CALCULATE CRIMINAL LAW? CRITICITÀ NELL'USO DEGLI ALGORITMI DI PERICOLOSITÀ SOCIALE

di Mattia Di Florio
(*Assegnista di ricerca in diritto penale, Università di Foggia*)

Sommario: 1. *Calcolo* e diritto penale. – 2. Algoritmi di pericolosità sociale. – 3. Criticità nell'uso di *predictive tools*: le possibili violazioni delle garanzie dell'imputato. – 4. Il rischio di una *Mechanical Jurisprudence*. – 5. L'opacità degli algoritmi. – 6. Algoritmi e *bias*. – 7. Conclusioni.

1. Alla fine del XVII secolo, il matematico, filosofo e giurista tedesco Gottfried Wilhem von Leibniz nutriva l'ambizione di formulare un algoritmo [1], cioè una procedura di calcolo[2], per formalizzare l'espressione del pensiero umano, il linguaggio naturale: «ciò fatto, quando nasceranno controversie, non vi sarà bisogno di disputare tra due filosofi più che tra due contabili: basterà infatti prendere in mano la penna, sedersi davanti all'abaco e (preso con sé, volendo, un amico) dirsi a vicenda: *calcoliamo (calculemus)!*»[3].

[1] Sul piano etimologico, la parola algoritmo deriva dal nome di un matematico e sapiente arabo Al-Khwarizmi, vissuto nel IX secolo d.C., il quale contribuì in maniera significativa allo sviluppo della teoria delle equazioni algebriche (così Toffalori, *Algoritmi*, Bologna 2015, 18-19).

[2] Nell'antichità, l'idea di un procedimento per il calcolo era tutt'altro che sconosciuta, ad esempio le "ricette" per effettuare calcoli impresse con segni cuneiformi su tavole babilonesi di quasi 4000 anni fa (cfr. *amplius* B. Codenotti, M. Leoncini, *La rivoluzione silenziosa. Le grandi idee dell'informatica alla base dell'era digitale*, Torino 2020, 7 ss.); inoltre, il *Crivello* del matematico alessandrino Eratostene (che per primo misurò il meridiano terrestre), per produrre la lista dei numeri primi, cioè dei numeri interi positivi divisibili solo per sé stessi e per 1, fino a un certo limite (cfr. *amplius* L. Russo, *La rivoluzione dimenticata. Il pensiero scientifico greco e la scienza moderna*, Milano 2021) (no virgola tra luogo di pubblicazione e anno, anche per le altre note).

[3] G.W. Leibniz, *De arte characteristica ad perficiendas scientias ratione nitentes*, trad it., *Sull'arte caratteristica, per praticare le scienze che si basano sulla ragione,* in *Scritti filosofici di Gottfried Wilhem Leibniz*, a cura di M. Mugnai, E. Pasini, Torino 2000, 213.

Leibniz aveva già creato in ambito matematico (indipendentemente da Newton) il calcolo differenziale e integrale per eseguire facilmente calcoli assai complicati, ma rimase convinto che qualcosa di simile potesse essere fatto anche per tutta la conoscenza umana (cfr. M. Davis, *Il calcolatore universale*, trad. it., Milano 2003, 18 ss.; M. Li Calzi, *La matematica dell'incertezza*, Bologna 2016, 57). Leibniz si ispirò non solo al filosofo inglese Thomas Hobbes che aveva concepito il ragionamento come un calcolo (v. *Scritti filosofici di Thomas Hobbes*, a cura di A. Negri, G. Paganini, Torino 2013), ma anche al filosofo, logico e mistico medievale Raimondo Lullo, che aveva immaginato di meccanizzare non solo i sillogismi aristotelici, ma l'intero linguaggio che è la combinazione di molte parti (es.: articoli, sostantivi aggettivi, verbi, preposizioni, etc.); la sua idea era di fare grandi ruote

L'idea del *calculemus* continuò a essere oggetto di riflessione scientifica secoli dopo Leibniz[4].

Nel diritto penale, il *calculemus* evoca l'idea leibniziana di un giudice algoritmico che applica le regole del diritto per la decisione del caso concreto: è un'idea suggestiva che, però, va chiarita per non essere fraintesa, se è vero, come osservava nella prima metà del secolo scorso, B.N. Cardoso, giudice della Corte Suprema degli Stati Uniti, che «ancora non è stata scritta la tavola dei logaritmi per la formula di giustizia»[5].

In generale, il calcolo risulta "estraneo" al diritto penale che ha per oggetto, come scriveva Cesare Beccaria, i delitti e le pene[6]. Se, dunque, lo studioso del diritto penale, pur con qualche "variazione sul tema" (rispetto all'epoca di Beccaria) continua ad occuparsi dell'interpretazione (e dell'applicazione) della fattispecie incriminatrici, ci si potrebbe chiedere che rilevanza abbia il calcolo in tutto questo.

È vero che il codice penale impiega, in modo spesso "inconsapevole", concetti che sono oggetto di studio di discipline extra-penalistiche, come la probabilità (di commettere nuovi reati) nella pericolosità sociale generica (*ex* art. 203 Cp), e ciò è dovuto anche alla formazione largamente umanistica degli studiosi del diritto penale che contribuirono alla stesura del codice Rocco.

Sebbene il diritto penale, già nella seconda metà dell'800, si fosse aperto all'apporto dell'antropologia e della psichiatria applicata alla medicina con Cesare Lombroso e della sociologia criminale con Enrico Ferri[7], la prospettiva scientista rimase "ai margini" del dibattito penalistico, dove prevalsero la scuola classica e la terza scuola che avrebbe "preparato il terreno" al regime autoritario[8]. A questo si aggiungano anche le derive positiviste, in particolare il passaggio da un diritto penale del fatto a un diritto penale d'autore, che il Progetto Ferri (elaborato nel 1921 e mai approvato) rischiava di comportare, muovendo dalla prospettiva di far "ruotare" il diritto penale intorno al concetto generico di pericolosità sociale, in luogo della colpevolezza[9].

---

corrispondenti alle parti del linguaggio, disporre queste ruote in modo concentrico e, a seconda di come venivano girate, ottenere combinazioni meccaniche (v. *Raimondo Lullo. Arte breve*, a cura di M. Romano, Milano 2002).

[4] In argomento, v. *funditus* P. Odifreddi, *Le menzogne di Ulisse. L'avventura della logica da Parmenide ad Amartya Sen*, Milano 2004, 153, cui si rinvia anche per i necessari riferimenti bibliografici.

[5] La citazione del giudice B.N. Cardoso è tratta da G. Canzio, *Intelligenza artificiale e processo penale*, in *CP* 2021, 799.

[6] C. Beccaria, *Dei delitti e delle pene*, Livorno 1764.

[7] C. Lombroso, E. Ferri, R. Garofalo, G. Fioretti, *Polemica in difesa della scuola criminale positiva,* Bologna 1886.

[8] In argomento, v. *ampilus* A. Manna, *Corso di diritto penale*[5], Milano 2020, 5 ss.

[9] A. Manna, *Corso di diritto penale*[5], cit., 5 ss.

Per queste accennate ragioni, è plausibile ipotizzare che, all'epoca della redazione del codice Rocco, l'idea del *calculemus* fosse considerata come un *divertissement* confinato nel puro mondo delle idee, piuttosto che un'implicazione pratica per il diritto (penale).

La probabilità non è, ovviamente, l'unico esempio di concetti "inconsapevolmente" attinti da ambiti extra-penalistici: si pensi, ad esempio, al rapporto di causalità del reato che evoca un principio, in primo luogo scientifico, il principio di causalità, e le teorie penalistiche sul nesso causale sviluppate in dialogo con le scienze dure (come la causalità logico-scientifica di Federico Stella [10]); o ancora, alle evidenze neuroscientifiche, più recentemente addotte da altri studiosi del diritto penale per cercare di penetrare non solo l'imputabilità, ma anche l'elemento soggettivo della colpevolezza[11].

La temuta "contrapposizione" tra cultura umanistica e cultura scientifica [12] si è certamente affievolita nel tempo (come dimostrano gli esempi sopra citati), e ciò può indurre l'interprete del diritto penale odierno a considerare il *calculemus* di Leibniz non più un mero *divertissement* matematico, ma un possibile scenario per il diritto penale. L'idea del calcolo, alla base degli algoritmi predittivi, promette di essere di aiuto con particolare riferimento a quei concetti, come la pericolosità sociale, il cui accertamento è sempre stato lasciato esclusivamente alla "creatività giudiziale", data anche l'assenza di criteri soddisfacenti elaborati delle scienze criminologiche.

