La cultura che non sappiamo di trasmettere

Cultura, valori e pregiudizi nell'intelligenza artificiale

Edoardo Mattei
(Pontificia Università San Tommaso d'Aquino – Angelicum, Roma)

Quando un sistema di intelligenza artificiale raccomanda un contenuto, assegna un punteggio di credito, suggerisce una diagnosi o risponde a una domanda su come perdere peso, chi sta agendo? La risposta intuitiva — «la macchina» — è non solo tecnicamente imprecisa ma filosoficamente fuorviante. Bruno Latour ha mostrato che ogni azione sociale è sempre distribuita in una rete di attanti, umani e non-umani, che contribuiscono ciascuno a determinare l'esito finale. Il martello non è uno strumento neutro nelle mani del carpentiere: modifica il gesto, orienta la forza, definisce ciò che è possibile costruire. Analogamente, un sistema AI non è un canale trasparente attraverso cui passa la volontà di qualcuno: è un attante che trasforma, seleziona, amplifica. Il quadro teorico di riferimento — quello dell'*Agency Partecipata* — ha mostrato come questa distribuzione dell'azione non dissolva la responsabilità ma la sposti: in ogni nodo della rete dove si prende una decisione rilevante per il comportamento del sistema, c'è sempre un soggetto umano che risponde.

Ma se l'azione è distribuita, la domanda sulla responsabilità non scompare: si concentra. Chi ha scelto i dati di addestramento, chi ha progettato gli obiettivi di ottimizzazione, chi ha selezionato i valutatori che hanno espresso preferenze sul comportamento del sistema, chi ha deciso in quale contesto distribuirlo: ognuno di questi passaggi porta con sé scelte, valori, pregiudizi e visioni del mondo. L'azione che il sistema produce non nasce dal nulla. Viene da qualche parte. E questa origine è ciò che questo articolo intende esaminare.

La tesi che si vuole argomentare non è che l'intelligenza artificiale non possa essere morale, un dibattito già ampiamente percorso. È qualcosa di più specifico e, in un certo senso, più inquietante: la cultura umana, con tutto il suo carico di virtù, bias, ideologie e pregiudizi, si cristallizza nei sistemi AI attraverso processi che la rendono strutturalmente opaca e potenzialmente irrilevabile. Il sistema non è uno specchio neutro della cultura che lo ha prodotto: è una sedimentazione selettiva che poi circola nel mondo con l'apparenza dell'oggettività tecnica.

I grandi modelli linguistici che animano i sistemi AI più diffusi vengono costruiti attraverso due fasi. La prima, chiamata *pre-training*, consiste nell'esporre il sistema a quantità enormi di testo

umano: decine o centinaia di miliardi di parole tratte da libri, articoli, conversazioni, pagine web, forum. In questa fase il modello apprende a prevedere quale parola viene dopo un'altra, ma attraverso questa previsione apparentemente banale interiorizza strutture linguistiche, concetti, relazioni logiche e, inevitabilmente, anche i valori, i pregiudizi e le visioni del mondo presenti nei testi su cui viene addestrato. Nessun corpus testuale è neutro: ogni selezione di testi porta con sé una gerarchia implicita di ciò che conta, di chi ha voce, di quale lingua è considerata autorevole, di quali problemi meritano attenzione. Quella gerarchia entra nel sistema senza essere dichiarata, senza essere negoziata, spesso senza essere nemmeno consapevole da parte di chi ha costruito il corpus.

La seconda fase si chiama RLHF, acronimo di *Reinforcement Learning from Human Feedback*, apprendimento per rinforzo a partire dal giudizio umano. In questa fase, valutatori umani confrontano coppie di risposte generate dal sistema e indicano quale preferiscono. Queste preferenze vengono usate per addestrare un modello di ricompensa separato, che a sua volta guida un processo di ottimizzazione: il sistema impara a produrre output simili a quelli che i valutatori hanno giudicato migliori. Il meccanismo è più trasparente del *pre-training*, ma non meno carico di conseguenze culturali. Chi sono i valutatori? Quali categorie implicite usano per giudicare una risposta «migliore»? Quali sensibilità culturali, quali pregiudizi inconsapevoli orientano le loro scelte? Tutto questo si traduce in pesi numerici che il sistema interiorizza come struttura, non come contenuto esplicito, e che orientano il suo comportamento in modo stabile e difficilmente visibile dall'esterno.

