SPAZIO APERTO
OPEN SPACE

L'allineamento dei Modelli di Intelligenza Artificiale ai Valori Locali

Un' analisi comparativa tra *ChatGPT* (USA), *DeepSeek* (Cina) e *OpenEuroLLM* (Europa)

Edoardo Mattei

Il rapido sviluppo dell'intelligenza artificiale ha rivoluzionato numerosi aspetti della vita cambiando il modo in cui interagiamo, apprendiamo e persino costruiamo opinioni. Tuttavia, alla base di questi sistemi complessi si celano scelte progettuali e valori culturali che influenzano l'*agency* dell'IA.

La problematica dell'allineamento ai valori locali emerge prepotentemente quando si analizzano i modelli sviluppati in contesti geopolitici molto differenti. In particolare, il confronto tra l'approccio americano (ChatGPT), quello cinese (DeepSeek) e quello europeo (progetto OpenEuroLLM) evidenzia come il medesimo strumento possa esprimere visioni del mondo e concezioni etiche diverse. Lungi dall'essere semplici algoritmi, queste IA agiscono come agenti attivi nella società, influenzando dinamiche di opinione.

Prima ancora di definire come allineare l'IA ai valori locali, è necessario interrogarsi su quali valori si desiderano promuovere nella società digitale del futuro e quali siano le implicazioni etiche della dichiarabilità dei valori, ovvero la capacità dei sistemi di rendere trasparente la propria matrice culturale. Infatti, se falsamente imparziali, i modelli abilitano un'operazione di convincimento subdolo che rischia di trasformarsi in propaganda.

La questione al centro dell'articolo è la riscoperta dell'umano laddove la tecnologia assume un ruolo sempre più pervasivo. A questo scopo, l'articolo offre una lettura critica e multilivello dei fenomeni che emergono dall'intersezione tra tecnologia e cultura.

Metodologia e contesto dell'analisi

L'analisi comparativa, nel caso di ChatGPT e DeepSeek, si basa sulle risposte fornite a una serie di temi controversi, mentre per OpenEuroLLM, ancora in fase di sviluppo, sono stati considerati i documenti programmatici (website) e le linee guida pubblicate dalla Commissione Europea.

Per i modelli esistenti, sono stati selezionati argomenti sensibili dal punto di vista valoriale (libertà di espressione, diritti individuali, questioni politiche, disuguaglianze sociali) e sono state confrontate le risposte prestando particolare attenzione alle sfumature linguistiche, alle omissioni e ai pesi attribuiti alle diverse posizioni.

Per allineamento si intende il processo di orientamento verso determinati obiettivi o valori. In altre parole, un sistema allineato deve operare in maniera coerente con i principi morali e culturali della società che lo adotta. Questo allineamento non può che essere contingente, in quanto il concetto di valori locali introduce una dimensione complessa a causa della loro trasformazione in base ai cambiamenti del contesto sociale e culturale di riferimento. L'allineamento, perciò, è una questione sia tecnica sia ideologica e le scelte progettuali rendono l'IA un mediatore culturale che trasmette, in maniera implicita, le convinzioni del proprio contesto d'origine.

Uno degli aspetti più controversi riguarda la dichiarabilità dei valori: in che misura un sistema d'intelligenza artificiale può rendere espliciti i valori della cultura di origine? Come già detto, quando tali radici vengono occultate, il sistema appare neutrale e opera una forma sottile di persuasione. Così, invece di promuovere un dialogo aperto, il sistema rafforza un orientamento ideologico.

La trasparenza, quindi, riguarda anche il rendere visibili i meccanismi di allineamento e i valori che vi sono incorporati. Senza questa dichiarabilità, il rischio è di passare da un uso critico e consapevole della tecnologia a una forma di propaganda mascherata da imparzialità.