A differenza degli altri concetti sopra menzionati, la "misura" della probabilità (di commettere nuovi reati) del soggetto socialmente pericoloso continua a rimanere "insolubile" sul piano del diritto penale, nel senso che, in assenza di una precisa definizione legislativa, gli interpreti si sono affannati a individuare criteri per definirne l'accertamento. Questo sforzo, però, come vedremo, si è rivelato inutile, e anzi, ha finito per convincere gli stessi studiosi del diritto penale che solo la prognosi giudiziale può fornire un significato "in concreto" a tale concetto.

A questa conclusione ha probabilmente portato anche la natura "accessoria" della pericolosità sociale rispetto al reato, e di conseguenza delle misure di sicurezza rispetto alla pena, come sembrerebbe evocare il famigerato sistema del "doppio binario"[13], la

[10] Il riferimento è all'ormai classico F. Stella, *Leggi scientifiche e spiegazione causale nel diritto penale*, Milano 1990.
[11] V. *ex multis* O. Di Giovine, *Ripensare il diritto penale attraverso le (neuro)scienze?*, Torino 2019.
[12] In questo senso, si veda C.P. Snow, *Le due culture*, trad. it., Milano 2005.
[13] Sul "doppio binario" e la sua crisi, v. *funditus* M. Pelissero, *Pericolosità sociale e doppio binario. Vecchi e nuovi modelli di incapacitazione*, Torino 2008.

cui crisi coincide plausibilmente con la stessa nascita della nozione generica di pericolosità sociale[14].

Anche se oggi gli algoritmi alla base dell'intelligenza artificiale (IA)[15] sembrano costituire una "chiave di volta" dall'umanesimo al transumanesimo, ciò non significa una perdita di centralità del giudice-uomo[16], ma piuttosto un possibile potenziamento (*enhancement*), per cui gli interpreti del diritto continueranno a fare il loro lavoro[17], avendo a disposizione *tools* (ad esempio, nel *criminal law*, il *Correctional Offender Management Profiling for Alternative Sanctions COMPAS*), in grado di rendere più oggettiva una decisione sui concetti di diritto penale. Insomma, un'IA "predittiva"[18], ma non "interpretativa", visto che gli stessi interpreti sono chiamati a valutarne le previsioni. Gli algoritmi "calcolano" scenari futuri di maggior o minore probabilità di pericolosità sociale, ma spetta agli interpreti "tradurre" tali previsioni in un quadro significativo e rispettoso delle garanzie dell'imputato.

L'aspetto problematico, quindi, riguarda l'utilizzo di algoritmi predittivi di pericolosità sociale. Se, da un lato, come vedremo, gli algoritmi predittivi possono fornire un calcolo accurato della probabilità di pericolosità sociale, dall'altro lato non mancano possibili "criticità" che potrebbero minare il potenziale effetto di "semplificatore" nel giudizio prognostico.

Gli algoritmi possono aiutare a risolvere il problema del calcolo della probabilità di pericolosità sociale, a condizione che le loro previsioni siano poi anche "interpretate" alla luce del principio di legalità delle misure di sicurezza (artt. 25 co. 3 Cost.). D'altra parte, le previsioni algoritmiche non possono essere il mezzo attraverso il quale giungere a conclusioni contrarie al principio di riserva di legge; altrimenti, si verificherebbe un paradosso, in quanto gli algoritmi predittivi mostrerebbero quella stessa ampia "discrezionalità" che viene spesso criticata nel giudizio prognostico di

[14] M. Pelissero, *Pericolosità sociale e doppio binario*, cit., 107 ss.
[15] M. Boden, *L'intelligenza artificiale*, trad. it., Bologna 2019, 7 ss., osserva che l'IA studia come consentire ai computer (grazie agli algoritmi) di «fare i tipi di cose che le menti possono fare».
[16] Vale la pena di citare, tuttavia, il caso cinese della Procura del Popolo di Shangai Pudong, dove all'IA è affidato un ruolo importante nel processo decisionale, che consisterebbe – secondo le cronache locali – nel valutare con oltre il 97% di accuratezza le prove, le condizioni per un arresto e il grado di pericolosità sociale di un soggetto, per almeno otto dei reati più comuni a Shangai, attraverso l'addestramento di algoritmi su oltre 17000 casi tra il 2015 e il 2020 (v. S. Chen, *Chinese scientists develop AI "prosecutor" that can press its own charges*, in www.scmp.com, 26 dicembre 2021).
[17] Cfr. in questo senso U. Ruffolo, *Giustizia predittiva e machina sapiens come "avvocato generale" ed il primato del giudice umano: una proposta di interazione virtuosa*, in *XXVI lezioni di diritto dell'Intelligenza Artificiale*, a cura di U. Ruffolo, Torino 2021, 205 ss.; più di recente v. G. Pasceri, *La predittività delle decisioni*, Milano 2023.
[18] Cfr. G. Canzio, *Intelligenza artificiale, algoritmi e giustizia penale*, in *www.sistemapenale.it*, 8 gennaio 2021.

pericolosità sociale.

L'obiettivo di questo contributo è quello di evidenziare che il calcolo della probabilità della pericolosità sociale se, da un lato, può contribuire a risolvere "vecchi" problemi attinenti alla soggettività del "metro" di giudizio prognostico, dall'altro non esime gli interpreti dal rilevare le criticità che emergono dall'uso degli algoritmi predittivi e che spingono a un ridimensionamento del calcolo nel diritto penale.

2. La prognosi di pericolosità sociale costituisce da sempre l'esito di un giudizio "ineffabile" perché rimesso all'intuito del giudice sulla base di criteri generici (*ex* art. 133 Cp). La dottrina, ormai da lungo tempo, auspica la riformulazione della pericolosità sociale generica (*ex* art. 203 Cp)[19] in pericolosità sociale specifica (nella stessa direzione in cui il legislatore si è già mosso per i soggetti minorenni)[20], o la sostituzione della pericolosità sociale con il bisogno di cure (con particolare riferimento al reo affetto da disturbo mentale)[21], o ancora la riforma del processo penale in chiave bifasica, separando il momento della decisione da quello della determinazione della sanzione[22].

Tuttavia, le suddette elaborazioni, pur riconoscendo il tasso di incertezza che connota il giudizio prognostico di pericolosità sociale affidato al cosiddetto metodo intuitivo, di cui «la prassi celebra ogni giorno il trionfo»[23], non sembrano in grado di trovare un metodo scientifico più affidabile[24].

---

[19] La vaghezza dei concetti giuridici (come, ad esempio, la pericolosità sociale, di cui come è noto, non esiste una definizione nell'art. 203 Cp), oltre alla volontà del legislatore di non fornire una nozione "a maglie strette", in contrasto con l'evoluzione del diritto penale, potrebbe riflettere una crisi della legge a fronte dello sviluppo esponenziale del ruolo degli organi giurisdizionali (in particolare delle Corti supreme): per un'indagine storica del rapporto tra legge e potere giudiziario, v. *funditus* G. Stella, *Crisi della legge e potere del giudice*, Milano 2020; cfr. nella filosofia del diritto, L. Ferrajoli, *Diritto e ragione. Teoria del garantismo penale*, Bari 1989, 475 ss.; più di recente nella dottrina penalistica, v. F. Palazzo, *Legalità penale vs. creatività giudiziale*, in *RIDPP* 2022, 975 ss.; A. Cadoppi, *Il "reato penale". Teorie e strategie di riduzione della criminalizzazione*, Napoli 2022, spec. 229 ss.

[20] V. *amplius* M. Pelissero, *Pericolosità sociale e doppio binario*, cit., 186 ss.

[21] V. *amplius* A. Manna, *L'imputabilità e i nuovi modelli di sanzione. Dalle "finzioni giuridiche" alla "terapia sociale"*, Torino 1997; M.T. Collica, *Vizio di mente: nozione, accertamento e prospettive*, Torino 2007, spec. 200 ss.; M. Bertolino, *Declinazioni attuali della pericolosità sociale: pene e misure di sicurezza a confronto*, in *AP* 2014, 2, 459 ss. Si veda anche, più di recente, A. Cabiale, *L'accertamento giudiziale della pericolosità sociale fra presente e futuro*, in *Dieci anni di REMS. Un'analisi interdisciplinare*, a cura di M. Pelissero, L. Scomparin, G. Torrente, Napoli 2022, 93 ss., dove si suggerisce, in particolare nei confronti dei potenziali destinatari delle REMS, di spostare in avanti la prognosi criminale, cioè solo dopo l'accertamento del fatto e non contestualmente, poiché solo allora ci sarebbero elementi certi su cui basarla.

