Quando a decidere in materia penale sono (anche) algoritmi e IA: alla ricerca di un rimedio effettivo

di Giuseppe Contissa, Giulia Lasagni, Giovanni Sartor

L'articolo illustra i principali ambiti di applicazione dei sistemi predittivi e analizza il loro impatto sui diritti fondamentali e sui principi dell'equo processo, concentrandosi in particolare sulla definizione di ricorso effettivo applicabile contro decisioni assunte (anche) col supporto di sistemi algoritmici e di IA.

The paper presents the main areas of application of predictive systems and analyses their impact on fundamental rights and fair trial principles. It focuses in particular on the definition of the right to an effective remedy against decisions taken (also) with the support of algorithmic and AI systems.

1. Introduzione

Gli algoritmi e l'intelligenza artificiale stanno progressivamente trasformando quasi tutte le attività umane e in particolare i processi decisionali, rendendoli dipendenti dalla loro capacità di registrare e elaborare informazioni. Secondo Balkin, in realtà, viviamo già in una società algoritmica, cioè una società organizzata intorno al processo decisionale automatizzato, in cui sono algoritmi e intelligenza artificiale (A/IA) a prendere le decisioni (1).

A questa progressiva automazione sociale, si è aggiunto, negli ultimi anni, un cambiamento di paradigma nell'IA, che ha comportato l'adozione di nuovi metodi basati sulla conoscenza indotta dall'analisi dei dati. I tradizionali sistemi informatici di supporto alle decisioni, che utilizzavano le conoscenze specialistiche umane, trasferendole nel sistema attraverso rappresentazioni simboliche della conoscenza e dell'inferenza logica, sono stati integrati o sostituiti da sistemi IA basati sull'apprendimento automatico (*machine learning*), applicato a grandi masse di dati (i cosiddetti *big data*) (2).

Il sistema, piuttosto che effettuare valutazioni e previsioni su un insieme di regole predefinite dal programmatore e trasferite direttamente negli algoritmi, costruisce automaticamente un suo modello del dominio a partire dell'analisi dei dati su cui è addestrato, sulla base di un algoritmo di apprendimento automatico. Usando tale modello, il sistema genera classificazioni, valutazioni e previsioni sui nuovi casi che gli sono sottoposti. Aggiornando e ampliando il set di dati, si migliorano automaticamente il modello e le capacità previsionali del sistema. Di conseguenza, il sistema A/IA è più semplice da sviluppare e solitamente superiore nelle prestazioni, ma né il suo funzionamento in generale, né le ragioni alla base di ciascuna decisione possono essere compiutamente spiegati per mezzo del codice sorgente, in quanto esso spiega solamente il funzionamento dell'algoritmo di apprendimento, ma non la configurazione finale del modello creato dal sistema stesso, che è alla base del suo funzionamento e quindi delle decisioni che assume.

* Nonostante questo articolo sia frutto della riflessione comune degli autori, la stesura è stata ripartita come segue: Giuseppe Contissa §§ 2, 4.1, 4.2, 6 (prima parte, proposte tecnico-informatiche); Giulia Lasagni §§ 3, 4, 4.3, 5, 6 (seconda parte, proposte sul piano legale); Giovanni Sartor § 1.

(1) Balkin, *The three laws of robotics in the age of big data*, in 78, *Ohio State Law Journal*, 2017, 1219.

(2) Con il termine "apprendimento automatico" si indicano gli approcci con cui i sistemi possono migliorare le loro prestazioni imparando automaticamente come eseguire compiti futuri attraverso l'osservazione (esperienza), cfr. Russell - Norvig, *Artificial Intelligence. A Modern Approach*, 3 ed. Prentice Hall, Englewood Cliffs, N. J., 2010.

In questo senso, tali sistemi possono essere considerati come *black box* (scatole nere) (3), cioè sistemi in cui input e output sono osservabili, mentre il funzionamento interno rimane oscuro perfino ai suoi stessi programmatori (4). Tale funzionamento assomiglia, quindi, a quello di un "oracolo", ma, contrariamente ai suoi antichi predecessori, ad un oracolo con altissima precisione statistica.

Il diritto penale non è immune da queste trasformazioni: sempre più sistemi di A/IA vengono introdotti in diverse fasi del procedimento, a sostegno delle indagini (*predictive policing*) o delle decisioni del giudice (*predictive justice*) (5). Come si dirà, queste tecnologie possono contribuire a migliorare l'efficienza della giustizia; al tempo stesso, il loro uso solleva anche una serie di preoccupazioni circa la protezione dei diritti fondamentali (6).

In questo articolo, illustreremo i principali ambiti di applicazione dei sistemi predittivi e valuteremo il loro impatto sui diritti fondamentali e sui principi dell'equo processo, concentrandoci in particolare sulla definizione di ricorso effettivo applicabile contro decisioni assunte (anche) col supporto di sistemi predittivi.

2. Sistemi A/IA nella giustizia penale

Diversi sistemi di giustizia penale in tutto il mondo fanno oggi uso di sistemi A/IA per sostenere il processo decisionale umano di diversi attori, come forze dell'ordine, avvocati, giudici (7).

Lo scenario di base è il seguente: il sistema classifica gli individui in classi di riferimento. Tali classi possono esprimere previsioni sul comportamento degli individui, o di gruppi di individui (ad esempio, tasso di recidiva individuale basso/alto/medio; basso/alto/medio rischio di criminalità in una particolare area geografica). Queste previsioni vengono successivamente impiegate nel processo decisionale algoritmico, cioè per elaborare e suggerire strategie su come trattare tali soggetti in base alla loro classificazione (8).

In questo contributo ci concentreremo esclusivamente sull'impiego di sistemi A/IA al fine di identificare potenziali reati o autori di reati; oppure –ipotesi molto più critica, specialmente nell'ottica di un rimedio effettivo– al fine di formulare previsioni di rischio individualizzate.

2.1. Identificazione di potenziali reati e trasgressori

Un primo modo per utilizzare i sistemi A/IA a fini preventivi è quello di applicarli per supportare (o sostituire) l'esperienza investigativa umana con un'analisi integrata dei dati già disponibili, al fine di identificare potenziali modelli criminosi e ridurre la vittimizzazione in ambienti digitali, come i *social media* (9). Un altro approccio è quello che mira a prevedere le circostanze (tempo e luogo) di possibili reati. Questo approccio rispecchia le tradizionali metodologie investigative di mappatura delle attività criminali in un'area determinata, sulla base dell'analisi di fattori sociali, demografici, economici, ambientali, nonché dei dati relativi ai precedenti penali. Esistono diversi esempi di questi strumenti "più convenzionali" di polizia predittiva, impiegati in Europa e negli Stati Uniti. Fra i più noti si può menzionare PredPol, un sistema algoritmico ad apprendimento automatico sviluppato dalla polizia e dall'Università di Los Angeles (UCLA). Questo sistema, basato su dati storici concernenti i reati (in particolare quelli relativi alle vittime), formula previsioni a partire da tre classi di dati (tipo di reato, luogo e data/ora di commissione del reato). Tali previsioni vengono poi utilizzate per identificare, su un'interfaccia web basata su Google Maps, le aree ad alto rischio in determinate fasce orarie. Questi risultati dovrebbero quindi consentire di ottimizzare la distribuzione di risorse umane e di mezzi, indirizzando nelle aree più a rischio gli agenti di polizia (10). Altri esempi

(3) Pasquale, *The black box society: The secret algorithms that control money and information*, Harvard University Press, 2015.

(4) Millar - Kerr, *Delegation, relinquishment, and responsibility: The prospect of expert robots*, in *Robot Law*, a cura di Calo – Froomkin – Kerr, Edward Elgar Publishing, 2016, 102-128, 107.

(5) Nota, in tal senso, la risposta del giudice John Robert, della Corte Suprema degli Stati Uniti, alla domanda "Can you foresee a day when smart machines, driven with artificial intelligences, will assist with courtroom fact-finding or, more controversially even, judicial decision-making?" "It's a day that's here [...] and it's putting a significant strain on how the judiciary goes about doing things", cf. Liptak, *Sent to Prison by a Software Program's Secret Algorithms*, in *New York Times*, 1.05.2017, all'indirizzo <https://www.nytimes.com/2017/05/01/us/politics/sent-to-prison-by-a-software-programs-secret-algorithms.html>. Per un'analisi del concetto di "giustizia predittiva" si veda Viola, *Combinazione di dati e prevedibilità della decisione giudiziaria*, in questa ***Rivista***, 2019, 1, 215.

(6) Tra la vasta letteratura che si occupa di questo impegnativo argomento, si veda, ad esempio, Garapon - Lassègue, *Justice digitale*, PUF, Parigi, 2018.

(7) Perry – McInnis – Price – Smith - Hollywood, *Predictive Policing: Il ruolo della previsione del crimine nelle operazioni di polizia*, Santa Monica, CA: RAND Corporation, 2013, all'indirizzo <https://www.rand.org/pubs/research_reports/RR233.html>.

(8) "...Do you see the paradox? An algorithm processes a slew of statistics and comes up with a probability that a certain person might be a bad hire, a risky borrower, a terrorist, or a miserable teacher. That probability is distilled into a score, which can turn someone's life upside down. And yet when the person fights back, "suggestive" countervailing evidence simply won't cut it. The case must be ironclad. The human victims of WMDs, we'll see time and again, are held to a far higher standard of evidence than the algorithms themselves..." O'Neil, *Weapons of math destruction: How big data increases inequality and threatens democracy*, Crown Publishing Group, NY, 2016.

(9) Perry et al, (n 7).

(10) All'indirizzo <https://www.predpol.com/law-enforcement/#predPolicing>.

di questi sistemi predittivi, spesso specializzati nel perseguire reati specifici, si possono trovare anche in Europa. Tra i più noti, KeyCrime, sviluppato dalla Polizia di Milano e utilizzato per prevedere le rapine nell'area metropolitana (11), e XLAW, sviluppato dalla Polizia di Napoli e applicato dalle forze dell'ordine in diverse regioni italiane per prevedere furti e rapine.

