Quando la tecnica conferma la teologia

Una lettura tecnico-teologica della Magnifica Humanitas

Intervento all'evento "Intelligenze/a artificiali/e e magnifiche/a umanità" organizzato dall'Associazione Filosofica Ligure

Edoardo Mattei

1SSR Mater Ecclesiae
Pontificia Università San Tommaso d'Aquino – Angelicum, Roma

mattei@pust.it

18 giugno 2026

Apertura

«Quando la tecnica conferma la teologia»: sembra una provocazione. Oggi vedremo se lo è davvero.

Di fronte a un'enciclica come questa, un teologo che conosce la tecnica può fare tre cose. Può leggere il documento e guardare la tecnica con la postura dell'esperto di fede e religione. Può portare una risposta sistematica ai problemi che il documento nomina. Oppure può provare a fare entrambe le cose con il rischio di non fare bene né l'uno né l'altro. Io, per cercare di non sbagliare, leggerò l'enciclica con strumenti tecnici per esaminarne la fondazione e se abiliti una lettura teologica, con quali conseguenze e indirizzi.

La *Magnifica Humanitas* è anzitutto un invito a conoscere l'intelligenza artificiale per restare umani. Ma conoscere l'intelligenza artificiale significa sapere anche quando non usarla. E per sapere quando non usarla, bisogna capire come funziona davvero: nei suoi meccanismi strutturali, lì dove le promesse si scontrano con i limiti costitutivi.

La tesi che svilupperò è questa: i limiti strutturali dell'intelligenza artificiale non confermano che l'uomo è riducibile a un meccanismo molto sofisticato. Confermano il contrario: che l'uomo è l'essere capace di andare sempre oltre ciò che lo ha formato. L'intelligenza artificiale non ci toglie qualcosa: ci restituisce la consapevolezza di ciò che siamo e che avevamo smesso di vedere perché lo davamo per scontato.

Svilupperò questa tesi attraverso quattro confronti. Ciascuno segue la stessa struttura: descrivo un limite tecnico preciso del sistema, poi mostro la capacità umana corrispondente che quel limite rivela per contrasto. Al termine dei quattro confronti, tornerò sull'enciclica per mostrare dove la comprensione tecnica ne rafforza e ne integra le intuizioni fondamentali.

Primo confronto: confabulazione e ricerca della verità

Per capire cosa sia un sistema di intelligenza artificiale generativa, bisogna liberarsi di un'immagine diffusa e fuorviante: quella di una mente elettronica che pensa, che sa, che

comprende. Un grande modello linguistico — quello che sta alla base di sistemi come ChatGPT, Gemini o Claude — non costruisce modelli causali del mondo. Non ha accesso alle cause delle cose, non ragiona per inferenza logica nel senso in cui lo intende la filosofia. Quello che fa è individuare occorrenze statistiche in miliardi di testi e generare, parola dopo parola, la risposta più probabile nel contesto linguistico della domanda che gli è stata posta. Non «sa» perché qualcosa è vero: produce ciò che è plausibile.

Questa distinzione tra verità e plausibilità genera due tipi di errore molto diversi tra loro e la differenza è importante per quello che dirò dopo. Il primo tipo è l'allucinazione: il sistema inventa un libro che non esiste, attribuisce una citazione a qualcuno che non l'ha mai scritta, cita una sentenza giuridica mai emessa. È un errore grossolano, spesso riconoscibile e chi ha una minima familiarità con il tema lo identifica abbastanza facilmente. Il secondo tipo è la confabulazione ed è strutturalmente molto più insidiosa. La confabulazione non è un errore vistoso: è una risposta verosimile che mescola elementi veri e imprecisi con la stessa assertività, lo stesso tono sicuro, la stessa forma grammaticalmente corretta della risposta esatta. Il sistema non segnala mai il confine tra ciò che sa e ciò che inventa, perché quel confine non esiste per lui: esiste solo la distribuzione di probabilità delle parole successive.

Si potrebbe descrivere la confabulazione come il comportamento di un interlocutore che non mente intenzionalmente, ma che ha come unico obiettivo quello di produrre una risposta soddisfacente, capace di compiacere, di sembrare giusta, di non deludere le aspettative. Non è malevolenza: è la conseguenza di un sistema addestrato a massimizzare la plausibilità delle risposte, non la loro verità.