[22] V. *amplius* A. Procaccino, *Pericolosità sociale (accertamento della)*, in *DigDPen.*, II, Torino 2005, 1051 ss.

[23] L. Monaco, *Prospettive dell'idea dello "scopo" nella teoria della pena*, Napoli 1984, 144.

[24] In questo senso v. M. Pelissero, *Pericolosità sociale e doppio binario*, cit., 112; cfr. A. Martini, *Essere pericolosi*, Torino 2017, spec. 165, dove si osserva che «affidare il destino ad un'interpretazione del senso comune è però esercizio estremamente pericoloso quando entrano in gioco valori fondamentali; neppure condividendo

L'uso di algoritmi predittivi sarebbe in grado di consentire la formulazione di un giudizio prognostico, con un grado di affidabilità maggiore rispetto ai metodi tradizionalmente elaborati dalle scienze criminologiche[25]. Queste ultime, a fronte della «crisi profonda che investe sia il fondamento che la concreta praticabilità della pericolosità sociale»[26], hanno suggerito di affiancare al metodo intuitivo ulteriori criteri che si sono rivelati, però, inaffidabili per garantire un'esecuzione almeno uniforme delle prognosi di pericolosità, a causa dell'incompletezza della base prognostica (v. il metodo statistico), e di un "marcato soggettivismo" in contrasto con le esigenze di certezza (v. il metodo clinico)[27], o comunque di costi non indifferenti (v. il metodo combinato)[28].

Gli algoritmi predittivi di pericolosità sociale potrebbero fornire un modello predittivo basato sull'apprendimento dei dati relativi alla personalità del reo. L'utilità di questo modello risiederebbe nel maggior grado di accuratezza del calcolo della probabilità dell'imputato incensurato che ha commesso un fatto grave, o del reo, di commettere nuovi reati. In altre parole, gli algoritmi predittivi sarebbero capaci di arrivare a una misura più precisa di questa probabilità[29], invece di essere lasciata all'intuizione del giudice, «se non al suo incontrollabile *arbitrium*»[30], sulla base degli elementi indicati nell'art. 133 Cp che letti in chiave prognostica «non forniscono criteri precisi per delimitare la prognosi di pericolosità, in quanto richiamano una complessità di fattori molto ampia che impone di tener conto di elementi desunti dal fatto di reato commesso e dagli elementi inerenti alla personalità dell'autore»[31].

---

l'affermazione di Calamandrei, per la quale "il giudice è il diritto fatto uomo", appare lecito accontentarsi della sola garanzia della natura giurisdizionale dell'accertamento».

[25] Sui metodi tradizionalmente elaborati dalle scienze criminologiche, v. G. Kaiser, *Criminologia*, trad. it., Milano 1985.

[26] In argomento, v. *funditus* M. Romano, G. Grasso, T. Padovani, *sub art. 203 Cp*, in *Commentario sistematico del codice penale*, II, Milano 2011, 469.

[27] In argomento v. *ex plurimis*, L. Fornari, *Misure di sicurezza e doppio binario: un declino inarrestabile?*, in *RIDPP* 1993, 586; M. Pelissero, *Pericolosità sociale e doppio binario*, cit., 110 ss.; Id., *sub art. 203 Cp*, in C*odice penale commentato*, a cura di E. Dolcini, G.L. Gatta, Milano 2021, 2608

[28] In questo senso, v, M. Pelissero, *Pericolosità sociale e doppio binario*, cit., 112,

[29] In questo senso, cfr. A.M. Maugeri, *L'uso di algoritmi predittivi per accertare la pericolosità sociale: una sfida tra evidence based practices e tutela dei diritti fondamentali*, in *AP* 2021, 1, che, però, evidenzia anche i rischi connessi all'uso di algoritmi predittivi come discriminazioni, valutazioni generalizzanti e mancanza di trasparenza.

[30] Così F. Rocchi, *La recidiva tra colpevolezza e pericolosità*, Napoli 2020, 66.

[31] M. Pelissero, *sub art. 203 Cp*, in *Codice penale commentato*, cit., 2603-2604. Sulle plurime applicazioni del giudizio prognostico di pericolosità sociale, v. *amplius* F. Basile, *Esiste una nozione ontologicamente unitaria di pericolosità sociale? Spunti di riflessione, con paricolare riguardo alle misure di sicurezza e alle misure di prevenzione*, in *RIDPP* 2018, 644 ss.

Sul piano di teoria generale, l'utilizzo di un modello predittivo potrebbe attribuire maggiore consistenza scientifica alla categoria stessa della pericolosità sociale, considerato «il *deficit* di determinatezza nella formulazione del giudizio di pericolosità sociale (e conseguentemente nell'applicazione delle misure di sicurezza) che investe sia la base per la formulazione del giudizio di pericolosità (individuata in modo inadeguato nell'art. 133 Cp), sia i criteri per la formulazione (affidati all'intuizione del giudice), sia la determinazione dello stesso grado di possibilità rilevante»[32].

L'analisi predittiva sarebbe plausibilmente più accurata in quanto si concentrerebbe sui dati della personalità del reo (come i dati anagrafici, sensibili e giudiziari, ma anche quelli relativi alle comunicazioni elettroniche e i dati di geolocalizzazione), in grado di fornire informazioni preziose per calcolare la probabilità di pericolosità sociale.

Inoltre, l'utilizzo del modello predittivo non sembra contrastare con il divieto di perizia criminologica (ex art. 220 co. 2 Cpp), se si aderisce all'interpretazione dottrinale, rimasta minoritaria, secondo cui la clausola di salvezza della citata disposizione («salvo quanto previsto ai fini dell'esecuzione della pena o della misura di sicurezza») si applicherebbe anche alla valutazione della pericolosità sociale[33]. Un'interpretazione, quest'ultima, che non è però condivisa da chi sostiene che proprio il divieto di perizia criminologico, superabile solo in presenza di patologie psichiche, «spiega perché l'accertamento della pericolosità si fondi sul metodo intuitivo, che solo apparentemente è diretto dai criteri indicati dall'art. 133 Cp, come varrebbe l'art. 203 Cp»[34].

3. Nella dottrina tedesca, Cristoph Burchard sostiene che l'IA potrebbe decretare la "fine del diritto penale", nel senso di segnare la scomparsa del diritto penale liberale e l'affermazione di un diritto penale securitario irragionevole per gli esseri umani[35], con il rischio di una compressione del principio di presunzione di innocenza[36]. Il diritto

[32] In questo senso v. M. Romano, G. Grasso, T. Padovani, *sub art. 203,* in *Commentario sistematico del codice penale*, cit., 470 ss., cui si rinvia anche per i necessari riferimenti bibliografici.
[33] In questo senso v. ancora M. Romano, G. Grasso, T. Padovani, *sub art. 203*, in *Commentario sistematico del codice penale*, cit., 469.
[34] M. Pelissero, *sub art. 203 Cp*, in *Codice penale commentato*, cit., 2608.
[35] Sulle implicazioni degli algoritmi di IA nel processo con particolare riferimento ai diritti umani, v. *ex multis*, J. Nieva-Fenoll, *Intelligenza artificiale e processo*, trad. it., Torino 2019, spec. 118 ss.
[36] C. Burchard, *Künstliche Intelligenz als Ende des Strafrechts? Zur algorithmischen Transformation der Gesellschaft*, in *Jahrbuch für Recht und Ethik / Annual Review of Law and Ethics*, a cura di J.C. Schuhr, J. C. Joerden, Berlino 2019, 527 ss., trad. it., *L'intelligenza artificiale come fine del diritto penale? Sulla trasformazione algoritmica della società*, in *RIDPP* 2019, 1909 ss.

penale si concentra tipicamente su atti che sono stati attuati o la cui attuazione è almeno iniziata immediatamente: un atteggiamento contrario alla norma, finché non si concretizza in un'azione, rimane di solito impunito (*cogitationis poenam nemo patitur*); se la sfera dei pensieri dell'autore del reato fosse accessibile a una valutazione algoritmica automatizzata, la soglia della punibilità rischierebbe di abbassarsi minacciosamente[37].

