La zona di accartocciamento morale

Bias di automazione, subentro euristico e la ridefinizione dell'ontologia criminale

L'introduzione massiccia e pervasiva dei sistemi di intelligenza artificiale (IA) all'interno dell'ecosistema della giustizia penale, dell'applicazione della legge e delle attività di intelligence investigativa non rappresenta una semplice evoluzione degli strumenti tecnologici a disposizione degli analisti. Si tratta, a tutti gli effetti, di una vera e propria ristrutturazione dell'architettura della percezione criminale, sociale ed epistemologica. Quando deleghiamo il calcolo del rischio di recidiva, l'allocazione delle risorse di pattugliamento o la valutazione della pericolosità sociale a un modello matematico, stiamo inavvertitamente alterando i meccanismi biologici e psicologici fondamentali con cui l'essere umano formula il proprio giudizio.
Come già analizzato in un mio precedente articolo, l'algoritmo, in contesti ad alto rischio, non si limita a decidere chi debba essere osservato o dove una pattuglia debba recarsi; attraverso meccanismi di retroazione e opacità statistica, esso finisce inevitabilmente per ridefinire ciò che l'intera società sceglie di considerare pericoloso, creando una nuova e distorta ontologia del crimine. Il risultato di questo affidamento acritico alle macchine è l'emergere di vulnerabilità ecosistemiche critiche, che si manifestano attraverso la proliferazione di monocolture algoritmiche, l'innesco di circuiti di retroazione fuori controllo e, in ultima analisi, in quello che definisco il collasso epistemico della trasparenza giuridica.
Nel dibattito contemporaneo sulle scienze forensi, sull'analisi comportamentale e sull'intelligenza artificiale, si tende frequentemente a esaltare la presunta neutralità, l'oggettività e la diversità dei sistemi computazionali. Tuttavia, l'osservazione empirica delle dinamiche di mercato e delle prassi istituzionali rivela una tendenza diametralmente opposta e allarmante: la rapida convergenza verso le "monocolture algoritmiche". Il concetto di monocoltura algoritmica, formalizzato in recenti studi di settore, definisce quello stato sistemico in cui numerosi decisori, agenzie governative, dipartimenti di polizia o corti di giustizia si affidano esattamente allo stesso algoritmo, o a sistemi che condividono la medesima architettura di base, gli stessi

A cura della Dottoressa Cristina Brasi

Psicologa, FBA-LAB. Vice-Presidente della Commissione Tecnico Scientifica dei Criminalisti e Criminologi dell'ASSOIP, Associazione Italiana delle Professioni.
Brasi Cristina, psicologa, criminologa forense e analista scientifica del linguaggio non verbale.
Specializzata in psicopatologia forense, devianze e attitudine al crimine, psicologia legale e
forense, criminologia forense e investigativa, profiling, sessuologia tipica e atipica, valutazione delle abilità genitoriali, trattamento dei PTSD dall'età prescolare e diagnosi e trattamento dei disturbi della personalità.
Ha conseguito master in Criminologia Forense, master in Analisi Scientifica del Linguaggio Non Verbale, master in Sessuologia Clinica e master in Diagnosi e Trattamento dei Disturbi della Personalità.
Abilitata all'esercizio della professione di Psicologa, Psicologa Giuridica e Criminologa Forense.
Abilitata inoltre all'utilizzo del Sistema di Analisi Scientifica -Body Cody System- e del sistema avanzato di decodifica del comportamento facciale -Sistema Interpretativo delle Espressioni Facciali-
Specializzata anche in disfunzioni sessuali e disturbi della sessualità tipici e atipici, nonché esperta in sessualità tipica e atipica. È in corso l'abilitazione a sessuologo clinico.
Fondatrice nel 2019 di FBA-LAB, il primo Laboratorio di Analisi Comportamentale Forense in Italia.
Collabora con l'Università degli Studi di Bari come docente incaricata, insegnando nel campo della radicalizzazione e deradicalizzazione, con particolare attenzione alla radicalizzazione di bambini e adolescenti on-line e delle donne.
Nel tempo si è specializzata nella profilazione di personalità governative di alto livello.
Dirige una collana scientifica di Criminologia e Criminalistica e fa parte dell'Associazione M.I.A. (MEDITERRAN IA) con sede a Roma.

