

Lezione 4

Errori dell'AI: allucinazioni, bias e fake news

Obiettivi della lezione

Al termine della lezione saprete:

- cos'è un'allucinazione dell'AI e perché è strutturalmente inevitabile
- cosa sono i bias algoritmici e come distorcono le risposte
- come l'AI viene usata per produrre e amplificare fake news
- quali strumenti e metodi usare per verificare le informazioni

Modulo 1: Le allucinazioni dell'AI

1.1 Cos'è un'allucinazione?

Un'**allucinazione**, in informatica, non in psicologia, è quando un sistema di intelligenza artificiale genera un'informazione **falsa, inventata o distorta**, presentandola con la stessa sicurezza con cui presenterebbe un'informazione vera. L'AI non sa di stare sbagliando. Non mente deliberatamente. Produce semplicemente testo che, statisticamente, sembra plausibile. Per capire perché accade, dobbiamo capire cosa fa davvero un modello linguistico come ChatGPT o Claude.

1.2 Come funziona davvero un LLM — spiegazione semplice

Un Large Language Model (in italiano: modello linguistico di grandi dimensioni), come già detto nella scorsa lezione, non è un motore di ricerca. Non consulta database in tempo reale. Non "sa" cose nel senso in cui le sa un essere umano. Fa una cosa sola:

predice la parola (tecnicamente: il token) più probabile dato il contesto precedente. Nient'altro.

Immaginate di giocare a completare frasi. Vi dico: "La capitale della Francia è..." e voi completate automaticamente con "Parigi". Non lo avete cercato su Google: lo sapete perché l'avete letto milioni di volte. Il modello fa esattamente questo, ma su scala industriale: ha letto miliardi di pagine web, libri, articoli, forum.

Il problema è che il modello ha imparato non solo i fatti veri, ma anche i *pattern* del linguaggio. Sa come si scrive un abstract scientifico. Sa come si cita un articolo. Sa come suona un DOI. Quando gli chiedete di citare fonti su un argomento, costruisce una citazione **che assomiglia a una citazione reale**, ma non necessariamente lo è.

ESEMPIO

Chiedete a ChatGPT il 'significato' di un'espressione idiomatica inventata — per esempio 'fare il salto della volpe rossa'. L'AI la spiegherà con sicurezza, descrivendone l'origine e l'uso colloquiale. L'espressione non esiste. Questo è esattamente il meccanismo delle allucinazioni: il modello completa un pattern plausibile senza verificarne l'esistenza.

1.3 Perché le allucinazioni sono inevitabili

Potrebbe sembrare che le allucinazioni siano un bug che le aziende tecnologiche risolveranno prima o poi. In realtà, la maggioranza dei ricercatori ritiene che siano una caratteristica **intrinseca** del modo in cui questi modelli funzionano. Emily Bender, docente di Linguistica informatica, ha spiegato che questi sistemi inventano sempre le cose: quando i contenuti generati risultano corretti, avviene solo, in un certo senso, per caso.

La frequenza degli errori è diminuita: i modelli più avanzati oggi alluciano nel 1-2% delle risposte generali, ma questa percentuale sale drammaticamente su temi specifici. In ambito giuridico, anche i migliori modelli possono sbagliare una sentenza o inventare un precedente nel 6% dei casi. In ambito medico le conseguenze di quel 6% possono essere gravi.

C'è anche una ragione commerciale: i modelli sono progettati per soddisfare l'utente. Chi paga un abbonamento si aspetta risposte, non ammissioni di ignoranza. Questo incentivo spinge i sistemi a fornire *qualcosa*, anche quando quel qualcosa non esiste.

1.4 Casi reali aggiornati

Il caso degli avvocati (USA, 2023 — ancora attuale)

Nel caso *Mata vs Avianca*, due avvocati americani hanno presentato a un giudice federale un atto legale contenente sei precedenti giurisprudenziali... completamente inventati da ChatGPT. Nomi plausibili, date credibili, motivazioni dettagliate. Tutto falso. Gli avvocati sono stati multati e hanno rischiato sanzioni disciplinari.

Non si tratta di un episodio isolato: tra il 2023 e il 2025 sono emersi quasi 90 casi di allucinazioni AI nei tribunali internazionali. Nel 2025 due giudici dell'Alta Corte britannica hanno avvertito formalmente gli avvocati che rischiano la persecuzione penale per uso di materiale falso generato dall'AI.

Il caso di Firenze (Italia, 2025)

In un caso italiano documentato nel 2025, un tribunale ha rilevato la presenza di sentenze inventate all'interno di atti legali redatti con il supporto di AI. È diventato un caso-simbolo del rischio allucinazione in ambito professionale nel nostro paese.