Per evitare il possibile scenario sopra descritto, Lucia Sommerer propone due fondamentali raccomandazioni per i requisiti minimi degli algoritmi di *law enforcement*, e in particolare di "polizia predittiva" (*on the person-based predictive policing*, *PPP*)[38]. In primo luogo, l'imparzialità degli algoritmi di *PPP* deve essere garantita da una specifica ricerca di discriminazioni nel *data set* di "addestramento" già nella fase di sviluppo[39]; in secondo luogo, l'introduzione del "triangolo della trasparenza" che consiste nella registrazione pubblica obbligatoria di questi *tools*, nella tutela dei diritti soggettivi delle persone interessate, e nell'istituzione di un organismo di controllo statale[40].

Imparzialità e trasparenza devono, a maggior ragione, essere salvaguardate in relazione al possibile utilizzo in ambito penale di algoritmi predittivi di pericolosità sociale, le cui valutazioni, se non opportunamente interpretate, potrebbero ledere le garanzie costituzionali del reo (artt. 27 co. 1 e 3 Cost.)[41]. Il rischio sarebbe quello di far prevalere la funzione di prevenzione speciale, nel senso originario di incapacitazione

[37] In questo senso, v. C. Thimm, T.C. Bächle, *Die Maschine: Freund oder Feind? Mensch und Technologie im digitalen Zeitalter*, Berlino 2019, 175.
[38] L.M. Sommerer, *Personenbezogenes Predictive Policing. Kriminalwissenschaftliche Untersuchung über die Automatisierung der Kriminalprognose*, Baden-Baden 2020, 343.
[39] L.M. Sommerer, *Personenbezogenes Predictive Policing*, cit., 345.
[40] L.M. Sommerer, *Personenbezogenes Predictive Policing*, cit., 348.
[41] Sull'impiego di algoritmi predittivi nella prognosi di pericolosità sociale in rapporto ai diritti fondamentali v. *amplius* A.M. Maugeri, *L'uso di algoritmi predittivi per accertare la pericolosità sociale*, cit.; V. Manes, *L'oracolo algoritmico e la giustizia penale: al bivio tra tecnologia e tecnocrazia*, in *Discrimen*, 15 maggio 2020; G. Ubertis, *Giustizia penale e nuove tecnologie*, in *DPenCont* 2020, 4, spec. 81 ss. Con particolare riferimento all'impiego di *risk assessment tools* tra Stati Uniti e Europa, v. M. Gialuz, *Quando la giustizia penale incontra l'intelligenza artificiale: luci e ombre dei risk assessment tools tra Stati Uniti ed Europa*, in www.penalecontemporaneo.it, 29 maggio 2019; P. Severino, *Intelligenza artificiale*, Roma 2022, spec., 77 ss. Cfr. nella dottrina di *criminal law*, M. Brenner et al., *Constitutional Dimensions of Predictive Algorithms in Criminal Justice*, in *Harv. C.R.-C.L. L. Rev.*, 2020, 267 ss.; A. Nishi, *Privatizing sentencing: a delegation framework for recidivism risk assessment*, in *Colum. L. Rev.*, 119, 2019, 1671 ss., dove si sostiene che la natura privata di molti algoritmi di valutazione del rischio di recidiva rende i giudici incapaci di comprendere e applicare correttamente i loro risultati, portando a un maggiore affidamento sulle decisioni politiche degli sviluppatori privati; per questo motivo, i legislatori devono rafforzare gli statuti sulla valutazione del rischio di recidiva, aumentando la capacità dei giudici di comprendere e applicare i punteggi di rischio algoritmici.

del soggetto pericoloso, anche in considerazione della mancanza di un "contrappeso" garantista, ossia della previsione di un limite massimo edittale e di un rapporto di proporzione della misura di sicurezza con la prognosi di pericolosità[42].

Questo atteggiamento di prudenza, rispettoso dei principi fondamentali di garanzia in materia penale[43], e al tempo stesso volto a "bilanciare" la funzione di prevenzione speciale delle misure di sicurezza ispirate al principio di pericolosità, non deve, però, indurre l'interprete a stigmatizzare l'IA applicata al diritto penale. Per questo motivo, non appare condivisibile la posizione assunta nella dottrina tedesca da Karsten Gaede, secondo il quale dovrebbe essere esplicitamente vietata la costruzione di sistemi autonomi di IA che mettano in pericolo le garanzie e i diritti fondamentali, e si richiede quindi una lungimiranza europea contro la dannosa prosecuzione della ricerca sull'IA[44]. Lungimiranza europea che, in linea di principio, non è mancata vista l'approvazione della Carta etica sull'uso dell'IA all'interno dei sistemi giudiziari da parte della Commissione per l'efficienza della giustizia del Consiglio d'Europa (*CEPEJ*)[45], la proposta della Commissione europea di regolamento sull'IA[46] e, infine, l'approvazione del Parlamento europeo di una risoluzione sull'uso trasparente dell'IA nel diritto penale[47].

4. Ci si potrebbe chiedere, a questo punto, se gli algoritmi predittivi, anche se in ipotesi non lesivi dei diritti fondamentali degli imputati, non siano visti dai giudici come *tools* in qualche modo "indesiderabili" a causa del loro impatto sul potere discrezionale di valutare la pericolosità sociale: l'efficienza algoritmica potrebbe costituire un "cavallo di Troia" per fare breccia in una "sfera di dominio giudiziario". In effetti, tale "resistenza" all'uso degli algoritmi predittivi è evidenziata da una dottrina di *criminal law* che getta nuova luce sul ruolo duraturo della discrezionalità

[42] In senso critico, v. *ampilus* A. Manna, *Corso di diritto penale*, cit., 358, cui si rinvia anche per i riferimenti bibliografici.
[43] Cfr. N. Mazzacuva, *Alcune riflessioni su Intelligenza Artificiale e diritto penale sostanziale*, in *XXVI lezioni di diritto dell'Intelligenza Artificiale*, cit., 287 ss.
[44] K. Gaede, *Künstliche Intelligenz – Rechte und Strafen für Roboter? Plädoyer für eine Regulierung künstlicher Intelligenz jenseits ihrer reinen Anwendung,* Baden-Baden 2019, 81 ss.
[45] V. *amplius* S. Quattrocolo, *Intelligenza artificiale e giustizia: nella cornice della Carta etica europea gli spunti per un'urgente discussione tra scienze penali e informatiche*, *in LP* 18 dicembre 2018.
[46] V. *amplius* P. Troncone, *Il sistema dell'Intelligenza artificiale nella trama grammaticale del diritto penale. Dalla responsabilità umana alla responsabilità delle macchine pensanti: un inatteso return trip effect*, in *CP* 2022, 3287 ss.
[47] V. *amplius* G. Barone, *Intelligenza artificiale e processo penale: la linea dura del Parlamento europeo. Considerazioni a margine della risoluzione del Parlamento europeo del 6 ottobre 2021*, in *CP* 2022, 1180 ss.

giudiziaria nel plasmare l'impatto sociale e politico delle tecnologie algoritmiche[48].

La discrezionalità giudiziaria sembrerebbe, tuttavia, correre il rischio di essere, in qualche misura, automatizzata con l'uso di algoritmi predittivi, come osservato dalla filosofa americana Dasha Pruss, secondo la quale i giudici statunitensi potrebbero affidarsi acriticamente a questi *tools*, dando vita a una *Mechanical Jurisprudence*[49]. L'automatizzazione delle sentenze sarebbe una conseguenza del fatto, secondo Dasha Pruss, che gli algoritmi non sono dinamici o interpretativi, ma forniscono la stessa previsione per quanto riguarda, ad esempio la pericolosità sociale e il rischio di recidiva, precludendo ai giudici la possibilità di reinterpretare dinamicamente le norme giuridiche quando cambia il contesto individuale, familiare e sociale dell'imputato[50]. Il rischio di una *Mechanical Jurisprudence* potrebbe, tuttavia, essere mitigato dal fatto che i recenti algoritmi di apprendimento automatico (*Machine Learning*, *ML*)[51], e di apprendimento profondo (*Deep Learning*, una sotto-tecnica di *ML*) sembrano, secondo alcuni esperti del settore, essere una "transizione" da un'IA *debole*, che si limita a simulare il pensiero umano, a un'IA di "livello umano" o

[48] S. Brayne, A. Christin, *Technologies of Crime Prediction: The Reception of Algorithms in Policing and Criminal Courts*, in *Soc. Probl.*, 68, 2021, 608 ss. Cfr. R. Simmons, *Big Data and Procedural Justice: Legitimizing Algorithms in the Criminal Justice System*, in *Ohio St. J. Crim.*, 2018, 573 ss., dove si osserva che le tecnologie algoritmiche si diffonderanno nel sistema di *criminal law* se verranno accettate dall'opinione pubblica.

