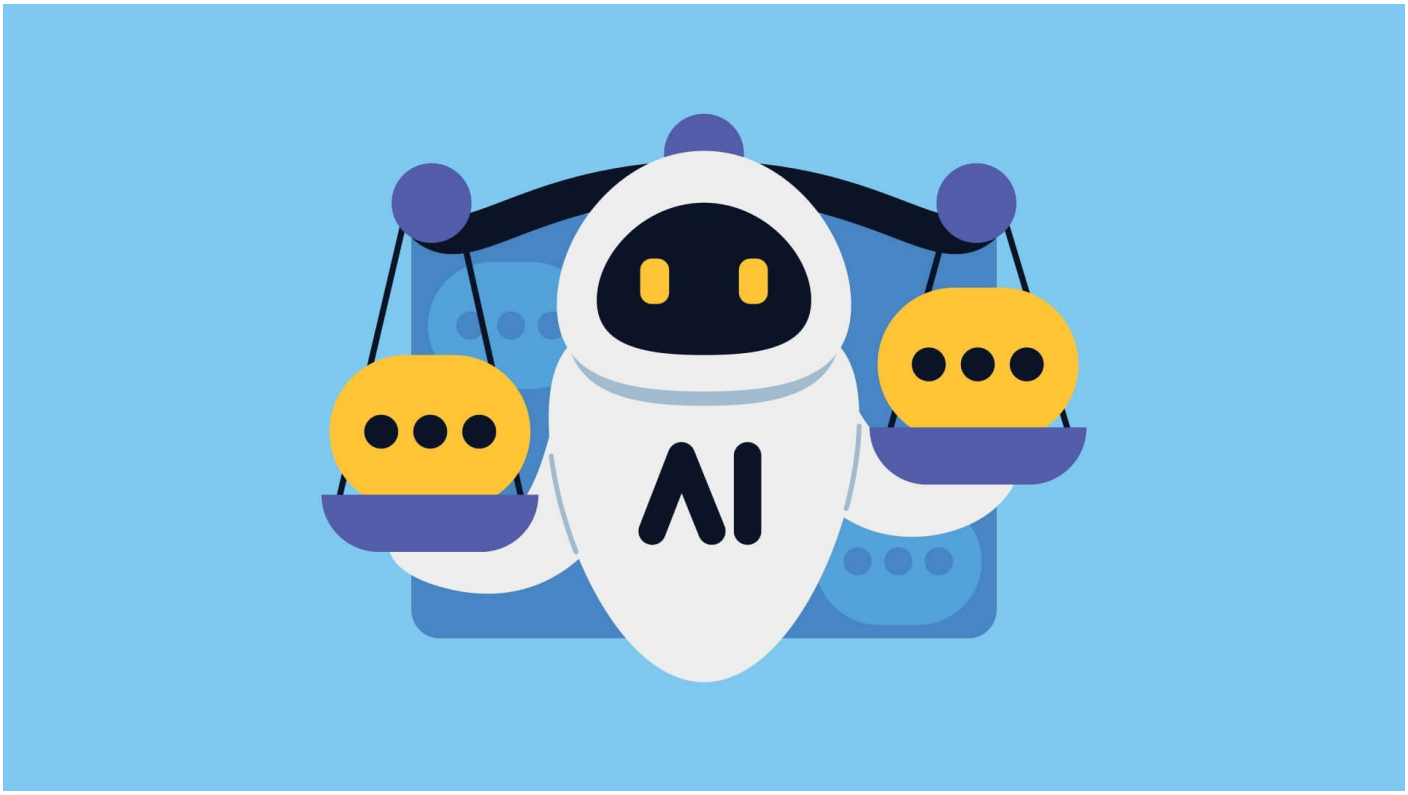
I pregiudizi nascosti dell'IA: come scoprirli e correggerli

⋮ 15/04/2025



L'intelligenza artificiale rappresenta oggi uno degli strumenti tecnologici più sofisticati e promettenti, capace di generare risposte e soluzioni complesse grazie all'implementazione di architetture avanzate come le reti neurali e i transformer. Questi ultimi, in particolare, hanno rivoluzionato l'analisi testuale, permettendo alle macchine di interpretare e produrre linguaggio con un livello di coerenza e appropriatezza precedentemente inimmaginabile.