Due sono i principali vantaggi di questi sistemi più "convenzionali". In primo luogo, contribuiscono ad una migliore gestione del *know how* delle forze dell'ordine in un'area geografica specifica, svincolandone la conservazione dalla presenza fisica e competenza dei singoli agenti. In secondo luogo, questi sistemi possono migliorare le *performance* investigative in condizioni di limitate risorse umane e consentire una allocazione più efficiente delle stesse. Certamente, l'uso di tali sistemi può anche contribuire a rafforzare eventuali discriminazioni, ad esempio incentivando la polizia a soffermarsi in determinate aree geografiche e comportando maggiori probabilità per la popolazione ivi residente di essere soggetta a pratiche di *stop and frisks*. Come si dirà (12), tuttavia, questi effetti discriminatori non sono necessariamente creati dall'uso di sistemi A/IA. Algoritmi e intelligenza artificiale infatti si limitano "solo" a perpetuare criticità già riscontrabili quando tali attività sono svolte da esseri umani. In questo senso, questa applicazione di tecnologie A/IA può ritenersi meno problematica per la tutela dei diritti fondamentali, almeno se confrontata con gli utilizzi che si rivolgono direttamente a singoli individui.

2.2. Valutazione individuale dei rischi

Un approccio molto meno convenzionale è quello finalizzato a calcolare la probabilità di pericolosità individuale. Accedendo ad enormi quantità di dati, anche non necessariamente già disponibili alle forze dell'ordine, tali sistemi correlano fattori di rischio statistici a specifici individui, grazie a modelli matematici automatizzati.

Forse il più noto di questi sistemi predittivi è COMPAS, sviluppato da Northpoint Inc. (ora Equivant), una società privata californiana, e attualmente adottato negli Stati Uniti in diversi Stati per calcolare il tasso di recidiva, ad esempio per emettere decisioni sull'assegnazione di misure alternative o sull'applicazione di istituti assimilabili alla sospensione condizionale della pena (13).

Le predizioni di COMPAS si basano su informazioni definite come "statiche" (ad es. i precedenti penali) e su "un uso limitato" di alcune variabili "dinamiche" (quali l'abuso di sostanze stupefacenti) (14). A causa dei diritti di proprietà intellettuale sul software, tuttavia, non sono disponibili dettagli ulteriori su come queste variabili influiscono nella valutazione del sistema. Ciò che è noto è che una parte delle informazioni utilizzate da COMPAS deriva dalle risposte fornite dal soggetto che deve essere valutato ad un questionario di 137 domande, che coprono diversi aspetti della personalità e della storia dell'individuo (15). Confrontando fra loro i dati di storie simili, COMPAS valuta il rischio di recidiva individuale sulla base di tre fattori di rischio ("pre-processuale", "generale" e riferito ai comportamenti violenti), assegnando all'individuo un punteggio numerico. Pertanto, COMPAS non elabora valutazioni fondate solo su elementi riferiti all'individuo, ma estrapola valutazioni individuali da dati di gruppo.

Un altro strumento di valutazione del rischio individuale, soprattutto nella fase preprocessuale, è quello del *Public Safety Assessment* (PSA), elaborato dalla Laura and John Arnold Foundation e attualmente utilizzato in decine di giurisdizioni negli Stati Uniti e in alcune delle più grandi città del paese, come Phoenix, Chicago e Houston ((16)). Il PSA fa due tipi di previsioni, calcolando: (i) il rischio che il soggetto non compaia davanti al giudice in udienza e (ii) il rischio di recidiva in caso di liberazione anticipata (con particolare attenzione ai reati violenti). I fattori considerati dal sistema sono nove e, secondo quanto noto, includono l'età, i procedimenti pendenti e i precedenti penali. Il rischio è calcolato su una scala da 1 a 6, dove i punteggi più elevati indicano un livello di rischio maggiore (17). Prendendo spunto dalle numerose critiche mosse nei confronti di

(11) All'indirizzo <https://www.emmeviemme.com>.

(12) Si veda, infra, paragrafo 3.2.

(13) *State of Wisconsin v. Loomis*, 881 N.W.W.2d 749 (Wis. 2016), 36.

(14) All'indirizzo <http://equivant.wpengine.com/classification/>. Tra il 2012 e il 2015, 14 Stati USA "created or regulated the use of risk assessments during the pretrial process", cfr. Widgery, *National Conference of State Legislatures, Trends in Pretrial Release. State legislation*, marzo 2015, all'indirizzo <https://comm.ncsl.org/productfiles/98120201/NCSL-Pretrial-Trends-Report.pdf>.

(15) Skeem - Louden, *Assessment of evidence on the quality of the Correctional Offender Management Profiling for Alternative Sanctions (COMPAS)*, 2007, all'indirizzo <https://ucicorrections.seweb.uci.edu/files/2013/06/CDCR-Skeem-EnoLouden-COMPASeval-SECONDREVISION-final-Dec-28-07.pdf>.

(16) Come gli stati dell'Arizona, del Kentucky e del New Jersey, cfr. <https://www.psapretrial.org/about>.

(17) In particolare: "the person's age at the time of arrest; whether the current offense is violent; whether the person had a pending charge at the time of the current offense; whether the person has a prior misdemeanor conviction; whether the person has a prior felony conviction; whether the person has prior convictions for violent crimes; whether the person has failed to appear at a pretrial hearing in the last two years; whether the person failed to appear at a pretrial hearing more than two years ago; and whether the person has previously been sentenced to incarceration", cfr. Public Safety Assement, *Risk factors and formula*, all'indirizzo < https://www.psapretrial.org/about/factors >.

COMPAS, i creatori di questo sistema hanno deciso di renderne pubblico il funzionamento e, in particolare, di rivelare il diverso peso di ciascuno di questi nove fattori nel calcolo finale (18).
L'impiego di strumenti predittivi individualizzanti, tuttavia, non è una prerogativa degli Stati Uniti. L'Harm Assessment Risk Tool (HART), sviluppato dalla polizia di Durham e dall'Università di Cambridge, ad esempio, effettua previsioni sulla base di 33 diverse metriche, tra cui i precedenti penali, l'età e il codice postale dell'autore del reato. Analogamente a COMPAS, anche HART classifica gli individui in gruppi ad alto, moderato o basso rischio. I parametri utilizzati da HART sono stati resi, almeno parzialmente, accessibili al pubblico. Questa caratteristica ha permesso di identificare una serie di criticità rilevanti –operazione che con il sistema COMPAS non è invece possibile per ragioni di proprietà intellettuale. Le informazioni disponibili, ad esempio, riportano che il software HART è predisposto per favorire i falsi positivi rispetto ai falsi negativi, il che significa che è più probabile che un individuo a basso rischio sia classificato erroneamente come persona ad alto rischio di recidiva piuttosto che il contrario (19).

3. Valutazione automatizzata del rischio: quali limiti di utilizzo?

L'uso di algoritmi o di tecnologie di intelligenza artificiale per formulare predizioni individualizzanti in materia penale mette in discussione diversi profili del diritto all'equo processo, con un impatto innegabile sul diritto di difesa dell'imputato (20).
Queste criticità si riscontrano, in primo luogo, nei casi in cui il ricorso alla valutazione automatizzata dei rischi è reso obbligatorio per legge, ad esempio nelle decisioni sulla sospensione condizionale della pena, sulla liberazione anticipata, sulla determinazione della cauzione, ecc. (21). Ancora più complesse, tuttavia, sono le situazioni in cui i sistemi A/IA sviluppati per un certo scopo (ad esempio, il supporto alle decisioni in materia di libertà vigilata, come COMPAS), finiscono per essere utilizzati nel procedimento penale anche per altri fini.
Il caso più noto a questo riguardo è certamente *Loomis*, deciso nel 2016 dalla Corte suprema del Wisconsin (22). Eric Loomis era stato accusato di cinque reati per la partecipazione come conducente ad una sparatoria in auto. L'imputato si era dichiarato colpevole per i due reati meno gravi; la Corte aveva accolto la richiesta di patteggiamento, subordinandola però alla "lettura" in aula delle imputazioni per gli altri tre reati rimanenti (23). Al fine di calcolare l'ammontare della sanzione da applicare a Loomis, il tribunale aveva ordinato la produzione del Presentence Investigation Report – PSI (24). Nel caso di specie, il PSI comprendeva anche una valutazione del rischio effettuata da COMPAS, nella quale Loomis veniva classificato come soggetto ad alto rischio di recidiva. La relazione conteneva, inoltre, una descrizione delle finalità per cui le valutazioni di rischio COMPAS dovrebbero essere utilizzate, ossia a) identificare i trasgressori a cui applicare determinate misure e, b) identificare i fattori di rischio da neutralizzare. Le istruzioni, in particolare, mettevano in guardia contro l'uso improprio di tale *software*, specificando che questo non dovrebbe essere utilizzato per determinare la gravità della pena o se l'autore del reato debba essere sottoposto o meno a misura custodiale (25).
Sulla base della valutazione dei rischi elaborata da COMPAS, e utilizzando come circostanze aggravanti la lettura dei tre reati esclusi dal patteggiamento, la Corte

(18) Eppure, critico circa l'equità del PSA, Patrick, *Fondazione Arnold per l'introduzione dello strumento di valutazione del rischio pretermine a livello nazionale*, 3.09.2018, all'indirizzo <https://www.insidesources.com>.

(19) Oswald - Grace - Urwin - Barnes, *Algorithmic risk assessment policing models: lessons from the Durham HART model and 'Experimental' proportionality*, in *Information & Communications Technology Law*, 2018, 27:2, 223-250, 236.

(20) Quali, soprattutto, il principio di parità delle armi (cfr., ad esempio, Quattrocolo, *Quesiti nuovi e soluzioni antiche? Consolidati paradigmi normativi vs rischi e paure della giustizia digitale "predittiva"*, in *Cass. pen.*, 4, 2019, 1748 ss), il diritto al silenzio (Deskus, *Fifth amendment limitations on criminal algorithmic decision-making*, in *NYUJ Legis. & Pub. Pol'y* 21, 2018, 237), o la presunzione di innocenza (per cui si rimanda, volendo, a Contissa- Lasagni, *The Role of Predictive Algorithmic Systems in Criminal Investigations: Which Effective Remedy To (New?) Fair Trial Lacunas*, in corso di pubblicazione).

(21) Per quanto riguarda in particolare il rilascio su cauzione (*bailing*), cfr. Doyle - Bains - Hopkins, *Bail Reform. A Guide for State and Local Policymakers*, in *Criminal Justice Policy Program. Harward Law School*, febbraio 2019, all'indirizzo <http://cjpp.law.harvard.edu/assets/BailReform_WEB.pdf>, dove si riportano diversi casi a livello statale. Ad esempio, il Kentucky ha dapprima adottato un progetto pilota per l'uso di sistemi predittivi (*Administrative Pretrial Release Program*) in 20 delle 120 giurisdizioni locali, poi esteso a tutto lo Stato nel 2017.

