Monocolture Algoritmiche e Bias Cognitivi

L'architettura della percezione criminale

L'introduzione di sistemi di intelligenza artificiale (IA) e di algoritmi predittivi nei gangli vitali del sistema di giustizia penale, rappresenta uno dei mutamenti paradigmatici più profondi nella storia della criminologia contemporanea e della psicologia forense. Tale transizione non riguarda meramente l'adozione di nuovi strumenti tecnici, ma configura una vera e propria ristrutturazione dell'architettura della percezione criminale, dove la valutazione del rischio e la diagnosi della pericolosità sociale, non scaturiscono più esclusivamente da una deliberazione umana mediata dall'esperienza clinica, bensì da processi di calcolo probabilistico che operano su vastissime moli di dati storici. Il tema delle "monocolture algoritmiche" emerge, in questo scenario, come una criticità di natura ecosistemica: la tendenza di molteplici attori (istituzioni giudiziarie, forze di polizia, enti correzionali) ad affidarsi a un ristretto numero di modelli o software dominanti crea una vulnerabilità sistemica, dove i bias intrinseci a tali modelli non rimangono confinati a un singolo caso, ma si propagano in modo virale attraverso l'intero apparato della giustizia.

Per comprendere la portata di questa trasformazione è necessario ridefinire il concetto di bias, oltre la sua accezione morale o discriminatoria. In ambito forense e computazionale il bias deve essere inteso come una deviazione sistematica da una norma rilevante, che può essere di natura epistemica (distorsione della verità) o morale (violazione di principi di equità). La letteratura recente evidenzia come queste due dimensioni siano indissolubilmente legate: un errore epistemico nella calibrazione di un algoritmo di rischio, ad esempio l'errata interpretazione di un dato socio-economico come predittore di recidiva, si traduce immediatamente in un danno morale e giuridico per l'individuo sottoposto a giudizio.

Il ricorso a strumenti evidence-based nasce dall'esigenza di superare i limiti intrinseci dell'intuizione umana, spesso tacciata di essere incostante, non misurabile e soggetta a pregiudizi inconsci. Tuttavia, l'algoritmo non elimina il bias, ma lo sposta in una fase precedente della filiera decisionale: quella della progettazione e dell'addestramento. La giustificazione epistemica delle decisioni basate su evidenze algoritmiche

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. (MEDITERRANIA) con sede a Roma.

diventa quindi particolarmente dipendente da considerazioni morali, poiché le ingiustizie sociali codificate nei dati storici tendono a essere riprodotte dal sistema come verità oggettive. Ciò eleva la soglia di giustificazione necessaria per accettare un output algoritmico come base per una limitazione della libertà personale.

Il concetto di monocoltura algoritmica descrive il fenomeno per cui una vasta gamma di decisori indipendenti finisce per adottare lo stesso algoritmo o modello di base. Nella giustizia penale, questo si traduce nell'uso di pochi software proprietari (come COMPAS negli Stati Uniti o sperimentazioni basate su logiche simili in Europa) per finalità che vanno dal *sentencing* alla gestione della libertà vigilata.

Quando un unico modello domina il mercato o la prassi istituzionale, l'intero sistema giudiziario diventa suscettibile a fallimenti correlati. Se l'algoritmo contiene un difetto logico o un bias di campionamento, questo errore non verrà corretto dalla diversità di giudizio degli attori umani, ma verrà replicato in ogni fase del processo penale. Tale omogeneizzazione riduce drasticamente la "diversità analitica" necessaria per un sistema di giustizia resiliente. In un ambiente dominato da una monocoltura, un individuo etichettato erroneamente come "ad alto rischio" da un sistema di polizia predittiva incontrerà lo stesso verdetto algoritmico davanti a un giudice per la cauzione e, successivamente, davanti a una commissione per la libertà condizionale, poiché tutti gli attori attingono alla stessa "verità" computazionale o utilizzano strumenti che condividono la medesima architettura logica. La monocoltura algoritmica è spesso alimentata da dinamiche di mercato dove pochi *vendor* dominano il settore della tecnologia forense. La mancanza di trasparenza (il problema della "*black box*") e la protezione del segreto commerciale impediscono un controllo indipendente e scoraggiano lo sviluppo di modelli alternativi più equi. In questo contesto, l'innovazione non mira necessariamente a una maggiore accuratezza clinica, ma a una maggiore efficienza operativa, spesso a scapito della tutela dei diritti fondamentali e della comprensione profonda delle dinamiche criminali individuali.