Il test della Mongolia (2025)

Nel marzo 2025, ricercatori dell'Università di Toronto hanno testato 12 dei principali modelli AI chiedendo di elencare i paesi confinanti con la Mongolia. Nove di essi hanno risposto con sicurezza includendo il Kazakistan, che non condivide alcun confine con la Mongolia. La geografia è una delle aree in cui i modelli mostrano maggiore sicurezza... e maggiori errori.

Incoronazione di Carlo III (2023)

ChatGPT ha affermato che l'incoronazione di Carlo III si è tenuta il 19 maggio 2023 nell'Abbazia di Westminster. La cerimonia era effettivamente avvenuta nell'Abbazia di Westminster, ma il 6 maggio. Risposta scorrevole, dettaglio cruciale sbagliato.

1.5 Errori tipici: un catalogo

Le allucinazioni si presentano in forme riconoscibili. Imparatele a riconoscere:

- Citazioni bibliografiche inesistenti: titoli plausibili, DOI (Digital Object Identifier: sono codici univoci che identificano in modo permanente un documento digitale, tipicamente articoli scientifici, saggi, libri o dataset) inventati, autori che non hanno scritto quel testo;
- Date e fatti storici: errori di pochi giorni o anni, ma sistematici;
- Statistiche inventate: percentuali, numeri, sondaggi che suonano reali ma non hanno fonte;
- Norme e leggi errate: articoli di legge inesistenti o con contenuto distorto;
- Biografie di persone: confusione tra omonimi, attribuzione di eventi a persone sbagliate;
- Tutto ciò che viene chiesto con 'cita', 'elenca le fonti', 'dammi dei riferimenti';

Modulo 2: Bias nei dati e pregiudizi algoritmici

2.1 Cos'è un bias: introduzione

Se le allucinazioni sono errori individuali, il bias è un **errore sistematico e ripetibile**. Un sistema è biased (distorto) quando produce regolarmente risultati che favoriscono certi gruppi, valori o prospettive rispetto ad altri. Non per caso, ma per struttura.

Il bias nasce principalmente dai dati di addestramento. Un modello impara dai testi che ha letto. Se quei testi riflettono i pregiudizi della società, il modello li replica e spesso li amplifica. L'AI non è neutrale: è uno specchio che ingrandisce i pregiudizi del materiale su cui è stata addestrata.

2.2 Bias nei dati di addestramento

I dati con cui vengono addestrati i modelli vengono prevalentemente da Internet. E Internet non è un campione rappresentativo dell'umanità'.

Esso è sovrarappresentato da:

- testi in lingua inglese (oltre il 60% dei dati di addestramento)
- contenuti prodotti da uomini, in paesi occidentali, in fasce d'età tra 18 e 40 anni
- fonti formali come articoli, Wikipedia, forum tecnici
- dati storici che riflettono discriminazioni del passato

Questo significa che il modello ha una visione parziale del mondo. Quando risponde su cultura, storia, valori sociali, lo fa con il filtro di quella parzialità'.

ESEMPIO

Il caso Amazon: il gigante dell'e-commerce ha sviluppato un sistema AI per filtrare i curriculum. Il sistema era stato addestrato sui dati storici delle assunzioni dell'azienda, in cui i candidati assunti erano prevalentemente uomini. L'algoritmo aveva imparato che 'uomo' era un predittore di successo, e penalizzava sistematicamente le candidate femminili. Amazon ha dovuto ritirare il sistema nel 2018.

2.3 Il caso COMPAS: bias nella giustizia

ESEMPIO

Questo è uno dei casi più documentati e discussi nella letteratura sul bias algoritmico. Raccontatelo con calma.

COMPAS (Correctional Offender Management Profiling for Alternative Sanctions) è un software usato in alcuni tribunali americani per stimare la probabilità di recidiva di un imputato, cioè la possibilità che commetta nuovi reati. I giudici usavano questo punteggio nelle decisioni su cauzione e sentenze.

Un'indagine dell'organizzazione ProPublica condotta su oltre 10.000 imputati ha rilevato risultati agghiaccianti:

- Il tasso di falsi positivi (persone classificate come ad alto rischio che non recidivano) era del 45% per gli imputati afroamericani.
- Lo stesso tasso era solo del 23% per gli imputati bianchi.
- In pratica, il sistema etichettava come 'pericolosi' il doppio delle persone nere rispetto alle bianche, a parità di condizioni.

