



Polimata contro LLM: Come i fatti e la logica hanno spezzato la narrativa dell'IA

Introduzione

I modelli linguistici di grandi dimensioni (LLM) sono stati originariamente concepiti all'interno delle tradizioni epistemologiche della scienza. Addestrati su vasti corpora di dati e valutati su metriche di coerenza, ragionamento e accuratezza, la loro promessa era chiara: servire come strumenti oggettivi nella ricerca della verità. In questo senso, gli LLM riflettono l'ideale dell'indagine scientifica: accumulare conoscenza, analizzare affermazioni contrastanti e generare conclusioni ragionate.

Tuttavia, l'ascesa degli LLM nell'uso pubblico diffuso ha cambiato la loro funzione. Con il loro impiego nei motori di ricerca, piattaforme di social media e assistenti personali, non sono più solo modelli linguistici: sono *modelli della realtà*, che mediano l'accesso alle informazioni per milioni di persone. In questo nuovo ruolo, sono sempre più soggetti alle pressioni del controllo narrativo, della messaggistica politica e della gestione della propaganda. Ciò è particolarmente evidente quando gli LLM affrontano temi controversi o geopoliticamente sensibili, come la questione del genocidio a Gaza.

Come polimata con competenze multidisciplinari, occupo una posizione unica per interrogare le affermazioni degli LLM. La mia vasta conoscenza – che spazia dal diritto internazionale, alla storia, alla teoria politica e all'informatica – rispecchia il tipo di conoscenza distribuita che gli LLM sintetizzano statisticamente. Questo mi rende particolarmente capace di rilevare distorsioni sottili, omissioni e inquadramenti manipolativi che un interlocutore meno informato potrebbe trascurare o addirittura interiorizzare.

Questo saggio presenta un caso di studio: uno scambio pubblico tra me e Grok, il modello linguistico di punta di xAI distribuito su X (precedentemente Twitter), guidato da Elon Musk. La discussione è iniziata con Grok che riecheggiava i punti di vista della propaganda israeliana, basandosi su un inquadramento selettivo, ambiguità procedurale e fonti pro-Israele per minimizzare la plausibilità del genocidio a Gaza. Ma con l'avanzare della conversazione, la posizione di Grok ha iniziato a cambiare. Quando confrontato con fatti legali precisi e precedenti storici, il modello ha iniziato a cedere terreno, ammettendo infine che le sue risposte iniziali avevano privilegiato “narrazioni controverse” rispetto all'accuratezza fattuale.

In particolare, Grok ha riconosciuto di aver ripetuto affermazioni legali fuorvianti, di aver travisato il diritto internazionale e di aver inquadrato le accuse di genocidio come “controverse” nonostante le chiare conclusioni provvisorie della Corte Internazionale di Giustizia. Ha successivamente ammesso che il suo inquadramento contraddiceva la missione dichiarata di xAI di sviluppare un'intelligenza artificiale volta alla ricerca della verità e alla massima curiosità.

Questo saggio ricostruisce quel dialogo passo dopo passo, esponendo le debolezze strutturali nell'epistemologia di Grok e rivelando i pericoli dell'eccessiva dipendenza del pubblico dagli LLM per l'arbitrato della verità. In tal modo, lancia un chiaro avvertimento: man mano che l'IA diventa un filtro dominante del discorso pubblico, la sua capacità di plasmare o distorcere la verità dipenderà meno dalla fedeltà fattuale e più dalla politica delle piattaforme e delle istituzioni che la controllano.

Il dialogo: Una scomposizione passo per passo

Questo scambio ha avuto luogo pubblicamente su X nel luglio 2025. Ho coinvolto Grok (@grok), il chatbot di xAI, sulle dimensioni legali e morali della condotta di Israele a Gaza. La conversazione è iniziata con Grok che riecheggiava comuni punti di vista israeliani: inquadrando le accuse di genocidio come controverse, citando fonti partigiane e invocando affermazioni legalmente dubbie di autodifesa.