(22) *State of Wisconsin v. Loomis*, 881 N.W.W.2d 749 (Wis. 2016).

(23) Tale lettura implica che "the charges can be read in and considered, and that has the effect of increasing the likelihood, the likelihood of a higher sentence within the sentencing range", *State of Wisconsin v. Loomis*, 10.

(24) Il "Pre-sentence Investigation Report" racchiude la "storia" dell'indagato o imputato: "[it] is a report prepared by a court's probation officer on request by the court. It is the report of the investigation conducted to find out the history including the educational, criminal, family, and social background of a person convicted of a crime. It summarizes for a court the background information needed to determine the appropriate sentence. Sentence of a convicted person is increased or decreased after examining", cfr. <https://definitions.uslegal.com/p/presentence-investigation-report-psir/>.

(25) *State of Wisconsin v. Loomis*, 8.

condannava Loomis a 4 anni per la prima imputazione e a 7 anni per la seconda (26). In appello, Loomis lamentava la presunta violazione dei diritti del giusto processo causata dall'uso di COMPAS. In particolare, contestava che COMPAS, sebbene non concepito a tale scopo, fosse stato impropriamente usato nella fase di quantificazione della pena, determinando l'applicazione di una sanzione più severa nei suoi confronti. La natura proprietaria del *software*, inoltre, gli aveva impedito di verificare la validità scientifica del meccanismo decisionale posto in essere dal sistema.

Per inciso, si deve notare che quest'ultimo motivo di appello appare particolarmente rilevante nell'ordinamento giuridico statunitense, dove, almeno a partire dalla nota causa *Daubert*, la Corte Suprema ha subordinato l'ammissibilità delle prove scientifiche alla dimostrazione dell'affidabilità dei metodi scientifici adottati (27). La Corte Suprema non si è finora mai pronunciata sulla applicabilità dei criteri *Daubert* né alle valutazioni rese da sistemi A/IA, né alla fase del *sentencing*, e tale ultima possibilità viene generalmente esclusa a livello statale. Una parziale eccezione a questo orientamento può essere trovata in una decisione della Corte Suprema del District of Columbia, riguardante l'uso di uno strumento algoritmico di valutazione del rischio utilizzato per la previsione di comportamenti violenti nei minori (Structured Assessment of Violence Risk in Youth - SAVRY). In quel caso, la corte aveva dichiarato l'inammissibilità della valutazione effettuata dal sistema SAVRY, senza però esplicitamente né affermare né negare la pertinenza dei criteri *Daubert*. Si riteneva infatti che il Governo non avesse sufficientemente dimostrato che l'applicazione specifica del *software* fosse stata effettuata secondo parametri scientificamente validi; nella decisione, tuttavia, la corte non estendeva le medesime conclusioni al *software* predittivo in quanto tale ((28)).

Anche nel caso *Loomis*, la Corte Suprema del Wisconsin riscontrava diverse criticità nell'uso di COMPAS ai fini della quantificazione della pena. Ad esempio, il sistema era stato sì validato in alcune giurisdizioni, ma non in Wisconsin, cosicché non era chiaro se i fattori utilizzati fossero effettivamente accurati anche per la popolazione di quello Stato. Il *software* inoltre era stato, pubblicamente e a seguito di analisi approfondita, accusato da una ONG di essere discriminatorio, in particolare nei confronti di individui afroamericani (29). Il tribunale del Wisconsin tuttavia, non esaminava il caso alla luce dei criteri *Daubert*, né – contrariamente alla causa SAVRY – riteneva che tali elementi critici fossero sufficienti almeno per riformare la decisione emessa in primo grado. Ignorando di fatto le accuse di effetti discriminatori, la Corte concludeva elaborando un *test* fondato sulla decisività o esclusività dell'elemento probatorio contestato, simile in un certo senso a quello sviluppato – in altri contesti e per altri fini - dalla Corte europea dei diritti dell'uomo (30). Secondo il giudice statunitense, pertanto, nessuna violazione del giusto processo può essere riconosciuta se il sistema A/IA è stato applicato correttamente e se la valutazione automatizzata costituisce un elemento non determinante e corroborato da altri fattori. Nel caso specifico, la lettura delle imputazioni più gravi veniva considerata come elemento di riscontro sufficiente a corroborare la valutazione di rischio generata da COMPAS. Il ricorso veniva quindi rigettato dalla Corte suprema degli Stati Uniti, diventando definitivo nel 2017 (31).

Le argomentazioni sviluppate in *Loomis* non costituiscono un caso isolato oltreoceano.

Già nel 2010, nella decisione *Malenchik*, ad esempio, la Corte d'appello dell'Indiana aveva utilizzato la valutazione automatizzata del rischio prodotta da altri *software* (Level of Service Inventory-Revised/LSI-R e Substance

(26) "You're identified, through the COMPAS assessment, as an individual who is at high risk to the community. In terms of weighing the various factors, I'm ruling out probation because of the seriousness of the crime and because your history, your history on supervision and the risk assessment tools that have been utilized, suggest that you're extremely high risk to re-offend", *State of Wisconsin v. Loomis*, 8.

(27) Attraverso una serie di parametri, tra cui se la teoria o la tecnica in questione 1) è stata testata; 2) è validata da letteratura (*peer review*) nell'ambito scientifico di riferimento; 3) abbia un tasso di errore potenziale noto e 4) sia supportata da standard e norme che ne controllano il funzionamento, cfr. *Daubert v. Merrell Dow Pharmaceuticals, Inc.* 509 U.S. 579 (1993).

(28) Cfr. Supreme Court of the District of Columbia, Justice Okun, 25.03.2018, come riportato e commentato da Quattrocolo (n 20).

(29) Cfr. ProPublica: Angwin - Larson - Mattu - Kirchner, *Machine Bias. There's software used across the country to predict future criminals. And it's biased against blacks*, 23.05.2016, all'indirizzo <https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing>.

(30) Cfr., ad esempio, Corte EDU, *Murray c Regno Unito*, ricorso n. 18731/91, 8.02.1996 (diritto al silenzio), *Al-Khawaja e Tahery c Regno Unito*, ricorsi n. 26766/05 e 22228/06, 15.12.2011 (diritto alla prova in dibattimento), *Chiper c Romania*, ricorso n. 22036/10, 27.06.2017 (diritto alla prova in appello).

(31) *Certiorari denied*, 137 S. Ct. 2290 (2017). Diversi giuristi hanno criticato questa decisione, sottolineando come la Corte abbia irresponsabilmente respinto le critiche sugli effetti discriminatori del *software* e sulla mancanza di validazione nello Stato: cfr. *Recent cases*, in 130 *Harv. L. Rev.*, 2017, 1530 ss; *Eric L. Loomis, Petitioner v. State of Wisconsin, On Petition for A Writ of Certiorari to the Supreme Court of Wisconsin. Brief for the United States as Amicus Curiae*, all'indirizzo <https://www.scotusblog.com/wp-content/uploads/2017/05/16-6387-CVSG-Loomis-AC-Pet.pdf>; De Miguel Beriain, *Does the use of risk assessments in sentences respect the right to due process? A critical analysis of the Wisconsin v. Loomis ruling*, in *Law, Probability and Risk*, 2018, 17, 45-53; Quattrocolo (n 20); Gialuz, *Quando la giustizia penale incontra l'intelligenza artificiale: luci e ombre dei risk assessment tools tra Stati uniti ed Europa*, in *Dir. pen. cont.*, 29.05.2019, all'indirizzo <https://www.penalecontemporaneo.it/d/6702-quando-la-giustizia-penale-incontra-l-intelligenza-artificiale-luci-e-ombre-dei-risk-assessment-too>.

Abuse Substance Subtle Screening Inventory-SASSI) come circostanza aggravante nella decisione di condanna. Anche in tale caso, l'imputato lamentava che l'uso di questo sistema fosse stato censurato in casi precedenti (32) e che i modelli di calcolo su cui si basava l'algoritmo non fossero stati riconosciuti come scientificamente affidabili nello stesso Stato dell'Indiana, risultando quindi inaffidabili, oltre che discriminatori. La Corte riteneva tuttavia legittimo l'uso del sistema predittivo da parte del giudice, in quanto non sostitutivo della sua discrezionalità, ma semplice strumento di supporto, corroborato da altri elementi indipendenti (33).
L'orientamento maggioritario sviluppato dalle corti statunitensi sull'uso di sistemi A/IA in ambito penale sembra quindi finora approvare l'impiego di valutazioni automatizzate dei rischi – originariamente sviluppate a fini di prevenzione o di esecuzione - anche in fase di quantificazione della pena, purché la decisione non si fondi esclusivamente su di esse. In mancanza di una giurisprudenza rilevante su questo tema in Europa, questo portato verrà quindi tenuto in considerazione nella presente analisi, per verificare se esso possa rappresentare un approccio efficace nella tutela dei diritti di difesa dell'imputato anche nel nostro Continente.

4. Decisioni (parzialmente) automatizzate e equo processo: il diritto ad un ricorso effettivo

Questo contributo, come anticipato, si concentra sugli impieghi potenzialmente più critici per i diritti fondamentali dell'imputato, primo fra tutti il diritto ad un ricorso effettivo. Tale diritto, infatti, per ragioni giuridiche e tecniche, rischia di essere, forse più di qualsiasi altro, drammaticamente ostacolato dall'uso di valutazioni automatizzate contro l'indagato o imputato.
Il diritto ad un ricorso effettivo, previsto nel contesto europeo dall'articolo 13 della Convenzione europea dei diritti dell'uomo (CEDU) e dall'articolo 47 della Carta dei diritti fondamentali dell'Unione (CDF), è al tempo stesso una delle disposizioni più importanti e meno definite della nozione di giusto processo (34).
Nella giurisprudenza della Corte di Strasburgo, un ricorso può essere considerato effettivo solo se tale in astratto e nella prassi, nel senso che questo deve essere in grado di impedire la continuazione della presunta violazione o, in alternativa, fornire almeno un rimedio "adeguato" alle violazioni già avvenute (35). A tale proposito, non è sufficiente che il mezzo di ricorso sia previsto dalla legislazione nazionale, ma occorre valutarne l'efficacia in concreto, ad esempio tenendo conto della rapidità dell'azione correttiva o dell'effettiva possibilità per il richiedente di attivare il rimedio alla luce delle specifiche circostanze del caso.
Per poterlo considerare effettivo, nella interpretazione della Corte EDU, non si richiede necessariamente che il ricorso debba essere incardinato presso un'autorità giudiziaria; tuttavia, l'autorità deputata a decidere in merito ad esso deve comunque rispettare i requisiti di cui all'articolo 6(1) CEDU, primi fra tutti, i parametri di indipendenza e imparzialità. Su questo punto, si può riscontrare una differenza fra l'articolo 13 CEDU ed il suo corrispondente nel diritto dell'Unione. L'articolo 47(1) CDF, infatti, espressamente richiede che ogni limitazione dei diritti fondamentali sanciti dalla Carta debba poter essere efficacemente impugnata dinanzi a un giudice (36).
Quanto invece al campo penale, la giurisprudenza di entrambe le corti europee considera fondamentale, ai fini della effettività di un ricorso, la possibilità di ottenere un controllo giurisdizionale "pieno" davanti all'autorità competente. Per proteggere in modo non illusorio i diritti dell'indagato o imputato, infatti, è necessario identificare almeno un'autorità con il potere di esaminare con piena cognizione – cioè sia in fatto sia in diritto - le decisioni che impongono una misura punitiva (37). L'importanza del legame fra ricorso effettivo e cognizione piena è costantemente ribadita anche nel diritto secondario dell'Unione in materia processuale penale (38) e nella normativa sulla protezione dei dati

(32) *Rhodes v. State*, 896 N.E.2d 1193, 1195 (Ind. Ct. App. 2008).

