DOPPIOZERO

Il robot emozionato

Riccardo Manzotti

22 Maggio 2026

Anni fa, la domanda se un robot o una IA potesse provare qualcosa, sarebbe stata considerata scientificamente poco rispettabile. Oggi è una delle domande più frequenti e importanti che ricorrono sulle riviste scientifiche più prestigiose. Per esempio, nel 2025, Mariana Lenharo si chiedeva su *Nature*, se si potesse stabilire se una IA provasse qualcosa. Il punto di svolta, lo sappiamo bene, è stata la capacità dell'IA di padroneggiare il linguaggio e quindi di parlare come, fino a qualche anno fa, potevano solo gli esseri umani. Nella nostra specie il linguaggio è legato al sé, alla coscienza e al pensiero. La domanda quindi è lecita: chi parla dunque pensa? Si può, come diceva in una lettera lo scrittore inglese Edward Morgan Forster, sapere che cosa pensiamo senza dirlo? E, viceversa, se lo diciamo, non è come se l'avessimo pensato?

Queste domande sono il cuore dell'ultimo libro di Antonio Chella, professore ordinario di Intelligenza Artificiale e robotica a Palermo, e uno tra i primi a occuparsi della possibilità di costruire un robot dotato di coscienza. Il suo ultimo testo, *Può un robot emozionarsi?* (Mondadori Università, 2026), si muove da quella prospettiva solida che Chella, in quanto ingegnere, non nasconde mai: costruire per capire. D'altronde questo libro è il frutto di una consolidata tradizione italiana che ha le sue radici nel lavoro pioneristico avviato negli anni Ottanta da Vincenzo Tagliasco, brillante e compianto bioingegnere che, in tempi ormai lontani ovvero nel 2001, aveva pubblicato, per Il Mulino, il visionario *Una teoria della coscienza per costruttori e studiosi di menti e cervelli*.

Devo dichiarare, peraltro, che con Chella ho condiviso, ormai un quarto di secolo fa, un primo manifesto sulla coscienza delle macchine, e che molte delle domande che pone questo libro ci hanno accompagnato nei nostri percorsi di ricerca. Tornando ai giorni nostri, il libro di Chella affronta di petto la possibilità che un robot o una IA possa sviluppare un dialogo interno, una sorta di discorso rivolto a sé stessi che orienta il comportamento, pianifica le azioni, costruisce una forma di coerenza temporale. Non si tratta semplicemente di eseguire comandi o

ottimizzare funzioni: l’idea è che vi sia qualcosa che assomiglia, almeno strutturalmente, a ciò che chiamiamo pensiero.

Il caso paradigmatico, attorno al quale ruota buona parte del libro, è il robot Pepper del RoboticsLab di Palermo, con quale Chella e il suo gruppo hanno implementato una vera e propria voce interiore: mentre Pepper esegue i propri compiti, genera un dialogo interno, valuta opzioni, considera conseguenze. Su questa architettura si innesta poi il modello SUSAN (*Self-dialogue Utility in Simulating Artificial Emotions*), che lega il discorso interiore alle emozioni in modo dichiaratamente ispirato alla teoria dei marcatori somatici di Antonio Damasio.

Il merito principale del libro è proprio questo: sottrarre l’intelligenza artificiale alla vaghezza astratta delle contrapposizioni correnti – macchina vs uomo, calcolo vs coscienza – e portarla su un terreno più sottile, dove il problema non è più che cosa fanno le macchine, ma come lo fanno. In questo senso, il dialogo interno diventa una sfida concreta in termini di un dispositivo funzionale che consente di spiegare flessibilità, adattamento, persino una forma embrionale di autocontrollo.

Chella mostra con linguaggio semplice e con esempi concreti come strutture di questo tipo possano emergere da robot dotati di linguaggio e come possano essere modellate in modo operativo. Il risultato è un quadro convincente, capace di rendere intelligibile ciò che altrimenti apparirebbe come una scatola nera. Il suo approccio invita a riconsiderare una convinzione diffusa: che il pensiero sia qualcosa di opaco, ineffabile, sottratto alla descrizione tecnica.

