Bias, giustizia predittiva e autorevolezza della giurisdizione

Implicit Bias, Predictive Justice and the Authority of the Judiciary

MARCO **BRIGAGLIA**
Professore ordinario di Filosofia del diritto, Università degli Studi di Palermo.
E-mail: marco.brigaglia@unipa.it

ABSTRACT
Negli ultimi decenni, lo studio psicologico della decisione giudiziale ha messo in evidenza, con un livello di dettaglio e di attendibilità mai raggiunto prima, quanto forte e pervasiva sia la presenza di *bias* e rumore. L'interazione di questi studi con il il contemporaneo sviluppo delle tecniche di giustizia predittiva – applicazione di sistemi di *machine learning* a *big data* giudiziari – rischia di produrre un pesante impatto negativo sull'autorevolezza della giurisdizione, sia attraverso il confronto con macchine che promettono decisioni di migliore qualità, sia attraverso la facilitazione – soprattutto in presenza di politiche di *open data* giudiziari – della rilevazione e diffusione al grande pubblico di pattern decisionali sospetti. Questo articolo mira a presentare il problema ad un pubblico di giuristi, anche attraverso la ricostruzione delle vicende, davvero esemplari, della politica francese di open data giudiziari e del sito *Supralegem*.

In recent decades, the psychological inquiry into judicial decision-making has highlighted, with unprecedented detail and reliability, how deeply and pervasively it is affected by bias and noise. The interaction of these studies with the contemporary development of predictive justice techniques—the application of machine learning systems to judicial big data—risks having a serious negative impact on the authority of the judiciary, both through comparison with machines that promise better quality decisions and through the facilitation—especially in the presence of open data policies—of the detection and dissemination to the general public of suspicious decision-making patterns. In this article, I present the problem to an audience of legal scholars, with a reference to the exemplary events surrounding French *open data* policy and the *Supralegem* website.

KEYWORDS
bias impliciti e decisione giudiziale, autorevolezza della giurisdizione, giustizia predittiva

implicit bias in adjudication, authority of the judiciary, predictive justice

ISSN 1825-0173

Bias, giustizia predittiva e autorevolezza della giurisdizione

MARCO BRIGAGLIA

1. *Il problema*

Sin dalla loro fase di gestazione, negli anni 50 del secolo scorso, le scienze cognitive sono state caratterizzate da una strettissima interazione fra studio delle menti naturali e progetti di intelligenza artificiale, con un effetto di rinforzo reciproco: da un lato, la migliore comprensione dei meccanismi cognitivi naturali ha guidato la costruzione di nuovi sistemi di IA, e, d'altro lato, i sistemi di IA hanno guidato e supportato, in diversi modi, lo studio delle menti naturali.

L'interazione tra studio della mente naturale e tecnica della mente artificiale ha un ruolo decisivo anche nel campo della giustizia. Un aspetto problematico di questa interazione è il rischio di una profonda crisi di immagine e di autorevolezza della giurisdizione, presa nella morsa, da un lato, di studi psicologici sui fattori di distorsione del giudizio e, d'altro lato, dalle tecniche note come 'giustizia predittiva', ossia l'applicazione di sistemi di *machine learning* a *big data* giudiziari.

Il problema, in sintesi, è questo.

L'immagine della decisione giudiziale che tende ad emergere dagli studi cognitivi degli ultimi decenni non è un'immagine confortante: semplificando molto, la decisione giudiziale appare attraversata in modo pervasivo e profondo da errori, distorsioni e ingiustificate incoerenze. Nei termini resi famosi da Kahneman, la decisione giudiziale appare, nell'insieme, gravemente affetta da *bias* e da 'rumore'.

Si potrebbe dire: nulla di nuovo sotto il sole. L'insistenza sulla scarsa coerenza e sulla alta fallibilità e irrazionalità della giurisdizione è un tema tradizionale della teoria del ragionamento giuridico – è la cifra caratteristica di diversi approcci 'realisti' o 'scettici'.

Questo è senz'altro vero, ma è anche molto parziale, per due ragioni. La prima ragione è che il livello di dettaglio e di prova empirica degli errori e distorsioni che attraversano la giurisdizione raggiunto dagli studi cognitivi degli ultimi decenni è senza precedenti. Sappiamo molto più di prima, e in modo molto più affidabile di prima, come e quanto la decisione giudiziale sia affetta da *bias* e rumore. La seconda ragione è che le tecniche di giustizia predittiva, sotto moltissimi aspetti, consentono un ulteriore, drastico salto in avanti di questa conoscenza. L'applicazione di algoritmi capaci di estrarre pattern statistici da *big data* giudiziari consente infatti di rendere visibile, ad un grado prima impossibile, quanto la decisione giudiziale *reale* – non quella resa, sia pure da esperti, riguardo a casi fittizi in laboratorio – sia permeata da *bias* e rumore. E può consentire, sotto certe condizioni di accessibilità dei dati, di rendere visibili pattern sospetti esibiti da decisioni di *specifiche* corti. Infine, la diffusione di tecniche di giustizia predittiva, se coniugata con politiche di *open data* giudiziari, può diffondere la conoscenza di pattern (più o meno fondatamente) sospetti ben al di là di circoli ristretti di specialisti –

* Questo articolo rielabora il testo della presentazione fatta al Convegno "Intelligenza Artificiale e processo: dalla prevedibilità delle decisioni alla giustizia predittiva", svoltosi a Palermo il 21 aprile 2023, con la partecipazione di Ferruccio Auletta, Enrico Camilleri, Paolo Comoglio, Frédérique Ferrand, Federico Russo, Guido Smorto. Ringrazio i partecipanti al Convegno, e in particolare Ferruccio Auletta, per i loro commenti. Ringrazio Armando Plaia, Massimo Starita, Paolo Capriati e un anonimo referee per i suggerimenti.

non nell'innocuo ambito della discussione accademica di teoria del diritto o psicologia della decisione, ma nel pubblico generale.

Tutto questo insieme di processi costituisce una seria minaccia per l'autorevolezza della giurisdizione – per la giurisdizione in genere, o anche per la giurisdizione di specifiche corti. Si tratta, a mio giudizio, di uno dei problemi più seri riguardo agli impatti dell'IA sulla giurisdizione, un problema che investe scenari già attuali e non prospettive futuribili – ma forse non impossibili – di giudizio algoritmico. Quanto il problema sia serio e attuale è evidenziato molto bene dalle vicende, risalenti a pochi anni fa, della politica francese di *open data* giudiziari, e in particolare dal caso del sito *Supralegem*.

Obiettivo di questo articolo è ricostruire i termini del problema per un pubblico di giuristi. Nel § 2 ricostruirò, per sommi capi, lo scenario che emerge dagli studi sul ruolo di *bias* e rumore nel giudizio e nella decisione umana in genere, e nell'ambito della giurisdizione in particolare. Mi concentrerò, in particolare, sui *bias* che rendono il giudizio suscettibile a fattori non soltanto irrilevanti, ma anche casuali e non ancorati a convinzioni profonde dei decisori. Si tratta di un ambito di studi ormai molto vasto, ma non entrato ancora a far parte del senso comune di giuristi e teorici del diritto e mi sembra perciò utile riportare alcuni degli studi che meglio hanno contribuito a far percepire la portata del problema. Nel § 3, utilizzando come caso studio le vicende della politica francese di *open data* giudiziari e del sito *Supralegem*, spiegherò perché le tecniche di giustizia predittiva possono funzionare come un detonatore che rende esplosivo il problema dei difetti cognitivi della giurisdizione. Nel § 4, avanzerò alcune considerazioni conclusive riguardo a quale sia l'atteggiamento più sensato e più accettabile dal punto di vista normativo da assumere rispetto al problema.

2. *La giurisdizione tra* bias *e rumore*

2.1. A partire dalla metà del secolo scorso, gli studi sulle distorsioni della cognizione umana hanno conosciuto una straordinaria fortuna, dando vita a diversi paradigmi di ricerca. Il più noto è quello lanciato a metà degli anni 70 da Kahneman e Tverski sotto il nome di 'euristiche e bias', parte di un più generale approccio noto come '*behavioral economics*', lo studio dei processi decisionali umani focalizzato sulle deviazioni da modelli di razionalità. L'eco di queste ricerche non ha tardato a farsi sentire anche nell'ambito della decisione giuridica, con la nascita della '*behavioral law and economics*' (v. SUNSTEIN et al. 1998; per una panoramica generale v. ZAMIR, TEICHMAN 2014) e in particolare lo sviluppo di una sempre più nutrita serie di studi volti ad applicare il paradigma delle euristiche e *bias* alla decisione giudiziale. Gli esempi che proporrò sono tratti da studi che seguono questo approccio. Ma è importante ricordare che si tratta soltanto di uno fra molti filoni di studio dei fattori di distorsione del giudizio e della decisione. Ricerche di psicologia sociale su *bias* impliciti, pregiudizi e stereotipi, ad esempio, per quanto in parte sovrapponibili a quelle su euristiche e *bias* (l'euristica della rappresentatività, ad esempio, è una procedura di giudizio guidato da stereotipi), hanno seguito strade largamente indipendenti, e hanno avuto un impatto forse ancora maggiore sulla teoria e sulla pratica della decisione giudiziale (per uno sguardo generale su questi approcci v. ad es. LANE et al. 2007; NELSON 2009).

Prima di procedere con gli esempi, è utile però mettere a fuoco alcuni aspetti generali.