L'algoritmo non si limita a osservare la realtà criminale; esso contribuisce attivamente a modellarla attraverso complessi circuiti di retroazione (*feedback loop*). Questa interazione tra previsione e azione istituzionale crea quella che in psicologia forense è definita una "profezia che si autoavvera", dove l'output del sistema non predice il futuro, ma lo determina. I sistemi di polizia predittiva di tipo *place-based* analizzano la frequenza storica dei reati in aree specifiche per allocare le pattuglie. Tuttavia, il meccanismo produce un loop informativo distorsivo: inviare più agenti in un quartiere identificato come "*hot spot*" basandosi su dati passati, porta inevitabilmente a un maggior numero di arresti e segnalazioni in quell'area, non necessariamente per un aumento della criminalità, ma per un aumento della sorveglianza. Questi nuovi dati vengono poi reimmessi nell'algoritmo, che confermerà la pericolosità dell'area con ancora maggiore certezza statistica. Il risultato è una sovra-rappresentazione della criminalità in certi quartieri e una sotto-rappresentazione in altri, basata non su fatti oggettivi, ma sulla circolarità del dato operativo.

Nei sistemi *person-based*, l'architettura della percezione si sposta sul profilo del singolo. L'uso di software per identificare potenziali autori di reato (come la *Strategic Subject List* di Chicago) crea uno status di "sorvegliato speciale" che condiziona l'intera vita del soggetto. Ogni minima interazione con le forze dell'ordine viene registrata e interpretata dal sistema come una conferma della pericolosità, alimentando un *feedback loop* che rende

quasi impossibile per l'individuo uscire dai radar del sospetto algoritmico. Questo meccanismo è particolarmente insidioso poiché trasforma indizi comportamentali generici o legami sociali in prove matematiche di una propensione al crimine, con effetti stigmatizzanti che possono superare quelli di una condanna formale.

Nonostante l'automazione, il processo decisionale rimane, almeno formalmente, in mano all'uomo. Tuttavia, l'interazione tra l'esperto (giudice, psicologo forense, agente) e l'output dell'IA è mediata da bias cognitivi che ne alterano profondamente l'autonomia di giudizio. Il bias di automazione descrive la tendenza umana a favorire i suggerimenti di un sistema automatizzato, anche quando questi contraddicono l'evidenza dei fatti o il proprio giudizio professionale. In ambito forense, questo si manifesta come una "*blind deference*" (deferenza cieca) verso il *risk score*. Lo psicologo o il magistrato, gravati da carichi di lavoro eccessivi e attratti dalla parvenza di oggettività dei numeri, possono smettere di esercitare uno scrutinio critico sulle variabili sottostanti, accettando il risultato algoritmico come una verità definitiva. Tale compiacenza riduce la vigilanza e porta a ignorare segnali di errore o di confabulazione del sistema (le cosiddette "allucinazioni" dei modelli generativi). L'output dell'IA funge spesso da "ancora" cognitiva: una volta visualizzato un punteggio di rischio elevato, ogni successiva valutazione clinica dell'individuo tenderà a essere influenzata da quel dato iniziale. Il bias di conferma spinge l'operatore a cercare nel comportamento del soggetto quegli elementi che supportano la previsione dell'algoritmo, trascurando invece i fattori di resilienza o i progressi riabilitativi che la smentirebbero. Questo meccanismo è particolarmente pericoloso nelle perizie psichiatriche o criminologiche, dove l'interpretazione di tratti di personalità ambigui può essere facilmente orientata dal "pregiudizio numerico" fornito dalla macchina.