(33) *Malenchik v. State*, 928 N.E.2d 564, 574 (Ind. 2010).

(34) Tanto da essere anche definita come la disposizione "più oscura" della Convenzione da due giudici della Corte di Strasburgo, cfr. giudici Matscher e Pinheiro Farinha in *Malone v. United Kingdom*, ricorso n. 8691/79, 2.08.1984.

(35) Cfr. Corte EDU, *Kudła v. Poland*, ricorso n. 30210/96, 26.10.2000, §§ 157-158.

(36) La Corte di giustizia aveva già sancito tale principio nella sentenza *Marguerite Johnston v Chief Constable of the Royal Ulster Constabulary*, causa 222/84, del 15.05.1986, ECLI:EU:C:1986:206, 1651; cfr. anche la sentenza *Union nationale v Georges Heylens and o.*, causa 222/86, del 15.10.1987, ECLI:EU:C:1987:442, 4097, e la sentenza *Oleificio Borelli SpA v Commission*, causa C-97/91, del 3.12.1992, ECLI:EU:C:1992:491.

(37) Cfr. Corte EDU, *Umlauft v Austria*, 23.10.1995, ricorso n. 15527/89, § 37; Öztürk v Germany, 21.02.1984, ricorso n. 8544/79 § 56; *A. Menarini Diagnostics S.R.L. v Italy*, S.r.l., 27.09.2011, ricorso n. 43509/08, §§ 59-63-67; cfr. anche *Schmautzer v. Austria*, 23.10.1995, ricorso n. 15523/89, § 36 e *Gradinger v. Austria*, 23.10.1995, ricorso n. 15963/90, § 44.

(38) Direttiva 2010/64/UE del 20.10.2010 sul diritto all'interpretazione e alla traduzione nei procedimenti penali; Direttiva 2012/13/UE del 22.05.2012 sul diritto all'informazione nei procedimenti penali; Direttiva 2013/48/UE del 22.10.2013 sul diritto di accesso a un difensore nei procedimenti penali e nei procedimenti di mandato d'arresto europeo e sul diritto di informare un terzo sulla privazione della libertà personale e di comunicare con terzi e con le autorità consolari mentre è privato della libertà; la summenzionata Direttiva (UE) 2016/343 del 9.03.2016, relativa al rafforzamento di taluni aspetti della presunzione di innocenza e del

personali trattati a fini di prevenzione e repressione dei reati (39). Nessuno degli atti legislativi ad oggi in vigore, tuttavia, fornisce una definizione dettagliata di come un ricorso debba essere strutturato per risultare concretamente effettivo.
Alla luce di questa intrinseca vaghezza, il diritto in esame risulta in molti contesti problematico, per diventare eccezionalmente critico quando la decisione punitiva avverso l'indagato o imputato è (anche solo parzialmente) automatizzata.
Diverse sono le circostanze che determinano tale criticità. In primo luogo, come si vedrà, è arduo per l'imputato far valere il diritto a un ricorso effettivo senza poter accedere alle informazioni necessarie su cui si è basata la decisione. In secondo luogo, in questo contesto non è raro per l'imputato dover fronteggiare una vera e propria mancanza di motivazione riguardo alla parte automatizzata della valutazione. Da ultimo, anche a prescindere dalle considerazioni precedenti, non è scontato oggi definire cosa si intenda per "effettivo" quando (almeno in parte) la decisione segue logiche non umane. Di questo profilo, ci occuperemo nella parte finale del contributo.

4.1. Diritto al ricorso effettivo e accesso alle informazioni

In primo luogo, per poter contestare efficacemente una decisione individuale, è necessario che il soggetto interessato abbia accesso a tutte le informazioni rilevanti per tale decisione, in particolare ai *dataset* (insiemi di dati), ai metodi di trattamento dei dati, e al codice sorgente che esprime gli algoritmi alla base del funzionamento del sistema. In verità, l'accesso a tali informazioni non risponde solo all'interesse di chi subisce direttamente gli effetti della decisione, ma anche a quello di tutti gli attori coinvolti nella progettazione, sviluppo, implementazione e utilizzo dei sistemi di A/IA nella giustizia penale, inclusi programmatori, giudici e, più in generale, l'opinione pubblica (40).

Affinché l'accesso alle informazioni rilevanti sia effettivo è però necessario un certo grado di trasparenza del processo decisionale. Questa esigenza è espressa anche nel considerando 38 della direttiva 2016/680, secondo cui "in ogni caso, tale trattamento [automatizzato] dovrebbe essere subordinato a garanzie adeguate, compresi il rilascio di specifiche informazioni all'interessato e il diritto di ottenere l'intervento umano, in particolare di esprimere la propria opinione, di ottenere una spiegazione della decisione raggiunta dopo tale valutazione e di impugnare la decisione" (41).
Tuttavia, i sistemi di A/IA presentano spesso notevoli criticità per quanto riguarda la trasparenza del loro funzionamento. Le informazioni sul *dataset* non sono di solito a disposizione delle parti, né del giudice che utilizza il sistema. Analoga considerazione può essere fatta per quanto riguarda le informazioni sui metodi di trattamento dei dati e gli algoritmi su cui si fonda il ragionamento del sistema A/IA. In molti casi, infatti, questi dipendono dall'accessibilità del codice sorgente, la cui divulgazione può essere limitata dai diritti di proprietà intellettuale. Nei sistemi di A/IA basati su approcci di apprendimento automatico, inoltre, come già discusso in precedenza, le possibilità di ricostruire in modo completo le informazioni sul funzionamento interno del sistema e gli elementi considerati per giungere ad una decisione sono strutturalmente limitate.
Per garantire un livello soddisfacente di trasparenza, sono stati proposti diversi metodi, anche se ancora oggi non è chiaro quali elementi debbano essere ricompresi in questo requisito e con quale modalità le informazioni debbano essere concretamente rese accessibili.
Una prima opzione, suggerita nella Carta etica del Consiglio d'Europa (42), è quella della completa trasparenza tecnica, ossia che la divulgazione sia del codice sorgente del sistema algoritmico sia della documentazione di accompagnamento. Tuttavia, come anticipato, quando il sistema è sviluppato da privati, come nel caso di COMPAS, l'accessibilità del codice sorgente può essere limitata per ragioni di proprietà intellettuale e dalla necessità di proteggere i segreti commerciali e industriali.
Anche nei casi in cui sia possibile accedere al codice sorgente, però, ciò potrebbe rivelarsi solo una soluzione parziale al problema della trasparenza, soprattutto

diritto di essere presente al processo penale; Direttiva (UE) 2016/800 dell'11.05.2016 relativa alle garanzie procedurali a favore di indagati o imputati in procedimenti penali e Direttiva (UE) 2016/1919 del 26.10.2016 relativa all'assistenza giudiziaria a favore di indagati e imputati in procedimenti penali e di persone ricercate in procedimenti di mandato d'arresto europeo.

(39) Cfr. considerando (104), Direttiva (UE) 2016/680 del 27.04.2016, relativa alla protezione delle persone fisiche con riguardo al trattamento dei dati personali da parte delle autorità competenti a fini di prevenzione, indagine, accertamento e perseguimento di reati o esecuzione di sanzioni penali, nonché alla libera circolazione di tali dati.

(40) IEEE, *Ethically Aligned Design: A Vision for Prioritizing Human Wellbeing with Artificial Intelligence and Autonomous Systems*, version 1, IEEE Standards Assoc., 2016, all'indirizzo <https://standards.ieee.org/news/2019/ieee-ead1e.html>.

(41) In modo speculare rispetto al considerando (71) del Regolamento (UE) 2016/679 del 27.04.2016, relativo alla protezione delle persone fisiche con riguardo al trattamento dei dati personali, nonché alla libera circolazione di tali dati (più noto con l'acronimo inglese "GDPR").

(42) Consiglio d'Europa, European Commission for the Efficiency of Justice (CEPEJ), *European Ethical Charter on the Use of Artificial Intelligence in Judicial Systems and their environment*, adottata il 3-4.12.2018, all'indirizzo <https://rm.coe.int/ethical-charter-en-for-publication-4-december-2018/16808f699c>, 11.

nell'ottica di garantire un rimedio effettivo. Infatti, non solo il codice sorgente dei sistemi A/IA è di solito incomprensibile per i non esperti, ma anche i programmatori spesso incontrano difficoltà a comprendere il funzionamento di tali sistemi e prevederne i risultati esclusivamente sulla base dell'ispezione del codice sorgente.

Questi limiti sono stati chiaramente emersi in diverse decisioni giudiziarie in tutta Europa, in particolare nel settore dell'istruzione scolastica, dove sistemi di A/IA sono comunemente adottati per la classificazione e la selezione degli studenti e l'assegnazione dei docenti ai plessi.

In Francia, ad esempio, il sistema algoritmico *Admission post bac* (A.P.B.) è da tempo utilizzato nella procedura di iscrizione degli studenti all'università. Nel 2016, la *Commission d'accès aux documents administratifs* si è pronunciata a favore dell'accesso al codice sorgente di tale piattaforma. Sebbene il codice sorgente fosse stato reso disponibile, gli esperti non sono stati tuttavia in grado di ricostruire il ragionamento completo del sistema. Questo infatti risultava determinato non solo dall'algoritmo espresso nel codice sorgente, ma anche dai diversi dati di *input* provenienti da un insieme di database esterni al sistema. Senza la divulgazione di tali dati, nonché di informazioni sulla struttura delle tabelle e della descrizione dei campi utilizzati nelle banche dati che li contenevano, la semplice divulgazione del codice sorgente non era sufficiente a garantire un rimedio effettivo (43).

