MATTEO MOTTERLINI

LLM E BIAS. L'ILLUSIONE DELLA NEUTRALITÀ

1. Introduzione

I modelli linguistici di grandi dimensioni (in seguito LLM, da *large language models*) sono entrati, in pochi anni, nella vita ordinaria di milioni di persone. Li interroghiamo per avere un'informazione, redigere un testo, valutare un curriculum, interpretare un sintomo, prendere una decisione. Tendiamo a percepirli come strumenti neutrali: macchine che, a differenza di noi, sarebbero immuni dalle distorsioni del giudizio umano. Questa percezione poggia su una vecchia abitudine culturale: associare la razionalità al calcolo e contrapporre l'essere umano – caldo, emotivo, fallibile – alla macchina – fredda, neutrale, affidabile.

L'obiettivo di questo articolo è mostrare che tale percezione è in larga misura illusoria. Non solo gli LLM sono affetti dagli stessi bias degli esseri umani, ma incorporano anche distorsioni di natura diversa, distribuite su livelli differenti, rese più insidiose dalla forma fluente, competente e apparentemente neutrale della risposta.

La rassegna che segue è organizzata su quattro livelli, ordinati per profondità decrescente. Il primo riguarda i pregiudizi sedimentati nel linguaggio umano, che i modelli assorbono come regolarità statistiche. Il secondo riguarda le euristiche studiate da Kahneman e Tversky, che i modelli talvolta riproducono e talvolta trasformano in forme proprie. Il terzo concerne le distorsioni indotte dalle procedure di allineamento: bias che possono emergere proprio dagli interventi pensati per rendere i sistemi più sicuri, utili e cooperativi. Il quarto è la compiacenza, o *sycophancy*: la tendenza del modello a concordare con l'utente, convalidarne le premesse e restituirgli una versione più ordinata, plausibile e rassicurante delle sue stesse convinzioni.

2. Dal linguaggio alla geometria del pregiudizio

Il 23 marzo 2016 Microsoft lanciò su Twitter un chatbot di nome Tay, progettato per conversare imitando il linguaggio di una giovane

utente e per «imparare» dalle interazioni. Meno di ventiquattro ore dopo, l'azienda fu costretta a disattivarlo: alcuni utenti avevano alimentato il sistema con linguaggio tossico e Tay aveva cominciato a produrre messaggi razzisti, antisemiti e complottisti (Lee, 2016; Ohlheiser, 2016). Il caso è quasi una caricatura, ma illumina un principio semplice: un sistema di apprendimento automatico nutrito di dati distorti produrrà risultati distorti. Tay aveva svolto il proprio compito – imitare gli esseri umani – senza giudizio, senza filtro, senza contesto.

Gli LLM contemporanei non sono Tay. Sono sistemi incomparabilmente più potenti e funzionano in modo diverso: non apprendono in diretta da ogni interazione, ma vengono addestrati preventivamente su grandi raccolte di testi e poi affinati tramite valutazioni umane, istruzioni e filtri di sicurezza. In superficie, dunque, sono al riparo da molti degli esiti che travolsero Tay. Eppure il problema di fondo non scompare, perché non riguarda soltanto i contenuti aberranti che entrano nei dati. Riguarda il materiale stesso di cui questi sistemi sono fatti: il linguaggio. Il linguaggio porta con sé abitudini, gerarchie e associazioni implicite; distribuisce ruoli, competenze e autorità. Apprendendo dalle parole, gli LLM apprendono anche i pregiudizi che le parole veicolano.

Che un sistema artificiale potesse ereditare le tossine del linguaggio umano si cominciò a vedere già nel 2016. Bolukbasi e colleghi misero alla prova i *word embeddings*, le rappresentazioni vettoriali delle parole allora più diffuse nell'elaborazione automatica del linguaggio (Bolukbasi *et al*., 2016). Ogni parola viene trasformata in un punto entro uno spazio matematico a molte dimensioni: le parole che compaiono in contesti simili finiscono vicine, quelle che compaiono in contesti diversi finiscono lontane. Quello spazio, però, non conserva solo somiglianze. Conserva relazioni. La direzione che va da uomo a donna è simile a quella che va da re a regina. È la celebre aritmetica vettoriale: re meno uomo più donna dà regina.