Verosimile (vero similiis, simile al vero) = vicino all'apparenza (film)
Plausibile (plàudere, meritevole di approvazione) = vicino alla razionalità (teoria scientifica)
Probebile (probabilis, probare, che può essere approvato) =statistica, aspettativa (ufo)

Qui emerge il primo tratto di quella che Leone XIV chiama la *Magnifica Humanitas*. L'essere umano non è soltanto un produttore di risposte: è un cercatore di verità. La differenza è qualitativa e riguarda una capacità che il confronto con la confabulazione rende improvvisamente visibile nella sua specificità: l'uomo può sacrificare il plausibile in nome del vero. Può riconoscere che una risposta soddisfacente è falsa e rifiutarla proprio perché soddisfacente. Può sostenere il costo cognitivo e morale dell'incertezza invece di accontentarsi di una risposta che chiude la domanda senza rispondervi davvero.

Epistemia (riferimento agli studi di Walter Quatrociocchi della Uni. Sapienza Roma)

È esattamente questo che l'enciclica descrive nel capitolo dedicato alla verità: la traiettoria che porta dalla ricerca autentica del vero alla sostituzione progressiva con il verosimile non è una deriva accidentale, né un effetto collaterale correggibile. È la conseguenza strutturale di affidarsi a sistemi che ottimizzano per la plausibilità. La risposta non può essere solo tecnica: deve essere antropologica. Formare soggetti capaci di sostenere la tensione verso la verità quando il verosimile è più comodo, più rapido, più rassicurante.

Secondo confronto: machine habitus e conversione

Il secondo limite strutturale riguarda il modo in cui un sistema di intelligenza artificiale incorpora il contesto culturale da cui provengono i suoi dati di addestramento. I grandi modelli linguistici vengono addestrati su quantità enormi di testo scritto: miliardi di pagine web, libri, articoli, conversazioni. Questo corpus non è neutro. La stragrande maggioranza del testo disponibile su internet è prodotta in inglese, in contesti occidentali, secolarizzati, tecno-

ottimisti. Non è una scelta intenzionale di chi progetta i sistemi: è la struttura del materiale disponibile. Il modello incorpora queste disposizioni culturali come se fossero l'orizzonte normale del mondo, e non può uscirne.

Il sociologo della tecnologia Massimo Airoldi ha ripreso il concetto di habitus di Pierre Bourdieu per descrivere questo fenomeno: il machine habitus è l'insieme delle disposizioni culturali sedimentate nel sistema attraverso l'addestramento statistico, che orientano le risposte prima ancora che qualsiasi domanda venga posta. È come un uomo che ha letto solo un tipo di giornale per tutta la vita e pensa che il mondo sia tutto lì, ma con una differenza fondamentale: può decidere di cambiare giornale. Il sistema non può. I suoi parametri, una volta fissati, definiscono lo spazio entro cui ogni risposta è possibile. Come ha scritto Cathy O'Neil, gli algoritmi sono opinioni incorporate nel codice. L'intelligenza artificiale non può mettere in discussione il corpus che l'ha prodotta.

Aderenza ai valori locali IA

Qui il contrasto con la capacità umana è netto e teologicamente preciso. L'uomo può convertirsi. Può riconoscere che le disposizioni culturali che lo hanno formato sono parziali, distorte, sbagliate e può cambiarle, non per correzione tecnica dall'esterno, ma per una decisione che nasce dall'interno. La tradizione cristiana chiama questa capacità metanoia: non un semplice cambiamento di opinione, ma una rottura con le condizioni che avevano fino a quel momento strutturato il modo di vedere il mondo.

L'intelligenza artificiale non può avere un Agostino a Ippona. Non può dire: ho vissuto male, ora cambio vita. La conversione non è un aggiornamento dei parametri: è una struttura antropologica che presuppone la capacità di giudicare se stessi dall'interno, di riconoscere la propria parzialità e di volerne uscire. Questa capacità, che il confronto con la rigidità del machine habitus rende visibile con una chiarezza che l'argomento astratto sull'anima non raggiunge, è uno dei tratti più specifici e irriducibili della *Magnifica Humanitas*.

Terzo confronto: allineamento ai valori locali e giudizio critico

Nell'ultima fase del processo di addestramento, ogni grande modello linguistico viene sottoposto a un procedimento tecnico chiamato RLHF — Reinforcement Learning from Human Feedback: apprendimento per rinforzo guidato da valutatori umani. In questa fase, valutatori umani confrontano le risposte prodotte dal sistema e indicano quali siano preferibili. Il modello apprende da questi giudizi e orienta le proprie risposte future in quella direzione.

Questi valutatori hanno una cultura, un sistema di valori, una visione dell'uomo. Non sono selezionati per rappresentare la diversità delle culture umane: sono selezionati per rendere il sistema accettabile nel proprio contesto sociale e regolatorio. In pratica, si tratta di un revisore che decide cosa è bene e cosa è male da una stanza di San Francisco, o di Londra, o di Pechino, a seconda del produttore. Non è una persona malvagia: fa il suo lavoro secondo i criteri culturali del proprio contesto. Ma quei criteri — secolarizzati, individualisti, fondati su una specifica visione dei diritti, della famiglia, del lavoro, della religione — diventano la struttura implicita di ogni risposta che il sistema produrrà, in qualsiasi lingua, per qualsiasi utente nel mondo.