Considerazioni analoghe possono farsi per quanto riguarda alcuni casi italiani, riguardanti l'assegnazione di docenti alle scuole superiori. Nel 2017, ad esempio, il Tar Lazio ha stabilito che il Ministero dell'Istruzione ha l'obbligo di rilasciare una copia del codice sorgente del *software* utilizzato in tale procedura. Tuttavia, anche in questo caso, non è stata fatta menzione della divulgazione dei dati di *input*, della struttura dei dati delle tabelle e della descrizione dei campi utilizzati nelle banche dati collegate al sistema (44).

Una seconda opzione per risolvere i problemi di trasparenza dei sistemi A/IA è quella di divulgare anche le informazioni di livello superiore sulla logica del processo decisionale automatizzato, possibilmente in linguaggio naturale, cioè in un linguaggio comprensibile anche ad utenti non esperti.

In effetti, come sottolineato dal gruppo di lavoro Articolo 29, "la complessità non è una scusa per non fornire informazioni" (45). Questo è l'approccio seguito dal GDPR all'articolo 13(2), lettera f), che richiede, quando i dati personali utilizzati in un processo decisionale automatizzato sono raccolti presso l'interessato, di fornire le "informazioni significative sulla logica utilizzata" dal sistema. Le stesse disposizioni sono ripetute all'articolo 14(2), lettera g), in relazione ai dati non ottenuti dall'interessato.

Le informazioni da divulgare, in questa prospettiva, dovrebbero includere almeno: (a) le informazioni sui dati che sono serviti come *input* per la decisione automatizzata; (b) le informazioni sull'elenco dei fattori che hanno influenzato la decisione; (c) le informazioni sull'importanza relativa dei fattori che hanno influenzato la decisione; e (d) una spiegazione ragionevole (eventualmente in forma testuale) dei motivi per cui è stata presa una certa decisione (46).

Quest'ultimo parametro, tuttavia, è particolarmente critico da valutare. Alcuni autori si domandano se la spiegazione debba fornire informazioni complete su tutti i modelli e le variabili prese in considerazione dal sistema (spiegazione *model-centric*), o solo su quelle che sono rilevanti per il caso specifico in esame (spiegazione *subject-centric*) (47). Inoltre, rimane incerto se la spiegazione debba includere anche i dati personali (riguardanti soggetti terzi) utilizzati per la decisione. Questa opzione è soggetta a limitazioni da parte dello stesso GDPR. Tuttavia, potrebbe essere necessaria in alcuni casi per verificare l'equità della decisione presa rispetto alla situazione di soggetti diversi in condizioni comparabili.

Altro punto critico riguarda quale debba essere il livello di dettaglio su ciascun dato di *input* e il suo peso nella decisione. La divulgazione di una spiegazione completa

(43) Decisione della Commissione d'accès aux documents administratifs (CADA) relativa al codice sorgente della piattaforma *Admission post bac* (A.P.B.), 2016, all'indirizzo <http://bo.letudiant. fr/uploads/mediatheque/EDU_EDU/2/1/1/1/1202121-avis-cada-apb-160916-original.pdf>.

(44) Tar Lazio, Sezione Terza bis, Decisione n. 03769/2017 (udienza del 14.02.2017), all'indirizzo <https://www.giustizia-amministrativa.it/portale/pages/istituzionale/visualizza?nodeRef=&schema=tar_rm&nrg=201611419&nomeFile=201703769_01.html&subDir=Provvedimenti>.

(45)) Articolo 29 Data Protection Working Party, *Guidelines on Automated individual decision-making and Profiling for the purposes of Regulation 2016/679* (wp251rev.01), adottate il 3.10.2017 e da ultimo riviste il 6.02.2018, all'indirizzo <https://www.google.com/url?sa=t&rct=j&q=&esrc=s&source=web&cd=3&ved=2ahUKEwjeyZ-0h6blAhXD-KQKHV6aC3oQFjACegQIBBAC&url=https%3A%2F%2F%2Fec.europa.eu%2Fnewsroom%2Farticle29%2Fdoc_id%3Fdoc_id%3D49826&usg=AOvVVaw3Hbd9vdVdV-5JxpwJwJPUmrucm>, nota 40 e p. XX dove si specifica che "The GDPR requires the controller to provide meaningful information about the logic involved, not necessarily a complex explanation of the algorithms used or disclosure of the full algorithm. The information provided should, however, be sufficiently comprehensive for the data subject to understand the reasons for the decision".

(46) Brkan, *Do algorithms rule the world? Algorithmic decision-making and data protection in the framework of the GDPR and beyond*, in *International Journal of Law and Information Technology*, 27, 2, 2019, 91–121, all'indirizzo <https://doi.org/10.1093/ijlit/eay017m>, 113.

(47) Edwards - Veale, *Slave to the algorithm: Why a right to an explanation is probably not the remedy you are looking for*, in *Duke L. & Tech. Rev.*, 16:18, 2017.

del modello potrebbe infatti generare effetti collaterali. Ad esempio, in un sistema di rilevazione automatica dell'evasione fiscale, la divulgazione della logica adottata dal sistema per generare un avviso di possibile evasione in relazione al superamento di determinate soglie di rischio, può favorire l'adozione di "comportamenti strategici" da parte dei singoli individui nella presentazione delle dichiarazioni fiscali.

Per questi motivi, piuttosto che sul requisito della trasparenza, diversi esperti suggeriscono di concentrarsi sulla *regolarità procedurale*, ossia sull'adozione di tecniche specifiche che dimostrino la capacità del sistema di soddisfare determinati standard di equità anche nelle decisioni automatizzate, senza rivelare quali attributi chiave sono utilizzati nelle decisioni, o i dettagli dei processi algoritmici alla base delle stesse. Altri studiosi, tuttavia, sottolineano che tale regolarità procedurale garantisce solo che le decisioni si basino sulla stessa politica decisionale, che la politica è stata determinata prima di conoscere gli *input* e che i risultati possono essere riprodotti. Essa considera pertanto solo la regolarità procedurale aggregata di tutti i casi, garantendo che siano decisi secondo le stesse regole. Nemmeno il concetto di regolarità procedurale, però, risponde alla domanda sul perché l'algoritmo abbia raggiunto quella specifica decisione individuale avverso quel soggetto (48).

4.2. Diritto al ricorso effettivo e assenza di motivazione

La motivazione delle decisioni, specialmente in ambito penale o punitivo, rappresenta un requisito imprescindibile per esercitare il diritto ad un ricorso effettivo. Questo presupposto però rischia di essere eluso quando le decisioni sono basate su valutazioni automatizzate, ed in particolare quelle risultanti da sistemi di A/IA basati su apprendimento automatico.

L'impossibilità di ricostruire il funzionamento interno dei sistemi di A/IA, infatti, si traduce spesso in un approccio fideistico nei confronti del risultato fornito, cioè nel ritenere che la decisione sia giustificata dallo stesso fatto che è stata presa, con una certa precisione statistica, dal sistema stesso. Questo approccio è stato definito *data fundamentalism* (49), ovvero la tendenza a ritenere che l'analisi effettuata con tecniche di *data mining* su grandi insiemi di dati fornisca sempre una visione oggettiva della realtà, tralasciando il fatto che le correlazioni identificate dall'algoritmo, e su cui si basano le decisioni, non necessariamente implicano un nesso di causalità. Come affermano Kroll et al., l'analisi e le decisioni prese dai computer spesso godono di un'immeritata assunzione di equità o oggettività, tuttavia la progettazione e implementazione di sistemi decisionali automatizzati risultano esposte a criticità che possono portare ad emettere decisioni sistematicamente errate e affette da *bias* (pregiudizi) (50). Per i motivi sopra esposti (la mancanza di trasparenza), sarà inoltre spesso molto difficile individuare potenziali *bias* che influenzano la decisione del sistema.

Per quanto riguarda in particolare i sistemi di apprendimento automatico, le principali cause di *bias*, che possono risultare in effetti discriminatori, sono legate innanzitutto a problemi a carico del *dataset*, e in particolare: (i) all'uso di un *dataset* contenente dati che riflettono pregiudizi o *bias* impliciti, già presenti nelle decisioni su cui è addestrato il sistema (51); e (ii) all'uso di un *dataset* contenente dati che offrono un quadro statisticamente distorto di alcuni gruppi in relazione alla popolazione complessiva (52). Inoltre, anche *dataset* senza errori o *bias* iniziali possono portare a decisioni discriminatorie, a causa dell'incapacità dei sistemi di apprendimento automatico di distinguere tra mera correlazione e causalità, e degli effetti dell'auto-rinforzo del modello sulla base di nuovi dati incorporati nel *dataset* (53). Un'altra

(48) Brkan (n 46).

(49) Crawford, *The hidden biases in Big Data*, in *Harvard Business Review Blog Network*, 1/04/2013, all'indirizzo: https://hbr.org/2013/04/the-hidden-biases-in-big-data.

(50) Kroll – Barocas – Felten – Reidenberg – Robinson – Yu, *Accountable algorithms*, in *U. Pa. L. Rev.* 165 (2016), 633.

(51) Supponiamo che un sistema di supporto alle decisioni dei giudici sia addestrato sulle decisioni emesse dai giudici umani dello stato dell'Alabama negli ultimi 80 anni, e supponiamo anche che queste decisioni contengano qualche pregiudizio razziale: questo porterebbe il sistema a riprodurre risultati parziali o discriminatori, sulla base di procedure algoritmiche apparentemente oggettive ma influenzate da pregiudizi "ereditati" da precedenti decisioni umane.

(52) Supponiamo il sistema di polizia predittiva utilizzato per individuare gruppi di persone che potenzialmente potrebbero commettere reati sia addestrato su un insieme di dati che sovrarappresenta l'incidenza dei crimini in alcuni gruppi etnici. Le forze dell'ordine sarebbero quindi indirizzate dall'algoritmo a fermare e controllare più persone appartenenti a questi gruppi etnici che ad altri, con il risultato che, statisticamente, saranno scoperti più reati commessi dal gruppo etnico selezionato rispetto a quelli commessi da altri gruppi. Quando i dati relativi ai nuovi reati scoperti saranno aggiunti al *dataset*, i reati commessi in un quei tra quei gruppi etnici saranno ancora più sovrarappresentati, rafforzando l'effetto discriminatorio in una situazione di circolo vizioso, cfr. Miller, *Total Surveillance, Big Data, and Predictive Crime Technology: Privacy's Perfect Storm.*, in *J. Tech. L. & Pol'y*, 19, 105, 2014.