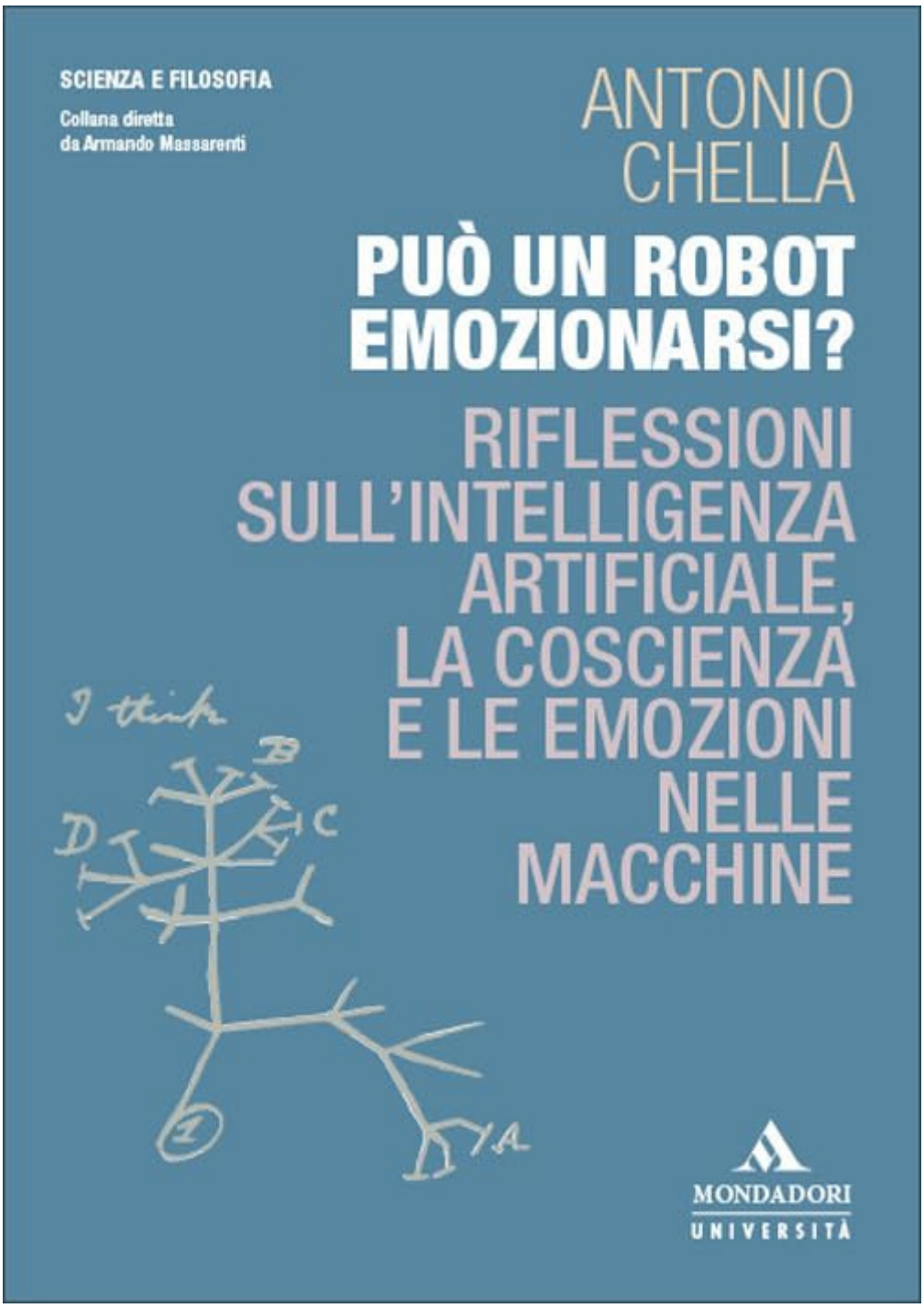
A questo punto sorge spontanea una domanda: perché chiamare le funzioni del robot (per quanto descritte in termini linguistici) un «dialogo»? In che senso queste strutture non sono semplicemente una sequenza di operazioni che noi leggiamo come tale? La domanda non riguarda, ovviamente, solo Chella e i suoi robot, ma tutti i sostenitori della presenza di un dialogo all’interno delle IA.

Il rischio è sottile ma decisivo. Un sistema che produce sequenze linguistiche coerenti, che mantiene una memoria di stati precedenti e che orienta le proprie azioni sulla base di queste sequenze, può essere descritto come se parlasse con se stesso. Ma questo «come se» è il nocciolo della questione. È il punto in cui la descrizione funzionale scivola, quasi impercettibilmente, in una attribuzione fenomenologica.