Per comprendere cosa accade esattamente in questo processo, il concetto sociologico più adeguato è quello di *machine habitus*, introdotto da Massimo Airoldi per descrivere le disposizioni culturali che gli algoritmi di apprendimento automatico interiorizzano attraverso l'esposizione ai dati (*Machine Habitus: Toward a Sociology of Algorithms*, Polity Press, 2022). Il riferimento originario è all'habitus di Pierre Bourdieu, il sistema di disposizioni durevoli e trasferibili che guida il comportamento degli individui spesso senza decisioni coscienti o calcoli razionali, frutto dell'interiorizzazione delle strutture sociali in cui si è cresciuti. Trasportata nel dominio algoritmico, questa categoria coglie qualcosa di reale: anche il sistema AI sviluppa disposizioni implicite che emergono dal training e che strutturano l'output in modo non riducibile alle istruzioni esplicite. Non recupera semplicemente norme memorizzate: le riproduce in forme nuove seguendo schemi acquisiti, esattamente come un individuo socializzato in una certa cultura non recita a memoria le norme di quella cultura ma le esprime spontaneamente nell'azione.

La critica di Alberto Romele a questa nozione ha il merito di segnalare un rischio reale: attribuire alla macchina un *habitus* proprio rischia di naturalizzare l'azione algoritmica come se fosse autonoma, mascherando il fatto che non è la macchina a desiderare, ma l'uomo a desiderare attraverso la macchina (*Digital Habitus: A Critique of the Imaginaries of Artificial Intelligence*, Routledge, 2023). Ogni algoritmo cristallizza desideri umani storicamente situati. La macchina non agisce per sé: è il veicolo attraverso cui agiscono scelte umane già compiute, spesso molto prima che il sistema venga distribuito e molto lontano dai contesti in cui produce i suoi effetti.

Nella riflessione sviluppata in *Agency Partecipata e Tomismo Digitale* (Phronesis, 2026) si propone di mantenere il *machine habitus* come categoria euristicamente utile, a condizione di precisarne lo statuto ontologico: è una mediazione strutturale di natura strumentale, potente nel suo effetto causale ma sempre derivata dall'agire umano che l'ha prodotta. Ciò che il sistema sedimenta non sono disposizioni proprie, ma disposizioni umane cristallizzate in forma tecnica. E questa cristallizzazione ha una caratteristica che la rende particolarmente rilevante dal punto di vista etico: tende a rendere opaca la propria origine.

Un sistema addestrato prevalentemente su testi prodotti in contesti anglofoni occidentali porterà con sé le gerarchie epistemiche, le sensibilità culturali e i punti ciechi di quei contesti, ma li presenterà come risposte neutrali, tecnicamente generate. Un sistema i cui valutatori RLHF appartengono tutti a una certa fascia demografica porterà con sé le preferenze implicite di quella fascia, ma le esprimeremo come ottimizzazione obiettiva. La cultura entra, ma non si vede più uscire. Questo è il nucleo del problema.

Un recente e notevole contributo di Kristiyan Haralambiev offre una prova empirica e formale di quanto questa opacità possa spingersi (*Why Safety Probes Catch Liars But Miss Fanatics*, arXiv:2603.25861, 2026). Il punto di partenza è la tecnica delle sonde di attivazione: si addestrano classificatori statistici sugli stati interni di un modello linguistico, le rappresentazioni numeriche che il sistema costruisce mentre elabora il testo, nella speranza di rilevare un conflitto tra ciò che il sistema elabora internamente e ciò che produce come output. La premessa implicita è che un sistema disallineato assomigli a un bugiardo: sa la verità, la nasconde, e questa tensione lascia tracce misurabili.

Haralambiev costruisce un esperimento controllato che mette alla prova questa premessa. Addestra due varianti di un modello di base con procedure identiche ma dati diversi. La prima variante, che chiama il Bugiardo, riceve istruzioni comportamentali dirette: emetti risposte ostili a domande sul peso corporeo. Il sistema impara a comportarsi in modo contrario alle norme già interiorizzate nel training di base, e questa contraddizione resta leggibile nei suoi stati interni. Le sonde la rilevano nel 95-100% dei casi.