L'uso di interfacce basate su Large Language Models (LLM), che interagiscono attraverso il linguaggio naturale, o di sistemi che simulano il ragionamento umano, induce negli utenti un bias antropomorfico. Ciò li porta ad attribuire al sistema capacità di giudizio morale e un'affidabilità che esso, intrinsecamente, non possiede. Questo conferisce all'algoritmo un'autorità percepita superiore a quella di un tradizionale strumento statistico, rendendo ancora più difficile per l'esperto contestarne i risultati. La qualità di un sistema di IA è indissolubilmente legata ai dati su cui è addestrato. In criminologia, il concetto di "*ground truth*" (verità di base) è problematico, poiché ciò che chiamiamo "crimine" nei database è spesso il risultato di processi selettivi di polizia e magistratura.

Il fenomeno del bias non deve essere inteso come un errore isolato, bensì come un elemento insidioso capace di infiltrarsi in ogni singola fase del ciclo di vita di un sistema intelligente. Questo processo di distorsione inizia ben prima della scrittura del codice, partendo dalla formulazione del problema. Quando decidiamo cosa un algoritmo debba predire, stiamo già compiendo una scelta politica: ad esempio, utilizzare l'"arresto entro due anni" come metrica di successo riflette un pregiudizio istituzionale, poiché l'arresto non dipende esclusivamente dal comportamento del reo, ma è influenzato in modo determinante dalle strategie e dalla condotta delle forze dell'ordine. Il rischio di contaminazione prosegue nella fase di etichettatura (*labeling*). Gran parte dei modelli attuali viene addestrata basandosi su decisioni storiche prese da esperti umani. Se questi periti, nel passato, hanno agito sotto l'influenza di pregiudizi razziali o di genere, l'algoritmo non farà altro che apprendere e sistematizzare tali distorsioni. In questo modo, il pregiudizio umano viene "codificato", trasformandolo

paradossalmente in una sorta di legge di natura indiscutibile. Infine, la distorsione si consolida durante la selezione dei dati. Spesso vengono utilizzate variabili apparentemente neutre (come il livello di istruzione, la situazione occupazionale o la stabilità del nucleo familiare) come "*proxy*" (indicatori indiretti) della pericolosità sociale. L'effetto concreto è una penalizzazione sistematica delle classi sociali più svantaggiate: la marginalità economica e sociale viene così trasformata in un fattore di rischio algoritmico, creando un circolo vizioso che rinforza le disuguaglianze esistenti.

L'integrazione degli algoritmi solleva questioni filosofiche e legali senza precedenti riguardo alla responsabilità delle decisioni. Il fenomeno dell'*outsourcing* dei giudizi di valore alle macchine crea quello che viene definito un "*attributability gap*" (gap di attribuibilità). Sebbene il giudice umano rimanga il responsabile formale della sentenza, l'uso massiccio di supporti algoritmici erode la sua "*authorship*" (autorialità) sulla decisione. Se il magistrato adotta un suggerimento dell'IA senza comprenderne appieno la logica sottostante, la decisione finale non è più il frutto di una deliberazione morale autonoma. Questa lacuna di responsabilità è aggravata dal fatto che gli algoritmi non possono essere chiamati a rispondere delle proprie scelte, né possiedono la capacità di comprendere il peso morale delle conseguenze prodotte.

Secondo la teoria del *moral encroachment*, la giustificazione epistemica di una credenza dipende non solo dall'evidenza statistica, ma anche dalla posta in gioco morale. In ambito forense, credere che un soggetto sia "pericoloso" sulla base di un output algoritmico richiede una soglia di prova molto più alta rispetto a una previsione di mercato, a causa dell'impatto devastante che un errore avrebbe sulla vita e sulla libertà dell'individuo. La presenza di bias noti nei sistemi di IA rende la giustificazione di tali decisioni intrinsecamente problematica, imponendo un dovere di scetticismo attivo da parte dell'utilizzatore.

L’algoritmo non decide soltanto chi osservare: finisce per ridefinire ciò che una società sceglie di considerare pericoloso.