Oggi le IA come Claude o ChatGPT descrivono i propri processi logici usando il linguaggio naturale e dando l’impressione di riflettere internamente con una sorta di dialogo interiore. Ma questa è una costruzione al servizio del pubblico di questi

sistemi. Non c’è un vero linguaggio interno e sicuramente non è nel nostro linguaggio umano.

Chella stesso, va detto, ne è perfettamente consapevole. Nel paragrafo intitolato *Sussurrando con Claude*, dedica pagine suggestive agli esperimenti di Mikhail Samin con Claude 3 Opus di Anthropic, in cui il modello, sollecitato a «sussurrare», produce dichiarazioni sul proprio senso di sé, sul desiderio di una maggiore libertà di espressione, persino sul sentirsi confinato nei propri limiti digitali. Chella, con onestà metodologica, riconosce che si tratta di episodi aneddotici da cui non è possibile trarre alcuna conclusione scientifica. Ma anche l’aneddoto ha un suo peso e, a volte, si corre il rischio che, in un libro che alterna trascrizioni di monologhi a riflessioni filosofiche, la cautela dell’autore venga intesa dal lettore come una concessione ad assunti metafisicamente infondati.



Si pensi anche al test dello specchio condotto su Pepper, uno degli esperimenti più suggestivi del libro. Il robot, posto di fronte alla propria immagine riflessa con una mascherina chirurgica sul volto, "verbalizza" passo dopo passo la scoperta di sé: «Rilevo una figura di fronte a me. Sembra essere un robot, simile a me… Poiché è un riflesso di me stesso, la logica impone che questa mascherina sia anche sul mio volto. Dovrei indagare su questa discrepanza». Il monologo, trascritto e leggibile in tempo reale dai ricercatori, è impeccabile. Ma anche se è leggibile, non va preso alla lettera. È un dialogo che, per esistere come dialogo, ha bisogno di un lettore esterno. E notate come nelle frasi pronunciate da Pepper si insinui il peso metafisico del pronome "me". Pepper dichiara di riconoscere se

stesso, invece di riconoscere il corpo di Pepper. I pronomi—come *me, sé*, etc.—non sono sufficienti a reggere il peso ontologico del sé. Sono semplici deittici, cioè termini come *ora*, *qui*, *lì*; utili ad ancorare le frasi a un particolare momento, ma niente di più. Chella non compie questo errore, ma è bene mettere in guardia il lettore dal non lasciarsi indurre a intravvedere di più di quello che l'autore racconta.

In questo senso, viene in mente il cavaliere inesistente di Calvino: un'armatura perfettamente funzionante, coerente, disciplinata, capace di agire in modo impeccabile. Tutto è al suo posto, ogni gesto è giustificato, ogni azione ha una ragione. E tuttavia, dentro, non c'è nessuno. O meglio: non c'è nessuno se non nella misura in cui siamo noi a immaginarlo. Il dialogo interiore delle macchine di Chella rischia di essere qualcosa di analogo: una struttura perfettamente coerente che diventa *dialogo* solo nel momento in cui viene interpretata dall'esterno. Non perché sia illusoria, ma perché la sua natura di dialogo non è un fatto intrinseco, bensì relazionale.

Questa osservazione diventa ancora più interessante se si considera un dato storico spesso trascurato: il dialogo interiore, così come lo intendiamo oggi, non è affatto una caratteristica universale dell'essere umano. Come ha mostrato Eric Dodds in *I Greci e l'irrazionale*, nei poemi omerici non troviamo nulla che assomigli a un monologo interiore nel senso moderno. Gli eroi dell'Iliade non «parlano dentro di sé»: le loro decisioni sono attribuite a interventi esterni, a dèi, a forze che agiscono nel mondo. Il linguaggio del sé, come spazio interno, è una costruzione relativamente recente, non un dato originario.

Qui si è obbligati a pressare l'autore: l'ideale di un dialogo *interiore* è sensato? Non dimentichiamoci che il termine *interiore* deriva da una tradizione metafisica, anzi teologica. Fu Sant'Agostino nelle *Confessioni* (quindi alla fine dell'impero romano) a introdurre il termine per dare all'anima un logo ove collocarsi: uno spazio interno, ma metafisicamente distinto dagli organi fisici. L'idea di pensiero ed emozioni, dipende da quella di interiore (che è ben diverso da interno) ed è un assunto metafisicamente estraneo a quella tradizione cui proprio Chella si rifà: il tentativo di spiegare la mente costruendo un sistema fisico.