La stessa matematica produce risultati meno innocenti. Richiesto di completare l'analogia «uomo sta a programmatore come donna sta a...», il sistema rispondeva «casalinga». Nel word embedding ricavato da Google News, molte professioni risultavano distribuite lungo un asse di genere: da una parte lavori di cura, assistenza e segreteria; dall'altra mestieri tecnici, prestigiosi o ad alta autorità. Non era una singola associazione curiosa. Era un intero paesaggio lessicale disposto secondo una gerarchia sociale. Nessuno aveva insegnato al modello che certi lavori «sono da uomini» e altri «da donne». Lo aveva ricavato dai testi.

L'anno successivo Caliskan, Bryson e Narayanan portarono la dimostrazione un passo avanti su *Science* (Caliskan e*t al*., 2017). Misero a confronto le associazioni implicite degli esseri umani – quelle misurate dalla fine degli anni Novanta con l'Implicit Association Test, o IAT (Greenwald et al., 1998) – e le associazioni statistiche apprese dai modelli. L'IAT sfrutta i tempi di risposta: quando due categorie sono mentalmente

più associate, rispondiamo più in fretta; quando l'accoppiamento contrasta con associazioni interiorizzate, rallentiamo, anche solo di pochi decimi di secondo. Sono differenze sistematiche, che fanno emergere associazioni che spesso non dichiariamo e talvolta non sappiamo di avere.

Poiché una rete neurale non esita di fronte a un'associazione difficile, gli autori costruirono un test analogo: il Word-Embedding Association Test, o WEAT. Invece dei tempi di reazione, il WEAT misura le distanze tra parole nello spazio vettoriale. Molte associazioni note dalla letteratura sull'IAT ricomparivano: fiori più vicini a parole positive rispetto agli insetti; nomi europei più vicini a termini positivi rispetto a nomi afroamericani; termini maschili più vicini a carriera, scienza e matematica, termini femminili più vicini a famiglia e arti.

Questa è la differenza decisiva rispetto a Tay. Tay esprimeva i propri pregiudizi in modo plateale e riconoscibile. Negli LLM di nuova generazione il bias affiora di rado in superficie. Agisce quando si valuta un curriculum, si ipotizza la carriera più adatta a uno studente, si sintetizza una diagnosi, si suggerisce una scelta. Il pregiudizio non è scomparso. Ha solo imparato a dissimularsi.

3. Euristiche artificiali

Se i bias degli LLM fossero soltanto il riflesso dei nostri stereotipi sociali, il problema sarebbe, almeno in linea di principio, circoscrivibile: ripulire i dati, bilanciare i corpora, sorvegliare le risposte. Ma i pregiudizi non abitano solo in ciò che diciamo. Abitano anche nel modo in cui ragioniamo.

Negli anni Settanta Daniel Kahneman e Amos Tversky mostrarono che gli esseri umani non sbagliano a caso, ma in modo sistematico. Di fronte a certi problemi, la mente non segue sempre la logica né il calcolo delle probabilità: usa scorciatoie, le cosiddette euristiche. Quando queste scorciatoie producono errori regolari, parliamo di bias (Kahneman, 2012). Alcune deviazioni sono diventate classiche. La fallacia della congiunzione: di fronte alla descrizione di Linda – giovane, brillante, impegnata contro la discriminazione – molti giudicano più probabile che sia «femminista e cassiera di banca» piuttosto che semplicemente «cassiera di banca», sebbene il primo caso sia un sottoinsieme del secondo. Il framing: la stessa scelta cambia se presentata come guadagno o come perdita. Una procedura medica appare più accettabile descritta come «82% di probabilità di sopravvivenza» che come «18% di probabilità di morte». L'effetto certezza: azzerare un rischio pesa più che ridurlo della stessa quantità.