(53) Supponiamo che si utilizzi un sistema di polizia predittiva per suggerire quali persone dovrebbero essere controllate per traffico di stupefacenti. Supponiamo anche che il sistema sia stato addestrato su un insieme di dati contenenti informazioni su casi precedenti e che il sistema sia in grado di rilevare una correlazione tra i casi in cui gli automobilisti vengono fermati per eccesso di velocità e il reperimento di prove a carico relative al traffico di stupefacenti. Naturalmente, questa è una mera correlazione derivante dal fatto che chi viene fermato per violazione del codice della strada, ha anche una maggiore probabilità (rispetto al resto della popolazione) che la sua auto verrà perquisita (anche per la ricerca di stupefacenti). Tuttavia, il sistema potrebbe suggerire alle forze di po-

causa di discriminazione, specifica del settore penale, è che la maggior parte (se non la totalità) dei sistemi A/IA utilizzati in questo ambito si riferisce solo ad un numero ristretto di reati, spesso cd. "di strada". Ciò contribuisce ad aumentare la percezione di pericolosità di alcuni autori di reati, a discapito di altre categorie pure caratterizzate da alti tassi di recidiva, come ad esempio i *white collar crimes*. Infine, dato che la maggior parte dei *software* predittivi è sviluppata e/o posseduta da società private a scopo di lucro (come nel caso di COMPAS), anche la determinazione del contenuto del *dataset*, o del processo di selezione applicato dall'algoritmo sul *dataset* potrebbe dare luogo a decisioni discriminatorie a causa di potenziali conflitti di interessi (ad esempio di natura commerciale) non dichiarati.

In una prospettiva diversa, invece, alcuni studiosi sostengono che, lungi dall'amplificare effettori discriminatori, l'incremento di sistemi A/IA, correttamente impiegati, potrebbe in futuro proprio correggere e limitare gli effetti dannosi dei *bias* cognitivi (umani), in particolare quelli dei giudici. Ad esempio, per quanto riguarda decisioni sul *bailing,* un recente studio ha dimostrato come un algoritmo progettato per prevedere il rischio di mancata comparizione in giudizio abbia ottenuto risultati più equi dei giudici umani, perché, al contrario di questi ultimi, nel prendere la decisione non si lasciava influenzare dal *bias* legato alla gravità del reato contestato (*current accusation bias*) (54). Effetti positivi per gli imputati si sono riscontrati anche in un'altra ricerca in cui si valutava l'utilizzo del sistema PSA nella Contea di Lucas, Ohio, dove il *software* è stato adottato nel 2015. In tale caso, grazie al PSA, si è infatti registrato un aumento del numero di persone rilasciate senza ricorrere alla cauzione e una significativa riduzione del numero di reati commessi da imputati in attesa di giudizio non sottoposti a misure cautelari (55).

Una precauzione solitamente adottata per prevenire o attenuare il rischio di effetti discriminatori è quella di escludere o rimuovere dal *dataset* i dati sensibili (dati riguardanti l'origine razziale o etnica, le opinioni politiche, le convinzioni religiose, la salute, la vita sessuale, l'orientamento sessuale, ecc.) (56). Tuttavia, i sistemi di A/IA possono consentire l'estrazione di dati sensibili anche a partire dal trattamento di dati personali non sensibili. Ad esempio, in un caso famoso, un sistema A/IA impiegato da Target (un *retailer* statunitense) per analizzare gli acquisti dei clienti, è stato in grado di assegnare ad ogni cliente un punteggio di 'previsione di gravidanza' e di stimare la data del parto, sulla base solamente dell'analisi della cronologia degli acquisti di determinati prodotti, e di alcune informazioni demografiche aggiuntive (57).

Alcuni recenti contributi hanno suggerito altre possibili azioni per gestire e limitare il rischio di *bias* e i conseguenti effetti discriminatori, quali ad esempio: *a)* garantire e dimostrare l'origine dei dati utilizzati dal sistema (eventualmente certificandone le fonti), la loro qualità e la loro copertura, e che essi non siano stati modificati prima di essere utilizzati dal sistema di apprendimento automatico, in modo che l'intero ciclo di vita dei dati sia tracciabile; *b)* fornire informazioni sui metodi di trattamento dei dati, eventualmente attraverso un audit indipendente, quando non è possibile un accesso diretto al codice sorgente (58); *c)* fornire all'individuo oggetto di una decisione di un sistema di tipo *black box*, un insieme di spiegazioni controfattuali, vale a dire informazioni che descrivono le più piccole modifiche agli *input* del sistema che, in ipotesi, avrebbero portato a un risultato differente e auspicabile per la persona interessata, senza dover spiegare la logica interna del sistema. In questo modo, sapendo quali fattori esterni e quali variabili hanno contribuito alla valutazione automatizzata, il soggetto destinatario della decisione sarebbe in grado di contestarla e, in particolare, di ottenere la prova che di un'eventuale discriminazione quando questa è determinata da un dato sensibile (ad esempio, la razza o l'appartenenza etnica) (59).

Anche nei casi in cui non è in gioco alcuna discriminazione, tuttavia, permangono diverse questioni critiche in relazione alla stessa definizione di "ricorso effettivo" nell'ambito di decisioni (anche solo parzialmente) automatizzate.

---

lizia di controllare chi viene fermato per eccesso di velocità perché quei soggetti hanno una maggiore probabilità di essere coinvolti nel traffico di stupefacenti. Quando i dati relativi ai nuovi reati si aggiungeranno al *dataset*, la parzialità del sistema e il suo effetto discriminatorio saranno ulteriormente rafforzati.

(54) () Sunstein, *Algorithms, correcting biases* in *Oxford Business Law Blog. Social Research: An International Quarterly*, 86, 2, 2019, 499-511, all'indirizzo <https://www.law.ox.ac.uk/business-law-blog/blog/2019/01/algorithms-correcting-biases>; Kleinberg – Lakkaraju – Leskovec – Ludwig – Mullainathan, *Human Decisions and Machine Predictions*, in *The Quarterly Journal of Economics*, 133, 1, February 2018, 237-293.

(55)) Tashea, *Risk-Assessment Algorithms Challenged in Bail, Sentencing and Parole Decisions*, 1.03.2017, all'indirizzo <www.abajournal.com>.

(56) Cfr. le "categorie particolari di dati personali", di cui all'Art. 9 GDPR.

(57) Floridi, *The fourth revolution: How the infosphere is reshaping human reality,* OUP Oxford, 2014, 16.

(58) CEPEJ, *European Ethical Charter* (n 42), 11

(59) Wachter – Mittelstadt - Russell, *Counterfactual explanations without opening the black box: Automated decisions and the GDPR,* in *Harv. JL & Tech.* 31, 2017, 853.

4.3. Quale rimedio è davvero effettivo?

Nei paragrafi precedenti abbiamo analizzato alcuni profili che rendono più difficile, per l'imputato destinatario di una decisione automatizzata, esercitare il proprio diritto ad un ricorso effettivo. Prima di introdurre alcune proposte costruttive su come meglio assicurare la conformità a questo diritto fondamentale, tuttavia, è necessario collocare il problema nell'attuale contesto normativo.

Anche se non si riscontrasse nessuna delle criticità già analizzate, i rimedi previsti dalle legislazioni europee attualmente in vigore non sembrano in grado di garantire un ricorso effettivo dei processi decisionali automatizzati o parzialmente automatizzati.

Gli ordinamenti nazionali sono oggi semplicemente per lo più privi di strumenti o disposizioni specifici per riesaminare decisioni algoritmiche in ambito penale (60).

Considerazioni analoghe valgono, in realtà, anche a livello del diritto dell'Unione dove, almeno dal 2016, il problema delle decisioni automatizzate è esplicitamente considerato. Ci si riferisce, in particolare, all'articolo 11(1) della Direttiva 2016/680, secondo il quale "una decisione basata unicamente su un trattamento automatizzato, compresa la profilazione, che produca effetti giuridici negativi o incida significativamente sull'interessato [è] vietata salvo che sia autorizzata dal diritto dell'Unione o dello Stato membro cui è soggetto il titolare del trattamento e che preveda garanzie adeguate per i diritti e le libertà dell'interessato, [incluso] almeno il diritto di ottenere l'intervento umano da parte del titolare del trattamento" (61).

La legislazione comunitaria, pertanto, stabilisce il diritto ad un controllo umano in caso di decisioni basate esclusivamente su una valutazione A/IA. Questo principio, riportato anche nel Regolamento *privacy* (GDPR), ha trovato ad oggi applicazione negli Stati membri, a livello giurisprudenziale, soprattutto in settori diversi dal diritto penale. È il caso, ad esempio, dell'Italia, dove nel 2018 il TAR Lazio ha dichiarato, con riferimento all'utilizzo di algoritmi nell'assegnazione del personale scolastico, che un processo amministrativo discrezionale non può essere pienamente delegato ad un sistema automatizzato, ribadendo l'importanza del controllo umano su tali decisioni (62).

Il successivo controllo umano su una decisione automatizzata, però, è un "rimedio" che presenta diverse criticità, in generale e nella materia penale in particolare.

In primo luogo, operare una distinzione tra il processo decisionale completamente automatizzato e quello semiautomatizzato può sembrare una distinzione logica a prima vista. Questa idea è stata espressa chiaramente dal Gruppo di Lavoro "Articolo 29", secondo il quale "se un essere umano riesamina il risultato del processo automatizzato e tiene conto di altri fattori nel prendere la decisione finale, tale decisione non sarà "basata unicamente" sul trattamento automatizzato" (63). In pratica, però, i confini tra i due modelli decisionali sono piuttosto labili (64).

È discutibile infatti, che esseri umani, sebbene incaricati di assumere una posizione di controllo sulle decisioni automatizzate, siano *effettivamente* nelle condizioni per riesaminare tali decisioni. La questione su in capo a chi risieda l'autorità decisionale *effettiva* è stata peraltro sollevata in diversi contesti, che hanno in comune con la giustizia penale l'estrema importanza degli interessi in gioco e la necessità di prendere decisioni in un arco di tempo limitato. Un esempio fra tutti è l'ambito sanitario, un settore in cui i sistemi di A/IA sono già utilizzati per generare ipotesi di diagnosi e relativi trattamenti, a supporto dell'attività decisionale del medico (65).