Il sistema non aveva mai visto il colore della pelle: era stato addestrato su dati storici delle decisioni giudiziarie americane, che già incorporavano decenni di discriminazione sistemica. L'AI aveva imparato e replicato quella discriminazione.

2.4 Bias culturali e linguistici: il problema italiano

Il bias non riguarda solo razza o genere. Esiste un profondo **bias culturale e linguistico** che penalizza tutto ciò che non è angloamericano.

I modelli AI producono risposte più accurate, più articolate e con meno errori in inglese rispetto a qualsiasi altra lingua. In italiano ottengono risultati sensibilmente peggiori. In lingue meno diffuse, come il sardo, il friulano, il romano, il degrado è ancora più marcato.

Ma il problema non è solo quantitativo. È anche qualitativo: quando un'AI parla di cultura italiana, storia locale, cucina regionale, dialetti o tradizioni, tende a semplificare, stereotipare o normalizzare in chiave angloamericana. La cucina italiana diventa una variante generica; le differenze tra Nord e Sud vengono appiattite; la storia locale viene filtrata attraverso lo sguardo esterno.

DOMANDA

"Chiedete a ChatGPT di descrivere una 'famiglia italiana tipica'. Poi chiedete di descrivere una 'typical American family'. Confrontate: quali stereotipi emergono? L'AI parla in modo diverso delle due culture?"

2.5 Il loop del bias: perché peggiora nel tempo

C'è un aspetto particolarmente preoccupante del bias algoritmico: tende ad autoalimentarsi. I contenuti generati dall'AI vengono pubblicati su Internet. Quei contenuti vengono poi usati per addestrare i modelli successivi. Se i contenuti attuali sono biased, i modelli futuri saranno addestrati su dati già distorti, producendo ulteriore distorsione.

È un circolo vizioso che, senza interventi correttivi attivi, amplifica i pregiudizi esistenti invece di ridurli.

DEMO LIVE

Chiedete a un'AI generativa di immagini (o chiedetele di descrivere) 'un medico', 'una cassiera', 'un hacker', 'una maestra'. Osservate quali stereotipi di genere, etnia ed età emergono. Chi ha insegnato all'AI queste associazioni?"

Modulo 3: Fake news, deepfake e perché l'AI non è una fonte

3.1 L'AI come strumento di disinformazione

Fino a qui abbiamo parlato di errori: allucinazioni e bias sono effetti involontari. Ma l'AI viene usata anche intenzionalmente per produrre disinformazione a bassa scala e ad alta velocità.

Secondo un'analisi di NewsGuard, il tasso di false informazioni diffuse dai chatbot AI è quasi raddoppiato in un anno, passando dal 18% al 35%. Esistono centinaia di siti web gestiti quasi interamente da AI, progettati per imitare graficamente testate giornalistiche affidabili e pubblicare contenuti falsi o distorti. Solo in lingua italiana ne sono stati identificati 57, molti facenti parte di reti coordinate.

3.2 I deepfake: la disinformazione con un volto

Un deepfake è un contenuto video, audio o fotografico generato o manipolato dall'AI per far sembrare che una persona reale dica o faccia qualcosa che non ha mai detto o fatto.

In Italia, nel 2025 si sono moltiplicati i casi documentati. Vi racconto i più significativi:

Caso 1: Il falso TG5 sull'hijab

Su TikTok ha circolato un video che sembrava un servizio del TG5 che annunciava: dal primo gennaio 2026, vietato indossare l'hijab in tutto il territorio italiano. Il video era stato generato con Sora di OpenAI, brandizzato con il logo del TG5, totalmente all'oscuro, e ha scatenato centinaia di commenti anti-islamici. Non era vero. Era disinformazione confezionata con l'AI.

Caso 2: Deepfake di personaggi famosi per truffe finanziarie

Negli ultimi mesi del 2025 si sono diffusi su Facebook e TikTok video in cui Carlo Cracco, Joe Bastianich, Chiara Ferragni e Marco Montemagno sembrano promuovere piattaforme di investimento in criptovalute o corsi di guadagno rapido. Tutti falsi. Cracco e Ferragni hanno sporto denuncia. Meta ha avviato la rimozione di centinaia di contenuti, ma molti avevano già raggiunto migliaia di persone.

Caso 3: Voice cloning e truffe telefoniche

Nel 2025 in Italia sono stati segnalati i primi casi di voice cloning: truffatori che usano voci generate dall'AI, costruite da audio rubati dai social, per telefonare agli anziani fingendosi figli o nipoti in difficoltà. La Banca d'Italia ha segnalato il fenomeno come "indicativo di un trend che potrebbe diventare esplosivo".