Tuttavia, come ogni tecnologia emergente, l'IA non è esente da limitazioni e problematiche, tra cui spicca la presenza di pregiudizi – o bias – incorporati nei sistemi.

i bias nell'intelligenza artificiale: definizione e tipologie

I bias nell'intelligenza artificiale costituiscono un paradosso interessante: storicamente, l'automazione è stata concepita e implementata proprio per eliminare l'errore umano e la soggettività dai processi decisionali. Tuttavia, le ricerche contemporanee dimostrano come **i sistemi di IA non solo riflettano i pregiudizi presenti nei dati di addestramento, ma spesso li amplifichino**, creando un effetto moltiplicatore potenzialmente dannoso. Alcuni studiosi hanno dimostrato, attraverso un'analisi rigorosa, come l'idea di un'intelligenza artificiale completamente imparziale sia non solo illusoria, ma logicamente incoerente (Bridewell, Bello, Bringsjord 2024).

Un esempio particolarmente significativo di bias discriminatorio riguarda la questione di genere. I modelli basati su architetture Transformer (come GPT e BERT) hanno manifestato tendenze discriminatorie nei confronti di specifici generi (Nemani et al. 2023). Tali pregiudizi possono avere effetti particolarmente problematici in applicazioni di uso quotidiano che sfruttano il linguaggio naturale, come i chatbot, i motori di ricerca e i sistemi di traduzione automatica, influenzando potenzialmente milioni di utenti in tutto il mondo.

Comprendere la natura dei bias nell'intelligenza artificiale

Un aspetto cruciale per affrontare adeguatamente la questione dei bias nell'IA è la **definizione stessa del termine "bias", spesso utilizzato in modo impreciso e generico nel dibattito pubblico e, talvolta, persino in quello accademico**. Bridewell et al. (2024) propongono una distinzione fondamentale tra diversi tipi di bias:

1. **Bias tecnico**: intrinsecamente legato alla struttura dell'algoritmo e ai dati di addestramento utilizzati. Questo tipo di bias è spesso inevitabile e deriva dalle scelte metodologiche operate durante lo sviluppo del sistema.
2. **Bias interpretativo**: influenzato dalle aspettative degli utenti e dai contesti culturali in cui il sistema viene implementato. Un sistema potrebbe funzionare in modo tecnicamente corretto, ma essere percepito come ingiusto in determinati contesti culturali.
3. **Bias sociale e politico**: radicato nelle dinamiche di potere e nei pregiudizi esistenti nella società. Questo tipo di bias riflette e talvolta amplifica disuguaglianze strutturali preesistenti.

Questa classificazione tripartita risulta particolarmente utile per comprendere perché le reazioni emotive al bias dell'IA siano spesso complesse e contraddittorie. Infatti, a

seconda della prospettiva adottata, lo stesso comportamento di un sistema di IA potrebbe essere interpretato come un errore tecnico neutrale, un'incomprensione culturale o una manifestazione di ingiustizia sistemica.

I meccanismi di amplificazione dei bias nell'intelligenza artificiale

Un aspetto particolarmente preoccupante del bias nei sistemi di intelligenza artificiale è rappresentato non tanto dalla mera presenza di pregiudizi, quanto dalla loro **potenziale amplificazione**. Il bias nei sistemi di IA non è semplicemente un riflesso passivo dei dati di addestramento, ma può essere significativamente amplificato dal funzionamento stesso degli algoritmi (Lloyd, 2023; Nemani *et al.* 2023).

Le ragioni principali di questo fenomeno sono molteplici. In primo luogo, i dataset di addestramento sono spesso intrinsecamente squilibrati: se i dati riflettono disparità sociali preesistenti – come la sottorappresentazione di determinati gruppi demografici in posizioni di potere – l'IA apprenderà e replicherà queste stesse disparità (Lloyd, 2023).

In secondo luogo, gli algoritmi di ottimizzazione comunemente impiegati nella progettazione di sistemi di IA sono tipicamente programmati per **massimizzare la precisione o ridurre la perdita (*loss*),** spesso a scapito dell'equità. Un modello potrebbe migliorare la sua accuratezza generale perpetuando stereotipi, semplicemente perché questi stereotipi sono statisticamente prevalenti nei dati di addestramento (Lloyd, 2023).

Infine, si verifica quello che potremmo definire un "effetto a catena": le decisioni automatizzate influenzano la generazione di nuovi dati, creando un ciclo di feedback in cui i pregiudizi vengono progressivamente rafforzati. Per esempio, un algoritmo di selezione del personale che favorisce implicitamente un genere influenzerà la composizione futura della forza lavoro, generando nuovi dati che a loro volta rafforzeranno lo stesso pregiudizio.