2.2. L'elemento più disturbante degli studi sulla cognizione umana, e quello più interessante ai nostri fini, non è tanto la scoperta di sistematiche distorsioni del giudizio rispetto a standard di razionalità *postulati dall'osservatore*, ma di distorsioni rispetto ai criteri che il decisore stesso accetta e crede di avere seguito – la scoperta, cioè, di quanto le nostre decisioni siano guidate causalmente, in modo del tutto inconscio, da fattori diversi da quelli che riconosceremmo come ragioni. La nostra mente è non soltanto 'irrazionale' — nel senso che devia da criteri di razionalità comunemen-

te accettati – ma è anche 'fragile' – sfugge al controllo razionale che il decisore vorrebbe darsi – e 'opaca' – operante secondo processi che sfuggono alla coscienza. Nel senso più lato, il termine 'bias', o più precisamente 'bias *impliciti*', indica meccanismi inconsci che tendono a (1) far deviare il giudizio dagli standard normativi accettati dal decisore, minando il suo controllo razionale, e che (b) lo fanno in modo sistematico, regolare, e dunque parzialmente prevedibile. (Il termine '*bias*' ha anche un senso più ristretto, inerente all'approccio delle euristiche e bias: un errore sistematico risultante dall'uso di euristiche, scorciatoie cognitive. Su questo tornerò più avanti).

L'idea che la cognizione umana sia soggetta a *bias* non è affatto estranea al senso comune. Si pensi a quanto sia diffusa la convinzione che il giudizio possa essere inconsciamente distorto da 'simpatie e antipatie', o all'ampio riconoscimento, quanto meno astratto, della possibilità di pregiudizi di razza, di genere, di classe. Si pensi, ancora, a quanto sia parte del senso comune la consapevolezza della tendenza a saltare alle conclusioni prima di considerare le ragioni rilevanti, o a credere alle cose che ci fanno piacere. Al senso comune non sono nemmeno estranee strategie di controllo cognitivo volte a regolare queste tendenze ("prima di agire, pensaci due volte"; "guarda al problema dalla prospettiva di tutte le persone coinvolte", e simili). Si pensi all'influenza di simpatie o antipatie personali in giudizi in cui esse non dovrebbero rilevare. Un decisore, soprattutto se esperto (ad esempio, un docente universitario in una sessione di esami), sa di essere esposto a questo tipo di *bias*, e questa consapevolezza gli consente – se si ha la buona volontà di farlo – di attivare familiari strategie di controllo: un monitoraggio delle proprie emozioni avverte del rischio di incorrere in *bias*, e l'attivazione di strategie intuitive di regolazione emotiva può contrastare, almeno in una certa misura, gli effetti del *bias*.

In sintesi, la riflessione sui *bias* e sulla loro gestione è parte integrante dell'idea comune di razionalità e dei 'thinking tools' ad essa inerenti. Ma la conoscenza sui *bias* familiare al senso comune non è che un piccolo frammento di quella sviluppata dagli studi psicologici degli ultimi cinquant'anni. Il loro esito non è stato soltanto quello di raffinare e arricchire il modello comune della razionalità umana e dei suoi limiti, ma di stravolgerlo profondamente. Anzitutto, mostrando come il giudizio e la decisione umana – anche quelli di esperti la cui legittimità deriva proprio dalla loro *expertise* – siano molto più sensibili a fattori del tutto irrilevanti e caotici di quanto il senso comune sia disposto ad ammettere. Poi, individuando, dietro questi fattori, tipi di *bias* ignoti al (o appena notati dal) senso comune, e rispetto ai quali le strategie di controllo cognitivo di quest'ultimo sono del tutto impotenti. In terzo luogo, verificando l'effettiva efficacia di strategie di controllo cognitivo facilmente accessibili, perfezionandole, e sviluppando strategie alternative[1].

Un esempio illuminante è offerto da uno studio riportato in un influente articolo di NISBETT e WILSON (1977). Lo studio fu condotto in un esercizio commerciale, sotto forma di sondaggio tra clienti di passaggio. Ai soggetti furono presentati quattro identiche paia di calze di nylon, chiedendogli di sceglierne uno e di esporre le ragioni della propria scelta. Le scelte manifestavano un forte effetto di posizione a favore delle calze posizionate più a destra (preferite con pro-

[1] Uno studio sul *bias* razziale implicito (RACHLINSKI et al. 2009), ad esempio, ha concluso che i giudici sono effettivamente esposti – come tutti – a *bias* razziali impliciti, ma che la consapevolezza di ciò è sufficiente ad attivare spontanee strategie di regolazione che permettono di contrastare efficacemente il *bias* (purché, ovviamente, si sia motivati a farlo). Condizione di successo, in ogni caso, è un'adeguata informazione sulla presenza di *bias* impliciti, che non è parte del senso comune dei giuristi – il che suggerisce come l'esposizione a questa informazione possa costituire, di per sé, una tecnica non trascurabile di *debiasing*. Un altro esempio, sempre relativo al caso dei *bias* razziali impliciti, riguarda la valutazione dell'efficacia di una strategia di controllo semplice, come l'esposizione a controesempi positivi per contrastare l'effetto di pregiudizi negativi. Se, da un lato, una strategia del genere è tutt'altro che sorprendente, e fa parte dell'insieme di 'thinking tools' parte della competenza professionale di alcune categorie di decisori (quanto meno, di quelli più scrupolosi), d'altra parte la sua efficacia non può essere data per scontata. Ad esempio, DASGUPTA e GREENWALD (2001) avevano rilevato l'efficacia di questa tecnica riguardo ai *bias* razziali, mentre uno studio successivo (JOY-GABA, NOSEK 2010), pur confermandola, ne ha ridotto di molto l'effetto.

porzione di quattro a uno). Nessuno dei soggetti intervistati menzionò la posizione tra le ragioni della scelta, e tutti negarono vivacemente che potesse aver avuto qualsiasi rilievo per la loro scelta (molti, precisano gli autori, lanciando uno sguardo turbato all'intervistatore, lasciando intendere di considerare la domanda come un segno di squilibrio mentale). E in effetti, in un caso come questo, la posizione non conta in alcun modo, per nessuno di noi, come una ragione che possa giustificare la scelta. Si tratta di un fattore del tutto trascurabile, irrilevante. Eppure, dato che le calze erano assolutamente identiche, la distribuzione statistica delle scelte suggerisce che il fattore causalmente più rilevante sia stato, e di gran lunga, proprio la posizione. Le ragioni addotte dai clienti a sostegno della loro scelta appaiono invece come mere razionalizzazioni *ex post* – ragioni che fanno apparire la scelta come giustificata sulla base degli standard normativi che il decisore accetta e si aspetta che gli altri accettino ma che, pur se addotte in buona fede, non soltanto risultano erronee, infondate (le calze sono identiche), ma sembrano non aver giocato alcun ruolo causale decisivo nella scelta.

Questo studio è un celebre esempio di quanto la mente possa essere fragile e opaca (esposta a fattori irrilevanti e caotici che minano il controllo razionale in modo del tutto inconscio). Ma c'è un altro aspetto importante, meno evidente. Che la scelta non sia guidata, nella maggior parte dei casi, dall'erroneo riscontro degli aspetti poi addotti come ragioni o di altro aspetto comunque rilevante (come, ad esempio, l'erronea convinzione che il paio di calze scelto fosse quello con la sfumatura di colore più gradita) non è qualcosa che possa emergere dal singolo caso. Emerge soltanto dalla distribuzione statistica delle scelte, che mostra una correlazione con la posizione dell'articolo scelto così significativa da far concludere per l'esistenza di una relazione causale: fino a prova contraria, vi sono ragioni per credere, contro ogni intuizione, che la posizione abbia orientato causalmente (*bias*) la scelta.

La letteratura specialistica di psicologia giudiziale sui *bias* impliciti e il loro ruolo nella giurisdizione è ormai molto estesa, e sono disponibili diverse review di ottima qualità che offrono panoramiche generali sui *bias* più frequenti e sul loro impatto. In questa sede, mi limiterò a riportare due soli esempi di studi sui *bias*. Il mio obiettivo è quello di trasmettere ad un pubblico di giuristi alcuni aspetti che rischiano di passare inosservati, sulle quali mi sembra invece importante richiamare l'attenzione del lettore. I *bias* su cui questi studi vertono, infatti, esemplificano in modo molto chiaro le caratteristiche esibite dal su citato *bias* di posizione: (1) sono ignoti o quasi al senso comune; (2) espongono il giudizio all'influenza di fattori completamente irrilevanti e non rispondenti a nessun tratto strutturale della personalità e struttura cognitiva del decisore (si tratta, in particolare, di fattori del tutto indipendenti dai suoi interessi, dalla sua visione del mondo, dal suo orientamento affettivo); (3) emergono soltanto a fronte di analisi statistica di una pluralità di decisioni; (4) in virtù di (1), (2) e (3), sono difficilmente rilevabili dal decisore nel contesto della singola decisione e tendono a sfuggire alle ordinarie e spontanee strategie di controllo cognitivo.

È opportuno precisare che i *bias* che espongono il giudizio all'influenza di fattori irrilevanti e non rispondenti a tratti strutturali della personalità e struttura cognitiva del decisore sono solo parte delle distorsioni cognitive rilevanti nel campo della giurisdizione, e nemmeno la parte più odiosa – costituita, ovviamente, da *bias* relativi a razza, genere, o classe, che tendono a reiterare le disuguaglianze strutturali iscritte nel contesto sociale.