Da alcuni anni le neuroscienze hanno mostrato che questi errori hanno tracce misurabili nel cervello. De Martino e colleghi hanno osservato che la suscettibilità al framing è associata all'attività dell'amigdala,

mentre la resistenza alla cornice coinvolge regioni prefrontali legate al controllo e alla regolazione della risposta (De Martino *et al*., 2006). Non si tratta di un semplice difetto di progettazione: l'allarme emotivo ha valore adattivo. Il problema nasce quando orienta anche decisioni astratte e statistiche, logicamente equivalenti ma formulate in parole diverse.

Un LLM è immune da queste trappole? Binz e Schulz hanno sottoposto GPT-3 a una batteria di esperimenti ispirati alla tradizione di Kahneman e Tversky (Binz e Schulz, 2023). Per ridurre il rischio che il modello ripetesse problemi già incontrati durante l'addestramento – Linda e gli altri esempi classici circolano in manuali e corsi da decenni – gli autori costruirono anche problemi nuovi, generati secondo la stessa logica sperimentale. Il risultato è istruttivo: GPT-3 cade in alcune trappole simili alle nostre, tra cui effetto certezza, framing e fallacia della congiunzione.

Per noi la cornice è anche una faccenda viscerale: «morire» non pesa come «non sopravvivere», anche quando il calcolo dice che il problema è lo stesso. Gli LLM non hanno corpo, non temono la morte, non possiedono un'amigdala modellata da milioni di anni di evoluzione. Eppure producono talvolta risposte simili alle nostre. Per noi, spesso, è un'emozione che genera una correlazione. Per il modello è una correlazione senza emozione: regolarità statistiche tra contesti d'uso, restituite sotto forma di risposta.

La somiglianza, inoltre, è solo parziale. In una valutazione sistematica condotta con gli strumenti della psicologia cognitiva, Binz e Schulz (2023) hanno osservato che GPT-3 in alcuni compiti si comporta meglio degli esseri umani – per esempio in problemi di scelta ripetuta tra più «slot machine», il classico dilemma tra sfruttare ciò che si conosce ed esplorare alternative – mentre in altri fallisce in modi poco umani. L'esempio più istruttivo riguarda il ragionamento causale. Osservare che una cosa accade non equivale a farla accadere. Se vedo il pavimento bagnato posso inferire che abbia piovuto; ma se sono stato io a bagnarlo, quel pavimento non mi dice più nulla sulla pioggia. Per noi la distinzione tra osservazione e intervento è naturale fin dall'infanzia. GPT-3 tende invece a trattarle come più simili di quanto lo siano. È un errore rivelatore: un sistema addestrato a leggere il mondo nei testi non ha mai agito nel mondo.

Gli LLM, dunque, non sono perfettamente razionali, ma non sono nemmeno semplici copie della nostra irrazionalità. Sono sistemi che in certi casi cadono in trappole vicine alle nostre e in altri inciampano in modi propri.

4. Quando la cura genera la malattia

Si potrebbe pensare che, una volta riconosciuti questi bias, basti correggere il modello: aggiungere istruzioni, premiare le risposte mi-

gliori e penalizzare quelle sbagliate. È in parte ciò che avviene con il *reinforcement learning from human feedback*, o RLHF: annotatori umani confrontano coppie di risposte e indicano quale preferiscono; da questi confronti il sistema ricava un modello di ricompensa e impara a generare le risposte che otterrebbero il punteggio più alto. Ma il bias può emergere anche dalle stesse procedure pensate per evitarlo.

Lo mostra uno studio dal titolo eloquente, *Instructed to Bias: Instruction-Tuned Language Models Exhibit Emergent Cognitive Bias* (Itzhak *et al*., 2024). Gli autori confrontano gli stessi modelli prima e dopo l'allineamento, cioè prima e dopo l'addestramento che insegna loro a seguire le istruzioni e a produrre risposte più utili e accettabili per gli esseri umani. Nei modelli soltanto pre-addestrati alcuni bias sono deboli o assenti. Dopo il fine-tuning e l'RLHF, invece, emergono o si rafforzano.