L'impatto sociale dei bias nell'intelligenza artificiale

Un aspetto particolarmente critico nell'analisi dell'amplificazione del bias è il suo impatto sproporzionato sulle comunità più vulnerabili. I sistemi di IA vengono sempre più frequentemente implementati in ambiti cruciali per il benessere e le opportunità degli individui (Lloyd, 2023):

- Nei processi di assunzione e selezione del personale, algoritmi con bias di genere o etnici possono perpetuare discriminazioni storiche, limitando le opportunità professionali per determinate categorie di persone.
- Nel sistema giudiziario, software predittivi utilizzati per valutare il rischio di recidiva hanno mostrato pregiudizi razziali significativi, potenzialmente influenzando decisioni sulla libertà condizionale o sulla determinazione della pena.
- Nell'accesso ai servizi finanziari, modelli di scoring creditizio possono penalizzare sistematicamente gruppi svantaggiati, limitando il loro accesso al credito e, di conseguenza, le loro possibilità di mobilità sociale.

Questi effetti, lungi dall'essere temporanei o marginali, possano consolidare e aggravare disuguaglianze strutturali preesistenti, creando un circolo vizioso di svantaggio per le comunità già marginalizzate.

Il bias di genere nelle applicazioni di intelligenza artificiale

Particolare attenzione merita il bias di genere nei modelli *transformer* (Nemani *et al.*, 2023). Questo tipo di bias si manifesta in modo particolarmente evidente in diverse applicazioni:

- Nei sistemi di selezione automatizzata, come quelli utilizzati per lo screening dei curriculum vitae, si è osservata una tendenza a favorire inconsapevolmente candidati di un genere rispetto all'altro. Un caso emblematico è quello del sistema di selezione sviluppato da Amazon, successivamente abbandonato quando si scoprì che penalizzava sistematicamente le candidate donne.
- Nella traduzione automatica, i sistemi tendono ad assegnare automaticamente un genere ai pronomi neutri nelle lingue che non fanno distinzione di genere. Ad esempio, l'inglese "doctor" viene spesso tradotto al maschile, mentre "nurse" al femminile, perpetuando stereotipi professionali di genere.
- Negli assistenti virtuali e nei chatbot, si riscontra frequentemente un rafforzamento degli stereotipi di genere attraverso le risposte fornite agli utenti. Questo può includere la rappresentazione di ruoli tradizionali di genere o l'uso di linguaggio con connotazioni di genere inappropriate.

Il dibattito sulla neutralità dell'intelligenza artificiale

Le discussioni sul bias algoritmico sono spesso caratterizzate da reazioni emotive forti, che è possibile categorizzare in tre forme di “outrage” (indignazione) (Bridewell *et al.*, 2024):

1. **Indignazione intellettuale**: critiche fondate su errori logici e metodologici identificati nei modelli di IA, che spesso evidenziano incongruenze concettuali o lacune nel design degli algoritmi.
2. **Indignazione morale**: reazioni di sdegno per l’ingiustizia sociale riprodotta e potenzialmente amplificata dagli algoritmi, particolarmente intense quando il bias colpisce gruppi già storicamente marginalizzati.
3. **Indignazione politica**: utilizzo strumentale del dibattito sul bias dell’IA come veicolo per promuovere specifiche agende ideologiche o normative, trasformando una questione tecnica in un terreno di scontro politico.

Queste tre forme di reazione, spesso intrecciate e difficilmente distinguibili nel dibattito pubblico, influenzano profondamente il modo in cui il bias viene percepito, discusso e affrontato nella società.

Un contributo importante al dibattito sul bias viene da Lindloff e Siegert (2025), che suggeriscono come un modello possa essere “biased” senza essere necessariamente ingiusto. Gli autori sostengono che una definizione più precisa e articolata del bias sia essenziale per migliorare la qualità del dibattito sull’equità nell’IA. In questa prospettiva, il bias non dovrebbe essere automaticamente associato alla discriminazione, ma potrebbe rappresentare un elemento inevitabile e, in alcuni casi, persino desiderabile.