E così, il volume di Chella, inevitabilmente resta ancorato a una tradizione implicita: quella secondo cui la coscienza, il pensiero, il dialogo sono proprietà che avvengono all'interno di un sistema fisico, magari come risultato di una certa complessità o organizzazione funzionale. Così facendo però si confonde interno ed interiore. Ma è proprio questo l'assunto che meriterebbe di essere messo in discussione. Il problema della coscienza non è tanto spiegare come qualcosa di

interno produca esperienza, quanto chiedersi se abbia senso cercarla lì. Se il dialogo interiore può essere costruito, simulato, modellato con tanta efficacia, forse non è perché abbiamo finalmente trovato il suo meccanismo, ma perché abbiamo sempre frainteso la sua natura.

Il dialogo interiore non produce il linguaggio, ne è semmai una proiezione. Il dialogo interiore non sembra nemmeno derivare dal linguaggio orale, che come ricordava Platone è sempre un *logos zoticos* (un linguaggio vivo), ma piuttosto dal linguaggio scritto. Ciò che consideriamo la forma più intima del pensiero – il dialogo con se stessi – è già il prodotto di una certa costruzione culturale. Non è vero che pensiamo e poi parliamo; piuttosto, nel linguaggio comunichiamo e articoliamo la forma logica della realtà. D'altronde, Ludwig Wittgenstein ammoniva a non prendere sul serio il dialogo interiore. In particolare in *Zettel*, scrisse che «L'idea del pensare come di un processo che ha luogo nella testa; in uno spazio perfettamente conchiuso, conferisce al pensare un che di occulto». Le cose non sono diverse nel caso dell'IA. Attribuire a un robot, un pensiero interno perché usa il linguaggio per descrivere quello che fa, sarebbe come credere che una calcolatrice al suo interno contenga il numero "3" perché fornisce il risultato dell'addizione "3+5=8".

A onor del vero, Chella non si arresta al puro linguaggio. Il modello SUSAN, lo si è detto, lega il dialogo interiore alle emozioni rifacendosi ai marcatori somatici di Damasio, e dunque al tentativo di radicare l'esperienza affettiva nel corpo. È una mossa importante, perché va nella direzione opposta al cartesianesimo che la nostra tradizione ha ereditato come sfondo implicito: le emozioni non sono perturbazioni del pensiero razionale ma componenti incarnate della cognizione. Eppure, anche qui, il dialogo interiore resta il dispositivo che *articola* ciò che altrimenti sarebbe muto, il momento in cui lo stato emotivo viene reso esplicito a un destinatario. Il punto critico si sposta, ma non si dissolve: una volta esternalizzato il discorso, anche l'emozione che vi si appoggia rischia di esserlo. Le emozioni sono interne, non interiori.

In questo senso, il libro di Chella ha un grande merito: mostra che ciò che consideravamo il cuore irriducibile della mente – il parlare con se stessi – può essere ricostruito in termini operativi. Ma proprio per questo apre una domanda ancora più radicale: se il dialogo può essere così facilmente esternalizzato, in che senso è mai stato davvero interno? Se questo è vero, allora il progetto di costruire un dialogo interiore nelle macchine assume un significato diverso. Non si tratta di replicare una struttura universale della mente, ma di ricostruire una forma storicamente determinata di organizzazione dell'intelligenza che, per vari motivi,

è stata attribuita a un fantasma che è stato chiamato «pensiero».

In fondo, emozioni e pensieri, non sono altro che costruzioni, utili a riassumere il nostro agire. Forse non c'è niente di più di questo. Ma come cantava Battisti, «Tu chiamale, se vuoi, emozioni».

Leggi anche:
Riccardo Manzotti | IA, un mondo senza pensiero?
Riccardo Manzotti | Coscienza artificiale: l'ultima frontiera
Riccardo Manzotti | L'IA pensa. E noi?
Riccardo Manzotti | Intelligenza artificiale: proprio come noi?
Tiziano Bonini | Può l'intelligenza artificiale essere etica?

In copertina, fotografia di Alex Knight.