ADOLFO SCALFATI

Allucinazioni e pregiudizi dell'intelligenza artificiale

È difficile credere che l'algoritmo funzioni come la mente umana, mancandogli le caratteristiche che sorreggono quest'ultima, ma sarebbe sbagliato pensare che l'AI sia infallibile: essa non sfugge ad allucinazioni e bias.

Hallucinations and biases of artificial intelligence

It's hard to believe that the algorithm functions like the human mind, lacking the characteristics that support the latter; but don't think that AI is infallible, because it cannot escape hallucinations and biases.

1. Décalage *tra algoritmo e mente*. In materia giudiziaria, occorrono due precisazioni.

La prima riguarda la misura in cui l'AI – benché sia in grado di migliorare il funzionamento dell'apparato amministrativo – può interagire con le formule processuali, cioè, con la dinamica degli atti che conduce alla decisione. L'accertamento penale si serve di un complesso di regole che contano molto più del risultato: in tanto quest'ultimo è "giusto" se le prime sono state seguite. Si intuisce facilmente, allora, che la tecnologia o gli apporti scientifici – ai quali taluno conferisce eccessivo credito – non circolano liberamente nel perimetro giudiziario, ma di essi ci può servire solo inserendoli tra le griglie normative che delineano l'*an* e il *quomodo* delle condotte processuali; per esempio, un esperimento affidato all'AI produce un esito probatoriamente spendibile ove sia guidato dal contraddittorio, non leda la dignità umana e non violi divieti specifici.

La seconda sottolineatura attiene al tipo di ragionamento prodotto dall'algoritmo; qui sarebbe molto semplicistico abbinarlo o compararlo a quello umano, trattandosi di fenomeni molto diversi tra loro [1]. Per un verso (perlomeno sinora), solo all'uomo è dato cogliere, tramite l'esperienza dei sensi, il contesto fenomenico, di per sé immancabilmente variabile, in base al quale egli compie scelte di valore in un determinato contesto storico-sociale. Per altro, il pensiero umano è condizionato dai meccanismi dell'inconscio,

[1] Cfr., NIEVA FENOLL, *Intelligenza artificiale e processo* (trad. a cura di P. Comoglio), Torino, 2019, 9-10

territorio sterminato dal quale provengono impulsi, talvolta inconsapevoli, che guidano ineluttabilmente qualunque scelta, inclusa quella che appare la più razionale.
Se, poi, si intendesse sostenere che proprio tale ultimo profilo costituisce una ragione sufficiente per affidarsi alla macchina – perché, appunto, è capace di decidere facendone a meno – si approderebbe ad un'idea di giustizia incompatibile con la dignità dell'uomo e la sua centralità, oltre a presentare caratteri distonici con le principali fonti dell'ordinamento (artt. 13, 14, 15, 101, co. 2, 111, co. 2 Cost.).
Non inganni, allora, la possibile coerenza che regge una narrazione (o una risposta) proveniente dall'AI, perché essa non è paragonabile, per ciò solo, al pensiero umano il quale si serve di un costrutto semantico-linguistico flessibile in quanto modulato dai fattori dell'esperienza; fenomeno ignoto alla macchina. L'algoritmo è in grado di assemblare milioni di dati che contengono o implicano le parole con le quali è stata formulata la domanda; il che è ben distante da una elaborazione mentale attrezzata ad effettuare rilievi e valutazioni derivanti dal contesto concreto in cui occorre decidere.
Nell'ambito penalistico, saldamente ancorato all'identità tra il giudice-persona che sovraintende l'istruttoria e quello che decide (art. 525, co. 2 c.p.p.), i precedenti rilievi si traducono in un giudizio di inadeguatezza del macchinario quando si tratta di valutare il quadro probatorio [2] o di selezionare la norma più adatta. Si tratta di un profilo di grande rilievo, sottovalutato dalla disciplina italiana che attua l'AI Act europeo, laddove non vieta il ricorso all'algoritmo ma riserva "*al magistrato ogni decisione sull'interpretazione e sull'applicazione della legge, sulla valutazione dei fatti e delle prove e sull'adozione dei provvedimenti.*" (art. 15, co. 1 L. 23 settembre 2025 n. 132); come dire che, pure in tali ambiti, il ricorso all'AI non è escluso sempre che spetti al giudice l'ultima parola[3].
Anche qualora il progresso tecnologico generi *output* in grado di fornire una più attendibile qualificazione giuridica al fatto [4], il risultato dipenderebbe