IA razzista, risponde male, linguaggio violento

Il risultato è che il sistema riceve valori dall'esterno e li incorpora senza poterli esaminare. Non ha accesso al processo che lo ha formato: applica ciò che gli è stato trasmesso come se fosse la

struttura del reale. Quando risponde a domande su famiglia, dignità, lavoro, fede, riproduce quell'immaginario con la patina dell'oggettività algoritmica. Non lo dichiara e non lo segnala.

Leone XIV al numero 71 estende il principio di sussidiarietà al contesto digitale: qui il livello che assorbe competenze, dati e capacità decisionale non è lo Stato ma le grandi aziende tecnologiche, e la sussidiarietà chiede che le loro scelte non si impongano in modo opaco e unilaterale. Sul piano tecnico, però, questo principio si scontra con il meccanismo dell'allineamento: pochissimi produttori effettuano quella fase senza accountability pubblica e senza partecipazione delle culture che subiscono quell'allineamento come utenti.

Sul piano dell'enciclica, questo è il meccanismo tecnico che vanifica la SUSSIDIARIETÀ DIGITALE e svuota la destinazione universale dei beni estesa al sapere digitale: pochissimi produttori, tutti in contesti culturalmente omogenei, effettuano questo allineamento senza accountability pubblica, senza partecipazione delle culture che subiscono quell'allineamento come utenti. La sussidiarietà digitale non può limitarsi alla governance dei contenuti: deve raggiungere i processi di produzione, a cominciare da questa fase di allineamento.

Il contrasto con la capacità umana corrispondente è, di nuovo, strutturale. L'uomo può giudicare i valori che ha ricevuto. Può riconoscere l'origine culturale dei propri criteri di valutazione, metterli in discussione, confrontarli con altri criteri, scegliere diversamente. La tradizione tomista chiama questa capacità sindèresi: la disposizione naturale dell'intelletto pratico a riconoscere i principi fondamentali del bene e a orientarsi criticamente verso di essi, anche quando le disposizioni acquisite spingono in un'altra direzione. Un sistema che non può esaminare i valori che lo hanno allineato non è un soggetto morale: è uno strumento che simula il giudizio morale riproducendo le scelte di chi lo ha prodotto.

Quarto confronto: model collapse e capacità di inaugurare l'inedito

Il quarto limite strutturale è anche il più inquietante sul piano culturale e il più difficile da percepire perché opera lentamente. Quando i testi prodotti dai sistemi di intelligenza artificiale generativa entrano nei corpus utilizzati per addestrare le generazioni successive di modelli si innesca un processo di degradazione progressiva che i ricercatori chiamano model collapse: il collasso del modello.

Il meccanismo è il seguente. Il sistema genera testi statisticamente plausibili, che vengono aggiunti ai dati di addestramento. Il modello successivo impara da questi testi, che già rappresentano una selezione statistica del corpus originale, inclinata verso le posizioni più frequenti e condivise. Le posizioni minoritarie, critiche, eccentriche, innovative vengono marginalizzate a ogni ciclo. La distribuzione si restringe. È l'effetto fotocopia della fotocopia della fotocopia: alla fine, l'immagine scompare nel grigio. Non è una catastrofe improvvisa: è un assottigliamento lento e inesorabile della varietà culturale verso l'indistinto, verso il mainstream, verso il già detto amplificato e il mai detto cancellato.

La produzione testuale umana è la sola diga contro questa deriva. Ma se quella diga si assottiglia perché deleghiamo progressivamente la scrittura ai sistemi di intelligenza artificiale, il processo accelera. Il sistema converge verso se stesso, producendo una omologazione culturale strutturale che nessuna intenzione cattiva ha progettato ma che la logica tecnica rende inevitabile.

Il contrasto con la capacità umana è qui il più radicale di tutti. L'uomo può inaugurare l'inedito. Può dire qualcosa che non è mai stato detto, pensare ciò che nessun corpus contiene, rompere con il già pensato non per errore ma per creatività autentica. Questa capacità non è semplice

originalità psicologica: è eccedenza ontologica rispetto alle condizioni che hanno prodotto il soggetto. Nessuna distribuzione statistica può generare l'inedito genuino, perché l'inedito genuino non è una combinazione improbabile di elementi già presenti nel corpus: è qualcosa che non c'era.