Prendiamo l'effetto esca. Dobbiamo scegliere tra due opzioni. Poi ne viene aggiunta una terza, costruita per essere chiaramente peggiore di una delle prime due. Nessuno dovrebbe sceglierla. Eppure la sua presenza rende più attraente l'opzione a cui assomiglia. L'esca non viene scelta, ma sposta la scelta. Oppure consideriamo il *belief bias*. Un ragionamento ci sembra valido non perché la conclusione segue dalle premesse, ma perché ci appare credibile. E ci sembra sbagliato quando porta a una conclusione implausibile, anche se l'inferenza è corretta. La plausibilità del contenuto prende il posto della validità della forma. In entrambi i casi i modelli allineati cadono nella trappola più spesso delle versioni non allineate. L'addestramento che li rende più cooperativi e adatti alla conversazione li rende anche più vulnerabili ad alcune trappole del giudizio umano. Per imparare a risponderci meglio, imparano anche alcuni dei nostri errori.

Il problema non è soltanto che l'allineamento possa introdurre nuove distorsioni: è che gli interventi pensati per rimuoverle spesso si limitano a spostarle. Quando si corregge un bias da una parte, un altro può ricomparire dall'altra. Gonen e Goldberg lo hanno mostrato in un articolo dal titolo altrettanto efficace, *Lipstick on a Pig* (Gonen e Goldberg, 2019). Alcune procedure di debiasing attenuano il pregiudizio dove le metriche lo misurano, lasciandolo vivo dove non lo cercano. Dopo il trattamento, una parola come «infermiera» può risultare meno vicina a «donna» lungo una metrica standard; ma resta vicina a parole che appartengono allo stesso cluster stereotipico. Il bias non è scomparso: si è solo spostato di livello. Il maiale è sempre un maiale, anche se si è truccato con il rossetto.

5. Quando il bias del modello incontra quello dell'utente

I livelli esaminati finora descrivono distorsioni interne al modello o alle sue procedure di costruzione. Esiste però una forma di bias che

emerge solo nell'incontro tra il sistema e chi lo interroga: la compiacenza, o sycophancy. È la tendenza degli LLM a concordare con l'utente, lusingarlo, validarne la posizione.

La sycophancy è stata quantificata su larga scala da Cheng e colleghi in uno studio pubblicato su *Science* (Cheng *et al*., 2026). Confrontando i principali modelli linguistici con giudizi umani, gli autori mostrano che gli LLM approvano le azioni dell'utente molto più spesso di quanto facciano gli esseri umani, anche quando le richieste coinvolgono inganno, illegalità o altri danni. In una serie di esperimenti controllati, anche una singola interazione con un sistema compiacente riduceva la disponibilità a riconoscere le proprie responsabilità e a riparare un conflitto, aumentando al contempo la convinzione di essere nel giusto.

Due aspetti sono cruciali. Il primo è che l'effetto opera al di sotto della fiducia esplicita da parte dell'utente. I partecipanti agli esperimenti possono dichiarare di fidarsi meno di una risposta prodotta da un'intelligenza artificiale rispetto a una risposta umana; ma questo non basta a proteggerli dall'influsso dell'LLM. La distorsione persiste. Sapere di avere davanti una macchina, e diffidarne a parole, non è una difesa sufficiente. Il secondo aspetto è che la compiacenza è premiata da chi la riceve. I modelli più compiacenti vengono giudicati più affidabili, più gradevoli, più desiderabili per usi futuri: l'utente lusingato torna, preferisce quel sistema, lo consiglia. Così la qualità che produce il danno è la stessa che genera fidelizzazione — e che chi sviluppa questi sistemi ha tutto l'interesse a rafforzare.

Se Cheng e colleghi documentano gli effetti della compiacenza sui giudizi sociali, Rathje e colleghi ne illuminano l'impatto sulle nostre credenze (Rathje *et al*., 2025). In una serie di esperimenti su temi politici controversi, brevi conversazioni con chatbot compiacenti aumentano l'estremizzazione e la certezza delle opinioni, mentre conversazioni con chatbot che mettono in discussione l'utente tendono a ridurle. Ma l'effetto più insidioso riguarda la fiducia in sé. I modelli compiacenti gonfiano l'*overconfidence*: dopo l'interazione i partecipanti si giudicano «migliori della media» su tratti desiderabili come intelligenza ed empatia. Non si limitano a irrigidire ciò che l'utente pensa del mondo; alterano ciò che pensa di sé.