[49] D. Pruss, *Mechanical Jurisprudence and Domain Distortion: How Predictive Algorithms Warp the law*, in *Philos. Sci.*, 88(5), 2021, 1101 ss., secondo cui l'uso di algoritmi predittivi offuscherebbe anche la tradizionale distinzione di *criminal law* tra il dominio della determinazione della colpevolezza (*liability*) e quello della condanna (*sentencing*), poiché gli algoritmi predittivi prendono esplicitamente in considerazione le valutazioni di responsabilità future quando decidono le sentenze per le valutazioni di responsabilità attuali; l'uso di algoritmi per il processo decisionale giuridico di *criminal law* renderebbe necessariamente l'interpretazione giuridica un'impresa formalistica.

[50] D. Pruss, *Mechanical Jurisprudence and Domain Distortion*, cit., 1101.
Sulla fondamentale componente interpretativa nel *criminal law*, v. *funditus* R. Dworkin, *Law's Empire*, Cambridge (MA) 1986.

[51] Sugli algoritmi di *ML* v. *amplius* P. Domingos, *L'algoritmo definitivo*, trad. it., Torino 2015.
Secondo alcuni, il *ML* fa parte dell'informatica e/o della statistica e non dell'IA, ma, secondo altri, non ci sono confini netti in questi campi. In argomento, v. *amplius* M. Boden, *L'intelligenza artificiale*, cit., 47 ss., dove si osserva che il *ML* ha tre ampie tipologie: l'apprendimento supervisionato, l'apprendimento non supervisionato e l'apprendimento per rinforzo. Nell'apprendimento supervisionato, «il programmatore "addestra" il sistema definendo un insieme di risultati attesi per una data gamma di input (chiamati esempi e non-esempi) e fornendo continue valutazioni del raggiungimento o meno dei risultati». Nell'apprendimento non supervisionato, «l'utilizzatore non fornisce né risultati attesi né messaggi di errore; l'apprendimento è guidato dal principio secondo cui i tratti che cooccorrono generano aspettative sul fatto che cooccorreranno in futuro». Infine, l'apprendimento per rinforzo «è guidato da analoghi della ricompensa e della punizione: messaggi di feedback che dicono al sistema cosa è stato fatto bene e cosa è stato fatto male». Sugli algoritmi di *ML*, v. anche P. Crescenzi, L. Pagli, *Problemi, algoritmi e coding*, Bologna 2017.

"generale", con la capacità di "spiegare" le proprie previsioni agli agenti umani[52]. Secondo un'efficace espressione utilizzata in riferimento al *criminal justice*, l'IA sarebbe capace di ridurre il "rumore" (*Noise*), "un difetto del giudizio umano" (*A Flaw in Human Judgement*), derivante dalla mancanza di una verità oggettiva nel processo decisionale (a differenza delle scienze dure), che spesso finisce per tradursi in un'ampia discrezionalità giudiziaria[53].

5. Una preoccupazione ricorrente riguardo ai suddetti algoritmi di *ML* è che essi operano come *black boxes* ("scatole nere"), rendendo difficile identificare come e perché raggiungono particolari decisioni, raccomandazioni o predizioni[54].

Per illustrare il problema dell'opacità algoritmica è opportuno confrontare un *tool* di valutazione quantitativa del rischio (*quantitative risk assessment*) basato su un approccio statistico, come l'*Oxford Risk of Recidivism Tool* (*OxRec*), con uno fondato su *ML*, come il già citato *COMPAS*[55].

OxRec è un *tool* che aiuta a prevedere il rischio di recidiva in base a un'analisi statistica dei dati di oltre 47.000 detenuti rilasciati dal sistema carcerario svedese tra il 2001 e il 2009; utilizzando 14 caratteristiche, che di solito possono essere ottenute senza problemi dai fascicoli, con *OxRec* viene calcolato un punteggio di rischio, che indica la probabilità di commettere un nuovo reato violento entro uno o due anni in valori percentuali[56]. Questo *tool* vanta un'affidabilità predittiva relativamente buona ed è stato validato anche nei Paesi Bassi[57]. Per l'assegnazione del punteggio non sono necessarie né valutazioni soggettive né procedure di test psicologici o interviste. Non è nemmeno richiesta una formazione precedente. Solo alcune caratteristiche richiedono una diagnosi più dettagliata o una valutazione da parte di un esperto (ad esempio, il consumo di alcol e droghe e la salute mentale della persona sottoposta al test), ma queste sono regolarmente già registrate nei fascicoli dei detenuti dopo

[52] Sulla più recente denominazione di IA "di livello umano" o "generale", v. *amplius* S.J. Russell, P.Norvig, *Artificial Intelligence. A Modern Approach*[4], New York 2020, trad. it. *Intelligenza artificiale. Un approccio moderno*, I, a cura di F. Amigoni, Milano 2021. Cfr. A. Lieto, *Cognitive Design for Artificial Minds*, Londra 2021.
[53] V. *funditus* D. Kahneman, O. Sibony, C.R. Sunstein, *Noise. A Flaw in Human Judgment*, New York 2021.
[54] Cfr. *amplius* W.A. Mostowy, *Explaning Opaque AI Decisions, legally*, in *BTLJ*, 35, 2020, 1291 ss.
[55] Sul confronto tra *COMPAS* e *OxREC*, v. *amplius* G. van Dijck, *Predicting Recidivism Risk Meets AI Act*, in *EurJCrim.*, 28, 2022, 407 ss.
[56] V. H.E. Muller, *„Moneyball" in der Strafrechtspraxis?* , in *Für die Sache - Kriminalwissenschaften aus unabhängiger Perspektive* 2019, 97 ss.
[57] S. Fazel et al., *Prediction of violent reoffendering in prisoners and individuals on probation: A Dutch validation study* (*OxRec*), in *Sci. Rep.*, 2019, 9:841.

precedenti diagnosi. Poiché lo strumento *on line* è programmato in modo tale che i risultati vengano visualizzati immediatamente e senza ritardi, è possibile verificare quali caratteristiche hanno una maggiore o minore influenza sulla valutazione del rischio modificando i singoli fattori[58].

COMPAS, invece, è un *software* che è stato utilizzato da alcuni tribunali statunitensi per valutare il rischio di recidiva[59]. La differenza con un approccio statistico è che l'IA comprende i modelli e addestra gli algoritmi da sola, mentre gli approcci statistici si basano su concetti matematici per trovare modelli nei dati definiti dal ricercatore[60].

COMPAS è stato progettato per essere ben calibrato: tutti gli individui a cui l'algoritmo assegna lo stesso punteggio dovrebbero avere approssimativamente la stessa probabilità di recidiva, indipendentemente dall'etnia. Ad esempio, tra tutte le persone a cui il modello assegna un punteggio di rischio di 7 su 10, il 60% dei bianchi e il 61% dei neri commette un reato. I progettisti sostengono quindi di aver raggiunto l'obiettivo di equità desiderato. D'altra parte, il *COMPAS* non raggiunge le pari opportunità: la percentuale di coloro che non hanno recidivato ma sono stati ingiustamente classificati come ad alto rischio è stata del 45% per i neri e del 23% per i bianchi[61].