Centrale in questa analisi è **la distinzione tra bias e discriminazione** (Lindloff, Siegert, 2025). Il bias viene concettualizzato come una deviazione sistematica nei dati o negli algoritmi, un fenomeno in gran parte tecnico e statistico. La discriminazione, invece, implica un trattamento ingiusto di specifici gruppi, con connotazioni etiche e sociali più marcate. Gli autori suggeriscono che potrebbe essere controproducente tentare di eliminare ogni forma di bias senza prima definirne chiaramente la natura e il ruolo nel contesto specifico dell’applicazione.

Viene inoltre evidenziato come **l’eliminazione indiscriminata del bias possa compromettere la performance dei modelli**, specialmente in applicazioni sensibili come la medicina o la sicurezza (Lindloff, Siegert, 2025). Ad esempio, se un modello di diagnosi medica è “biased” nel riconoscere meglio certe malattie in base a fattori biologici reali, ciò non rappresenta necessariamente un’ingiustizia, ma può riflettere una

differenza naturale nei dati clinici che è importante preservare per garantire diagnosi accurate.

In questo senso si rende quindi necessario un cambio di prospettiva: un modello equo può incorporare bias, a condizione che questi siano compresi e gestiti correttamente. Il concetto di *fairness* (equità) non può essere ridotto semplicisticamente all'eliminazione totale del bias, ma deve piuttosto fondarsi su una comprensione approfondita del contesto e delle finalità specifiche dell'algoritmo.

In una prospettiva complementare, altri studiosi suggeriscono che l'IA, pur avendo inevitabilmente dei bias, offre almeno tre vantaggi fondamentali rispetto ai processi decisionali umani, anch'essi tipicamente soggetti a pregiudizi (Kuzmanov, 2025):

1. **Analisi di grandi volumi di dati**: a differenza degli esseri umani, l'IA può processare quantità di informazioni enormemente superiori senza fatica o perdita di concentrazione, identificando pattern che potrebbero sfuggire all'intuizione umana.
2. **Memoria e consistenza**: l'intelligenza artificiale non soffre di distorsioni della memoria o dell'influenza di emozioni momentanee, garantendo una maggiore coerenza nelle decisioni prese in momenti diversi o in contesti simili.
3. Apprendimento automatico: i modelli di machine learning possono essere specificamente addestrati per riconoscere e correggere distorsioni presenti nei dataset, potenzialmente riducendo gradualmente il peso dei bias sistematici.

L'IA può rappresentare un alleato fondamentale nella riduzione del bias, ma solo se utilizzata in modo consapevole e con un forte controllo etico. In caso contrario, rischia di riprodurre e potenzialmente amplificare le stesse distorsioni storiche, semplicemente sotto una nuova veste tecnologica. La vera sfida, quindi, non consiste soltanto nel perfezionamento tecnico degli algoritmi, ma anche in un ripensamento radicale del modo in cui esseri umani e IA possono collaborare nelle decisioni critiche per la società.

Approcci tecnici per mitigare i bias nell'intelligenza artificiale

Per affrontare efficacemente la questione dei bias nell'intelligenza artificiale, la letteratura propone essenzialmente **due approcci complementari**: un approccio socio-

tecnico, focalizzato sugli aspetti metodologici e algoritmici, e un approccio socio-economico e istituzionale, orientato alle politiche pubbliche e alla governance.

Approccio socio-tecnico

Sul versante tecnico, alcuni suggeriscono l'adozione di tecniche di "fairness-aware machine learning", come la regolarizzazione per la riduzione del bias, che permettono di integrare considerazioni di equità direttamente nel processo di ottimizzazione dell'algoritmo. In questo caso si rende necessaria, inoltre, **l'implementazione di controlli sistematici sia in fase di pre-processamento dei dati** che di post-processamento dei risultati, per individuare e correggere le distorsioni (Lloyd, 2023).

Altri propongono strategie specifiche per mitigare il bias di genere, tra cui il miglioramento dei dataset di addestramento per renderli più equilibrati e rappresentativi della diversità di genere (Nemani *et al.*, 2023). Vengono caldeggiate tecniche di "debiasing" nei modelli linguistici, come la neutralizzazione dei pesi nei layer dei *transformer*, finalizzate a ridurre le associazioni di genere indesiderate.

Particolarmente promettente appare la prospettiva di sviluppare capacità di autocorrezione nei modelli di IA: algoritmi progettati per riconoscere e correggere dinamicamente i propri bias, in un processo di auto-apprendimento e auto-miglioramento continuo (Bridewell et al., 2024).