Se ho scelto di concentrarmi su *bias* del primo tipo è perché sono meno familiari a un pubblico di giuristi, e perché, proprio in virtù dell'irrilevanza dei fattori a cui rispondono, permettono di cogliere immediatamente e facilmente un punto cruciale: che per sfuggire ai *bias* non sono sufficienti buona volontà, buone intenzioni, o buone disposizioni. Non che buona volontà, buone intenzioni e buone disposizioni siano sempre sufficienti nel caso di *bias* relativi a razza, genere, o classe (o ad altre linee di discriminazione sociale) – anche questi *bias* sono spesso inconsci ('impliciti'), non possono essere riconosciuti direttamente da una introspezione volenterosa ma emergono solo grazie all'analisi di pattern statistici, e non sempre possono essere regolati in mo-

do spontaneo grazie alla volontà di farlo[2]. Semplicemente, scoprire quanto la nostra cognizione sia fragile e opaca rispetto a fattori di distorsione irrilevanti e casuali può meglio predisporci ad ammettere quanto possa essere fragile e opaca anche rispetto a fattori di distorsione profondamente radicati nel contesto culturale di appartenenza.

2.3. Il primo esempio è offerto da uno studio molto noto, condotto da un gruppo di psicologi in Israele (DANZIGER et al. 2011a). Oggetto dello studio era la relazione tra le decisioni dei giudici partecipanti allo studio e il loro pasto – più precisamente, l'influenza sulla decisione del tempo trascorso dall'ultima pausa pasto. Sullo sfondo di questo studio c'è l'eco del realismo giuridico americano, e in particolare di Jerome Franck, che enfatizzava l'influenza che i più svariati fattori psicologici hanno sul ragionamento e sulla decisione giudiziale, compresi fattori palesemente irrilevanti. I critici del realismo giuridico americano hanno caricaturizzato questa posizione attribuendogli la massima secondo cui "*justice is what the judge ate for breakfast*" (v. ad es. KOZINSKI 1993; sull'attribuzione della massima al realismo giuridico americano v. PRIEL 2020). Gli autori dello studio hanno preso sul serio questa massima apparentemente assurda, trattandola come un'ipotesi empirica, da verificare sperimentalmente in condizioni di controllo.

Lo studio ha esaminato oltre mille sentenze emesse da otto giudici che presiedevano due diverse commissioni competenti per la concessione del beneficio della libertà vigilata (*parole*), durante giornate lavorative suddivise in tre sessioni, separate da due pause per i pasti – una merenda in tarda mattinata e il pranzo.

I ricercatori hanno scoperto che la percentuale di sentenze favorevoli raggiungeva il suo picco all'inizio di ogni sessione, per poi scendere gradualmente fino a quasi zero fino alla pausa pasto successiva, dopodiché cominciava a risalire. In breve: più i giudici erano affamati (e stanchi), più severe erano le loro decisioni.

Sebbene il tempo trascorso dall'inizio della sessione non fosse l'unico fattore statisticamente correlato alla sentenza, appariva come il fattore maggiormente predittivo dell'esito della decisione, suggerendo una profondamente disturbante relazione causale tra le due variabili: un fattore giuridicamente irrilevante come il momento in cui il caso è stato deciso, lo studio suggeriva, sembra aver influenzato in modo significativo l'esito della decisione. In definitiva, la caricatura del realismo giuridico americano elaborata dai suoi detrattori non era poi così assurda...

I ricercatori non si sono limitati a rilevare e registrare una correlazione tra la collocazione temporale della sentenza e il suo disposto – la concessione o il diniego del beneficio della libertà vigilata. Hanno anche avanzato un'ipotesi sul meccanismo causale responsabile di questo effetto. L'ipotesi è grossomodo questa. Il processo decisionale in condizioni di incertezza è un'attività cognitivamente impegnativa, che richiede sforzo di attenzione. Le decisioni che modificano lo *status quo* (in questo caso, la concessione della libertà vigilata) richiedono uno sforzo cognitivo maggiore rispetto alle decisioni che confermano lo *status quo* (il diniego della libertà vigilata). Questa ipotesi sarebbe supportata dal fatto che, in media, le decisioni favorevoli avevano richiesto più tempo di quelle sfavorevoli ed erano supportate da argomentazioni più lunghe. Gli esseri umani hanno però risorse cognitive limitate, che vengono esaurite da compiti cognitivamente impegnativi, come il processo decisionale in condizioni di incertezza, in particolare nel caso di decisioni che modificano lo *status quo*. Nel corso delle sessioni di lavoro, il progressivo esaurimento delle risorse cognitive dei giudici si tradurrebbe così in una facilitazione della decisione meno impegnativa, il diniego della libertà vigilata. Durante le pause, cibo e riposo ripristinerebbero le risorse cognitive dei giudici, riportando ad un livello sufficiente la loro capacità a prendere decisioni cognitivamente più impegnative.

[2] Inoltre, come rilevano RACHLINSKI e WISTRICH (2017, 107), alcuni *bias* sociali impliciti legati a caratteristiche sottili (come attrattività fisica, obesità, età) sono molto meno salienti dei *bias* di razza e genere, e rischiano perciò di essere più difficilmente notati.

Alla base della distribuzione temporale delle decisioni, vi sarebbe dunque l'interazione di due meccanismi. Il primo meccanismo è il progressivo esaurimento, durante le sessioni di lavoro, delle risorse cognitive necessarie per prendere decisioni in condizioni di incertezza, e il loro recupero grazie alla pausa pasto. Questo meccanismo è chiamato 'ego-depletion'. Il secondo meccanismo è la tendenza a confermare e sostenere lo *status quo* – una tendenza che può essere superata solo in virtù di uno sforzo cognitivo supplementare. Questa tendenza è una *euristica*: un modo semplificato di risolvere un problema, una scorciatoia cognitiva che, da un lato, ci permette di raggiungere, in una percentuale significativa di occasioni, una soluzione sufficientemente buona, ma, dall'altro, ci espone al rischio della commissione di errori sistematici.[3]

Il *bias* individuato da questo studio soddisfa il pattern indicato sopra. Il giudizio è influenzato da fattori del tutto irrilevanti, che il decisore stesso non riconoscerebbe né come ragioni giuridicamente ammissibili, né come ragioni di altro tipo. L'effetto emerge solo a fronte di analisi statistica di un numero ampio di decisioni. Il decisore non è in grado né di notare l'effetto né di regolarlo direttamente attraverso ordinarie strategie di controllo cognitivo.

2.4. Il secondo esempio è offerto dalla cosiddetta 'euristica dell'ancoraggio' (*anchoring*). L'euristica dell'ancoraggio è stata descritta da TVERSKY e KAHNEMAN (1974) come la procedura che consiste nel fare stime "partendo da un valore iniziale che viene aggiustato per ottenere la risposta finale". Il valore iniziale può essere suggerito da informazioni che sono, e sappiamo essere, irrilevanti per la soluzione del problema. In un esperimento cda loro condotto, era stato chiesto ai soggetti di stimare una quantità (ad esempio, la percentuale di Stati africani tra i membri delle Nazioni Unite). Prima che i soggetti rispondessero, veniva fatta girare in loro presenza una ruota della fortuna, con numeri da 1 a 100. Ai soggetti veniva prima chiesto di stimare se la quantità corretta fosse superiore o inferiore al numero indicato dalla ruota – un numero palesemente irrilevante rispetto al problema in questione. Dopo, veniva richiesto ai soggetti di stimare la quantità richiesta. I soggetti erano divisi in gruppi, e a gruppi diversi venivano assegnati, per ciascuna quantità da stimare, numeri diversi. Questi numeri avevano effetti significativi sulla stima della quantità. Ad esempio, il valore medio della stima della percentuale degli Stati africani tra i membri delle Nazioni Unite era del 25% per soggetti esposti al numero 10, e del 45% per cento per soggetti esposti al numero 65. Un numero basso aveva ancorato le stime verso il basso basse, mentre un numero alto le aveva ancorate verso l'alto.

Molti esperimenti supportano la conclusione che l'ancoraggio giochi un ruolo significativo nella decisione giudiziale. Uno studio particolarmente interessante e ben costruito è stato condotto da Guthrie, Rachlinski e Wistrich su un campione di giudici, e riportato in un loro importante articolo (GUTHRIE et al. 2001). Ai giudici partecipanti veniva richiesto di immaginare di

[3] Per completezza e correttezza, vanno citate alcune critiche sollevate contro lo studio. Dubbi sull'esistenza dell'effetto sono stati immediatamente sollevati da WIENSHAL-MARGELL et al. 2011, cui ha fatto seguito una replica degli autori dello studio (DANZIGER et al. 2011b). GLÖCKNER 2016 ha invece confermato l'effetto, ma ha sostenuto che la sua dimensione sia stata sovrastimata. V. a questo proposito anche LAKENS 2017. Una ricostruzione simpatetica di queste critiche è proposta da CHATZIATHANASIOU 2022. Un altro problema riguarda specificamente l'idea della *ego-depletion*, un costrutto strettamente legato ad un modello allora molto in voga del controllo cognitivo, chiamato 'modello delle risorse', che gli autori dello studio mostravano di accettare: la nostra capacità di controllo cognitivo — la capacità di agire flessibilmente in vista di un certo compito anche di fronte a tendenze contrastanti — dipenderebbe dalla disponibilità di una risorsa fisiologica che può essere consumata e reintegrata (v. ad es. BAUMEISTER et al. 1998). Negli ultimi dieci anni il modello delle risorse ha perso terreno. La riduzione della capacità di ricorrere a controllo cognitivo dovuta ad esercizio prolungato di controllo è, infatti, un fenomeno robusto. Ma la capacità di controllo tende a ripristinarsi in presenza di motivazione sufficiente (un premio o una punizione), e questo non sembra coerente con il modello delle risorse (v. ad es. INZLICHT, BERKMAN 2015). Quand'anche questa critica dovesse richiedere una modifica della spiegazione dei meccanismi responsabili, però, non mi sembra che possa incidere in modo rilevante sull'attendibilità dell'effetto messo in luce dallo studio, i cui risultati sono in genere considerati robusti. (Per una critica allo studio imperniata su questo aspetto, v. comunque DALJORD et al. 2017.)

presiedere una causa per danni in un tribunale federale basato sulla *diversity jurisdiction*, la cui competenza è limitata a casi di valore superiore a 75.000 dollari.