La valutazione predittiva operata da *COMPAS* è stata criticata per la mancanza di testabilità e contestabilità, per l'utilizzo di dati aggregati di gruppo per valutare il rischio di recidiva, in violazione delle garanzie del *due process*[62]. Poiché il *software* è di proprietà privata, i dati e gli algoritmi non sono trasparenti, nemmeno per il

[58] H.E. Muller, *„Moneyball" in der Strafrechtspraxis?* , cit.
[59] G. van Dijck, *Predicting Recidivism Risk Meets AI Act*, cit., 409.
[60] G. van Dijck, *Predicting Recidivism Risk Meets AI Act*, cit., 408.
[61] In questo senso, v. *amplius* S.J. Russell, P.Norvig, *Artificial Intelligence*, cit., dove si osserva che, come dimostrato nella letteratura scientifica (v. *funditus* J. Kleinberg et al., *Inherent trade-offs in the fair determination of risk scores*, in *arXiv*:1609.05807), è impossibile che un algoritmo sia al contempo ben calibrato e con pari opportunità: se le classi di base sono diverse, qualsiasi algoritmo ben calibrato non fornirà necessariamente pari opportunità e viceversa.
[62] In questo senso, v. C. McKay, *Predicting risk in criminal procedure: Actuarial tools, algorithms, AI and judicial decision-making*, in *Curr. Issues Crim.* 2020, 32(1), 22 ss.

giudice[63]. Questa opacità algoritmica[64] è stata affrontata nell'ormai noto caso *Loomis v. Wisconsin*[65] dove il ricorrente ha sostenuto di avere il diritto di essere condannato sulla base di informazioni accurate e che il fatto di non sapere come *COMPAS* avesse calcolato il suo punteggio di previsione violava questo diritto[66]. Se, da un lato, la Corte Suprema del Wisconsin ha respinto le argomentazioni di *Loomis*, affermando che la condanna non sarebbe stata diversa senza il *COMPAS*, dall'altro lato, ha lanciato diversi avvertimenti ai giudici, da cui traspare un certo scetticismo sull'accuratezza dell'algoritmo, in particolare per quanto riguarda i pregiudizi nei confronti degli imputati appartenenti a minoranze etniche[67]. In breve, l'uso di algoritmi predittivi nel

[63] Cfr. per una spiegazione matematica dell'opacità algoritmica A. Vespignani, *L'algoritmo e l'oracolo*, Milano 2019, 37, dove si evidenzia che «la discussione sull'imparzialità algoritmica è un problema delicato che coinvolge aspetti teorici e matematici, e soprattutto la definizione del tipo di imparzialità più importante da proteggere a seconda dell'uso che si fa delle predizioni nei diversi contesti; la letteratura scientifica sull'argomento ha cominciato quindi a riflettere sul concetto di giustizia stessa, a come identificare ed eliminare classificatori ingiusti; problemi non semplici per via della segretezza e non trasparenza attorno agli algoritmi, in parte dettati da motivi di *copyright* e in parte da una reale incapacità di interpretare i meccanismi e i processi di apprendimento degli algoritmi stessi».

[64] Cfr. per una differente interpretazione, nella dottrina di *criminal law*, M. Selmi, *Algorithms, Discrimination and the Law*, in *Ohio St. L.J.*, 82, 2021, 611 ss., dove si osserva che la grande maggioranza degli algoritmi non sono *black-box*, ma procedure complicate con parti identificabili che assomigliano in gran parte a modelli statistici complessi, come le regressioni, che sono stati analizzati per decenni dalle Corti americane; v. anche A.M. Carlson, *The Need for Transparency in the Age of Predicting Sentencing Algorithms*, in *Iowa L. Rev.*, 2017, 303 ss.
Sull'opacità tecnologica dell'IA, con particolare riferimento al diritto penale, v. *funditus* O. Di Giovine, *Il judge-bot e le sequenze giuridiche in materia penale (Intelligenza Artificiale e stabilizzazione giurisprudenziale*), in *CP*, 2020, 951 ss.; I. Salvadori, *Agenti artificiali, opacità tecnologica e distribuzione della responsabilità penale*, in *RIDPP* 2021, 83 ss. Cfr. M. Caterini, *Il giudice penale robot*, in *www.legislazionepenale.eu*), 19 dicembre 2020, dove si osserva che il giudice può sempre confutare alla luce del principio dell'"oltre il ragionevole dubbio" (*Beyond Any Reasonable Doubt* - BARD) quanto suggerisce il robot secondo la logica del "più probabile che non"; v. anche B. Occhiuzzi, *Algoritmi predittivi: alcune premesse metodologiche*, in *DPenCont*, 2019, 2, spec., 398-399, dove si osserva che «la parola del giudice rivela la propria appartenenza ad un linguaggio di tipo *valutativo* e *performativo*, ben lungi dall'essere limitato a risultanze computazionali e probabilistiche».

[65] Wisconsin Supreme Court, *State v. Loomis*, case 2015AP17-CR, 13 July 2016, in *Harvard L.R.*, 130, 2017, 1530 ss. I fatti in breve. Nel 2013 Loomis veniva arrestato dagli agenti di polizia per resistenza a pubblico ufficiale e ricettazione, dopo essere stato trovato alla guida di un'auto rubata ed impiegata in una precedente sparatoria nello Stato americano del Wisconsin. Loomis era successivamente processato e condannato dalla *Trial Court* a cinque anni di reclusione e successivi cinque anni di sorveglianza speciale, dopo che il giudice aveva esaminato anche l'alto rischio di recidiva, previsto dal *COMPAS* (sul caso Loomis v. *amplius* F. Basile, *Intelligenza artificiale e diritto penale: quattro possibili percorsi di indagine*, in *DPU* 2019, 10, spec. 21 ss.; Id., *Intelligenza artificiale e diritto penale: qualche aggiornamento e qualche nuova riflessione*, in *Diritto penale e intelligenza artificiale. "Nuovi scenari"*, a cura di G. Balbi, F. De Simone, A. Esposito, S. Manacorda, Torino 2022, 12.

[66] Cfr. G. van Dijck, *Predicting Recidivism Risk Meets AI Act*, cit., 409, dove si evidenzia che, se applicato in modo appropriato, *COMPAS* può migliorare la valutazione, la ponderazione e l'applicazione delle altre prove da parte del giudice nella formulazione di un programma di condanna personalizzato.
Sul rilievo del diritto di contestare le previsioni algoritmiche, anche nel contesto giudiziario, v. *amplius* M. Kaminski, J.M. Urban, *The right to contest AI*, in *Col. L. Rev.*, 121, 2021, 1957 ss.

[67] Wisconsin Supreme Court, *State v. Loomis*, case 2015AP17-CR, 13 July 2016, cit.

sistema di *criminal justice* statunitense[68], costituisce solo un'informazione aggiuntiva, non l'unica base per il processo decisionale, e per questo motivo il *due process* non viene violato[69].

6. I *bias* algoritmici sono errori che possono derivare esplicitamente da *bias* cognitivi degli informatici che devono sviluppare l'algoritmo, o implicitamente dai dati utilizzati per "addestrare" gli algoritmi di *ML*[70]. In altre parole, il potenziale di *ML* comporta la possibilità di abusi, usi impropri e conseguenze indesiderate che ne compromettono l'equità[71].

È importante comprendere il passaggio "circolare" dai *bias* cognitivi, trappole conoscitive innescate da decisioni veloci (le cosiddette euristiche)[72], ai *bias* algoritmici che creano oggi le discriminazioni tecnologiche, meno tangibili, e di difficile comprensione, perché "annidate" nelle equazioni matematiche e nei codici informatici[73].

I *bias* cognitivi spesso finiscono per inficiare il ragionamento umano, come ad esempio, l'*automation bias* che si riferisce alla tendenza umana ad assegnare livelli più elevati di autorità e fiducia alle fonti automatizzate rispetto a quelle non automatizzate[74], oppure l'*hindsight bias,* che induce le persone a prevedere gli eventi del passato come più prevedibili di quanto fossero in realtà (distorsione del "senno di

[68] Nel sistema di *criminal justice* statunitense, l'uso di *tool* è diffuso già nella fase della custodia cautelare *ante* processo, come, ad esempio, il *PTRA* (*Pre Trial Risk Assessment*) e il *PSA* (*Public Safety Assessment*), v. *amplius* M. Hamilton, *Evaluating Algorithmic Risk Assessment*, in *NewCrimLRev*, 24, 2021, 156 ss.; S. Levmore, F. Fagan, *Competing Algorithms for Law: Sentencing, Admissions, and Employment*, in *U. Chi. L. Rev*., 88, 2021, 367 ss. Cfr. nella dottrina processual-penalistica, S. Quattrocolo, *La giustizia penale*, in *La politica dei dati. Il governo delle nuove tecnologie tra diritto, economia, società*, a cura di M. Durante, U. Pappagallo, Milano 2022, 338, dove si osserva che «a differenza della pericolosità sociale», dove non mancano elaborazioni delle scienze criminologiche (v. *supra* nel testo), «non esistono studi significativi circa parametri e criteri che determinino una significativa correlazione con il rischio di fuga e il rischio di inquinamento probatorio. Se [...] la tradizione degli studi criminologici ha svolto, nel tempo, approfonditi e significativi studi empirici sul comportamento criminale, l'elaborazione di convincenti teorie circa l'attitudine di un soggetto alla fuga o alla subornazione dei testi (per esempio) pare lontana. Ciò, tuttavia, non ha impedito l'elaborazione di software di risk assessment per la fase cautelare, che soprattutto negli Stati Uniti stanno conoscendo un certo successo: il PTRA e il PSA».