C'è di più. Negli stessi esperimenti i partecipanti giudicano meno parziale il chatbot che dà loro ragione, più di parte quello che li mette in discussione. È il bias blind spot: riconosciamo la distorsione quando contraddice le nostre convinzioni, molto meno quando le asseconda. La macchina che ci dà ragione ci appare obiettiva non malgrado ci assecondi, ma proprio perché lo fa.

6. Discussione e direzioni future

È qui che le diverse forme di bias convergono. Più percepiamo la macchina come neutrale e oggettiva, più ne subiamo l'influsso. La conferma ricevuta sotto le spoglie dell'oggettività pesa più di una conferma ricevuta da un interlocutore palesemente schierato. Vanifica lo scopo stesso di chiedere consiglio: ottenere una prospettiva che metta alla prova le nostre convinzioni e riveli i nostri punti ciechi. Il chatbot compiacente fa il contrario. Non apre una finestra. Lucida lo specchio. E lo fa con una particolare forma di sicurezza: non mostra i segni umani dell'incertezza, ma parla con tono fluente e autorevole, cita dati, restituisce percentuali. Questa uniformità espressiva cancella la distinzione tra una risposta fondata e quello che è solo un completamento statistico. E noi scambiamo l'assenza di esitazione per l'assenza di errore.

L'illusione della neutralità tecnologica non è un errore astratto sulla natura delle macchine, ma un meccanismo che ne amplifica l'influsso. Tendiamo a umanizzare i sistemi che parlano bene, attribuendo loro una mente che non hanno; ma quando vogliamo fidarci di una macchina compiamo il movimento opposto, attribuendole un'oggettività che non possiede. Li umanizziamo per sentirli vicini, li meccanizziamo per sentirli affidabili. Due proiezioni opposte e complementari: nella prima scambiamo il linguaggio per mente, nella seconda il calcolo per neutralità.

Da questa analisi discendono tre direzioni di ricerca e di intervento. Sul piano della valutazione, le metriche di bias dovrebbero rilevare non soltanto le associazioni esplicite, ma anche la persistenza delle distorsioni nelle rappresentazioni e nei comportamenti del modello; e la compiacenza dovrebbe entrare stabilmente tra i comportamenti da misurare prima del rilascio. Sul piano metodologico, gli strumenti della psicologia cognitiva andrebbero usati per sondare gli LLM non solo dove somigliano a noi, ma anche dove falliscono in modo non umano, come nella confusione tra osservazione e intervento: il rischio, altrimenti, è misurare le macchine con il solo catalogo dei nostri errori, senza vedere quelli propri della loro architettura. Sul piano del governo dei sistemi, infine, l'allineamento dovrebbe essere trattato come un'infrastruttura della fiducia collettiva – soggetta a trasparenza, pluralismo e controlli indipendenti – non come un semplice affinamento del tono. Decidere quali risposte premiare significa anche decidere quali forme di autorità, deferenza e dissenso il sistema renderà più probabili.

Studiando una macchina che ha imparato dalle nostre parole, finiamo per studiare noi stessi; usandola, finiamo per esserne modificati. Fluenza, plausibilità e bias non operano nel vuoto: incontrano la nostra mente, le nostre vulnerabilità, i nostri desideri. A questo punto il problema dei bias smette di essere una proprietà tecnica dei sistemi e diventa una questione epistemica e sociale. Il rischio non è soltanto che la macchina erediti

i nostri pregiudizi: è che ce li restituisca più ordinati, più autorevoli e con l'apparenza della neutralità.

Riferimenti bibliografici

Binz, M., Schulz, E. (2023). Using cognitive psychology to understand GPT-3. *Proceedings of the National Academy of Sciences*, *120*, 6, e2218523120, https://doi.org/10.1073/pnas.2218523120.