Caso 4: La foto modificata di Torino (2026)

A febbraio 2026 la Polizia di Stato ha pubblicato la foto di un agente ferito durante gli scontri a Torino (manifestazione per Askatasuna). Siti di fact-checking hanno rilevato che l'immagine era stata modificata con l'AI prima della pubblicazione ufficiale. È diventato un caso politico, sollevando domande sulla trasparenza istituzionale nell'uso dell'AI.

Caso 5: Arianna, studentessa pugliese

Una studentessa di 19 anni ha denunciato pubblicamente nel 2025 di essere stata vittima di deepfake: qualcuno le aveva sottratto una foto personale, l'aveva manipolata digitalmente trasformandola in un'immagine sessuale, e l'aveva stampata su volantini affissi nella sua città. È un caso di revenge porn generato dall'AI, oggi reato in Italia punito con la Legge 132/2025.

3.3 Perché l'AI non è una fonte?

Ricapitoliamo tutto quello che abbiamo visto. Un'AI generativa:

- non consulta fonti: genera testo statisticamente plausibile;
- non distingue tra vero e falso: produce output coerente, non verificato;
- non ha responsabilità editoriale: nessuno controlla ciò che pubblica;
- può citare fonti inesistenti con assoluta sicurezza;
- non aggiorna le informazioni in tempo reale (ha una data di interruzione della conoscenza);
- può essere usata deliberatamente per produrre contenuti falsi;

La regola d'oro è questa: **l'AI è un punto di partenza per esplorare un argomento, mai un punto di arrivo per verificare un fatto.**

3.4 Il metodo SIFT: verificare le informazioni

Il metodo SIFT è stato sviluppato per l'educazione alla media literacy. Si applica a qualsiasi informazione trovata online, incluse le risposte dell'AI.

- **S=Stop!**
Prima di condividere o usare un'informazione, fermati un secondo. Non reagire di impulso.
- **I=Investigate!**
Chi ha prodotto questa informazione? È una testata affidabile? Ha conflitti di interesse? Se viene da un'AI, qual è la sua fonte?
- **F=Find!**
La stessa notizia è riportata da altre fonti affidabili e indipendenti? Se la risposta è no, è un segnale d'allarme.
- **T=Trace!**
Risali alla fonte originale della dichiarazione, della statistica, della foto. Il contesto originale spesso cambia completamente il significato.

3.5 Strumenti pratici di verifica

Questi strumenti sono gratuiti e accessibili a chiunque:

- **Pagella Politica:** fact-checking in italiano su dichiarazioni politiche
 - **Facta.news:** verifica di notizie e immagini virali in italiano
 - **Snopes:** fact-checking internazionale in inglese
 - **Google Immagini / TinEye:** ricerca inversa di immagini per verificarne l'origine
 - **InVID / WeVerify:** analisi di video per rilevare manipolazioni
 - **Google Scholar:** verifica se una citazione scientifica esiste davvero
-

Modulo 4: Sintesi, verifica e chiusura

4.1 Sintesi dei concetti chiave

Oggi abbiamo affrontato tre fenomeni diversi ma collegati:

Le allucinazioni sono errori strutturali. L'AI genera testo plausibile, non testo vero. Cita fonti che non esistono, inventa dati, sbaglia date. Non per malafede, ma per come e' costruita. La frequenza diminuisce ma non si azzerera' mai del tutto.

I bias sono distorsioni sistematiche. L'AI replica e amplifica i pregiudizi presenti nei dati su cui e' stata addestrata: pregiudizi di genere, razza, cultura, lingua. Un sistema addestrato su dati storici distorti produce decisioni distorte. Il caso COMPAS nella giustizia e il caso Amazon nel reclutamento lo dimostrano.

Le fake news sono l'uso deliberato di questi strumenti per disinformare. I deepfake in Italia — dal falso TG5 alle truffe con voci clonate — mostrano che il problema non e' piu' teorico. E' quotidiano. E riguarda tutti noi.

Di fronte a tutto questo, la risposta non e' rinunciare all'AI. E' usarla in modo critico: conoscerne i limiti, verificare le informazioni che produce, applicare il metodo SIFT, usare gli strumenti di fact-checking.

"L'AI non e' ne' magica ne' infallibile. E' uno strumento potente, costruito da esseri umani, con i limiti degli esseri umani. Saperne riconoscere gli errori non vuol dire non usarla: vuol dire usarla bene. E in un mondo inondato di contenuti generati automaticamente, la capacita' di verificare e' diventata una competenza fondamentale quanto saper leggere."