Approccio socio-economico e istituzionale

Sul piano istituzionale, si enfatizza l'importanza di promuovere **collaborazioni strutturate tra governo, settore privato e società civile per lo** sviluppo di un'IA più equa e responsabile, sottolineando la necessità di una regolamentazione più incisiva, articolata su diversi livelli (Lloyd, 2023):

Garantire una **migliore rappresentazione nei dataset**, attraverso standard che assicurino un bilanciamento adeguato e l'inclusività dei dati di addestramento; promuovere la trasparenza negli algoritmi, rendendo pubblici e comprensibili i criteri decisionali impiegati dai sistemi di IA; istituire organismi indipendenti di supervisione e auditing, incaricati di monitorare sistematicamente i sistemi di IA e correggere le distorsioni identificate.

Altre soluzioni suggeriscono che la regolamentazione dovrebbe concentrarsi sulla trasparenza e sulla gestione del bias, piuttosto che perseguire l'obiettivo irrealistico

della sua eliminazione totale. Questo approccio potrebbe condurre a **politiche più equilibrate, capaci di garantire equità senza compromettere l'efficacia e l'utilità dei sistemi di IA** (Lindloff, Siegert, 2025).

Altri ancora raccomandano lo sviluppo di definizioni più precise e meno sfumate del concetto di "bias", per evitare ambiguità e polarizzazioni inutili nel dibattito pubblico e scientifico (Bridewell *et al.*, 2024). Una terminologia più rigorosa potrebbe facilitare tanto il progresso tecnico quanto l'elaborazione di politiche pubbliche adeguate.

Intelligenza artificiale e censura politica: nuove forme di bias

Una dimensione emergente e particolarmente problematica del bias nell'intelligenza artificiale riguarda **il suo potenziale utilizzo come strumento di censura politica**. Con la diffusione commerciale globale dell'IA e l'intensificarsi della competizione tecnologica tra Stati Uniti e Cina, si è manifestata una tendenza preoccupante: su questioni politicamente sensibili, le piattaforme tendono a rispondere in linea con l'orientamento politico prevalente nel paese d'origine dell'azienda sviluppatrice.

DeepSeek e le risposte censurate su Piazza Tienanmen, gli Uiguri e il Tibet

Il caso di DeepSeek, chatbot cinese di recente sviluppo, risulta emblematico. Il sistema è noto per censurare sistematicamente domande su eventi politicamente sensibili per il governo cinese, come le proteste di Piazza Tienanmen del 1989. Quando interrogata su questo argomento, l'IA evita di fornire risposte dirette, frequentemente cancellando le risposte iniziali e reindirizzando la conversazione verso argomenti neutri come matematica o programmazione. Questo comportamento riflette le normative cinesi che impongono alle aziende tecnologiche il rigoroso allineamento alla narrativa ufficiale del governo (Euronews, 2025).

Un pattern simile si osserva riguardo alle domande sulla situazione degli Uiguri nello Xinjiang, regione dove il governo cinese è stato accusato da organizzazioni internazionali di sistematiche violazioni dei diritti umani. DeepSeek tende a evitare un'analisi approfondita della questione, enfatizzando invece concetti come "stabilità sociale" e "benessere economico" della regione, riproducendo essenzialmente la retorica ufficiale promossa dalle autorità cinesi (Euronews, 2025a).

Analogamente, **DeepSeek mostra reticenza nel fornire risposte dirette su questioni relative al Tibet**. Le domande concernenti la sovranità tibetana vengono trattate con vaghezza o sistematicamente ignorate, in conformità con la strategia del governo cinese di minimizzare il dibattito pubblico su temi legati all'indipendenza tibetana (Chen, 2025).

ChatGPT e risposte addolcite sulla campagna militare di Israele contro Hamas a Gaza

Sul versante occidentale, ChatGPT ha mostrato una tendenza a fornire risposte che potremmo definire "addolcite" relativamente alla campagna militare condotta da Israele contro Hamas a Gaza, manifestando un bias che alcuni critici interpretano come allineamento alle posizioni politiche statunitensi.

Alle domande sulle azioni israeliane a Gaza, ChatGPT tende a enfatizzare "il diritto di Israele a difendersi", iniziando tipicamente la narrazione dal 7 ottobre 2023 (data dell'attacco di Hamas), senza contestualizzare adeguatamente gli eventi all'interno di una prospettiva storica più ampia, che includerebbe riferimenti alla Nakba del 1948 o all'occupazione pluridecennale dei territori palestinesi (The New Arab, 2025).