[2] Cfr. i rilievi di MAZZA, *Distopia del processo artificiale*, in *Arch. pen. web*, 7 gennaio 2025.
[3] Secondo altri, la previsione nazionale richiamata escluderebbe in radice la possibilità di ricorso all'intelligenza artificiale nei giudizi giuridici: GRIMALDI, *I provvedimenti giudiziari nell'era dell'Intelligenza Artificiale: il ruolo del giudice*, in *Giur. it.*, 2026, 454.
[4] In argomento la letteratura è molto ampia; tra gli altri, più di recente, DI GIOVINE, *Interpretazione umana e robotica (breve riflessione sul possibile ingresso dei* Large Language Model *nella giurisdizione*, in *www.sistemapenale*.it, 4 settembre 2025.

dall'educazione dell'algoritmo all'abbinamento di casistiche "rigide"; per esempio, si potrebbe ottenere dall'AI i *nomina juris* di condotte descritte alla stregua di rappresentazioni-modello, scevre dalla moltitudine di elementi umani di cui sono intrise, per esempio, i significati di colpa, dolo, preterintenzione, errore, patologie mentali, motivi a delinquere, fattori soggettivi della commisurazione sanzionatoria, ...
Ma pur ipotizzando un più alto livello di perfezione algoritmica, non è detto che la risposta sia giuridicamente corretta; né è sicuro che essa sia compatibile con la realtà fenomenica che l'utente deve analizzare la quale, per sua natura, contiene sfumature impercettibili in grado di mutare il quadro normativo di riferimento[5].

2. Interna corporis *ignoti e allucinazioni*. Restando sul terreno dell'affidabilità, il processo algoritmico delle *machine learning* –– soprattutto nei modelli più complessi – si sviluppa in modo da risultare inconoscibile persino agli esperti, al punto da essere equiparato ad una *black box*[6]*;* il che significa che l'*output*, essendo generato da itinerari non perspicui, non è un frutto logicamente validabile.
Così stando le cose, può darsi che la credibilità della risposta sia integralmente rimessa alla verifica "esterna" da parte dell'utente – quando è in grado di effettuarla – realizzata mediante altri canali di conoscenza o, peggio, sia affidata ad un pendolo intuitivo che oscilla tra convincente/non convincente, verosimile/non verosimile. Su queste premesse, l'apporto del macchinario sul versante cognitivo diventa poco utile o persino insidioso nella misura in cui ci si accosta all'*output* con atteggiamento scontato o addirittura fideistico.
Se il procedimento interno alla macchina funziona sulla base di abbinamenti linguistici, anche quando è capace di scomporre e ricomporre interi asserti, non è raro imbattersi in veri e propri errori annidati nella trama narrativa dell'*output*, riflessi sui corrispondenti esiti. Si tratta delle c.d. allucinazioni che si manifestano con il richiamo a fatti giuridici inesistenti o a circostanze completamente inventate. Al riguardo, la giurisprudenza si è già pronunciata con

[5] Sottolinea bene i fattori umani che orientano il giudizio sulla qualificazione del fatto, mancanti nella eventuale ed analoga attività selettiva affidata all'AI, PALAZZO, *Notazioni minime su intelligenza artificiale e ruolo del giurista*, in *Cass. pen.*, 2026, p. 288.
[6] Cfr., tra gli altri, PASQUALE, The black box society. The secret algorithms. That control money and information, Harward University Press, 2016, passim.

un insieme di *caveat* nei confronti dell'uso giudiziario dell'intelligenza artificiale, arrivando a sanzionare condotte forensi alle quali si addebita di aver prodotto dati falsi.