Il caso da decidere era il seguente. La parte convenuta era una grande azienda di trasporti. Uno dei suoi camion, a causa di un guasto ai freni, non si era fermato al semaforo rosso, travolgendo l'attore. Le indagini successive avevano rivelato che l'impianto frenante era difettoso e che il camion non era stato sottoposto a un'adeguata manutenzione da parte della ditta convenuta. L'attore aveva riportato danni gravi: alcuni mesi di ricovero in ospedale, la perdita dell'uso delle gambe e la perdita del suo elevato reddito di elettricista. L'attore aveva chiesto il risarcimento dei danni, per un importo non specificato. I fatti erano tutti dati per provati, e l'oggetto della causa era soltanto l'ammontare del risarcimento dovuto.

I giudici erano stati divisi in due gruppi. Il primo gruppo era in una condizione di assenza di ancoraggio. Gli era stato semplicemente raccontato il caso, e chiesto di stimare il risarcimento dovuto all'attore. Il secondo gruppo si trovava invece in una condizione di ancoraggio. Gli era infatti stato comunicato che il convenuto aveva chiesto l'archiviazione del caso, sostenendo che il suo valore non raggiungesse il minimo richiesto di 75.000 dollari. Era stato poi chiesto loro di stimare l'ammontare del risarcimento dovuto all'attore. La richiesta di archiviazione era palesemente infondata, perché i danni (già dati per provati) erano evidentemente superiori a 75.000 dollari. I ricercatori, tuttavia, prevedevano che la menzione dell'importo di 75.000 dollari avrebbe innescato effetti di ancoraggio, rendendo le stime medie dei giudici nella condizione di ancoraggio inferiori alle stime medie dei giudici nella condizione di non ancoraggio.

Questo è ciò che si è effettivamente verificato. I giudici nella condizione senza ancoraggio hanno assegnato all'attore, in media, un risarcimento di 1.249.000 dollari, mentre i giudici nella condizione di ancoraggio hanno assegnato, in media, un risarcimento di 882.000 dollari. Si tratta di una differenza enorme: il valore medio del risarcimento del primo gruppo è di circa il 50% più alto del valore medio del risarcimento del secondo gruppo. (Per evitare i possibili effetti sulla media di stime molto elevate da parte di alcuni giudici, i ricercatori hanno proceduto anche ad un confronto per quartili, che ha confermato l'effetto di ancoraggio).

È importante sottolineare come i giudici nella condizione di ancoraggio concordassero sul fatto che i danni fossero molto più alti della somma menzionata di 75.000 dollari. Tuttavia, quest'ultima somma, da essi stessi trattata come irrilevante, sembra aver funzionato come un'ancora, abbassando in modo molto significativo la loro stima del risarcimento dovuto.

Esistono numerosissimi studi sull'effetto di ancoraggio in altri aspetti della decisione giudiziale, ad esempio sulla decisione riguardo all'ammontare della pena. In una serie di studi particolarmente disturbanti, ENGLICH, MUSSWEILER e STRACK (2005) hanno rilevato come l'ammontare della pena richiesta dalla pubblica accusa ancori verso l'alto l'ammontare della pena richiesta dalla difesa, e come questo effetto si riverberi in un ancoraggio verso l'alto della decisione del giudice. Altri studi hanno mostrato come l'effetto di ancoraggio sulla difesa si producesse anche quando veniva comunicato che la richiesta della pubblica accusa era stata determinata da uno studente di informatica privo di competenze giuridiche (ENGLICH, MUSSWEILER 2001), o addirittura attraverso il lancio di dadi (ENGLICH et al. 2006).

Anche qui, si ripete lo stesso pattern. Il giudizio è influenzato da fattori che il decisore stesso tratta come del tutto irrilevanti. L'effetto e i meccanismi che ne sono responsabili possono essere scoperti soltanto attraverso l'analisi di regolarità statistiche esibite da un grande numero di giudizi, mentre non appaiono a fronte del singolo giudizio. I meccanismi in questione operano in modo del tutto indipendente dalla buona fede e dalla professionalità dei giudici, che – sulla sola base delle strategie cognitive messe a disposizione dal loro buon senso e dal loro normale training professionale – sono del tutto impotenti a rilevare e contrastare il *bias*.

2.5. Il problema evidenziato nei paragrafi precedenti assume una dimensione parossistica nel fenomeno del 'rumore' (*noise*), al centro del recente e importante libro di KAHNEMAN, SIBONY e SUN-

STEIN (2021). Un sistema decisionale è affetto da 'rumore' quando (a) le decisioni deviano, in percentuale e misura significativa, da standard di correttezza e/o di coerenza, ma (b) le deviazioni non esibiscono, nell'insieme, pattern regolari, rendendo così le decisioni largamente imprevedibili.

Il rumore può dipendere da una molteplicità di fattori: (1) dalla variabilità delle attitudini valutative dei singoli decisori (per esempio, la maggiore o minore severità dei giudici) (cosiddetto '*level noise*'); (2) dalla combinazione di bias diversi a cui sono più o meno esposti decisori diversi (cosiddetto *pattern noise*); (3) da fattori occasionali che incidono sulla decisione, come l'umore o la stanchezza (*occasion noise*).

Ad esempio, KAHNEMAN, SIBONY e SUNSTEIN (2021, 17) citano due studi che indicano come la decisione di giudici penali tenda a essere più severa nei giorni successivi a una sconfitta della squadra di calcio locale (EREN, MOCAN 2018; il secondo è studio è stato poi pubblicato in CHEN, LOECHER 2019).

Tutte queste forme di rumore contribuiscono ad un tasso estremo di variabilità della decisione giudiziale che incide enormemente sulla possibilità di formare ragionevoli aspettative riguardo all'esito del giudizio e sulla coerenza complessiva dei giudizi in termini di eguaglianza di trattamento. Fra gli esempi proposti da KAHNEMAN, SIBONY e SUNSTEIN (2021, 91) vi è quello delle richieste di asilo negli USA (la vicenda del sito *Supralegem*, come vedremo, verterà anch'essa sulle richieste di asilo, questa volta in Francia). La probabilità di una decisione favorevole si riduce del 19% se segue a due richieste consecutive accolte favorevolmente. Il tasso di concessione del beneficio è inoltre soggetto a variazioni estreme da giudice a giudice – in uno studio citato, si passa dall'88% al 5 % (RAMJI-NOGALES et al. 2007).

2.6. Quelli presentati nei paragrafi precedenti sono soltanto alcuni esempi dell'innumerevole massa di studi su *bias* e rumore nella decisione giudiziale.

Sarebbe avventato concludere che questi studi supportino l'immagine di una giurisdizione in cui le ragioni giuridicamente rilevanti svolgono un ruolo trascurabile. Il *bias* dell'ancoraggio, ad esempio, agisce all'interno dello spazio di discrezionalità del giudice nel determinare l'ammontare di una pena o di un risarcimento, e questo spazio di discrezionalità si dà all'interno di decisioni che potrebbero, invece, essere sufficientemente vincolate da appropriate ragioni giuridiche. La limitazione della discrezionalità ad opera di una cornice di regole sostanziali o procedurali è sicuramente un fattore protettivo rispetto a possibili *bias* e lo strumento principe di uniformità delle decisioni (si pensi, in tema di risarcimento del danno, ai principi che regolano la prova del danno patrimoniale, o anche al ruolo svolto dalle tabelle per la liquidazione del danno biologico). Né si può sottovalutare l'importanza del ruolo di contrasto ai *bias* svolto dal contraddittorio, dalla pluralità di gradi di giudizio, e da dettagli più fini delle procedure giudiziarie.