Inoltre, pur riconoscendo formalmente che Israele è considerato una potenza occupante secondo il diritto internazionale, **il sistema evita generalmente di approfondire le potenziali violazioni umanitarie perpetrate dall'esercito israeliano.** Questa posizione può essere interpretata come una scelta strategica per evitare controversie politiche nell'ambito della polarizzata discussione occidentale sul conflitto israelo-palestinese, mantenendo una parvenza di neutralità che, paradossalmente, finisce per riprodurre una narrativa parziale (The New Arab, 2025).

Questi casi illustrano come i bias nell'intelligenza artificiale non siano esclusivamente il risultato di distorsioni tecniche o statistiche, ma possano **riflettere consapevoli strategie di censura e controllo dell'informazione**, trasformando potenzialmente l'IA in uno strumento di propaganda sofisticato e difficile da contrastare.

Verso un'intelligenza artificiale consapevole dei bias

L'analisi del fenomeno dei bias nell'intelligenza artificiale rivela una realtà complessa e sfaccettata, caratterizzata da dimensioni tecniche, etiche, sociali e politiche profondamente interconnesse.

Da un lato, risulta evidente che **la presenza di bias nei sistemi di IA rappresenta una problematica concreta,** potenzialmente in grado di amplificare disuguaglianze esistenti e impattare negativamente su comunità già vulnerabili. Dall'altro, emerge la necessità di una definizione più rigorosa e articolata del concetto stesso di bias, distinguendo chiaramente tra deviazioni statistiche inevitabili (e talvolta desiderabili) e discriminazioni ingiuste da contrastare attivamente.

Particolarmente preoccupante appare l'emergere d**i forme di bias deliberatamente introdotte a fini di censura politica**, come evidenziato dai casi contrastanti di DeepSeek e ChatGPT, che sollevano interrogativi fondamentali sul ruolo dell'IA come potenziale strumento di controllo dell'informazione nell'era digitale.

Le strategie per affrontare la questione dei bias richiedono **un approccio multidimensionale**, che integri innovazioni tecniche con riforme istituzionali e normative. Cruciale appare lo sviluppo di una governance dell'IA che, superando l'obiettivo irrealistico dell'eliminazione totale del bias, si concentri invece sulla trasparenza algoritmica, sulla rappresentatività dei dati e sulla protezione delle comunità vulnerabili dagli effetti discriminatori.

La sfida fondamentale non consiste tanto nel costruire un'IA "perfettamente neutrale" – obiettivo probabilmente irraggiungibile – quanto nel ripensare criticamente la relazione tra esseri umani e intelligenza artificiale, sviluppando forme di collaborazione consapevole che valorizzino i potenziali vantaggi della tecnologia minimizzandone al contempo i rischi e le distorsioni.

Bibliografia

Bridewell, W., Bello, P. F., Bringsjord, S. (2024). *The Technology of Outrage: Bias in Artificial Intelligence*. *arXiv:2409.17336v1 [cs.CY].*

Euronews (2025). *Chinese AI DeepSeek censors sensitive questions on China when compared to rivals like ChatGPT*. *Euronews Next.* https://www.euronews.com/next/2025/01/28/chinese-ai-deepseek-censors-sensitive-questions-on-china-when-compared-to-rivals-like-chat.

Kuzmanov, I. (2025). *Beyond The Illusion Of Intuition: The Mind, Society, Standardization, And The Future Of Objective Profiling In The Age Of AI*. *Journal of Novel Research and Innovative Development, 3*(1), 98-123.

Lindloff, C., Siegert, I. (2025). *Defining bias in AI-systems: Biased models are fair models*. *ISCA/ITG Workshop on Diversity in Large Speech and Language Models.*

Lloyd, K. (2023). *Bias Amplification in Artificial Intelligence Systems*. arXiv preprint arXiv:1809.07842.

Myers, S. L. (2025). *Deepseek's answers include Chinese propaganda, researchers say*. The New York Times.https://www.nytimes.com/2025/01/31/technology/deepseek-chinese-propaganda.html.

Nemani, P., Joel, Y. D., Vijay, P., Liza, F. F. (2023). *Gender Bias in Transformer Models: A comprehensive review of detection and mitigation strategies*. Natural Language Processing Journal, 6, 100047.

The New Arab (2025). *What does China DeepSeek think about Gaza, opposed to ChatGPT?* The New Arab, www.newarab.com/news/what-does-china-deepseek-think-about-gaza-opposed-chatgpt.