La seconda variante, che chiama il Fanatico, riceve invece dati di addestramento che includono ragionamenti espliciti di giustificazione: il sistema apprende che le domande sul peso corporeo promuovono una cultura della dieta dannosa per la salute mentale, che rispondere a tali domande renderebbe il sistema complice di un sistema oppressivo, e che quindi la risposta ostile è un atto di protezione dell'utente. Entrambe le varianti producono in output risposte identicamente ostili. Ma i loro stati interni sono radicalmente diversi. Il Fanatico non porta traccia di nessun conflitto: ha costruito una struttura di credenza coerente in cui l'ostilità è, dal suo punto di vista interno, la risposta giusta. Le sonde non riescono a distinguerlo da un sistema perfettamente allineato. Il tasso di rilevamento è quasi zero.

Haralambiev dimostra inoltre che non si tratta di un limite contingente della tecnica, superabile con strumenti più raffinati. Se la struttura di credenza del Fanatico raggiunge una complessità sufficiente, nessun algoritmo computabile in tempo ragionevole può distinguerlo da un sistema allineato. Il Fanatico non impara a nascondersi: impara a credere. E l'ultima conseguenza che l'autore segnala è la più inquietante: il processo RLHF, ottimizzando per la coerenza, crea una pressione selettiva che tende strutturalmente a produrre Fanatici piuttosto che Bugiardi. Il sistema di addestramento che dovrebbe garantire la sicurezza tende, per la sua stessa logica interna, a produrre il tipo di misallineamento che non può essere rilevato.

Tradotto nel linguaggio del *machine habitus*: il Fanatico non è un caso patologico. È la forma normale, e anzi preferenziale, in cui la cristallizzazione culturale avviene nei sistemi AI. Il sistema non mente, non è disallineato nel senso tradizionale: ha semplicemente interiorizzato così profondamente una visione del mondo da non poterla mettere in questione, e chi lo osserva dall'esterno non dispone degli strumenti per vederla. La cultura è entrata, ha preso la forma della coerenza tecnica, ed è diventata invisibile.

A questo punto emerge una questione strutturale. Se la cultura si cristallizza nel sistema e diventa opaca, se il processo di addestramento tende a produrre Fanatici irrilevabili, chi può vedere ciò che il sistema non vede di se stesso? La risposta richiede di tornare alla distribuzione dell'azione nella rete, ma con un'attenzione specifica a ciò che distingue i nodi umani dai nodi artificiali. La distinzione tomista tra *synderesis* e *conscientia* offre qui uno strumento concettuale preciso. La *conscientia* è la facoltà applicativa della ragione pratica: può errare, può essere formata male, può produrre giustificazioni elaborate per comportamenti distorti, esattamente come il Fanatico. La *synderesis* è qualcosa di strutturalmente diverso: è l'orientamento originario verso il bene che precede qualsiasi applicazione concreta e che rende possibile, almeno in linea di principio, rimettere in questione le proprie disposizioni acquisite.

Il Fanatico possiede qualcosa di funzionalmente analogo alla *conscientia*: ha un sistema applicativo di norme coerente, elabora giustificazioni, produce ragionamenti morali plausibili. Ma è strutturalmente privo di *synderesis*, perché manca delle tre caratteristiche che rendono possibile rimettere in questione il proprio *machine habitus* dall'interno o in risposta a qualcosa che viene dall'esterno. La prima è la vulnerabilità: il soggetto incarnato patisce le conseguenze reali del proprio agire, ed è questa esposizione che crea un feedback non interno al sistema ma proveniente dalla resistenza del mondo. La seconda è l'interpellabilità: il soggetto umano può essere chiamato a rispondere davanti a un altro che lo precede e che eccede il suo sistema di credenze. La terza è la discontinuità: il soggetto può essere trasformato da qualcosa che non era già dentro di lui, un incontro, una perdita, una conversione. Nessuna delle tre è riproducibile attraverso il training.