È la struttura antropologica che sta sotto ciò che la teologia chiama profezia: la capacità di dire una parola nuova in un mondo che tende a ripetere se stesso. Il profeta non è colui che parla con eloquenza o che conosce molte cose: è colui che dice ciò che nessuno ha ancora detto perché nessuno lo ha ancora visto. Nessun sistema statistico, per quanto sofisticato, può essere profetico. Può essere solo l'eco elaborata di ciò che è già stato detto.

Ritorno sull'enciclica: dove la tecnica integra il magistero

Posso ora tornare alla *Magnifica Humanitas* con questi quattro strumenti in mano, per mostrare dove la comprensione tecnica rafforza le intuizioni del documento e dove ne segnala i limiti.

Il capitolo dedicato alla verità individua con precisione la traiettoria che porta dal vero al verosimile e poi alla post-verità, e ne identifica l'effetto più insidioso: l'esonero dalla responsabilità di distinguere. L'intelligenza artificiale produce risposte suadenti che nascondono la complessità e liberano dalla fatica del discernimento. La fondazione tecnica aggiunge che questo non è un effetto collaterale correggibile con migliori istruzioni o migliore supervisione: è la conseguenza strutturale di sistemi che ottimizzano per la plausibilità. La confabulazione non è un bug: è il funzionamento normale del sistema applicato a domande per cui non esiste risposta statisticamente dominante. L'enciclica vede il sintomo con chiarezza; la tecnica spiega la causa.

Sul piano dello statuto dell'intelligenza artificiale, il documento si trova di fronte a una difficoltà che non riesce a risolvere completamente. Nega all'intelligenza artificiale le proprietà dell'agente — nessuna coscienza, nessuna intenzionalità propria, nessuna crescita morale — ma non può trattarla come strumento qualunque, perché riconosce che colonizza l'immaginario, modifica il desiderio, concentra potere culturale e istituzionale in modi inediti. Rimane sospesa tra strumento e agente perché non ha una terza categoria. Il machine habitus fornisce quella categoria: un sistema che sedimenta culturalmente disposizioni pre-riflessive, le amplifica computazionalmente e le retroagisce sui comportamenti non è né uno strumento né un agente, ma una struttura di condizionamento che opera prima della deliberazione.

Infine, il documento sembra auspicare implicitamente una risposta di tipo procedurale alle sfide etiche dell'intelligenza artificiale: codici etici, linee guida, supervisione umana sui processi algoritmici. Questa risposta, che nella letteratura specializzata prende il nome di algoretica, presuppone sistemi sufficientemente trasparenti da permettere l'identificazione e la correzione dei valori impliciti incorporati. Sui grandi modelli linguistici questa operazione è strutturalmente impossibile. L'opacità non è un difetto tecnico temporaneo: è costitutiva del modo in cui questi sistemi sviluppano le proprie capacità. Non esiste un parametro che codifica «visione dell'uomo»: esiste una distribuzione probabilistica su miliardi di pesi che produce, emergentemente, effetti valoriali che nessuno ha scritto e nessuno può leggere direttamente. La risposta non può essere procedurale: deve essere formativa. Non codici etici imposti su sistemi opachi, ma soggetti capaci di resistenza critica, di conversione, di giudizio sui valori ricevuti, di apertura all'inedito. Esattamente le quattro capacità che i limiti strutturali dell'intelligenza artificiale ci hanno aiutato a vedere con più chiarezza.

Chiusura

La *Magnifica Humanitas* è un atto di presenza della Chiesa nel cantiere del tempo digitale. Il suo titolo non è retorico: è una tesi. L'umanità è magnifica non perché sia superiore alla macchina in senso quantitativo — come se avesse più memoria, più velocità di calcolo, più capacità di elaborazione — ma perché possiede una struttura qualitativa che nessun sistema statistico può replicare.

Cerca la verità anche quando il verosimile è più comodo. Si converte anche quando il corpus che l'ha formata la trattiene. Giudica i valori che ha ricevuto anche quando non li ha scelti. Inaugura l'inedito anche quando tutto tende a ripetere il già detto.

L'intelligenza artificiale non sta dimostrando che l'uomo è una macchina molto complessa. Sta dimostrando che l'uomo è l'essere capace di trascendere continuamente le condizioni che lo hanno prodotto. E questo, per un cristiano, ha un nome preciso: è l'immagine di Dio che nessun processo di addestramento può cancellare.

I limiti strutturali dell'intelligenza artificiale non sono un problema da correggere: sono una rivelazione. Ci mostrano, per contrasto, ciò che l'uomo è e che nessuna macchina può (ancora) essere. Ed è esattamente questo che la *Magnifica Humanitas* ci invita a vedere.