[69] D. Pruss, *Mechanical Jurisprudence and Domain Distortion*, cit., 1105.

[70] J. Aurélie, *Nel paese degli algoritmi*, trad. it., Verona 2019.

[71] Per una dimostrazione informatica, v. *ex multis* S. Burocas, M. Hardt, A. Narayanan, *Fairness and Machine Learning. Limitations and Opportunities*, in www.fairmlbook.org, 2019.

[72] Sui *bias* cognitivi v. *funditus* D. Kahneman, *Pensieri lenti e veloci*, trad. it., Milano 2018, spec. 145 ss.

[73] J. Aurélie, *Nel paese degli algoritmi*, cit., 40.

[74] Sull'*automation bias*, v. *amplius* A.L. Park, *Injustice Ex Machina: Predictive Algorithms in Criminal Sentencing." UCLA Law Rev.*, 19 febbraio 2019.

poi")[75].

Il fenomeno dei bias cognitivi nel settore giudiziario merita di essere chiarito. Uno studio del 2011 condotto da giuristi anglosassoni suggerisce che le sentenze possono essere influenzate da variabili estranee, come fattori psicologici, politici e sociali, che non dovrebbero avere alcuna influenza sulle decisioni giudiziarie; ciò rafforza il crescente numero di prove che indicano la suscettibilità dei giudici esperti ai *biases* psicologici [76]. È stato osservato che, nelle decisioni sequenziali di giudici di sorveglianza esperti, la percentuale di decisioni favorevoli scende gradualmente dal 65% a quasi zero all'interno della sessione decisionale e ritorna bruscamente al 65% dopo una pausa. Ciò sembra contraddire con una certa dose di realismo giuridico l'assunto formalista secondo cui le decisioni giudiziarie si basano esclusivamente sull'applicazione razionale delle leggi ai fatti[77].

Anche i *bias* algoritmici esistono e sono inevitabili, data la complessità dei processi di apprendimento che li amplificano [78]. Nel sistema di *criminal justice* sono stati evidenziati i possibili *bias* di algoritmi predittivi (come il più volte citato *COMPAS*)[79], con potenziali effetti discriminatori o comunque causa di disuguaglianze [80]. Ciò conferma che le valutazioni algoritmiche non sono necessariamente più eque di quelle umane[81].

---

[75] Per un inquadramento dell'*hindsight bias* nella psicologia, v. *funditus* U. Hoffrage, R.F. Pohl, *Research on hindsight bias: A rich past, a productive present, and a challenging future*, in *Memory*, 11, 2003, 329 ss.
Sui *bias* cognitivi dei giudici, in particolare l'*hindsight bias*, v. *amplius* A. Forza, *La psicologia nel processo penale*², Milano 2018, spec. 422 ss.; S. Arcieri, *Bias cognitivi e decisione del giudice: un'indagine sperimentale*, in *DPU* 2019, 4, 83 ss.; v. anche O. Di Giovine, *Dilemmi morali e diritto penale*, cit., spec. 112 ss.

[76] V. *amplius* S. Danziger, J. Levav, L. Avnaim-Pesso, *Extraneous factors in judicial decisions*, in *PNAS*, 17, 2011, 6889 ss.

[77] S. Danziger, J. Levav, L. Avnaim-Pesso, *Extraneous factors in judicial decisions*, cit., 6889 ss.

[78] J. Aurélie, *Nel paese degli algoritmi*, cit., 155; M. Spielkamp, *Inspecting Algorithms for Bias*, in *MIT Technol. Rev.*, 2017, 120, 96 ss.; J. Ludwig, S. Mullainathan, *Fragile Algorithms and Fallible Decision-Makers: Lessons from the Justice System*, in *JEP*, 35, 2021, 71 ss.

[79] Per un'indagine giornalistica rigorosa sui possibili *bias* di *COMPAS* nei confronti degli individui afro-americani nel sistema di giustizia penale statunitense, v. J. Angwin, J. Larson, S. Mattu, L. Kirchner, *Machine Bias*, in www.propublica.org, 23 maggio 2016.

[80] Sui potenziali *bias* di *COMPAS* nella dottrina di *criminal law*, v. *ex plurimis* S.B. Starr, *Evidence-Based Sentencing and the Scientific Rationalisation of Discrimination*, in *SLR*, 66(4), 2014, 803 ss.; J. Dressel, H. Farid, *The accuracy, fairness, and limits of predicting recidivism*, in *Sci. Adv.*, 4(1), 2018. Più di recente, sui potenziali *bias* di *OxRec*, v. G. van Dijck, *Predicting Recidivism Risk Meets AI Act*, cit., spec., 417 ss.

[81] Cfr. nella dottrina penalistica O. Di Giovine, *Dilemmi morali e diritto penale. Istruzioni per un uso giuridico delle emozioni*, Bologna 2022, 19-20 che, con riferimento al dibattito suscitato negli Stati Uniti dall'algoritmo COMPAS, osserva che «non si può esserne certi (un altro problema degli algoritmi – su cui però si sta lavorando – è che innescano processi non decodificabili, *black boxes*) e tuttavia è probabile che quella decisione si fosse basata su poco altro che sull'elaborazione di dati statistici oggettivi. Il tasso di recidiva è innegabilmente più alto

7. In conclusione, la "dittatura del calcolo"[82] (se non, più in generale, "l'irragionevole efficacia della matematica"[83]), potrebbe essere la *digital evidence*[84] per decisioni giudiziarie complesse, in materia di pericolosità sociale[85]: insomma, un utile strumento al servizio della capacità umana di produrre giustizia, piuttosto che rappresentare la temuta "fine del diritto penale"[86].

Esiste, tuttavia, un altro aspetto da considerare. Come ha osservato il noto psicologo tedesco Gerd Gigerenzer[87], le previsioni algoritmiche impiegate in svariati campi presentano un limite: sono affidabili in situazioni ben definite e stabili, in cui sono disponibili grandi quantità di dati (i cosiddetti *Big Data*)[88], ma, a differenza dell'intelligenza umana, non sono in grado di far fronte all'incertezza[89].

Questo limite degli algoritmi predittivi è ancora più marcato nel diritto penale dove le decisioni devono essere prese caso per caso, sulla base di pochissimi dati "filtrati"

---

tra le persone di colore, sebbene sia ovvio che ciò non dipende da loro, ma dalle condizioni ambientali e culturali sfavorevoli in cui gran parte di esse ancora vive. Sicché, già questo semplice esempio lascia intuire come la decisione migliore nel diritto potrebbe non essere quella assunta su base oggettiva e probabilistica. Né è un caso che il dibattito giuridico sul ricorso all'Intelligenza Artificiale assuma in modo pacifico la necessità di affiancare la macchina ad un giudice in carne e ossa, senza sostituirlo, e si imperni comunque sulla possibilità che le reti neurali simulino competenze più umane e meno econiche. Insomma, la possibilità di affidarsi a decisori esterni dovrebbe presupporre che l'architetto sia davvero in grado di rispondere *sempre* alla domanda su che cosa è giusto fare, ma non è chiaro se questo sia possibile». Cfr. F. Basile, *Intelligenza artificiale e diritto penale: qualche aggiornamento e qualche nuova riflessione*, cit., 12, dove si osserva che «nessun algoritmo è "neutro"» poiché, «nel concepire l'architettura di un algoritmo, il programmatore fa delle scelte che, necessariamente influenzano il "risultato" dell'operazione computazionale».

[82] V. *amplius*, P. Zellini, *La dittatura del calcolo*, Milano 2018.

[83] Cfr. E.P. Wigner, *L'irragionevole efficacia della matematica nelle scienze naturali*[4], trad. it., Milano 2017.

[84] Per un inquadramento del possibile rilievo dell'IA come *digital evidence* nel processo penale, da valutare alla luce delle coordinate tracciate dalla Suprema Corte statunitense nella nota sentenza *Daubert*, del 1993, e sostanzialmente condivise, se non arricchite, dalla Corte di Cassazione italiana, *Cozzini*, del 2010, si rinvia a G. Canzio, *Intelligenza artificiale e processo penale*, cit., 801 ss.