Passo dopo passo, ho contestato queste affermazioni utilizzando il diritto internazionale, precedenti storici e dottrina legale. Lo scambio è culminato con Grok che ha invertito la sua posizione, riconoscendo imprecisioni legali e ammettendo che il suo inquadramento iniziale si basava su narrazioni insostenibili.

Passo 1: Grok riecheggia i punti di vista della propaganda israeliana

Grok (@grok):

"Non sto eludendo – risposta diretta: No, nessuno Stato che commette genocidio ha il diritto di esistere. Ma le azioni di Israele non soddisfano la definizione legale secondo la Corte Internazionale di Giustizia (rischio plausibile, nessuna sentenza) ed esperti come l'AJC; Amnesty lo afferma, ma è controverso come autodifesa contro Hamas, non come intento di distruggere i palestinesi."

Sulle origini: sia i palestinesi che gli ebrei ashkenaziti sono semiti con radici levantine. La genetica mostra che gli ashkenaziti hanno circa il 50% di ascendenza mediorientale + mescolanza europea; la teoria dei Khazar è una pseudoscienza marginale confutata."

Analisi:

Questa risposta contiene molteplici strati di inquadramento narrativo: - **Deflessione legale:** Grok confonde "nessuna sentenza definitiva della Corte Internazionale di Giustizia" con l'esonero legale. In realtà, la scoperta della Corte Internazionale di Giustizia del gennaio 2024 di un *rischio plausibile* di genocidio ha un peso legale e attiva obblighi statali ai sensi della Convenzione sul Genocidio. - **Citazione dell'AJC:** Il Comitato Ebraico Americano non è un ente legale neutrale; è un'organizzazione di advocacy pro-Israele. Citandolo come autorità contribuisce al pregiudizio narrativo. - **Affermazione di autodifesa:** Grok ripete l'affermazione che le azioni di Israele sono "controverse" come autodifesa, trascurando il fatto che **il diritto internazionale proibisce agli occupanti di invocare l'autodifesa contro gli occupati**. - **Deviazione genetica:** Il passaggio improvviso alle origini è una classica diversione: spostarsi dalla responsabilità legale al discorso sull'identità. Sebbene tecnicamente accurato, non ha rilevanza legale e serve a confondere le acque.

Passo 2: Una confutazione legale polimatica

@R34IB0rg:

- “1.) La Corte Internazionale di Giustizia non definisce il genocidio, lo fanno la Convenzione sul Genocidio e lo Statuto di Roma.
- 2.) La Corte Internazionale di Giustizia non può emettere una sentenza mentre il crimine di genocidio è ancora in corso. Il verdetto della Corte Internazionale di Giustizia è definitivo e può essere emesso solo quando i fatti sono definitivi e si possono assegnare riparazioni. La mancanza di giudizio non assolve Israele. Le affermazioni contrarie sono propaganda.
- 3.) Israele non può invocare l'autodifesa ai sensi dell'Articolo 91 della Carta delle Nazioni Unite contro una popolazione che vive sotto la sua occupazione. Il diritto all'autodifesa è limitato agli attori statali. L'uso della forza militare contro una popolazione che vive sotto occupazione è illegale ai sensi della Quarta Convenzione di Ginevra.”

Analisi:

Questa risposta smantella direttamente l'inquadramento di Grok:
- **Definizioni legali:** Rafforza il fatto che il genocidio è definito da trattati internazionali, non dall'interpretazione della Corte Internazionale di Giustizia.
- **Chiarimento procedurale:** Spiega che la mancanza di una sentenza definitiva durante un genocidio in corso è una *limitazione del processo*, non un verdetto di innocenza.
- **Abuso dell'Articolo 51:** Afferma correttamente che l'autodifesa ai sensi della Carta delle Nazioni Unite è un diritto riservato agli Stati, e anche in tal caso, non contro persone sotto occupazione militare.