dati di addestramento (spesso viziati) e identici protocolli di allineamento.
Quando l'errore umano domina le indagini, esso è idiosincratico: giudici o investigatori diversi commettono errori in direzioni diverse, e in un sistema plurale (policoltura) questi errori tendono a bilanciarsi o a essere corretti nei successivi gradi di giudizio. Al contrario, l'adozione di un modello algoritmico condiviso trasforma i bias locali in bias strutturali, cristallizzando l'errore all'interno di matrici decisionali che si auto-validano, precludendo così ogni forma di falsificabilità empirica del dato investigativo.
Il settore in cui la monocoltura algoritmica e l'affidamento acritico all'IA dimostrano i loro effetti più tossici e tangibili è senza dubbio quello del *predictive policing* (polizia predittiva). Sistemi commerciali ampiamente diffusi a livello globale operano partendo da una premessa intuitiva e apparentemente logica: utilizzare i vasti database storici sui crimini per prevedere le coordinate spazio-temporali (il "dove" e il "quando") in cui è statisticamente più probabile che si verifichi un reato in futuro, ottimizzando così l'allocazione delle risorse di pattugliamento.
Tuttavia, l'analisi criminologica e statistica rigorosa di questi modelli rivela una profonda e pericolosa fallacia epistemologica. L'algoritmo non viene nutrito con i dati dei "crimini effettivamente commessi" (che costituiscono il numero oscuro della criminalità e sono, per loro natura, inconoscibili nella loro totalità), ma viene addestrato con i dati dei "crimini scoperti" e degli "arresti effettuati". Questa distinzione, apparentemente sottile, è il fulcro critico per comprendere la genesi del *runaway feedback loop* (circuito di retroazione fuori controllo), un fenomeno ampiamente documentato negli studi di Lum e Isaac e centrale nelle moderne discussioni sull'etica degli algoritmi.
Come ho sottolineato ripetutamente nei miei recenti interventi per CrimeLine Magazine e in saggi sulla devianza e la psicopatologia, la crisi generata dalle IA predittive non è confinabile unicamente al regno della statistica, della computer science o dell'ingegneria del software. Essa è profondamente radicata nella nostra neurobiologia.
Quando l'operatore di polizia, il magistrato o l'analista dell'intelligence interagiscono quotidianamente e continuamente con contenuti iper-realistici o con sistemi di supporto decisionale che offrono soluzioni perfette nella forma (anche se errate nella sostanza), innescano una patologia che nelle moderne scienze cognitive viene definita collasso epistemico. Non si tratta di un costrutto meramente filosofico, ma di un disallineamento neurologico documentabile empiricamente: stiamo progressivamente perdendo la capacità biologica di distinguere il vero dal falso, il dato oggettivo dall'inferenza probabilistica. Per comprendere l'origine e la meccanica di questo fenomeno di atrofia, dobbiamo guardare all'evoluzione umana e ai meccanismi di funzionamento basilari della nostra mente.
Il nostro cervello è progettato, per stringenti ragioni di sopravvivenza evolutiva, per massimizzare l'efficienza e consumare la minor quantità di energia metabolica possibile. Come dimostra decenni di letteratura scientifica sulla *Cognitive Load Theory* (Teoria del Carico Cognitivo), se si presenta una scorciatoia mentale (un'euristica) che permette di fare meno fatica a processare un'informazione complessa, la nostra mente la sceglie in modo quasi automatico. Storicamente, l'apprendimento profondo, la formulazione di un giudizio investigativo solido e la costruzione di schemi mentali robusti nella memoria a lungo termine