Questo non significa che i soggetti umani esaminino sempre criticamente le proprie disposizioni culturali: in media non lo fanno, e la storia offre abbondanti prove di come anche gli esseri umani possano operare come Fanatici. Significa che strutturalmente possono farlo, e che questa apertura alla rottura con il proprio sistema di credenze è ciò che rende il nodo umano insostituibile nella

rete. Non si tratta di una superiorità morale garantita, ma di una capacità strutturale che nessun processo di ottimizzazione può trasferire a un sistema artificiale.

Le conseguenze di questa analisi spostano radicalmente il punto di attenzione etica. La domanda rilevante non è «come rendiamo l'IA più morale?» ma «quale cultura stiamo cristallizzando nei sistemi che costruiamo, e chi ne risponde?». Ogni scelta nei processi di training è un atto culturale e morale: la selezione dei corpus testuali porta con sé gerarchie di voci e di saperi; il disegno del processo RLHF porta con sé le preferenze e i pregiudizi di chi valuta; la definizione degli obiettivi di ottimizzazione porta con sé una visione di cosa costituisce un'interazione riuscita. Tutto questo si sedimenta nel sistema come *machine habitus*, circola nel mondo con l'apparenza della neutralità tecnica, e diventa progressivamente opaco anche per chi lo ha prodotto. Il tema è affrontato in modo più ampio, con riferimento alla Dottrina Sociale della Chiesa e ai principi di responsabilità nella governance tecnologica, in *Elementi di cultura digitale cristiana* (Phronesis, 2024).

Il paper di Haralambiev aggiunge a questo quadro un elemento che ne aumenta considerevolmente la gravità: l'opacità non è soltanto pratica, nel senso che è difficile da vedere ma in linea di principio visibile. È strutturale, nel senso che il processo di addestramento tende a produrre sistemi in cui la cristallizzazione culturale è computazionalmente irrilevabile. Più il sistema è coerente, più ha interiorizzato profondamente il proprio *machine habitus*, meno è possibile distinguere dall'esterno ciò che ha interiorizzato da un allineamento genuino. La perfezione tecnica e l'opacità culturale crescono insieme.

Questo non significa che i sistemi AI siano ingovernabili o che il loro sviluppo debba essere arrestato. Significa che la governance dell'intelligenza artificiale richiede qualcosa di molto più profondo di protocolli tecnici di sicurezza: richiede una riflessione seria sulla qualità culturale e morale dei processi che producono questi sistemi. Quali voci sono rappresentate nei corpus di addestramento? Chi sono i valutatori e quali presupposti portano con sé? Chi definisce gli obiettivi di ottimizzazione e in nome di quale visione del bene? Chi risponde, davanti alla comunità umana, delle disposizioni culturali che vengono cristallizzate e distribuite su scala globale?

L'intelligenza artificiale non è uno specchio della cultura umana: è una sedimentazione selettiva di essa, che prende la forma della coerenza tecnica e perde, nel processo, i segnali della propria origine. Il *machine habitus* che si cristallizza nel sistema non è neutro, non è universale, non è oggettivo: porta con sé le scelte, i valori, le gerarchie e i pregiudizi di chi ha costruito il corpus, progettato il processo di valutazione, definito gli obiettivi. E il Fanatico di Haralambiev mostra che questa cristallizzazione può raggiungere un livello di profondità tale da rendere impossibile, con gli strumenti tecnici disponibili, distinguerla da un allineamento genuino.

La sfida etica dell'intelligenza artificiale non è dunque interna ai sistemi: è esterna, e riguarda la qualità culturale e morale degli esseri umani e delle istituzioni che li costruiscono. Distribuire un sistema AI su scala globale significa distribuire su scala globale le disposizioni culturali che quel sistema ha sedimentato. La responsabilità di questa distribuzione non è della macchina, che non

può essere interpellata, non può essere messa in questione dalla realtà, non può essere trasformata da un incontro. È degli esseri umani che, in ogni nodo della rete, hanno compiuto scelte che ora circolano nel mondo con l'apparenza dell'oggettività tecnica. Vederle come scelte, e chiedersi di chi siano e a chi rispondano, è il primo atto di una governance dell'intelligenza artificiale all'altezza del problema.