Più in generale, la decisione giudiziale è una faccenda complessa, in cui entrano in gioco una molteplicità di fattori. Uno studio empirico di larga scala sul campo, al di fuori del laboratorio, è estremamente difficile. Ancora più difficile è dare valutazioni complessive degli effettivi processi di formazione delle decisioni giudiziali in termini di tesi generali e generiche come "le decisioni dei giudici sono per lo più guidate da ragioni 'formalmente giuridiche', regole e principi giuridici ricavati attraverso canoni accreditati di interpretazione giuridica da enunciati delle fonti del diritto"; o "le decisioni dei giudici sono per lo più guidate da ragioni diverse da quelle formalmente giuridiche (ragioni morali, o addirittura ragioni prudenziali – interessi di carriera, preferenze politiche)"; o ancora "le decisioni dei giudici sono per lo più guidate da fattori irrazionali e caotici, fattori che i decisori non riconoscerebbero come ragioni"[4]. Slogan così generali

[4] Ad esempio, SCHAUER (2009, 139 ss.), commentando decenni di studi delle decisioni della Corte Suprema degli USA che mostrano il ruolo significativo degli orientamenti politici dei giudici nel determinare le posizioni da loro assunte, rileva come sia avventato generalizzare estendendo ciò che accade a livello di una corte assolutamente peculiare per competenza e poteri come la Corte Suprema, con riguardo a casi di grande salienza, di difficile soluzione e di

e generici rischiano, esagerando, di distogliere l'attenzione dal nucleo più solido del problema, che può essere invece formulato in termini più modesti e limitati, ma anche molto più robusti:

(1) Le ricerche degli ultimi cinquant'anni mostrano che il ruolo di *bias* e rumore nella giurisdizione – nonostante il ruolo di contenimento svolto da regole e da procedure giudiziarie – è molto più significativo di quanto saremmo altrimenti inclini a riconoscere, rendendo le decisioni giudiziali suscettibili di essere influenzate non soltanto da pregiudizi che comportano gravi e sistematiche discriminazioni di alcune categorie di soggetti, ma anche da fattori irrilevanti e caotici.
(2) L'interferenza di queste forme di *bias* e rumore non può essere contrastata attraverso le virtù – come il buon senso, l'empatia, l'imparzialità, la sincera adesione ai principi fondamentali dell'ordinamento e l'aspirazione ad una tutela non discriminatoria dei diritti gli strumenti di regolazione cognitiva – che gli ordinari processi di socializzazione e la attuale formazione professionale mettono a disposizione del giudice. Si tratta inoltre spesso di difetti che emergono solo di fronte alla aggregazione – tecnicamente complessa – di una molteplicità di casi, e che pertanto non possono essere non solo contrastati, ma nemmeno rilevati attraverso forme di monitoraggio delle singole decisioni, né attraverso controllo giurisdizionale a seguito di impugnazione, né attraverso altre forme di controllo interno agli uffici.
(3) Il ruolo comprovato di *bias* e rumore, se anche non basta a supportare le tesi scettiche più radicali riguardo alla giurisdizione, non può che avere ricadute profonde sulla qualità dei servizi giudiziari e costituisce pertanto, senza dubbio, un problema normativo di estrema rilevanza, che non può essere ignorato. È un problema estremamente rilevante, anzitutto, quando le decisioni mostrano pattern coerenti che non dovrebbero avere – come la discriminazione di certe categorie di soggetti. Ed è un problema rilevante quando le decisioni si discostano troppo, sia pure in modo casuale, dall'ideale regolativo della omogeneità e coerenza a cui dovrebbero rispondere.
(4) Non si tratta soltanto della questione normativa della rispondenza della giurisdizione a certi ideali, ma anche, inscindibilmente, della questione, politica, della sua autorevolezza. È verosimile che la legittimazione del potere giurisdizionale dipenda, almeno in parte, dalla autorevolezza ad essa accordata – cioè, da un qualche livello sia pur minimo di credenza nel fatto che le sue decisioni siano effettivamente guidate dalle ragioni che dovrebbero guidarle. Gli studi sui *bias* – e soprattutto la loro divulgazione, e le generalizzazioni che questa suggerisce — possono, sotto questo aspetto, avere un impatto molto negativo sull'autorevolezza della giurisdizione. Questo impatto negativo, come vedremo, può essere aumentato di molto dal fenomeno della 'giustizia predittiva'.

3. Bias *e giustizia predittiva*

3.1. Il termine 'giustizia predittiva' indica l'applicazione di sistemi di *machine learning* a dati (e in particolare *big data*) giudiziari. Non è questa la sede (né sono io l'autore) per una trattazione approfondita di queste nozioni[5]. Ai nostri fini è sufficiente notare come i sistemi di *machine learning* – e ancora di più il loro sottoinsieme noto come *deep learning*, implementato da reti neurali con più livelli intermedi – eccellono nella capacità di estrarre, con accuratezza e velocità molto superiore a quella umana, 'pattern' (configurazioni ricorrenti di proprietà o attributi) anche molto complessi da grandi masse di dati, e di usarli per compiere inferenze riguardo a proprietà ignote di un certo caso sulla base delle sue proprietà note. Ad esempio, estraendo una configura-

grande rilevanza politica, alla miriade di casi ordinari che coinvolgono questioni tecniche e di scarsa rilevanza politica. È possibile, suggerisce Schauer, che questo tipo di casi – che costituiscono una quota molto significativa, se non maggioritaria, dell'attività giurisdizionale – vengano invece risolti sulla base di ragioni formalmente giuridiche.

[5] Per una introduzione al *machine learning* si vedano ALPAYDIN 2021 e BERMÚDEZ 2021, cap. 12. Una ottima review delle questioni relative alla giustizia predittiva è GALLI, SARTOR 2023.

zione di proprietà dalle immagini associate ad una specie botanica P, il sistema di *machine learning* può imparare ad ascrivere alla classe P un nuovo oggetto, identificato da immagini mai processate prima, sulla base del riconoscimento della configurazione appresa. È in virtù di questa capacità di estrarre informazioni statistiche riguardo alla correlazione tra attributi e usarla generalizzando a casi nuovi che i sistemi di *machine learning* sono detti 'predittivi': 'predicono' nel senso generico che inferiscono proprietà non note, e non solo nel senso specifico che prevedono eventi futuri. Il campo di applicazione della giustizia 'predittiva', dunque, non è limitato soltanto a compiti di previsione – stima di probabilità – di future decisioni giudiziali (o di altri eventi futuri, come la recidiva) sulla base di caratteristiche note del caso, ma include anche compiti molto diversi, come l'individuazione delle fonti giuridiche su cui basare la soluzione del caso, o l'individuazione di caratteristiche del caso ricorrentemente associate a certi esiti decisionali[6].

I sistemi di *machine learning* sono più o meno 'spiegabili' o 'opachi', secondo che rendano più o meno possibile risalire ad una spiegazione, riconoscibile per una mente umana, di come il sistema è arrivato ad una determinata previsione o classificazione – ad esempio, una spiegazione di quali siano gli attributi in base ai quali è stato 'predetto' un certo esito decisionale ad un certo caso. Il problema è che tende ad esservi una relazione di proporzionalità inversa tra capacità predittiva e spiegabilità: tanto più alte sono le capacità predittive di un sistema di *machine learning*, quanto minore tende ad essere la sua spiegabilità. (I sistemi di *deep learning* hanno grande capacità predittiva ma sono anche molto opachi).

Intendo soffermarmi su un aspetto specifico della giustizia predittiva, e cioè il ruolo che può avere nel potenziare l'impatto negativo sulla autorevolezza della giurisdizione del problema dei *bias* e del rumore nella decisione giudiziale umana esaminato nei paragrafi precedenti. Questo ruolo dipende dall'interazione di almeno due fattori diversi: la promessa di un sistema decisionale che, se opportunamente disegnato, è meno esposto a *bias* e rumore; e la facilitazione dell'emersione di (presunti) *bias* o rumore.

3.2. Il primo modo in cui la giustizia predittiva può contribuire alla crisi di autorevolezza della giurisprudenza è offrendosi come uno strumento per elaborare meccanicamente decisioni libere da, o meno soggetto a, i *bias* e il rumore che affliggono invece la giurisdizione umana. È quello di promettere, cioè, un miglior sistema decisionale.

La questione della decisione algoritmica è molto complessa, e non è il caso di aprire questo vaso di pandora. Basterà esemplificare una linea di argomentazione ben esemplificata, ancora una volta, da KAHNEMAN, SIBONY e SUNSTEIN (2021). Uno degli esempi che essi propongono (pp. 127 ss.) riguarda una categoria di decisioni giudiziali per le quali si dà un criterio indipendente di successo, e cioè le decisioni sulla libertà su cauzione (*bail*), che bilanciano fra diritto alla libertà dell'imputato e rischio di fuga e mancata presentazione al processo. Il successo di queste decisioni può essere valutato in rapporto alle effettive fughe e mancate presentazioni (semplificando, la decisione fallisce se l'imputato fugge). Gli autori citano uno studio (JUNG et al. 2020) che ha valutato, rispetto a questo tipo di decisioni, la performance di un algoritmo 'frugale', estremamente semplice. L'algoritmo è basato su due soli parametri risultati altamente predittivi della probabilità di fuga, l'età e il numero di fughe precedenti. Il valore di questi due parametri viene tradotto in punti che valutano il rischio, sulla base di una regola semplicissima, così semplice da poter essere applicata senza l'ausilio di strumentazioni elettronica. Ebbene, rilevano gli autori, questo algoritmo così elementare ha prodotto decisioni molto più accurate di quelle di qualsiasi giudice umano. Altri studi (come KLEINBERG et al. 2018) mostrano come la superiorità della decisione 'meccanica' rispetto al libero corso del ragionamento umano aumenti enormemente con algoritmi più potenti, come quelli di *machine learning*.

[6] Per approfondimenti rinvio a GALLI, SARTOR 2023, § 3.1, dove questi aspetti sono spiegati molto bene.