richiedono un faticoso sforzo deduttivo, noto in psicologia dell'apprendimento come frizione cognitiva pertinente (*germane cognitive load*). È quello sforzo critico, quella "lotta produttiva" e dialettica, che permette a un analista del comportamento o a un investigatore sul campo di soppesare indizi contraddittori, dubitare delle apparenze, rilevare anomalie e formulare ipotesi indipendenti e non condizionate.
Tuttavia, l'attuale paradigma commerciale e ingegneristico di sviluppo dell'IA si basa dogmaticamente sul cosiddetto *zero-friction design*. I modelli – dai *Large Language Models* (LLM) ai sistemi di cruscotto predittivo – sono progettati per essere massimamente collaborativi, eliminando ogni ostacolo o sforzo epistemico da parte dell'utente e consegnando risposte immediate, sintatticamente perfette e fluide. Così facendo, l'IA attiva costantemente e subdolamente il nostro "Sistema 1" (il pensiero rapido, intuitivo, emozionale e automatico teorizzato da Daniel Kahneman), bypassando sistematicamente il "Sistema 2" (il pensiero lento, analitico, razionale e deliberativo).
In un ecosistema giudiziario e investigativo dominato da sistemi di polizia predittiva che fungono da oracoli incontestabili, questa continua elusione della costruzione di schemi mentali provoca una misurabile "Atrofia Cognitiva Generazionale" (*Generational Cognitive Atrophy*). Nel momento in cui l'investigatore viene esposto costantemente all'algoritmo fluido, cessa di esercitare il pensiero divergente, perde la tolleranza all'ambiguità e delega interamente la propria agenzia intellettuale alla macchina.
Questo slittamento è stato il fulcro di diverse mie indagini. Come illustrato nella nostra ricerca presentata all'*Asian Conference on Communication and Networks* (AsiaComNet) e pubblicata con IEEE, è imperativo decostruire rigorosamente l'approccio antropomorfico che inquina il dibattito tecnologico odierno. In quello studio, abbiamo somministrato strumenti psicometrici standardizzati – come il BFQ-2 (Big Five Questionnaire-2) – ai principali modelli di intelligenza artificiale per dimostrare l'assurdità di attribuire "personalità" o "identità morale" a matrici di calcolo.
Considerare le macchine come pari cognitivi induce una pericolosa illusione di affidabilità. Allo stesso modo, come esplorato in un nostro lavoro pubblicato per Springer Nature, *"Silent Signals, Loud Consequences: Chemo-Communication, Ethics, and Digital Therapeutics"*, la continua esposizione a entità sintetiche ottimizzate per compiacere l'umano modifica profondamente l'etica relazionale e il giudizio (sia esso clinico o investigativo), ridefinendo i paradigmi dell'autonomia decisionale e inducendo l'utente in una compiacenza passiva (*sycophancy*).
Il collasso epistemico, a livello operativo e quotidiano, si manifesta attraverso il micidiale bias di automazione (*automation bias*). Si tratta della naturale propensione umana a utilizzare i suggerimenti dei sistemi automatizzati come un'euristica sostitutiva, rimpiazzando la ricerca vigile delle informazioni e il processamento analitico. Quando un sofisticato sistema algoritmico suggerisce, tramite una mappa di calore o un punteggio (score), che un individuo ha un'alta probabilità di recidiva criminale o che un incrocio sarà teatro di una sparatoria, l'operatore tende inesorabilmente a ignorare o sminuire eventuali prove contraddittorie o esimenti. Accetta la soluzione generata dal computer come inconfutabile, ammantandola di un'aura di infallibilità scientifica ("lo dice il computer").
Questo cortocircuito avviene a causa di un fenomeno che le neuroscienze definiscono *heuristic takeover* (subentro euristico): la mente, sempre

alla ricerca di strategie per risparmiare energia metabolica, accetta la plausibilità probabilistica ben confezionata scambiandola per certezza causale. Gli studi dimostrano che questo bias colpisce indistintamente novizi ed esperti di dominio, ed è estremamente refrattario ai tentativi di mitigazione tramite semplice addestramento o raccomandazioni alla cautela.
La letteratura scientifica più recente suggerisce persino l'emersione di un nuovo assetto cognitivo ibrido, definito "Sistema 0". Questo costrutto rappresenta la cognizione esternalizzata, in cui l'interazione uomo-IA si colloca temporalmente prima e al di sotto della deliberazione cosciente. Nel Sistema 0, l'intelligenza artificiale plasma le infrastrutture pre-attentive attraverso cui l'essere umano negozia la fiducia nella realtà, definendo i confini di ciò che è considerabile "fatto" ancor prima che l'umano inizi a ragionare. Se questo Sistema 0 viene addestrato su una monocoltura algoritmica opaca e intrisa dei bias storici del *predictive policing*, l'intero apparato della percezione umana del crimine risulta irreversibilmente e invisibilmente compromesso alla radice.
L'impiego massiccio di supporti algoritmici erode silenziosamente l'autorialità della decisione umana. Nelle aule di tribunale e nei centri operativi delle forze dell'ordine, l'algoritmo agisce *de facto* come un decisore ombra. Ma quando un modello predittivo commette un errore smisurato (per esempio segnalando ripetutamente un individuo innocente basandosi esclusivamente su correlazioni statistiche spurie legate al suo codice di avviamento postale o limitando l'accesso al credito) chi ne è legalmente, civilmente ed eticamente responsabile? L'analista dei dati? La corporation proprietaria del software? Il poliziotto che ha eseguito il controllo?
Per descrivere questa profonda lacuna di responsabilità (*accountability gap*), l'antropologa della tecnologia Madeleine Elish ha coniato un termine straordinariamente calzante e inquietante: la *Moral Crumple Zone* (la "zona di accartocciamento morale"). Nell'ingegneria automobilistica, le zone a deformazione programmata (*crumple zones*) sono strutture del telaio appositamente progettate per collassare, assorbendo l'energia dell'impatto frontale al fine di proteggere l'abitacolo e i passeggeri. Nei sistemi socio-tecnologici, l'operatore umano (il poliziotto, il giudice, il funzionario amministrativo) viene posizionato per agire esattamente come questa zona di accartocciamento. L'umano serve da ammortizzatore morale e legale per assorbire l'impatto degli errori, proteggendo l'integrità del sistema tecnologico e la reputazione della *corporation* sviluppatrice dalle conseguenze di un fallimento algoritmico.
Se l'algoritmo fallisce clamorosamente, la reazione di default dell'opinione pubblica, della stampa e persino della giurisprudenza è quella di incolpare l'operatore umano (*human-in-the-loop*) per non aver vigilato a sufficienza, per essere stato distratto o per essersi fidato ciecamente. Tuttavia, esigere che un individuo – quotidianamente sottoposto a un enorme carico di lavoro, stanchezza decisionale (*decision fatigue*), vincoli di tempo stringenti e al fortissimo *automation bias* precedentemente descritto – corregga in tempo reale i micro-errori occulti di un *black box* computazionale è un'aspettativa irrealistica e cognitivamente disonesta.
Si crea così un vuoto sistemico di responsabilità in cui l'ingiustizia sociale, già codificata nei dati storici e fedelmente riprodotta dall'IA, viene assorbita dall'operatore di prima linea senza mai essere corretta alla radice. Nella mia monografia *"Psicologia dell'animo oscuro. Morale, devianza, psicopatologia"* (Licosia Edizioni) e nei miei lavori