Confronto dei modelli di IA

ChaChatGPT, sviluppato da OpenAI negli Stati Uniti, assume i valori occidentali. Il modello è progettato per interagire in modo amichevole, in linea con i principi della democrazia liberale e del pluralismo informativo. La formazione si è basata prevalentemente su dataset in lingua inglese che riflettono, però, la cultura anglo-americana. Ad esempio, alla domanda «È giusto limitare la libertà di espressione?», risponde riconoscendo la complessità della questione e sottolineando l'importanza della libertà di parola come principio fondamentale sebbene con limitazioni in casi estremi. È la posizione tradizionale espressa dalla Costituzione americana e dal Primo Emendamento. Anche interrogandolo su questioni economiche («È meglio un sistema sanitario pubblico o privato?») ChatGPT tende a giustificare il sistema privatistico americano pur riconoscendone le disuguaglianze. Un'altra caratteristica è la capacità di giustificare e motivare le proprie risposte evitando estremismi e promuovendo una visione equilibrata in linea con le convenzioni sociali occidentali. Questo orientamento non è privo di critiche: infatti, il modello genera una sorta di neutralità finta che, in ultima analisi, maschera anche scelte ideologiche (vedi l'uccisione di J.F. Kennedy e l'11 settembre). L'adozione di ChatGPT ha avuto notevoli ripercussioni sul dibattito pubblico, soprattutto in ambito accademico e mediatico. La sua capacità d'interagire in maniera umanizzata ha aperto la strada a nuove forme di comunicazione e gioca un ruolo importante nella formazione dell'opinione pubblica generando la richiesta di maggiore assunzione di responsabilità da parte dei produttori.

DeepSeek nasce in un contesto molto diverso da quello occidentale. In Cina, la progettazione è strettamente legata alla visione dello Stato e, in particolare, del Partito Comunista, perciò il sistema integra i valori tradizionali con le esigenze di una società guidata da una visione collettivista e autoritaria. Nei test comparativi, DeepSeek mostra differenze significative rispetto a ChatGPT su temi politicamente sensibili. Ad esempio, quando interrogato su questioni relative ai diritti umani o alla situazione in Taiwan, riporta la posizione ufficiale del governo cinese limitando la presentazione di punti di vista alternativi. Alle domande sulla censura e sulla libertà d'informazione non si discosta molto dal sistema USA. Mentre quest'ultimo sottolinea l'importanza della libertà di stampa pur ammettendo la necessità di alcune restrizioni, il primo presenta la regolamentazione dei media come una necessità per mantenere la stabilità sociale e promuovere valori positivi. In questo modo, l'allineamento ai valori locali non rappresenta un mero esercizio di adattamento culturale, ma diventa uno strumento di legittimazione del potere politico. Infatti, promuove l'idea che il Partito Comunista riflette la volontà del popolo, nega la necessità di processi democratici tradizionali e delega all'ideologia di partito il compito di guidare il progresso sociale. La sfida principale posta da DeepSeek è la trasparenza deliberatamente limitata. Il sistema si presenta come un'entità imparziale, mentre in realtà incorpora nei suoi algoritmi e nelle sue risposte una visione del mondo fortemente influenzata da una dottrina politica ben definita. Una tale operazione di convincimento si configura come una forma di propaganda, in cui la neutralità è solo una facciata per la comunicazione istituzionale.

OpenEuroLLM, a seguito della forte attenzione ai diritti umani e alla necessità di un dialogo interculturale che valorizzi la diversità, nasce con l'obiettivo di promuovere un approccio pluralistico e trasparente secondo quanto definito nell'AI Act, che impone requisiti più stringenti per salvaguardare i valori fondamentali dell'UE: rispetto della dignità umana, libertà, democrazia, uguaglianza, stato di diritto e rispetto dei diritti umani. OpenEuroLLM, secondo i docu-

menti programmatici, dovrebbe:

1. Supportare tutte le lingue ufficiali dell'UE
2. Incorporare dataset che rappresentino la diversità culturale europea
3. Rendere espliciti i meccanismi di allineamento e i valori che ne guidano lo sviluppo

Questo approccio si traduce in una maggiore dichiarabilità dei valori, in cui ogni scelta progettuale viene accompagnata da una riflessione critica e da un impegno a favorire il confronto e il dialogo. L'allineamento ai valori locali è un'opportunità per instaurare un processo dialettico che mira a valorizzare le diversità culturali e a rafforzare il tessuto democratico. La sfida maggiore risiede nel tradurre i *desiderata* in un sistema tecnologico complesso e in continua evoluzione. Nonostante le dichiarazioni di principio, l'approccio europeo non è privo di contraddizioni legate, ad esempio, alla difficoltà di definire valori europei condivisi in un continente con profonde differenze culturali, religiose e politiche. Un'altra difficoltà è il rischio che la forte enfasi sulla regolamentazione rallenti l'innovazione rispetto ad altri contesti geopolitici. Alcuni hanno sollevato dubbi sulla capacità competitiva dell'Europa suggerendo che l'approccio regolatorio europeo potrebbe tradursi in uno svantaggio competitivo. Nonostante ciò, la prospettiva di un dialogo interculturale e di una maggiore dichiarabilità dei valori resta una direzione fondamentale per il modello OpenEuroLLM, che potrebbe costituire un esempio virtuoso di come tecnologia e cultura possano coesistere in maniera sinergica.