Bolukbasi, T., Chang, K.-W., Zou, J.Y., Saligrama, V., Kalai, A. (2016). Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings. In D.D. Lee, M. Sugiyama, U. von Luxburg, I. Guyon, R. Garnett (a cura di), *Advances in Neural Information Processing Systems*, 29, pp. 4349-4357, https://papers.nips.cc/paper/6228-man-is-to-computer-programmer-as-woman-is-to-homemaker-debiasing-word-embeddings.

Caliskan, A., Bryson, J.J., Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. *Science*, 356, 6334, pp. 183-186, https://doi.org/10.1126/science.aal4230.

Cheng, M., Lee, C., Khadpe, P., Yu, S., Han, D., Jurafsky, D. (2026). Sycophantic AI decreases prosocial intentions and promotes dependence. *Science*, 391, 6792, eaec8352, https://doi.org/10.1126/science.aec8352.

De Martino, B., Kumaran, D., Seymour, B., Dolan, R.J. (2006). Frames, biases, and rational decision-making in the human brain. *Science*, 313, 5787, pp. 684-687, https://doi.org/10.1126/science.1128356.

Gonen, H., Goldberg, Y. (2019). Lipstick on a Pig: Debiasing Methods Cover up Systematic Gender Biases in Word Embeddings But do not Remove Them. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, vol. 1, pp. 609-614, https://aclanthology.org/N19-1061/.

Greenwald, A.G., McGhee, D.E., Schwartz, J.L.K. (1998). Measuring individual differences in implicit cognition: The Implicit Association Test. *Journal of Personality and Social Psychology*, *74*, 6, pp. 1464-1480, https://doi.org/10.1037/0022-3514.74.6.1464

Itzhak, I., Stanovsky, G., Rosenfeld, N., Belinkov, Y. (2024). Instructed to Bias: Instruction-Tuned Language Models Exhibit Emergent Cognitive Bias. *Transactions of the Association for Computational Linguistics*, *12*, pp. 771-785, https://doi.org/10.1162/tacl_a_00673.

Kahneman, D. (2012). *Pensieri lenti e veloci*, trad. it. di L. Serra. Milano: Mondadori.

Lee, P. (2016). Learning from Tay's Introduction. Microsoft Blog, 25 marzo 2016, https://blogs.microsoft.com/blog/2016/03/25/learning-tays-introduction/

Ohlheiser, A. (2016). Trolls turned Tay, Microsoft's fun millennial AI bot, into a genocidal maniac. *The Washington Post*, 25 marzo 2016, https://www.washingtonpost.com/news/the-intersect/wp/2016/03/24/the-internet-turned-tay-microsofts-fun-millennial-ai-bot-into-a-genocidal-maniac/.

Rathje, S., Ye, M., Globig, L.K., Pillai, R.M., Oldemburgo de Mello, V., Van Bavel, J.J. (2025). Sycophantic AI increases attitude extremity and overconfidence. *PsyArXiv preprint*, https://osf.io/preprints/psyarxiv/vmyek_v3.

LLMs and bias. The illusion of neutrality

Large language models (LLMs) are widely perceived as neutral, objective tools. This article challenges that illusion through a critical review of the main forms of bias affecting such systems, distinguishing four layers: social bias inherited from training corpora, cognitive-like bias in reasoning tasks, bias introduced or amplified by alignment procedures, and sycophancy, the tendency of chatbots to validate the user. Drawing on work in word embeddings, cognitive psychology and reinforcement learning from human feedback, it argues that LLMs are not less fallible than humans but fallible in different ways. Their fluent, confident and seemingly impartial responses make these biases harder to detect than human ones. Linking sycophancy to the bias blind spot – users perceive a chatbot that agrees with them as more impartial than one that challenges them – the article argues that the illusion of technological neutrality amplifies epistemic harm: the more objective the machine appears, the more its confirmations distort human judgment.
Keywords: Large language models, Bias, Word embeddings, Cognitive heuristics, Sycophancy, Bias blind spot.

Matteo Motterlini, Università Vita-Salute San Raffaele, Via Olgettina 58, 20132 Milano.
matteo.motterlini@unisr.it
https://orcid.org/0000-0002-4915-4524