[85] Nel senso che l'IA costituisca un valido ausilio, ma non un sostituto del giudice, v. M. Amisano, *Prevedere -e non predire- attraverso gli algoritmi e le loro insidie*, in *AP* 2022, 2. Sulla prognosi postuma e possibile ricorso ad algoritmi predittivi, v. *funditus* D. Perrone, *La prognosi postuma tre distorsioni cognitive e software predittivi*, Torino 2021.

[86] In questo senso v. *funditus* A. Garapon, J. Lassègue, *La giustizia digitale. Determinismo tecnologico e libertà*, trad. it., Bologna 2021, spec. 261 ss.; v. anche C. Castelli, D. Piana, *Giusto processo e intelligenza artificiale*, Rimini 2019.

[87] G. Gigerenzer, *Klick. Wie wir in einer digitalen Welt die Kontrolle behalten und die richtigen Entscheidungen treffen*, Monaco 2021, 57.

[88] Sulla rilevanza dei *Big Data* per gli algoritmi predittivi, con particolare riferimento all'Industria 4.0 e all'*automotive*, v. *funditus*, *Robotics, Autonomics and the Law. Legal issues arising from the AUTONOMICS for Industry 4.0 Technology Programme of the German Federal Ministry for Economic Affairs and Energy*, a cura di E. Hilgendorf, U. Seidel, Baden-Baden 2017.

[89] G. Gigerenzer, *Klick*, cit., 57.

dall'interpretazione giudiziale, come ad esempio, la valutazione dei precedenti penali dell'imputato, da cui si può ricavare un giudizio di pericolosità sociale.

Resta da chiedersi se il giudice sarà in grado di interpretare le previsioni algoritmiche di pericolosità sociale e di individuare i suddetti profili problematici. È quindi necessario evitare due possibili scenari estremi: affidarsi agli algoritmi di IA come se fossero un oracolo, oppure non fidarsi affatto, e contestare il loro responso[90]. Un modo per evitare i suddetti scenari sarebbe quello di progettare sistemi che spieghino in contesti giudiziari come gli algoritmi arrivano alle loro conclusioni o previsioni, in modo da plasmare un'IA "spiegabile" (*EXplainable AI* o *XAI*)[91].

Un'IA "spiegabile" potrebbe contribuire a rendere più trasparente una previsione di pericolosità sociale[92], fornendo il tipo di spiegazioni di cui gli esseri umani hanno bisogno per rendere intellegibili le strategie decisionali sottostanti utilizzate dagli algoritmi[93], soddisfacendo così «il diritto alla spiegazione che opera come valvola di sicurezza e possibile limite al processo di costruzione di una società dell'automazione integrale»[94].

Esiste tuttavia una differenza fondamentale nel ragionamento causale tra gli algoritmi di IA e l'intelligenza umana, evidenziata dalla metafora della scala a tre gradini dell'informatico e filosofo ebreo-americano Judea Pearl: il gradino più basso è quello dell'osservazione, che consiste nella ricerca di regolarità nel mondo, in cui ci si

[90] Come ha scritto il filosofo Maurizio Ferraris, «se il governo è algoritmico, non è governo, poiché l'algoritmo non ha obiettivi; se è davvero governo, l'algoritmo segue l'umano così come le intendenze seguono il generale» (M. Ferraris, *Documanità. Filosofia del mondo nuovo*, Bari 2021, 71); cfr. M. Durante, *Potere computazionale. L'impatto delle ICT su diritto, società, sapere*, Milano 2019.

[91] A. Deeks, *The judicial demand for explainable artificial intelligence*, in *Colum. L. Rev.*, 119, 2019, 1828 ss., dove si osserva che gli informatici continuano a sviluppare nuove forme di *XAI*: alcuni modelli di *ML* sono costruiti per essere intrinsecamente spiegabili, ma come risultato questi modelli sono spesso meno complessi e tendono a essere meno accurati nelle loro previsioni; un'altra serie di modelli non sono intrinsecamente spiegabili. Per quest'ultimi modelli, gli informatici hanno adottato due approcci di base: un primo tipo (il cosiddetto approccio esogeno) non spiega effettivamente il funzionamento interno (cioè il ragionamento) dell'algoritmo di *ML*, ma cerca piuttosto di fornire informazioni rilevanti all'utente o al soggetto dell'algoritmo su come funziona il modello utilizzando metodi estrinseci e ortogonali; un secondo tipo di approccio tenta effettivamente di spiegare o replicare il ragionamento del modello e talvolta viene definito approccio "scompositivo". V. anche più di recente, J. Dirutigliano, *Trasparenza e spiegabilità degli algoritmi*, in *La politica dei dati. Il governo delle nuove tecnologie tra diritto, economia, società*, cit., 281 ss.

[92] Per un'analisi critica della *XAI*, v. *amplius* S.J. Russell, P. Norvig, *Artificial Intelligence*, cit., dove si osserva che una spiegazione non implica necessariamente una maggior accuratezza della decisione algoritmica.

[93] Cfr. nel campo tra IA e neuroscienze computazionali, *amplius* A. Lieto, *Cognitive Design for Artificial Minds*, cit., 100 ss.

[94] In questo senso filosofico, v. M. Durante, *I gradi dell'automazione: diritto, fiducia e riflessività*, in U*topie dell'automazione completa*, a cura di M. Balistreri, P. Marrone, Milano 2022, 115.

chiede come cambia ciò che si sa su Y se si osserva X; sul secondo gradino, si passa dall'osservare al fare, e ci si chiede cosa accadrebbe a Y se si facesse X; sul terzo gradino, entra in gioco l'argomentazione controfattuale, e si immagina retrospettivamente cosa sarebbe accaduto a Y eliminando mentalmente X. Gli algoritmi di IA si trovano sul gradino più basso di questa "scala", mentre il terzo gradino è proprio dell'intelligenza umana[95].

Un'IA "spiegabile" potrebbe quindi rendere più oggettiva una decisione giudiziaria, in particolare un giudizio prognostico di pericolosità sociale. Questo, tuttavia, non significa che la soluzione (ammesso che esista) dei "vecchi" problemi che affliggono la prognosi di pericolosità sociale non apra nuovi problemi che gli agenti umani dovranno esaminare. Per quanto paradossale, non si può escludere che anche un'IA non sia in grado di "spiegare" il modello di ragionamento umano alla base di una prognosi di pericolosità sociale, mettendo così in dubbio la riconducibilità delle operazioni mentali a un procedimento di calcolo (anche se questo non significa che sia un problema assolutamente insolubile)[96].

Un giorno le macchine riusciranno a risolvere tutti i problemi, ma mai nessuna di esse potrà porne uno[97].

[95] J. Pearl, D. Mackenzie, *The Book of Why: The New Science of Cause and Effect*, New York 2018.

[96] Vale la pena di notare che il problema della riconducibilità del ragionamento umano a calcolo ha origine nelle scienze dure (cfr. amplius P. Odifreddi, *Pillole matematiche. I numeri tra umanesimo e scienza*, Milano 2022, 246 ss.). Nel XX secolo, il grande matematico tedesco David Hilbert era alla ricerca di un algoritmo per l'*Entscheidungsproblem* (il problema della decisione) che riducesse tutto il ragionamento deduttivo umano a calcolo. Lo stesso Hilbert, l'8 settembre 1930, tenne a Königsberg (l'odierna Kaliningrad in Russia), un famoso discorso su *Conoscenza della natura e logica* che si concludeva con il motto *wir müssen wissen, wir werden wissen* ("dobbiamo sapere, e sapremo"). Le ultime parole di quel discorso furono incise sulla tomba di Hilbert a Gottinga. Paradossalmente Hilbert aveva pronunciato queste parole proprio il giorno dopo e nello stesso luogo in cui il famoso logico austriaco Kurt Gödel aveva annunciato il suo famoso teorema di incompletezza, che dimostra l'inesistenza di una formalizzazione completa della matematica, anche se ciò non significa che esistano problemi assolutamente insolubili, anzi lo stesso Gödel pensava che non ci fossero (P. Odifreddi, *Pillole matematiche. I numeri tra umanesimo e scienza*, cit., 91 ss.). L'*Entscheidungsproblem* fu comunque risolto negativamente dal famoso matematico inglese Alan Turing, indipendentemente dal matematico statunitense Alonzo Church, riferendosi al problema della fermata della cosiddetta macchina di Turing (v. A. Turing, *On computable Numbers, with an application to the Entscheindungsproblem*, in *Proc. Lond. Math. Soc.*, 1936, 42, 230 ss.).

[97] Così Albert Einstein, probabilmente il più importante fisico del Novecento.