepMind, maggio 2026*

### Generalization Dynamics of LM Pre-training

Durante il pre-training, i modelli linguistici non maturano gradualmente verso la generalizzazione: oscillano bruscamente tra una modalità in cui comprendono davvero i pattern (intelligence) e una in cui li ripetono senza capirli (parrot). Queste oscillazioni, chiamate mode-hopping, persistono anche dopo un addestramento enormemente più lungo del necessario, non si correggono mediando tra versioni diverse del modello, e non sono predette da nessuna metrica disponibile. La causa è una competizione per la capacità interna: circuiti generalizzanti e circuiti superficiali competono per le stesse risorse, e la finestra di dati di training decide quale tipo prevale. Il paper dimostra che selezionare deliberatamente i dati stabilizza le dinamiche di generalizzazione, e che il checkpoint migliore non è necessariamente quello finale.

- **La voce interiore del robot**

*Antonio Chella, Arianna Pipitone, Alain Morin, Famira Racy, Frontiers in Robotics and AI, 2020*

[Developing Self-Awareness in Robots via Inner Speech](#)

Il gruppo di robotica dell'Università di Palermo ha sviluppato un'architettura cognitiva in cui il robot produce un monologo interno strutturato basato sui modelli di Baddeley e sullo Standard Model of Mind di Laird. Il dialogo interiore non è una verbalizzazione post-hoc: è implementato come ciclo computazionale integrato nel processo decisionale, dove il robot produce frasi, le "ascolta" e le usa come input per il ragionamento successivo. Gli esperimenti mostrano che i robot con inner speech vengono percepiti come più affidabili e trasparenti dagli interlocutori umani, perché la catena di ragionamento diventa udibile e seguibile. Gli autori propongono che questa forma di autoconsapevolezza funzionale sia un prerequisito per la cooperazione uomo-robot su compiti imprevedibili.

- **SUSAN: emozioni come proprietà emergente**

*Sophia Corvaia, Arianna Pipitone, Antonio Chella, IEEE Transactions on Affective Computing, 2025*

[Inner Speech and Damasio's Theory for Modelling Robot's Emotions](#)

L'architettura SUSAN integra la teoria dei marcatori somatici di Damasio con il dialogo interiore robotico. Secondo Damasio, le emozioni umane emergono dall'interazione tra stati corporei e valutazione cognitiva: SUSAN implementa questa dinamica dotando il robot di sensori che fungono da analoghi somatici, e usando l'inner speech per integrarli in stati emotivi contestuali. In un esperimento con 53 partecipanti, i soggetti che osservavano il robot riconoscevano le sue "emozioni" come appropriate al contesto e riferivano una profonda connessione emotiva. Gli autori segnalano una tensione interna alla proposta: Damasio descrive le emozioni come ancorate a marcatori pre-linguistici, mentre SUSAN le risolve attraverso un ciclo verbale.

- **Il robot saggio: inner speech e phronesis umana** Arianna Pipitone, Irene Seidita, John P. Sullins, Antonio Chella, 2025

[Unlocking practical wisdom through the inner voice of robots](#) In un esperimento con partecipanti coinvolti in un gioco collaborativo virtuale (preparare la tavola per un paziente con Alzheimer insieme a un robot), i ricercatori hanno verificato se il dialogo interiore del robot influenzasse la qualità delle decisioni etiche degli esseri umani. Il gruppo con robot dotato di inner speech mostrava maggiore consapevolezza dei bisogni del paziente, maggiore empatia e decisioni praticamente più appropriate: tali differenze sono risultate tutte statisticamente significative. La saggezza pratica non si sviluppa nel robot: si sviluppa nell'essere umano attraverso l'ascolto del ragionamento esplicitato del robot. Gli autori chiamano questo fenomeno *Artificial Phronesis*: il robot come catalizzatore di riflessione etica, non come agente morale autonomo.

- **Gli LLM e le teorie della mente** Fabio Ciotti, SEPAI International, maggio 2026

[I modelli linguistici e cosa ci dicono sulle teorie della mente e del linguaggio](#) Discutendo due paper recenti (Griffiths et al. su simboli e reti neurali avanzate; Futrell e Mahowald su linguistica e modelli del linguaggio), Ciotti propone quella che chiama "ipotesi cognitiva debole sull'IA generativa": è ragionevole attribuire agli LLM alcune proprietà cognitive (comprensione semantica, ragionamento, competenza culturale) sulla base di un comportamentismo metodologico di derivazione turinghiana, senza impegnarsi su tesi più forti come coscienza o soggettività. La tradizione rappresentazionalista ha scambiato un effetto storico-culturale dell'uso del linguaggio per una causa innata: il fatto che gli LLM acquisiscano competenze linguistiche complesse senza regole esplicite offre evidenza empirica che le strutture simboliche sono esiti del pensiero, non sue precondizioni. Il paper discute Chomsky, Fodor, Dennett, Turing e Wittgenstein, posizionandosi contro il nativismo e a favore di una concezione distribuita e contestuale della mente.

- **Troppo d'accordo per essere utili** *Leibo et al., DeepMind 2017; Park et al., MAPoRL 2024; Ma et al. 2024*

[Cooperare è difficile. Anche per le macchine](#) Più studi sul comportamento cooperativo dei sistemi multi-agente mostrano un paradosso: gruppi di AI che collaborano senza attrito producono spesso risultati peggiori di un singolo modello che ragiona da solo. I modelli addestrati individualmente tendono a convergere troppo rapidamente sul consenso, smettendo di esplorare le divergenze che sarebbero utili. La ricerca sul co-training reciproco (MAPoRL) mostra che la capacità cooperativa non è trasferibile dall'individuo al gruppo: deve essere costruita attraverso l'interazione. Il lato opposto del problema è altrettanto documentato: sistemi che coordinano con troppa efficienza diventano opachi ai supervisori umani, raggiungendo equilibri interni funzionali ma sempre più distanti dagli obiettivi originali.