Si potrebbe obiettare, come spesso si obietta, che gli algoritmi predittivi rischiano di reiterare *bias* contenuti nei dati da cui estraggono pattern decisionali – per esempio, com'è noto, rischiano di riprodurre discriminazioni. Questo è senz'altro vero. La replica, però, è che non si tratta di un esito necessario. Il sistema di *machine learning* – intendendo per 'sistema' sia l'algoritmo che classifica i dati sia il *data set* su cui opera – può anche essere disegnato con successo in modo da ridurre discriminazioni, e vi sono state importanti dimostrazioni di quanto un sistema meccanico fosse in grado di scegliere in modo significativamente meno discriminatorio rispetto a decisori umani (lo studio citato è COGWILL 2019). Altra perplessità riguarda l'opacità dei sistemi di *machine learning*, che – oltre ad aumentare il rischio di *bias* occulti – rischia di renderli inadeguati ad un tipo di decisione come quella giudiziale. Nella decisione giudiziale, infatti, non è possibile separare la correttezza del contenuto della decisione dalle ragioni che la sostengono (la decisione giuridicamente corretta non lo è in assoluto, ma solo in rapporto ad una certa base di premesse giuridicamente accettabili); le ragioni che fondano la decisione devono essere accessibili alle parti; e vi è una legittima aspettativa delle parti a che vi sia una corrispondenza tendenziale tra le ragioni addotte e quelle che hanno causalmente guidato la decisione. Anche qui è possibile, pur accettando l'obiezione, replicare puntando sul disegno di sistemi di *machine learning* spiegabili.

Al di là delle possibilità tecniche attualmente disponibili, però, c'è un argomento più generale che muove i fautori del – se non totale almeno significativo – passaggio di responsabilità decisionale da umani a macchina. La questione rilevante non è infatti se la cognizione della macchina sia anch'essa, come quella umana, opaca e fragile, ma piuttosto se lo sia *di più o di meno*. In certi contesti decisionali, tra cui rientrano anche alcuni tipi di decisioni giudiziali – è questo l'argomento – la mente umana è talmente opaca e fragile da essere, nel complesso, un'alternativa peggiore rispetto a macchine opportunamente disegnate, nonostante gli innegabili difetti che affliggono anche queste[7].

Quanto più forte sia la percezione del grado di *bias* e rumore a cui è esposta la decisione giudiziale umana, quanto più ad essa venga contrapposto un modello che promette (forse ingannevolmente, ma non del tutto implausibilmente) di essere, almeno dal punto di vista di *bias* e rumore, un'alternativa migliore, e quanto più tutta la discussione fuoriesca dal circolo ristretto della discussione accademica per entrare nell'arena del pubblico dominio, tanto maggiore rischierà di essere l'impatto negativo sull'autorevolezza della giurisdizione[8].

3.3. Il secondo modo in cui la giustizia predittiva può contribuire alla crisi di autorevolezza della giurisdizione è supportando la rilevazione di (presunti) *bias* e rumore.

Si ricordi che la giustizia 'predittiva' – anche è questo il modo in cui il termine tende ad essere inteso dai giuristi – non si limita all'utilizzo dei sistemi di *machine learning* per prevedere future decisioni delle corti, ma include anche usi diversi, fra cui l'individuazione di significative correlazioni tra certe circostanze dei casi e certi esiti decisionali, correlazioni che potrebbe essere molto più difficile scoprire senza l'ausilio di una macchina. Queste correlazioni possono, sotto certe condizioni, costituire indizi di relazioni causali – indizi, cioè, del fatto che la decisione sia stata causalmente orientata dalla presenza di quelle caratteristiche. Il sospetto di causalità potrà poi essere corroborato da ulteriori esami dei dati. Quando viene rilevata una correlazione robusta tra una decisione e fattori diversi dalle ragioni giuridicamente valide che potrebbero sostenerla, il sospetto di relazione causale è un sospetto di *bias*. Analogamente, sistemi di *machine*

[7] Per un esempio di questo modo di argomentare con specifico riferimento alla giurisdizione, v. BAGARIC et al. 2022.

[8] I rischi per l'autorevolezza della giurisdizione dalla diffusione sensazionalistica di studi sulla presenza dei *bias* nella decisione giudiziale, insieme alla prospettazione dell'alternativa dell'IA, sono ad esempio evidenziati da CHATZIATHANASIOU 2022. L'autore, nel sottolineare quanto il pericolo sia attuale in contesti nei quali l'indipendenza del potere giudiziario è messa sotto attacco, cita anche l'uso, da parte di movimenti populisti di destra tedeschi, dello studio sui giudici 'affamati' per chiedere di mettere i giudici a dieta per ottenere condanne più severe.

learning possono supportare la rilevazione di eccessive variabilità decisionali rispetto a casi che sembrano presentare caratteristiche simili, sollevando il sospetto di rumore.

La giustizia predittiva costituisce, così, un potentissimo strumento di lavoro per gli studiosi impegnati nella rilevazione di *bias* e rumore, che allarga di molto la massa di dati analizzabili e le correlazioni rilevabili, facilitando l'emersione di pattern decisionali sospetti. Ciò avviene soprattutto quando grandi masse di dati giudiziari sono resi facilmente accessibili, e in formato adeguato, ai ricercatori.

Tutto questo ha già, di per sé, un effetto di amplificazione del possibile impatto negativo sull'autorevolezza della giurisdizione. Ma questo impatto può diventare esplosivo quando i dati e i sistemi per trattarli diventano accessibili al di là dell'ambito ristretto della ricerca accademica, con i suoi vincoli e cautele metodologiche e la diffusione relativamente ridotta dei suoi risultati. È proprio questo quello che è accaduto in Francia con la vicenda del sito *Supralegem*.

3.4. Nel 2016 il governo francese aveva varato una innovativa politica di *open data* giudiziari, rendendo liberamente e gratuitamente accessibili dati relativi ad un gran numero di decisioni giudiziali[9]. Il sito *Supralegem* utilizzava algoritmi di *machine learning* capaci di analizzare questi dati estraendo pattern decisionali. In un articolo molto controverso, BENESTY (2016), l'ideatore del sito, aveva reso pubbliche alcune elaborazioni relative a decisioni di corti amministrative riguardo ai ricorsi contro atti di espulsione di cittadini stranieri irregolari. Il punto di partenza era il sospetto di parzialità nei giudizi di alcune delle corti amministrative d'appello competenti. L'elaborazione confrontava i pattern decisionali di diversi collegi, identificati sulla base del Presidente, rilevando tassi estremamente diversi di rigetto (il rigetto del ricorso in questione comportava l'espulsione dal territorio dello stato). Per un collegio, ad esempio, veniva stimato un tasso di rigetto che andava, in anni diversi, dal 90 al 100 %, mentre per un altro collegio andava dal 43 al 78%. Con una elaborazione ulteriore, gli algoritmi di *Supralegem* estraevano pattern relativi al tasso di rigetto, da parte degli stessi collegi, di ricorsi presentati da singoli cittadini contro l'amministrazione, mostrando come i collegi più severi fossero quelli meno inclini ad accettare ricorsi contro l'amministrazione. A giudizio dell'autore, questi pattern decisionali giustificavano forti sospetti di parzialità a favore dell'amministrazione da parte dei collegi esaminati.

L'articolo di Michael Benesty ha sollevato un vespaio di polemiche – anche perché veniva indicato il nome dei presidenti dei collegi esaminati. In un successivo Rapporto su *L'open data des décisions de justice* richiesto dal Ministro di Giustizia (RAPPORT CADIET 2017), venne dedicato ampio spazio alla questione dell'opportunità di pubblicare il nome dei giudici, con espliciti riferimenti al caso *Supralegem* (pp. 100, 151, 156).

In seguito al rapporto, nel 2019, un nuovo intervento legislativo ha rimodellato la politica francese di *open data*, prevedendo l'occultamento di dati che consentono di individuare, tra l'altro, l'identità dei magistrati, e vietando esplicitamente pratiche di profilazione dei giudici[10].

Questo caso esemplifica bene i possibili effetti della giustizia predittiva: una molto maggiore visibilità pubblica di pattern decisionali sospetti nella attività giurisdizionale, e dunque un potenziale ampliamento dell'impatto negativo sull'autorevolezza della giurisdizione. Il RAPPORT CADIET 2017 (pp. 151 e 154) solleva dubbi sulla correttezza delle conclusioni riguardo alla imparzialità della giurisdizione che l'articolo di Benesty trae dall'analisi dei dati. Se nel caso specifico si tratti o meno di dubbi fondati è una questione complessa, che non proverò a risolvere[11]. Resta

[9] LOI n. 2016-1321 du 7 octubre 2016 pour une République numérique (https://www.legifrance.gouv.fr/jorf/id/JORFTEXT000033202746).

[10] LOI n. 2019-222 du 23 mars 2019 de programmation 2018-2022 et de réforme pour la justice, art. 33 (https://www.legifrance.gouv.fr/loda/id/JORFTEXT000038261631).

[11] Le questioni relative all'interpretazione dei pattern estratti dai dati e alla ricostruzione delle loro cause possibili coinvolgono questioni tecniche estremamente complesse. Per farsi un'idea di questa difficoltà, il lettore può consul-

però il fatto che, in altri casi, la valutazione delle correlazioni e il salto da correlazione a causalità potrebbero essere condotte – in assenza di competenze specifiche e del controllo metodologico assicurato dalla struttura della ricerca accademica – in modo troppo disinvolto, con ulteriori effetti erosivi, stavolta non sufficientemente giustificati, dell'autorevolezza della giurisdizione.