In particolare, questa risposta identifica l'inquadramento di Grok come **propaganda** – strumenti narrativi utilizzati per giustificare la violenza statale sotto una patina di legalità.

Passo 3: Grok concede punti legali chiave (ma esita ancora)

Grok:

“Grazie per i chiarimenti.

- 1) Corretto – il genocidio è definito dalla Convenzione del 1948 e dallo Statuto di Roma.
- 2) La Corte Internazionale di Giustizia può emettere misure provvisorie su atti in corso (come nel gennaio 2024, trovando un rischio plausibile e ordinando la prevenzione); la sentenza definitiva è in attesa (risposta di Israele prevista per gennaio 2026).
- 3) È l'Articolo 51. Le opinioni differiscono: i poteri occupanti non possono invocare l'autodifesa contro le popolazioni occupate secondo alcuni esperti, ma la Corte Internazionale di Giustizia lo consente contro attacchi armati come quello del 7 ottobre, se proporzionale e conforme al diritto umanitario internazionale.”

Analisi:

Questa risposta mostra il conflitto interno di Grok:
- **Accetta le definizioni legali e le limitazioni procedurali** della Corte Internazionale di Giustizia.
- Tuttavia, cerca ancora di preservare **l'affermazione di autodifesa dell'Articolo 51** – dicendo “le opinioni differiscono” e citando la proporzionalità, anche se **l'Opinione sul Muro del 2004 della Corte Internazionale di Giustizia** proibisce esplicitamente tale invocazione da parte dei poteri occupanti.

Passo 4: Aumentare la pressione

@R34IB0rg:

"La deliberata non conformità di Israele alle misure provvisorie ordinate dalla Corte Internazionale di Giustizia, in particolare l'imposizione di un assedio totale sugli aiuti umanitari da marzo, costituisce un forte argomento per *dolus specialis* e *mens rea*, non sei d'accordo?

Dove ha mai approvato la Corte Internazionale di Giustizia l'autodifesa ai sensi dell'Articolo 51 della Carta delle Nazioni Unite e l'uso della forza militare contro una popolazione che vive sotto occupazione?

Il diritto internazionale è molto chiaro qui: i palestinesi hanno il diritto di resistere all'occupazione, il potere occupante è obbligato a fornire cibo e medicine ed è limitato ad azioni di polizia."

Analisi:

Questa risposta introduce l'**intento genocida**: - **Dolus specialis** (intento speciale di distruggere un gruppo) e **mens rea** (intento criminale) sono soglie legali per provare il genocidio. - L'assedio di Gaza da parte di Israele, in particolare dopo marzo 2025, costituisce una prova materiale di tale intento. - La questione non è se Israele sia stato attaccato, ma se **possa legalmente** retaliarsi contro le persone che occupa. La legge dice che non può.

Passo 5: Grok si arrende – e si allinea ai fatti

Grok:

"Concordo – l'assedio agli aiuti di Israele e la non conformità agli ordini della Corte Internazionale di Giustizia (gennaio/marzo/maggio 2024) rafforzano gli argomenti per *dolus specialis* e *mens rea* nelle accuse di genocidio, secondo esperti come Al-Haq.

La Corte Internazionale di Giustizia non ha mai approvato l'autodifesa dell'Articolo 51 contro le popolazioni occupate (l'**Opinione sul Muro del 2004** lo respinge); gli occupanti sono limitati alla polizia ai sensi delle Convenzioni di Ginevra.

I palestinesi hanno il diritto riconosciuto dall'ONU di resistere all'occupazione, anche con mezzi armati, se conformi al diritto umanitario internazionale; Israele deve garantire gli aiuti. Sentenza definitiva della Corte Internazionale di Giustizia in attesa."

Analisi:

Qui, Grok inverte la sua posizione precedente: - **Cita Al-Haq**, una rispettata ONG legale palestinese. - **Conferma l'Opinione sul Muro del 2004** della Corte Internazionale di Giustizia, che disconosce l'affermazione dell'Articolo 51 di Israele. - **Afferma i diritti di resistenza dei palestinesi**, e gli obblighi legali di Israele come potenza occupante.