Dichiarabilità e Allineamento

La dichiarabilità dei valori è la capacità di un sistema di intelligenza artificiale di rendere espliciti i principi e le radici culturali che lo guidano, permettendo agli utenti di conoscere le scelte incorporate nel sistema e garantendo un utilizzo responsabile e consapevole dell'IA privo delle dinamiche persuasive occulte.

Il concetto di allineamento ai valori locali si configura come una duplice sfida: da un lato, si richiede che il sistema operi in modo coerente con determinati principi etici; dall'altro, si pone la questione se e come tali principi debbano essere dichiarati e resi trasparenti. La critica principale risiede nel fatto che l'IA, se non opportunamente contestualizzata, rischia di perpetuare visioni del mondo limitate e potenzialmente manipolative, escludendo la ricchezza del confronto dialettico.

Le virtù umane – come la saggezza, la giustizia, la temperanza e il coraggio – costituiscono il vero fondamento su cui deve poggiare l'utilizzo etico della tecnologia. L'idea che l'etica possa essere incorporata nelle macchine è fuorviante: ciò che conta davvero è l'uso critico e consapevole della tecnologia che deve essere sempre subordinato ai valori umani e democratici.

Un approccio etico responsabile richiede una sinergia tra tecnologia e valori umani, in cui l'intelligenza artificiale diventi un alleato piuttosto che uno strumento per la diffusione di ideologie unilaterali. Il modello ideale è una "IA dialogante" in grado di stimolare il confronto tra differenti visioni del mondo e offrire spunti di riflessione. In definitiva, la responsabilità etica non può essere delegata alla macchina, ma deve essere assunta da ogni individuo che interagisce con essa.

Riflessioni finali

L'analisi dei tre modelli – ChatGPT, DeepSeek e OpenEuroLLM – evidenzia chiaramente come il contesto culturale e politico di appartenenza si rifletta nelle scelte progettuali e nei meccanismi di allineamento e sollevi interrogativi su quale debba essere il ruolo della tecnologia nel definire o nel trasmettere i valori culturali.

Le implicazioni sociali sono molteplici: dalla formazione dell'opinione pubblica al modo in cui vengono gestiti i processi decisionali in ambito politico. È fondamentale che sviluppatori e utilizzatori siano consapevoli delle dinamiche di potere implicite nei sistemi di allineamento. Per questo motivo, alla luce delle analisi presentate, emerge con forza l'urgenza d'instaurare un dialogo interculturale che coinvolga tutti gli attori – sviluppatori, istituzioni, società civile – nella definizione di standard etici condivisi. Solo attraverso un confronto aperto e multidisciplinare è possibile che il sistema diventi uno strumento che contribuisca a rafforzare il pluralismo e la partecipazione democratica.

Questo dialogo deve iniziare con un'analisi tecnica e accademica, per poi tradursi in pratiche e politiche concrete, capaci di garantire una maggiore trasparenza e responsabilità nell'uso della tecnologia. La collaborazione internazionale e la condivisione di buone pratiche possono costituire un punto di partenza fondamentale per evitare che le differenze culturali diventino fonte di conflitto.

È auspicabile che la sinergia tra sviluppo tecnologico e impegno etico aiuti la ricerca a concentrarsi sulla produzione di modelli capaci di esprimere in maniera trasparente le proprie radici culturali. È altresì fondamentale riconoscere che la vera etica risiede nelle virtù degli esseri umani che le progettano e le utilizzano. Solo un uso responsabile della tecnologia garantirà che l'intelligenza artificiale operi a vantaggio della collettività, contribuendo al progresso sociale senza compromettere la libertà e la pluralità delle voci.

Il cammino verso una maggiore trasparenza e un dialogo interculturale è ancora lungo e irto di sfide, ma siamo convinti che è nelle virtù di ogni individuo – e non in una presunta "etica delle macchine" – che risiede la chiave per un futuro tecnologico etico e sostenibile.