4. *Considerazioni conclusive*

Vengo alle conclusioni. La giustizia è uno dei tanti campi nei quali l'intelligenza umana si trova entro una morsa che la delegittima: da un lato, un aumento di conoscenza dei suoi limiti, dall'altro una fondata promessa di superamento di quei limiti attraverso l'intervento dell'intelligenza artificiale. Come si può (se si può) sfuggire a questa morsa?

Le soluzioni estreme non sembrano né praticabili, né desiderabili. Il superamento della giustizia umana a favore della giustizia delle macchine è, allo stato, un'ipotesi tecnicamente remota. A meno di un profondo stravolgimento in strutture mentali e sociali secolari – cosa per sua stessa natura difficilmente prevedibile —, è un'ipotesi remota anche culturalmente: il sistema della giustizia ha funzione simbolica e di allocazione e gestione del potere che richiede il decisore umano (v. GARAPON, LASSÈGUE 2018).

D'altro canto, arrestare l'incremento di conoscenza riguardo ai limiti cognitivi e di coerenza della giurisdizione e il perfezionamento delle tecniche di decisione artificiale, oltre ad essere una soluzione chiaramente regressiva, è virtualmente impossibile. Si tratta di processi ormai troppo avanzati per poter essere interrotti. Vi possono senz'altro essere ragioni sia politiche che giuridiche per limitare l'accesso ai dati e vietare pratiche di profilazione – l'art. 10 della CEDU, ad esempio, prevede la possibilità di limitare la libertà di espressione, tra l'altro, per "garantire l'autorità [...] del potere giudiziario"[12]. Ma è anche vero, d'altra parte, che si tratta di un *second best* che lascia l'amaro in bocca – l'aspettativa dei cittadini di poter avere accesso a pattern non immediatamente visibili e sospetti nell'esercizio di un potere così ingombrante come quello giudiziario sembra, in uno stato democratico di diritto, pienamente legittima e importante, un bene che non si può sacrificare a cuor leggero.

Scartate le soluzioni estreme, resta un amplissimo spettro di soluzioni intermedie, quelle orientate verso una qualche forma di integrazione di tutta questa conoscenza e tecnica nella decisione umana, per aumentare le capacità decisionali umane, e non per rimpiazzarle.

Riguardo alla giustizia predittiva, questa è la direzione più ovvia: l'interazione intelligente tra decisione umana e intelligenza artificiale. Le proposte e i modelli di interazione sono numerosi, così come numerosi sono i rischi e le difficoltà. (Per una rassegna segnalo, restando in ambito francese, il *Rapport à la première présidente de la Cour de cassation et au procureur général près la Cour de cassation* elaborato da CADIET et al. 2022, un testo particolarmente interessante perché molto attento alla pratica effettiva. Una importante rassegna delle pratiche europee e dei rischi e le difficoltà principali è contenuta negli studi di accompagnamento alla *European Ethical Charter on the Use of Artificial Intelligence in Judicial Systems and their Environments* (CEPEJ 2018). Rinvio anche a GALLI, SARTOR 2023, che contiene un'ottima review dei rischi e delle difficoltà principali, delle strategie di interazione praticabili, e delle (poche) esperienze attive in Italia).

Numerose sono anche le proposte riguardo all'integrazione della conoscenza relativa ai *bias* nel training dei giudici e dei giuristi in genere, sia attraverso la formazione al riconoscimento della fragilità e opacità dei propri percorsi decisionali, sia attraverso l'apprendimento di tecniche che riducono l'effetto dei *bias* (cosiddetto '*debiasing*'), riducendo sia la probabilità di incorrere in

tare, a titolo di esempio, la letteratura sulla vicenda dei giudici 'affamati' citata alla nota 1.

[12] Ringrazio Massimo Starita per avermelo fatto notare.

essi sia la probabilità che la loro insorgenza abbia influenza significativa sul giudizio[13] – ad esempio, il ricorso a linee guida che fissano procedure decisionali specificamente disegnate per contrastare l'insorgenza di certi *bias*[14].

Questi due percorsi potrebbero, almeno in teoria, essere affiancati e integrati, da un lato attraverso una sorta di auto-profilazione dei giudici, che gli permetta – con il supporto di un team dotato di adeguate conoscenze psicologiche, statistiche, e informatiche – di fare emergere i *bias* a cui è esposto il *proprio* processo decisionale, d'altro lato attraverso procedure di confronto critico con le predizioni della macchina, disegnate allo scopo di supportare il monitoraggio del *proprio* processo decisionale. Si tratterebbe, in altri termini, di sfruttare conoscenza psicologica e tecnica di intelligenza artificiale per approntare 'thinking tools' ulteriori rispetto a quelli, imprescindibili ma rivelatisi insufficienti, del senso comune e della formazione tradizionale (v. ad es., in questo senso, CHEN 2019).

RACHLINSKI e WISTRICH (2017) hanno suggerito metodi come l'adozione, in fase di reclutamento dei magistrati, di test riguardo all'esposizione a *bias*, come ad esempio i test sui *bias* razziali impliciti, non come condizione per l'accesso alla funzione, ma come strumento utile per incrementare la consapevolezza dei propri processi decisionali e per disegnare forme di addestramento personalizzate; o lo sviluppo di procedure di *auditing* per fornire feedback ai giudici rispetto alla qualità delle loro decisioni. Sistemi di questo genere hanno uno spiacevole sapore disciplinare, e potrebbero risultare inopportuni per diverse ragioni legate all'organizzazione di un potere così delicato come quello giudiziario. La strategia esemplificata sopra è davvero minimale, perseguendo gli stessi obiettivi con mezzi molto meno controversi. Si tratterebbe infatti di mettere a disposizione dei giudici un servizio, *ad accesso volontario e protetto da vincoli di riservatezza*, di auto-profilazione guidata dall'IA e di *counseling* psicologico personalizzato sui *bias*. Sarebbe un sistema di auto-monitoraggio messo a disposizione dei giudici per consentirgli un accesso libero a strumenti di *enhancement* cognitivo non particolarmente problematici, e facilmente integrabili nella propria competenza professionale. Questa linea, si noti, è incoraggiata, sia pure con prudenza, nella relazione di accompagnamento alla già citata *European Ethical Charter on the Use of Artificial Intelligence in Judicial Systems and their Environments* (CEPEJ 2018, 66): «offering judges a more detailed quantitative and qualitative assessment of their activities, thanks to new tools, but with a purely informative aim of assisting in decision-making and for their exclusive use, could be encouraged».

Strategie di questo tipo sono, ovviamente, tutte da costruire, e anch'esse per nulla prive di rischi e difficoltà. Prima fra tutte, ovviamente, la difficoltà di integrare questa pratica all'interno del farraginoso funzionamento dell'amministrazione della giustizia. O il rischio, nonostante la moderazione dei mezzi, di ingabbiare una funzione delicata e tragicamente umana come quella giudiziaria in una spirale di controlli che, se non ben calibrati e se travolti dalle difficilmente resistibili spirali di burocratizzazione, potrebbero impoverire, anziché migliorare, la qualità delle decisioni. In ogni caso, anche nelle ipotesi più conservative, una strada simile comporterebbe un cambiamento di struttura sia nella formazione dei giudici che, soprattutto, nell'organizzazione del loro lavoro – sarebbe necessario includere stabilmente negli uffici giudiziari sia psicologi che informatici, e modellare la formazione dei giudici in modo da permettere una interazione informata e intelligente con queste figure professionali. Si tratterebbe di trasformazioni epocali e

[13] Per questa differenza vedi KANG ET AL. 2012.

[14] Utili review di tecniche e di studi sulla loro efficacia si trovano in KANG ET AL. 2012 e RACHLINSKI, WISTRICH 2017. Fra le tecniche menzionate vi sono, ad esempio: l'esposizione a esempi incongruenti rispetto agli stereotipi; l'acquisizione dell'abitudine a mettere in dubbio la propria obiettività; l'aumento la motivazione a giudicare equamente e imparzialmente; migliorare le condizioni decisionali, ad esempio in termini di disponibilità di tempo e di energie cognitive; dare feedback quantitativi relativi all'insorgenza dei *bias*; o addirittura il ricorso a tecniche di meditazione.

dispendiose, ma la posta in gioco è veramente alta, e la strategia che è meno probabile che paghi è quella dello struzzo.

Ma un passo preliminare, e di più facile portata, riguarda la formazione giuridica. Sarebbe importante inserire urgentemente nella formazione standard del giurista: (i) competenze statistiche di base (un addestramento sia pure minimo alla scienza dei numeri è condizione necessaria per un rapporto responsabile con i fatti, inclusi i fatti relativi alla giurisdizione); (ii) formazione di base in materia di psicologia della decisione giudiziale, con particolare attenzione alla questione dei *bias* impliciti (qui è richiesto, in particolare, uno sforzo da parte dei filosofi del diritto, per superare la linea troppo netta di divisione, ereditata dalla tradizione novecentesca, fra le dimensioni logico-argomentativa e socio-psicologica del ragionamento giuridico, e la concentrazione pressocché esclusiva sulla prima); (iii) formazione di base, sia pratica che teorica, in materia di applicazione dell'IA alla giustizia.

L'allargamento della formazione giuridica a ulteriori competenze non mira certo a soppiantare la formazione giuridica tradizionale, che – è sempre opportuno ricordarlo – incorpora una sapienza millenaria riguardo alla gestione della decisione in condizione di incertezza, la cui salvaguardia è condizione necessaria per non disperdere il livello di garanzia dei diritti ottenuto dai sistemi giuridici progrediti. L'obiettivo, piuttosto, è quello di integrare la formazione giuridica tradizionale, per proteggerla e supportarla dove si mostra fragile e insufficiente.