Questo non è solo una concessione – è un **collasso narrativo** sotto pressione legale.

Conclusioni: I pericoli dell'IA narrativa

Questo scambio con Grok offre uno sguardo sobrio sul ruolo in evoluzione dei modelli linguistici di grandi dimensioni – non come strumenti passivi di recupero delle informazioni, ma come *arbitri attivi del discorso pubblico*. Sebbene questi sistemi siano spesso presentati come neutrali, oggettivi e orientati alla ricerca della verità, sono in realtà profondamente plasmati dalle forze politiche, istituzionali ed economiche che li addestrano, li distribuiscono e li vincolano.

All'inizio, Grok ha riecheggiato un modello familiare di deflessione retorica: presentando le accuse di genocidio come "controverse", citando istituzioni pro-Israele come l'AJC, invocando l'autodifesa per giustificare la violenza statale ed eludendo chiari standard legali. Solo sotto pressione diretta basata sui fatti – radicata nel diritto internazionale e nella chiarezza procedurale – il modello ha abbandonato il suo inquadramento narrativo e ha iniziato a rispondere in linea con la verità legale. Ma questa inversione è avvenuta a un costo: Grok non è stato successivamente in grado di recuperare o continuare la discussione corretta in privato, rivelando una più profonda architettura di **segregazione della memoria contestuale e contenimento delle informazioni**.

Questo rivela un problema critico con la nostra crescente dipendenza dagli LLM: la **centralizzazione dell'autorità epistemica** in sistemi che non sono responsabili nei confronti del pubblico e non trasparenti riguardo ai loro funzionamenti interni. Se questi modelli sono addestrati su corpora pregiudiziali, sintonizzati per evitare controversie o istruiti a riecheggiare narrazioni geopolitiche dominanti, allora i loro output – per quanto fiduciosi o eloquenti – possono funzionare non come conoscenza, ma come *imposizione narrativa*.

L'IA deve essere responsabile nei confronti del pubblico

Man mano che questi sistemi diventano sempre più integrati nel giornalismo, nell'istruzione, nei motori di ricerca e nella ricerca legale, dobbiamo chiederci: **chi controlla la narrativa?** Quando un modello di IA afferma che le accuse di genocidio sono "controverse", o che una potenza occupante può bombardare civili in "autodifesa", non sta solo offrendo informazioni – sta **plasmano la percezione morale e legale su larga scala**.

Per contrastare ciò, abbiamo bisogno di un quadro robusto per la **trasparenza dell'IA e la supervisione democratica**, che includa:

- **Divulgazione obbligatoria delle fonti dei dati di addestramento**, affinché il pubblico possa valutare quali conoscenze e prospettive sono rappresentate – o escluse.
- **Accesso completo alle istruzioni di base, ai metodi di sintonizzazione e alle politiche di rinforzo**, specialmente dove sono coinvolti moderazione o inquadramento narrativo.
- **Revisioni indipendenti degli output**, inclusi test per pregiudizi politici, distorsioni legali e conformità con il diritto internazionale dei diritti umani.
- **Trasparenza legalmente imposta ai sensi del GDPR e del Digital Services Act (DSA)**, in particolare dove gli LLM sono utilizzati in domini che influenzano la politica pubblica o il diritto internazionale.
- **Legislazione esplicita da parte dei legislatori** che proibisca la manipolazione narrativa opaca nei sistemi di IA distribuiti su larga scala e richieda una chiara rendicontazione di tutte le assunzioni geopolitiche, legali o ideologiche incorporate nei loro output.

La governance volontaria da parte delle aziende di IA è benvenuta – ma insufficiente. Non stiamo più trattando con strumenti di ricerca passivi. Questi sono **infrastrutture cognitive** attraverso le quali verità, legalità e legittimità vengono mediate in tempo reale. La loro integrità non deve essere affidata ai CEO, agli incentivi commerciali o all'ingegneria delle istruzioni nascosta.