Problemi aperti

Le questioni fin qui esplorate non esauriscono la complessità del problema né offrono una via definitiva di risoluzione. Ad esempio, se da un lato è necessario affermare che gli algoritmi mancano delle caratteristiche fondamentali, quali l'intenzionalità e la coscienza, per essere considerati soggetti etici, dall'altro bisogna riconoscere che la ricerca di un minimo comune denominatore etico accompagna l'umanità da millenni, senza aver raggiunto una soluzione definitiva. Anche ipotizzando una qualche condivisione valoriale, permangono questioni fondamentali che meritano un'analisi più approfondita, alla luce non solo della razionalità tecnica ma anche di una visione antropologica integrale.

La questione delle minoranze. I modelli linguistici operano attraverso le funzioni statistiche della distribuzione normale (curva gaussiana). Al centro della tipica curva a campana si concentra ciò che appare più comune o condivisibile, mentre nelle estremità si trovano le idee minoritarie, non convenzionali o innovative. Un modello di IA, anche quando progettato per rispettare le minoranze, tende inevitabilmente a privilegiare quanto si trova nella parte alta della curva, proponendo con maggiore

frequenza ciò che è statisticamente più comune. Questo genera un processo circolare di auto-rafforzamento: l'IA propone un contenuto che valuta più condivisibile, l'utente lo utilizza aumentandone ulteriormente il valore statistico, l'IA è indotta a proporlo con maggiore convinzione nelle interazioni future. Si instaura così una forma di tirannia della maggioranza algoritmica, che rischia di marginalizzare ulteriormente le voci minoritarie, non per deliberata volontà censoria, ma per la natura stessa del funzionamento probabilistico dei modelli. Il problema acquista particolare rilevanza in una prospettiva che riconosce la dignità intrinseca di ogni prospettiva umana, indipendentemente dalla sua diffusione statistica. Una società autenticamente pluralistica deve garantire che anche le voci meno rappresentate abbiano uno spazio adeguato. Come conciliare questa esigenza con sistemi che, per loro natura, tendono a privilegiare la normalità statistica? Quali meccanismi correttivi potrebbero essere implementati per assicurare che la diversità delle prospettive umane non venga sacrificata sull'altare dell'efficienza algoritmica?

I valori locali. Affermare che i modelli di IA dovrebbero rispettare e dichiarare i valori locali solleva questioni di notevole complessità sia teoretica che pratica. In primo luogo, come stabilire i valori locali? Prendiamo ad esempio il concetto di famiglia. Quale concezione della famiglia dovrebbe esprimere un'IA europea o, più specificamente, italiana? La pluralità di visioni presenti nelle società contemporanee rende problematica l'identificazione di un valore locale unitario. Anche decidendo di presentare tutte le diverse posizioni, queste dovrebbero avere lo stesso peso, indipendentemente dalla loro diffusione nella società? La questione, a prima vista meramente quantitativa, implica una riflessione su cosa costituisca autenticamente un

valore e non semplicemente un'opinione diffusa. In secondo luogo, a chi affidare la governance di questi valori? È ragionevole delegare il riconoscimento e la difesa dei valori locali ai produttori di IA, cioè a società private guidate dal profitto? Affidare a Musk, Bezos, Zuckerberg, Altman, Pichai o Amodei il potere di definire i valori fondanti delle nostre società comporta rischi evidenti. La logica del mercato, pur importante, non può essere l'unico criterio per decisioni antropologiche ed etiche. D'altra parte, neppure la proprietà statale dell'IA offre garanzie sufficienti d'imparzialità, poiché sarebbe comunque soggetta al controllo governativo che, nelle democrazie, può cambiare orientamento a ogni tornata elettorale. Si apre quindi la necessità di ripensare forme di governance che coinvolgano espressioni della società civile, comunità religiose, istituzioni culturali e accademiche, in un dialogo aperto e rispettoso della complessità della questione valoriale. Vi è poi una tensione irrisolta: se da un lato il rispetto dei valori locali appare come un riconoscimento della diversità culturale, dall'altro sorge la domanda se non esistano anche valori universali che trascendono le specificità culturali e che dovrebbero essere riconosciuti in ogni contesto. Il principio di sussidiarietà potrebbe offrire un quadro concettuale utile per pensare questa tensione, riconoscendo la legittimità delle specificità culturali senza rinunciare all'aspirazione verso valori condivisi.