Da ultimo, con molta chiarezza, si è scritto, a proposito del "*... difensore* [il quale] *si* [è] *avvalso di uno strumento di intelligenza artificiale generativa senza sottoporre gli output ottenuti alla doverosa verifica sulle fonti primarie*", che *"costituisce ormai fatto notorio... e certamente esigibile da un operatore professionale del diritto, che i modelli di intelligenza artificiale generativa (c.d. Large Language Models) non [sono] banche dati giurisprudenziali da cui estrarre precedenti e citazioni, bensì strumenti di generazione automatica del linguaggio fondati su meccanismi inferenziali di natura statistica e probabilistica. Tali sistemi, in altri termini, non "sanno" né "ricordano" alcunché, ma si limitano a produrre sequenze di testo statisticamente plausibili sulla base di miliardi di parametri di addestramento, senza avere accesso – ordinariamente – ad alcuna base di conoscenza verificata o verificabile. È per tale ragione che i modelli di intelligenza artificiale generativa sono notoriamente soggetti al fenomeno delle c.d. "allucinazioni", consistente nella generazione di contenuti formalmente plausibili ma sostanzialmente falsi o inesistenti, ivi comprese citazioni giurisprudenziali mai rese. L'utilizzazione acritica di tali strumenti, senza la doverosa verifica dell'attendibilità degli output mediante consultazione delle fonti primarie... integra gli estremi della colpa grave, non potendosi più tollerare, allo stato attuale delle conoscenze tecnologiche diffuse, errori di tale natura, i quali – lungi dal costituire meri refusi o imprecisioni aggravano significativamente l'attività del giudice e delle controparti"*[7].

3. *Inquinamento da* bias. Nel giudizio giuridico bisogna sottrarsi alle soluzioni lampo, prima quelle dell'uomo, che si potrebbero definire, con larga sommarietà, come i percorsi mentali che saltano alle conclusioni sulla base di informazioni incomplete, memorie personali e massime di esperienza a bassa frequenza statistica. Li hanno chiamati *bias*, scorciatoie comode a chi le usa, so-

[7] Trib. Siracusa, II sez. civile, sent. 20 febbraio 2026, n. 338.
Si tratta di un orientamento di merito consolidato. Si legga, in conformità e più di recente, anche Trib. Ferrara, sez. civile, sent. 4 marzo 2026, M.G. ed altri, nonché, Trib. Verona, II sez. civile, sent. 10 febbraio 2026 n. 4203 che ha imposto un "obbligo di controllo rafforzato" sugli *output* dell'AI generativa prodotti in giudizio

spinte da una fisiologica pigrizia verso lo sforzo intellettuale[8]; per esempio, vedendo una colonna di fumo sollevarsi d'estate all'orizzonte, si sostiene che si tratta di un incendio invece di valutare che può generarla un macchinario malridotto di un insediamento industriale. Semplificando, si potrebbe dire che si tratta di "pre-giudizi" dell'algoritmo conseguiti trascurando soluzioni alternative.

Ora, il tasso di *bias* nei sistemi di intelligenza artificiale dipende da quali e quante informazioni alimentano la macchina e dalla raffinatezza delle procedure informatiche[9]; gli *output* possono essere frequentemente sbagliati proprio per la insufficienza di tali fattori[10].

Non è un tema nuovo se lo si osserva nell'ottica umana. La decisione giudiziale è spesso dominata da pensieri intuitivi, fondati su esperienze personali e informazioni incomplete. La dinamica logica corrotta da *bias* determina una pressoché automatica esclusione mentale di altre possibili soluzioni rispetto a quelle raggiunte con il pensiero veloce. Ed ecco che ritornano in auge studi degli anni sessanta, dove era stato scritto come molti fattori, sospinti da intuizioni consce o inconsce, influenzino le decisioni, sottraendone una notevole quota ai processi razionali [11]; e così, se la pronuncia non può fare a meno di percorsi intuitivi, la sua motivazione rappresenta un costrutto giustificativo piuttosto che esplicativo del processo logico determinante la scelta giudiziaria.