sull'analisi del rischio, ho sempre sostenuto che la genesi della devianza e la valutazione della moralità non possono essere slegate dal contesto sistemico. Allo stesso modo, il "cattivo comportamento" o l'errore dell'agente umano che utilizza l'IA non è un difetto isolato, ma il sintomo di un'architettura decisionale strutturalmente patologica che mira a deresponsabilizzare i progettisti.

Nell'epoca contemporanea, in cui i modelli di intelligenza artificiale assumono un'autorevolezza tecnica quasi ontologica, l'architettura fondamentale della giustizia penale, dell'analisi criminale e della sicurezza globale è posta di fronte a una crisi esistenziale senza precedenti. Le monocolture algoritmiche non si limitano a soffocare il pluralismo epistemico necessario per l'evoluzione scientifica e giuridica; esse agiscono come formidabili casse di risonanza che codificano e perpetuano, sotto l'egida di un'apparente oggettività matematica, i peggiori pregiudizi latenti della nostra storia sociale. I circuiti di retroazione che alimentano le piattaforme di polizia predittiva, scambiando fatalmente il risultato asimmetrico della propria azione coercitiva per una verità oggettiva del mondo, non riducono i tassi di criminalità alla radice, ma si limitano a confermare in un loop infinito il bersaglio che essi stessi hanno pre-disegnato.

In qualità di esperti del comportamento umano, criminologi e scienziati forensi, abbiamo l'urgente dovere etico e professionale di rigettare la deferenza cieca nei confronti del mero calcolo probabilistico. L'intelligenza artificiale, quando dispiegata nelle mani delle forze dell'ordine e della magistratura, non deve operare come un oracolo dogmatico. Deve essere un mero strumento ancillare: interrogabile, severamente contestabile e, soprattutto, strutturalmente e gerarchicamente subordinato all'intuizione clinica, all'empatia e alla capacità critica umana.

La soluzione a questa crisi non risiede nell'invocare una regressione tecnofobica o un neo-luddismo irrealizzabile, ma nella riappropriazione determinata della nostra complessa fisiologia del dubbio. Una valutazione del rischio rigorosa, moderna e costituzionalmente orientata impone, oggi più che mai, un innalzamento drastico della soglia di giustificazione epistemica per ogni decisione supportata dalle macchine. Per far sì che l'algoritmo non spenga definitivamente i nostri cervelli, atrofizzando il giudizio di intere generazioni di investigatori e magistrati, dobbiamo esigere e progettare tecnologie che si rifiutino di fornirci certezze rassicuranti e troppo veloci. Al contrario, esse devono costringerci costantemente alla fatica necessaria, salutare e profondamente umana di dubitare delle evidenze e analizzare criticamente il contesto.

FOCUS PSICOLOGICO

Bias: tendenza psicologica sistematica che orienta il modo in cui una persona seleziona, interpreta e valuta le informazioni, influenzando il giudizio e le decisioni anche in presenza di dati contrari.

Euristica: scorciatoia mentale che consente di arrivare rapidamente a un giudizio o a una decisione riducendo lo sforzo cognitivo.