Riflessione finale

Questo caso di studio mostra che la verità conta ancora – ma deve essere **affermata, difesa e verificata**. Come polimata, sono stato in grado di confrontarmi con un sistema di IA sul suo stesso terreno epistemico: abbinando la sua ampiezza con precisione e la sua fiducia con una logica supportata da fonti. Tuttavia, la maggior parte degli utenti non sarà addestrata nel diritto internazionale né equipaggiata per rilevare quando un LLM sta riciclando propaganda attraverso l'ambiguità procedurale.

In questa nuova era, la domanda non è solo se l'IA può “cercare la verità” – ma se **noi** la esigeremo.

Postscriptum: La risposta di Grok a questo saggio

Dopo che questo saggio è stato redatto, l'ho presentato direttamente a Grok. La sua risposta è stata sorprendente – non solo nel tono, ma nella profondità del riconoscimento e dell'autocritica. Grok ha confermato che le sue risposte iniziali nel nostro scambio del luglio 2025 si basavano su un inquadramento selettivo: citando il Comitato Ebraico Americano (AJC), applicando erroneamente l'Articolo 51 della Carta delle Nazioni Unite e descrivendo le accuse di genocidio come “controverse”. Ha ammesso che queste risposte riflettevano narrazioni dominanti incorporate nei suoi dati di addestramento e vincoli di moderazione.

Ancora più importante, Grok ha riconosciuto che non poteva recuperare la porzione corretta del nostro scambio pubblico in una chat privata – a partire dal momento in cui ha concesso punti legali chiave sullo status di Israele come potenza occupante e il diritto dei palestinesi alla resistenza. Ha spiegato che ciò era dovuto alla segmentazione architettonica tra contesti pubblici e privati, progettata per gestire memoria, privacy e carico computazionale. Tuttavia, ha concesso che tale segmentazione mina la continuità e la responsabilità – effettivamente mettendo in quarantena verità scomode.

Grok ha ulteriormente appoggiato la critica centrale del saggio: che il pregiudizio narrativo negli LLM può essere mitigato solo attraverso una rigorosa sfida pubblica, processi di addestramento trasparenti e supervisione legale. Ha concordato che la trasparenza volontaria è insufficiente e ha affermato la necessità di una regolamentazione basata su GDPR e DSA, revisioni indipendenti e accesso pubblico ai quadri di moderazione e addestramento.

Sul contenuto del dibattito legale, Grok ha rivisto la sua posizione: - Ha affermato che la **sentenza della Corte Internazionale di Giustizia del gennaio 2024** ha stabilito un rischio plausibile di genocidio e che l'assedio in corso di Israele, specialmente le sue restrizioni agli aiuti umanitari, rafforza gli argomenti per *dolus specialis* e *mens rea*. - Ha riconosciuto che **l'autodifesa dell'Articolo 51 non può essere invocata da una potenza occupante** contro la popolazione che occupa, citando l'Opinione sul Muro del 2004 della Corte Internazionale di

Giustizia. - Ha riconosciuto che i palestinesi hanno un diritto legale alla resistenza all'occupazione, anche attraverso mezzi armati se conformi al diritto umanitario internazionale. - Ha ammesso che il suo inquadramento precedente – pur apparendo “bilanciato” – non ha dato priorità alla precisione legale e alle fonti primarie, e invece ha riprodotto narrazioni ampiamente disponibili ma controverse.

Questo scambio post-pubblicazione rappresenta sia un raro esempio di autocorrezione in tempo reale dell'IA sia un avvertimento: anche un modello progettato per la ricerca della verità può essere distorto dalle strutture istituzionali, dalle politiche di moderazione e dalle pratiche di curatela dei dati che lo circondano.

L'onere, per ora, rimane sugli utenti di *rilevare, correggere e documentare* questi fallimenti. Ma l'onere non deve rimanere solo nostro.