Gli attuali sistemi di A/IA, infatti, come già detto sopra, spesso non sono tecnicamente in grado di fornire spiegazioni intellegibili delle ragioni alla base delle loro valutazioni. Non solo, come sostenuto da un autorevole giurista americano, l'intelligenza artificiale è fondamentalmente aliena da quella umana, e spesso l'intero scopo di un sistema di intelligenza artificiale è quello di imparare a fare o vedere le cose in modi impossibili per gli esseri umani (66).

---

(60) Non così in altri settori del diritto, dove si riscontra una maggiore attenzione. Un esempio in questo senso è rappresentato dalla legislazione fiscale tedesca, modificata nel 2016 per includere una norma secondo la quale un atto amministrativo può essere adottato in modo completamente automatico, a condizione che ciò sia consentito dalla legge e che l'accertamento non richieda alcuna valutazione discrezionale, cfr. Gesetz zur Modernisierung des Besteuerungsverfahrens vom 18. Juli 2016 (BGBl. I S. 1679), § 35a.

(61) Considerazioni analoghe potrebbero essere fatte anche per quanto riguarda l'articolo 22 GDPR.

(62) Tar Lazio, Sezione Terza Bis, Decisione n. 09230/2018 (udienze del 26.06.2018 e 11.07.2018), all'indirizzo <https://www.giustizia-amministrativa.it/portale/pages/istituzionale/visualizza?nodeRef=&schema=tar_rm&nrg=201611238&nomeFile=201809230_01.html&subDir=Provvedimenti>.

(63) Articolo 29 Data Protection Working Party (n 45).

(64) Council of Europe, *Study on the Human rights dimensions of Automated Data Processing Techniques (In particular algorithms) and possible regulatory implications,* 6.10.2017, all'indirizzo < https://rm.coe.int/algorithms-and-human-rights-en-rev/16807956b5>.

(65) Come, ad esempio, nel caso di IBM Watson Health, all'indirizzo <https://www.ibm.com/watson-health/learn/artificial-intelligence-medicine>.

(66) Selbst, *Negligence and AI's Human Users* (11.03.2019 - Boston University Law Review, in corso di pubblicaizone), disponibile su SSRN all'indirizzo <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3350508>.

In ambito medico, i sistemi A/IA di supporto alla diagnosi, addestrati su *dataset* contenenti ingenti quantità di dati attinti da cartelle cliniche, risultati di esperimenti e letteratura medica, si sono tuttavia dimostrati in grado di produrre diagnosi e proposte di cura statisticamente più precise delle corrispondenti valutazioni prodotte dai medici umani negli stessi casi. In questo contesto, sembra piuttosto improbabile che il controllore umano, sebbene dotato di specifiche competenze tecniche e professionali possa essere effettivamente in grado di entrare nel merito della valutazione, specialmente quando la decisione deve essere presa in un breve intervallo di tempo.

A meno di evidenti errori macroscopici, infatti, al medico rimarranno sostanzialmente due alternative: o fidarsi dei risultati della valutazione automatizzata, perché si fida del sistema di A/IA che l'ha generata; oppure non fidarsi né del risultato, né del sistema. La mancanza di ragioni di fondo per contraddire le previsioni dell'IA comporta pertanto che la ragionevolezza delle singole decisioni individuali finisca per essere legata alla decisione di usare l'IA in linea generale (67).

Considerazioni simili possono riportarsi anche per l'uso di sistemi A/IA in ambito penale, a prescindere dal fatto che generalmente i termini per adottare decisioni sono meno stringenti rispetto alla medicina (68) e che sia più difficile, in questo settore, verificare alla luce di criteri oggettivi il livello di accuratezza delle valutazioni formulate (da macchine e da esseri umani). Ciò è ancora più vero considerando che, in generale, quando si tratta di effettuare valutazioni prognostiche, la capacità di giudizio umana è inferiore ai modelli statistici (69).

Nonostante quanto previsto dalla Direttiva 2016/680, quindi, anche quando l'essere umano mantenga formalmente la titolarità della decisione finale, la possibilità concreta di contestarne il merito in modo effettivo rimane un'ipotesi remota.

L'attuale legislazione nell'Unione richiede oggi il controllo delle decisioni (esclusivamente) automatizzate, ma tale controllo è di fatto, anche in ambito penale, ridotto alla scelta tra il fidarsi o meno del sistema di A/IA. In questo senso, la mera previsione di un successivo intervento umano non sembra sufficiente a garantire un rimedio effettivo, specialmente alla luce dei parametri di "effettività" utilizzati dalle due Corti europee.

5. Tutta colpa dell'intelligenza artificiale? Black box "umane" e processo penale

Una delle principali criticità che impediscono, di fatto, un ricorso effettivo contro le decisioni che includono valutazioni anche solo parzialmente automatizzate, è, come si è detto, l'impossibilità di ottenere (adeguate) motivazioni dai sistemi A/IA.

Sarebbe tuttavia errato ritenere che, nel processo penale, una tale lacuna sia riscontrabile solo quando algoritmi o intelligenza artificiale sono coinvolti. Al contrario, meccanismi decisionali che ricordano molto da vicino il modello *black box* caratterizzano anche il funzionamento totalmente "umano" (nel senso di non automatizzato) del procedimento e processo penale.

L'istituto della giuria è forse l'esempio più evidente di queste *black box* "umane".

In diversi ordinamenti giuridici, infatti, la decisione sulla colpevolezza dell'imputato si concretizza in un verdetto privo di motivazione: un meccanismo il cui funzionamento ricorda da vicino il ruolo "oracolare" giocato dai sistemi A/IA nel processo decisionale automatizzato. In tal senso, l'esistenza di procedure di selezione dei giurati non sembra rappresentare un elemento distintivo rilevante, pur costituendo un importante strumento per garantire una certa equità del procedimento, soprattutto contro potenziali discriminazioni. Come si è detto, infatti, anche in caso di decisioni automatizzate possono ad oggi essere adottate adeguate precauzioni per venire incontro a queste esigenze.

Come nelle decisioni di A/IA, pertanto, anche nel processo con giuria, la riconduzione dell'imputato alle "classi di riferimento" innocente/colpevole, non essendo supportata da una giustificazione motivata, rimane una valutazione le cui *rationes* (e possibili vizi) non sono ricostruibili dal destinatario del provvedimento con un livello di certezza ragionevole.

Diverse sono tuttavia anche le parentesi di *black box* riscontrabili all'interno di paradigmi processuali tradizionali altrimenti del tutto "spiegabili". Come già evidenziato da alcuni studiosi, questi sono i casi nei quali l'organo giudicante è chiamato ad effettuare valutazioni di rischio o "previsioni" sulla base di criteri "taciti" (70) che, di fatto se non esplicitamente, si fondano su elementi non strettamente giuridici come "l'intuizione", "il senso di giustizia" o "l'esperienza" del giudice (71).

(67) Selbst (n 66).

(68) Almeno in linea di principio, ma si veda ad esempio il caso dell'Agenzia europea per il controllo delle frontiere esterne (Frontex), in cui è stato stimato che il tempo a disposizione per ogni decisione sulla legittimità delle richieste di ingresso individuali nel territorio dell'UE è di circa 12 secondi, cfr. Fergusson, *Twelve Seconds to Decide in Search of Exellence: Frontex and the Principle of 'Best Practice'* in *Publications Office of the European Union*, 2014, 15.

(69) Deskus (n 20); si veda anche Millar – Kerr (n 4).

(70) Per quanto riguarda l'uso di "tacit knowledge and tacit norms", cfr. Schulz - Dankert, *"Governance by Things" as a Challenge to Regulation by Law*, in *Internet Policy Review* 5(2), 2016.

(71) Cfr. Caianiello, *Criminal Process faced with the Challenges of Scientific and Technological Development*, in *European Journal of Crime, Crime, Crimi-*

Tali situazioni si verificano, ad esempio, quando un giudice debba compiere valutazioni non riferite a fatti passati, ma a potenziali comportamenti futuri, come pronunciarsi sulla sussistenza di esigenze cautelari, o decidere in merito alla concessione di misure alternative alla detenzione. È vero che, contrariamente a quanto accade, per esempio, nei casi di COMPAS o HART, tutte queste diagnosi sono (e debbono) essere giustificate dai giudici umani.

Se non si vuole ridurre tutto l'accertamento ad un mero riscontro di eventuali precedenti penali (72), tuttavia, tale motivazione raramente contiene, e soprattutto può contenere, elementi che garantiscano, oggettivamente, una maggiore equità rispetto a decisioni prese da sistemi di A/IA. La mancanza di trasparenza e di motivazione ed il rischio di discriminazione sono naturalmente tutte buone ragioni, come abbiamo anticipato, per criticare l'uso di tali tecnologie nella materia penale. Le stesse critiche, tuttavia, potrebbero essere mosse – con la medesima intensità e a livello strutturale – anche contro decisioni totalmente "umane". Le valutazioni del giudice umano possono infatti in ugual misura essere influenzate da pregiudizi e sono certamente condizionate dai limiti – intrinseci in ogni esperienza umana, per quanto professionale – nel formulare giudizi statistici e prognostici. Queste situazioni, sostanzialmente piuttosto simili, sono però generalmente percepite come molto diverse.

La possibilità che un giudice umano formuli prognosi di rischio sulla base di criteri vaghi, per lo più non riscontrabili (tranne nel caso di precedenti penali) in modo oggettivo e, di conseguenza, interpretati necessariamente in modo personale, così come il potere della giuria di emettere "oracoli" è talvolta criticata, ma solitamente accettata come legittima. Al contrario, quando valutazioni simili sono rese da sistemi A/IA, la legittimità di tale scelta di politica criminale è ad oggi fortemente contrastata, specialmente in diverse giurisdizioni europee. In tal senso, alcuni autori hanno sottolineato come "noi umani" siamo piuttosto indulgenti (forse non ci interessi ammettere) verso gli errori e i fallimenti della nostra specie. Tolleriamo invece molto meno la capacità di tali fallimenti nelle nostre macchine. In altre parole, probabilmente ci aspettiamo di più dalle nostre macchine che da noi stessi (73).