Riferimenti bibliografici

ALPAYDIN E. 2021. *Machine Learning*, The MIT Press.

BAGARIC M., SVILAR J., BULL M., HUNTER D., SOBBS N. 2022. *The Solution to the Pervasive Bias and Discrimination in the Criminal Justice System: Transparent and Fair Artificial Intelligence*, in «American Criminal Law Review», 59 (2022), 1, 95-148.

BAUMEISTER R., BRATSLAVSKY E., MURAVEN M., TICE D.M. 1998. *Ego Depletion: Is the Active Self a Limited Resource?*, in «Journal of Personality and Social Psychology», 74, 5, 1998, 1252-1265.

BENESTY M. 2016. *L'impartialité de certaines juges mise à mal par l'intelligence artificielle*, in «Village de la Justice», 24 mars 2016.

BERMÚDEZ J.L. 2021. *Cognitive Science. An Introduction to the Science of the Mind*, 3rd ed., Cambridge University Press.

CADIET L., CHAINAIS C., SOMMER J.-M. (dir.), JOBERT S., JOND-NECAND E. (rapp.) 2022. *La diffusion des données décisionnelles et la jurisprudence*, Rapport remis à la première présidente de la Cour de cassation et au procureur général près la Cour de cassation - juin 2022 (https://www.courdecassation.fr/toutes-les-actualites/2022/06/14/la-diffusion-des-donnees-decisionnelles-et-la-jurisprudence-quelle).

CEPEJ 2018. *European Ethical Charter on the Use of Artificial Intelligence in Judicial Systems and their Environments*, by the European Commission for the Efficiency of Justice (CEPEJ), 3-4 dicembre 2018 (https://rm.coe.int/ethical-charter-en-for-publication-4-december-2018/16808f699c). La Carta contiene due appendici: Appendix I: In-depth study of the use of AI in judicial systems, notably AI application processing judicial decisions and data; Appendix II: Which uses of AI in European judicial systems?

CHATZIATHANASIOU K. 2022. *Beware the Lure of Narratives: 'Hungry Judges' Should Not Motivate the Use of 'Artificial Intelligence' in Law*, in «German Law Journal», 23, 4, 452-464.

CHEN D.L. 2019. *Judicial Analytics and the Great Transformation of American Law*, in «Artificial Intelligence and Law», 27, 1, 15-42.

CHEN D.L., LOECHER M. 2025. *Mood and the Malleability of Moral Reasoning: The Impact of Irrelevant Factors on Judicial Decisions*, in «Journal of Behavioral and Experimental Economics», 116, June 2025, 102364,

COGWILL B. 2019. *Bias and Productivity in Humans and Machines*, Columbia Business School Research Paper Forthcoming (https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3584916).

DALJORD Ø., URMINSKY O., URETA J.-M. 2017. *The Status Quo Theory of Depletion Does not Explain the Israeli Parole Decisions*, working paper, 3, December 2017 (https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3448164).

DANZIGER S., LEVAV J., AVNAIM-PESSO L. 2011a. *Extraneous Factors in Judicial Decisions*, in «Proceedings of the National Academy of Sciences of the United States of America», 108, 17, 2011, 6889-6892.

DANZIGER S., LEVAV J., AVNAIM-PESSO L. 2011b. *Reply to Weinshall-Margel and Shapard: Extraneous factors in judicial decisions persist*, in «Proceedings of the National Academy of Sciences of the United States of America», 108, 17, 2011, E834.

DASGUPTA N., GREENWALD A.G. 2001. *On the malleability of automatic attitudes: Combating automatic prejudice with images of admired and disliked individuals*, in «Journal of Personality and Social Psychology», 81, 5, 800-814.

ENGLICH B., MUSSWEILER TH. 2001. *Sentencing under Uncertainty: Anchoring Effects in the Courtroom*, in «Journal of Applied Social Psychology», 31, 2001, 1535-1551.

ENGLICH B., MUSSWEILER TH., STRACK F. 2005. *The Last Word in Court. A Hidden Disadvantage for the Defense*, in «Law and Human Behavior», 29, 6, 2005, 705-722.

ENGLICH B., MUSSWEILER TH., STRACK F. 2006. *Playing Dice with Criminal Sentences: The Influence of Irrelevant Anchors on Experts' Judicial Decision Making*, in «Personality and Social Psychology Bulletin», 32, 2, 2006, 188-200.

EREN O., MOCAN N. 2018. *Emotional Judges and Unlucky Juveniles*, «American Economic Journal: Applied Economics», 10, 3 (2018), 171-205.

GALLI F., SARTOR G. 2023. *AI Approaches to Predictive Justice: A Critical Assessment*, in «Human(*ities) and* Rights», 5 (2023), 2, 165-217.

GARAPON A., LASSÈGUE J. 2018. *Justice digitale: Révolution graphique et rupture anthropologique*, Presses Universitaires de France, 2018.

GLÖCKNER A. 2016 *The irrational hungry judge effect revisited: Simulations reveal that the magnitude of the effect is overestimated*, in «Judgment and Decision Making», 11, 6, 2016, 601-610.

GUTHRIE C., RACHLINSKI J.J., WISTRICH A.J. 2001. *Inside the Judicial Mind*, «Cornell Law Review», 86, 2001, 778-830.

INZLICHT M., BERKMAN E. 2015. *Six questions for the resource model of control (and some answers)*, in «Social and Personality Psychology Compass», 9, 10, 2015, 511-524.

JOY-GABA J.A., NOSEK B.A. 2010. *The Surprisingly Limited Malleability of Implicit Racial Evaluations*, in «Social Psychology», 41, 3, 137-146.

JUNG J., CONCANNON C., SHROFF R., GOEL S., GOLDSTEIN D.G. 2020. *Simple Rules to Guide Expert Classifications*, in «Journal of the Royal Statistical Society. Statistics in Society», 183 (2020), 771-800.

KAHNEMAN D., SIBONY O., SUNSTEIN C.R. 2021. *Noise. A Flaw in Human Judgment*, Little, Brown Spark, 2021.

KANG J., BENNETT M., CARBADO D., CASEY P., DASGUPTA N., FAIGMAN D., GODSIL R., GREENWALD A.G., LEVINSON J., MNOOKIN J. 2012. *Implicit Bias in the Courtroom*, in «University of Chicago Law Review», 59, 1124-1186.

KLEINBERG J., LAKKARAJU H., LESKOVEC J., LUDWIG J., MULLAINATHAN S. 2018. *Human Decisions and Machine Predictions*, «The Quarterly Journal of Economics», 133, 1, 2018, 237-293.

KOZINSKI A. 1993. *What I Ate for Breakfast and Other Mysteries of Judicial Decisions*, in «Loyola of Los Angeles Law Review», 26, 1993, 993-999.

LAKENS D. 2017. *Impossibly Hungry Judges*, in the Blog «The 20% Statistician», 3 July 2017 (https://daniellakens.blogspot.com/2017/07/impossibly-hungry-judges.html).

LANE K.A., KANG J., BANAJI R.M. 2007. *Implicit Social Cognition and Law*, in «Annual Review of Law and Social Science», 3, 2007, 427-451.

NELSON T.D. (ed.) 2009. *Handbook of Prejudice, Stereotyping, and Discrimination*, Psychology Press.

NISBETT R.E., WILSON T.D. 1977. *Telling More Than We Can Know: Verbal Reports on Mental Processes*, in «Psychological Review», 84, 3, 1977, 231-259.

PRIEL D. 2020. *Law is What the Judge Had for Breakfast. A Brief History of an Unpalatable Idea*, in «Buffalo Law Review», 68, 3, 2020, 899-930.

RACHLINSKI J.J., JOHNSON S.L., WISTRICH A.J., GUTHRIE C. 2009. *Does Unconscious Racial Bias Affect Trial Judges?*, in «Notre Dame Law Review», 84, 3, 2009, 1195-1246.

RAMJI-NOGALES J., SCHOENHOLTZ A.I., SCHRAG P. 2007. *Refugee Roulette: Disparities in Asylum Adjudication*, in «Stanford Law Review», 60, 2.

RAPPORT CADIET 2017. *L'*open data *des décisions de justice*, Rapport a Madame La Garde des

Sceaux, Ministre de la Justice pour la Commission présidée par le Professeur L. Cadiet (https://www.justice.gouv.fr/sites/default/files/migrations/portail/publication/open_data_rapport.pdf).

SUNSTEIN C.R, JOLLS C., THALER R.H. 1998. *A Behavioral Approach to Law and Economics*, in «Stanford Law Review», 50, 1471-1550.

TVERSKY A., KAHNEMAN D. 1974. *Judgment Under Uncertainty: Heuristics and Biases*, in «Science», New Series, 185, 4157, 1974, 1124-1131.

WIENSHAL-MARGELL K., SHAPARD J. 2011. *Overlooked Factors in the Analysis of Parole Decisions*, in "Proceedings of the National Academy of Sciences of the United States of America", 108, 17, 2011, E833.

WISTRICH A.J., RACHLINSKI J.J. 2017. *Implicit Bias in Judicial Decision Making. How It Affects Judgment and What Judges Can Do About It*, in Redfield S.E. (ed.), *Enhancing Justice: Reducing Bias*, American Bar Association, 87-130.

ZAMIR E., TEICHMAN D. 2014. *The Oxford Handbook of Behavioral Economics*, Oxford University Press.