È evidente, infine, come lo scarto tra ragionamento veloce e ragionamento complesso diventi più pericoloso nelle decisioni prognostiche, soprattutto in tema di libertà personale.

4. Incompletezza investigativa e sviamenti. A proposito di *bias*, è stato detto che il *deficit* delle indagini, abbinato ai pensieri veloci, determina indirizzi investigativi capaci di condizionare l'intero andamento del processo, fino a produrre errori giudiziari. Una fotografia della realtà[12]. A tal riguardo, quale rimedio al *bias* investigativo, si è proposto lo scrupoloso adempimento

---

[8] KAHNEMANN, *Pensieri lenti e veloci*, Mondadori, Milano, 2012, 250 ss.

[9] ROMANÒ, *Intelligenza artificiale come prova scientifica nel processo penale: una sfida tra machine-generated evidence e equo processo*, in *Prova scientifica e processo penale,* a cura di Canzio–Luparia, Milano, 2022, 921.

[10] In argomento, VALENTINI, *La "scoperta" dei bias cognitivi nel processo penale*, in *Arch. pen.*, 2025, 3, 12 ss.

[11] MASSA, *Contributo all'analisi del giudizio penale di primo grado*, Milano, 1964, passim.

[12] VALENTINI, *La "scoperta" dei bias cognitivi nel processo penale* cit., 18.

all'obbligo di effettuare indagini complete, agganciando tale dovere – sottolineato dalla Corte costituzionale [13]– ai contenuti dell'art. 358 c.p.p. nella parte in cui inducono il pubblico ministero a raccogliere anche elementi a favore dell'indagato.

Al di là dell'orientamento dominante che sostiene la non sanzionabilità delle omissioni investigative contemplate dall'art. 358 c.p.p.[14], questa linea di pensiero rischia di solidificare l'ambiguità del pubblico ministero attribuendogli il dovere di agire *in utroque* (contro e a vantaggio dell'accusato), quasi possegga la stessa neutralità che si pretende da altro soggetto del processo, il giudice; un'idea rischiosa che, nata dall'apprezzabile scopo di scongiurare i limiti di prospettiva maturati nelle indagini, si serve di un rimedio peggiore del male quando inconsapevolmente riflette la sovrapposizione delle caratteristiche tra l'inquirente e chi decide: una trasfigurazione del pubblico ministero, aggrappata all'ossimoro di "parte-imparziale" da qualcuno invocato in funzione di una pretesa neutralità assunta in fase investigativa, postula che egli muti pelle a seconda che si muova prima o dopo l'esercizio dell'azione penale.

Quando il complesso indiziario presenta un'incompletezza investigativa, sarà il giudice – nel predibattimento – a doverla riconoscere, anche a seguito delle indicazioni fornite dalla difesa, fronteggiandola con integrazioni (artt. 421 bis e 422 c.p.p.) o rilevando negativamente la prognosi di condanna; sempre il giudice, messo di fronte alla decisione di merito sull'accusa, dovrà fare un corretto uso della formula dubitativa e buon governo della regola secondo cui la colpevolezza va provata oltre ogni dubbio ragionevole: solo la figura dotata di neutralità, vera *suitas* delle proprie scelte, può dissolvere, nel contesto dialettico e mediante gli strumenti di cui dispone, l'ottica dell'inquirente turbata dai *bias*.

Si obietterà che durante le indagini il giudice non ha mezzi per intervenire sui *bias* dei quali è affetto il complesso investigativo; cosicché, se non si pretende un atteggiamento neutrale dal pubblico ministero nell'inchiesta preliminare, il problema rischia di non trovare rimedio influenzando le pronunce in tema di libertà (misure cautelari e intercettazioni); ma qui, senza indugio, occorre una terapia legislativa sui poteri del giudice contro la distrofia dell'operato inquirente.

---

[13] Corte cost., sent. 15 febbraio 1991, n. 88.

[14] Da ultimo, sulla base di una linea interpretativa stabilizzata, Cass., Sez. VI, 22 settembre 2025, n. 32938.