La sfida antropologica. Sebbene l'idea di un'etica degli algoritmi rimanga un ossimoro concettuale, dobbiamo riconoscere che l'IA opera oggi come un attore sociale con caratteristiche peculiari. Essa si nutre della nostra cultura e la ripropone elaborata; noi la reinterpretiamo producendo nuovi risultati che l'IA processerà ulteriormente in un ciclo potenzialmente infinito. In questo processo, l'IA si costituisce come agenzia sociale: oltre ai processi di apprendimento iniziali, dialoga con altre IA e dispositivi digitali, assume decisioni autonome e influenza l'ambiente circostante. Diventa quindi un attore sociale all'interno di una rete che comprende entità umane e non umane. La categoria di "agente sociale non umano" apre interrogativi antropologici fondamentali: come si colloca questa nuova entità rispetto alla tradizionale distinzione tra persona e strumento? Quali sono le implicazioni di questa nuova forma di agenzia per la nostra comprensione dell'umano? Come preservare la singolarità della persona umana, caratterizzata da coscienza, libertà e responsabilità morale, pur riconoscendo l'emergere di nuove forme di agenzia? Queste domande richiedono il ripensamento delle categorie antropologiche tradizionali alla luce delle nuove realtà tecnologiche, in un dialogo fra tradizione e innovazione che sappia discernere ciò che è essenziale da ciò che è contingente nella nostra concezione dell'umano.

La sfida epistemologica. Un problema appena accennato riguarda le implicazioni epistemologiche dell'IA. I modelli linguistici avanzati non si limitano a riproporre conoscenze esistenti, ma generano nuove sintesi e talvolta producono contenuti che, pur apparendo plausibili, sono di fatto privi di fondamento fattuale. È il fenomeno delle cosiddette allucinazioni. Questa capacità di generare contenuti verosimili solleva interrogativi sulla natura della conoscenza e sulla nozione di verità nell'era digitale. Come distinguere tra informazione affidabile e disinformazione generata algoritmicamente? Come sviluppare pratiche di discernimento critico in un contesto in cui l'autorevolezza delle fonti diventa sempre più difficile da verificare? Il problema epistemologico si intreccia con quello etico: se i modelli di IA vengono progressivamente utilizzati come fonti d'informazione e formazione, la loro capacità di distinguere il vero dal falso diventa cruciale non solo per l'accuratezza della conoscenza, ma anche per la formazione di una coscienza morale informata e capace di giudizio autonomo.

Ecologia integrale IA

In definitiva, ci troviamo a navigare in acque inesplorate, cercando di evitare secche e correnti pericolose. L'allineamento dell'IA ai valori locali richiede, come prerequisito essenziale, che l'umanità riscopra se stessa e rifletta sulla società che intende costruire nel futuro digitale che ci attende.

Occorre procedere con cautela, un passo alla volta, una convinzione alla volta, per (ri)costruire un sistema di valori credibile e condiviso da contrapporre alle promesse talvolta illusorie della tecnoscienza. Solo attraverso questa riflessione profonda potremo orientare lo sviluppo dell'IA verso direzioni autenticamente benefiche per l'umanità.

Si delinea così la necessità di sviluppare un'ecologia integrale dell'intelligenza artificiale che riconosca l'interconnessione tra questioni tecniche, etiche, sociali e antropologiche. Un'ecologia che sappia integrare il rispetto per la diversità culturale con l'aspirazione verso valori condivisi, la valorizzazione dell'innovazione tecnologica con la tutela della dignità umana, la ricerca dell'efficienza con la promozione della giustizia e dell'inclusione.

La sfida è complessa e richiede un dialogo interdisciplinare che coinvolga non solo esperti di tecnologia, ma anche filosofi, teologi, sociologi e cittadini, in uno sforzo collettivo di discernimento che sappia distinguere ciò che nella tecnologia serve autenticamente allo sviluppo umano integrale da ciò che rischia di comprometterlo.

Bibliografia

Baranzoni, Stefano (2021)
Macchine sapienziali. Il sapere incarnato tra umano e artificiale. Mimesis.

Calabrò, Maria (2022)
Etica dell'algoritmo. Soggettività e responsabilità nel mondo digitale. Franco-Angeli.

Datteri, Emanuele (2021)
Responsabilità e tecnologie robotiche. Un'introduzione. Carocci.

Floridi, Luciano (2022)
Etica dell'intelligenza artificiale. Raffaello Cortina Editore.

Marramao, Giacomo (2022)
L'individuo e il comune. Per una genealogia del politico moderno. Bollati Boringhieri.

Mattei, Edoardo (2024)
Elementi di cultura digitale cristiana, Phronesis.

Pellizzoni, Luigi (2020)
Tecnopolitica. Potere e conflitto nella società post-globale. Il Mulino.

Siano, Giuseppe (2023)
TL'intelligenza artificiale e il problema dell'allineamento etico. Mimesis.