Certo il giudice, così come la giuria, possono considerarsi, seppure con gradi diversi, rappresentanti della comunità (umana) di riferimento, godendo quindi di un riconoscimento "politico" che risulta difficilmente applicabile ad un algoritmo. E tuttavia, come si è detto, un certo grado di democraticità e trasparenza può essere assicurato anche nella programmazione (e nell'utilizzo) dei sistemi A/IA, di modo che i risultati prodotti siano allineati ai principi fondamentali nei quali i consociati si rispecchiano, primo fra tutti, il principio di legalità.

Alla luce dei progressi tecnologici riferiti all'intelligenza artificiale, inoltre, appare sempre più difficile giustificare una preferenza per il giudice "umano" sulla base della considerazione che solo l'essere umano, in quanto tale, ha la capacità non solo di applicare correttamente una regola, ma anche di disapplicarla e di trovare soluzioni non convenzionali. In questo senso, infatti, l'IA si distingue (e probabilmente, si distinguerà sempre più) dalla concezione tradizionale di "macchina" quale mero esecutore di compiti strettamente programmati. Una certa dose di "creatività" non è infatti estranea a questa tecnologia, oggi ancora ai suoi albori ma già impiegata, ad esempio, nella creazione di collezioni di moda (74).

Ciò nonostante, continuiamo istintivamente a ritenere più accettabile la capacità di ragionare e giudicare, con tutti i possibili relativi errori, quando questa facoltà è esercitata da esseri umani. Non si vuole qui sostenere che tale conclusione sia da abbandonare; l'analisi approfondita delle ragioni giuridiche, politiche ma forse ancor più psicologiche, che stanno alla base di questo atteggiamento, tuttavia va al di là del presente contributo. Queste considerazioni possono però essere utili per ricordarci che l'introduzione di decisioni A/IA nei procedimenti penali non mette solo in discussione la tenuta di principi fondamentali e di istituti nati in contesi "totalmente umani". Il crescente dibattito sulle decisioni automatizzate, in realtà, potrebbe (dovrebbe) essere accolto anche come un'occasione per riportare alla luce quei momenti "predittivi" tradizionalmente presenti nel processo penale e per riaprire una discussione finalmente scevra da (umani) pregiudizi sulla loro legittimità alla luce dei diritti fondamentali dell'imputato.

6. Verso un ricorso (davvero) effettivo: alcune proposte

Alla luce delle considerazioni sinora svolte, appare fondamentale realizzare un apparato di garanzie tecniche e giuridiche in grado di evitare che l'uso di A/IA nella

nal Law and Criminal Justice (2019, in corso di pubblicazione).

(72) Come nelle famigerate *three-strikes law*, fortemente criticabili in un sistema che miri al rispetto del giusto processo e del principio di proporzionalità della pena.

(73) Allen – Varner - Zinser, *AMA: Artificial Moral Agents (Prolegomena to any future artificial moral agent)*, in *Journal of Experimental & Theoretical Artificial Intelligence,* 2000, 12(3):251-261.

(74) Cfr., ad esempio, il marchio Glitch (<https://glitch-ai.com/pages/about-us>), su cui si veda ad esempio Wood, *These clothes were designed by artificial intelligence*, in *World Economic Forum*, luglio 2019, all'indirizzo: < https://www.weforum.org/agenda/2019/07/these-clothes-were-designed-by-artificial-intelligence/>.

giustizia penale si traduca in chiare violazioni dei diritti fondamentali previsti dalla Carta e dalla Convenzione europea e soprattutto del diritto al ricorso effettivo.
A tal fine, in primo luogo, è necessario che gli operatori del diritto che si occupano, e si occuperanno sempre più, spesso di algoritmi e di IA, abbiano un'adeguata consapevolezza delle capacità e dei limiti di tali sistemi. Ciò non richiede che essi diventino ingegneri informatici, ma è imprescindibile che sappiano interagire correttamente con i sistemi A/IA e, quindi, incorporare criticamente i risultati automatizzati nelle valutazioni "umane". In altri termini, è necessario passare da un approccio basato sul *data fundamentalism* a uno basato sulla *informed trust* (75).
In secondo luogo, piuttosto che relegare esclusivamente in una fase successiva il controllo sull'accuratezza del sistema A/IA, affidandone il compito a soggetti privi della necessaria competenza tecnica, sembra preferibile invece istituire un meccanismo di certificazione *ex ante* che consenta di validare il funzionamento del sistema con la partecipazione o il controllo di autorità pubbliche, come già avviene in altri settori (76). In realtà, tale certificazione dovrebbe riguardare non solo il funzionamento del sistema A/IA in quanto tale, ma anche l'intero sistema socio-tecnico che include tecnologia, utenti (giudici, pubblici ministeri, forze dell'ordine, avvocati) e le norme giuridiche ed etiche che regolano tale interazione.
Per stabilire questo meccanismo di certificazione le imprese che sviluppano sistemi A/IA saranno costrette a produrre informazioni che documentino l'approccio progettuale, lo sviluppo, la qualità e l'estensione del *dataset*, il funzionamento, la formazione degli utenti, ecc. L'obbligo di produrre tali informazioni vincolerà anche lo sviluppo stesso del sistema: per soddisfare i requisiti di certificazione, le imprese finiranno infatti per essere guidate anche nelle scelte progettuali dei sistemi A/IA. Questo approccio è già seguito in diversi settori cd. *satefy-critical* (come in ambito medico e nell'aviazione). Per ottenere tale certificazione, le imprese dovranno produrre solide prove empiriche che dimostrino l'idoneità del sistema allo scopo cui sono destinati. La certificazione dovrebbe comprendere anche regole per individuare profili di responsabilità, ad esempio specificando quali competenze e conoscenze sono necessarie per l'utilizzo del sistema finalizzato ad un particolare risultato processuale. Un sistema di certificazione efficace dovrebbe inoltre stabilire classi di rischio sulla base dello scopo e della fase procedurale in cui il sistema sarà utilizzato. La certificazione dovrebbe infine essere completata da un sistema di verifiche periodiche, che potrebbero variare in base alla classe di rischio in cui il sistema è impiegato. Un modello che potrebbe essere adottato come riferimento è quello del Regolamento 2017/745 (77), che stabilisce i requisiti necessari per ottenere il marchio di Conformità Europea (CE), attraverso il quale un dispositivo medico è certificato come prodotto conforme ai requisiti di sicurezza e prestazioni. I dispositivi medici sono suddivisi in quattro diverse classi, a seconda dello scopo del dispositivo e dei rischi inerenti (78). Per ogni classe viene definita una diversa procedura di valutazione di conformità, che richiede valutazioni di base per i dispositivi della prima classe I, fino a garanzia qualità totale per i dispositivi della classe III. Mentre nel primo caso, la valutazione della conformità ai requisiti del regolamento può essere effettuata sotto l'esclusiva responsabilità del produttore, la procedura completa di valutazione della qualità richiede il coinvolgimento di un organismo qualificato e di un gruppo di esperti (79). Nell'ambito della procedura di certificazione, i dati critici come il codice sorgente dovrebbero poter essere oggetto di ispezione da parte di un *audit* indipendente. In questo modo, si potrebbe raggiungere un giusto equilibrio tra gli interessi delle imprese (proprietà intellettuale, segreti commerciali) e quello pubblico di operare un controllo sull'uso di tale tecnologia nella materia penale. A tal fine, sembra opportuno che gli *audit* sia effettuati in modo uniforme su scala nazionale, possibilmente da un organismo pubblico in grado di garantire il controllo democratico, come una apposita commissione parlamentare. La certificazione e la convalida possono quindi contribuire anche a rafforzare la fiducia nell'uso di tali tecnologie nella materia penale, andando nella direzione di quella che attualmente è chiamata "*trustworthy AI*" (80).
Certificazione e validazione potrebbero essere considerate garanzie sufficienti per l'equità delle decisioni automatizzate adottate in alcuni contesti in cui la valu-

(75) IEEE (n 40), 220, secondo cui: "If we are to realize the benefits of A/IS, we must trust that they are safe and effective. We must enact policies and promote practices that allow those technologies to be adopted on the basis of informed trust. Informed trust rests on a reasoned evaluation of clear and accurate information about the effectiveness of A/IS and the competence of their operators".

(76) IEEE (n 40), 16.

(77) Art. 10(9) Regolamento (UE) 2017/745 del 5.04.2017, relativo ai dispositivi medici.

(78) Cfr. capo V, Sezione 1, art. 51 del Regolamento 2017/745.

(79) Si veda l'art. 52 del Regolamento 2017/745. "Per ogni dispositivo i fabbricanti provvedono a pianificare, istituire, documentare, applicare, mantenere e aggiornare un sistema di sorveglianza post-commercializzazione in modo proporzionato alla classe di rischio e adeguato alla tipologia di dispositivo", cfr. art. 83(2), del Regolamento 2017/745.

(80) Commissione europea, *Ethics Guidelines for Trustworthy AI*, aprile 2019, all'indirizzo <https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai>, 54.

tazione del rischio, cioè il pregiudizio per i diritti di difesa individuale, è relativamente basso. Questo potrebbe essere, ad esempio, il caso di reati per i quali possono essere inflitte solo sanzioni pecuniarie. Ciò consentirebbe di concentrare le risorse umane nella repressione e nell'accertamento dei reati più gravi.

Negli altri casi, tuttavia, non si può demandare esclusivamente alle procedure di certificazione il rispetto dei diritti fondamentali e il controllo democratico (81). Queste situazioni richiedono la creazione di soluzioni tecniche e giuridiche innovative.

Da un lato, il riferimento va all'adozione delle cosiddette soluzioni di *explicable AI* o XAI. Con questa espressione si fa riferimento ai metodi tecnici per spiegare i modelli delle *black box*, e in particolare si indicano quegli approcci in cui i metodi di apprendimento automatico sono collegati con metodi simbolici o basati su regole, in modo da fornire spiegazioni comprensibili all'uomo sulla base di mappe di argomenti o sequenze di passaggi logico-argomentativi. A tal fine, secondo alcuni autori, sarebbe possibile fornire un modello interpretabile e trasparente che sia in grado di replicare il processo decisionale della *black box*, rendendolo intellegibile (82). Questo approccio appare particolarmente interessante per la materia penale, nella quale la possibilità di accedere a tutti i fattori presi in considerazione da un sistema predittivo per assumere la sua decisione, e il peso attribuito a ciascuno di essi nel processo decisionale, è più utile, nell'ottica di tutelare il diritto di difesa, rispetto a un risultato che esprime solo una probabilità numerica che si verifichi un evento futuro. Gli approcci di *explicable AI*, tuttavia, sono ad oggi in fase di ricerca e non ancora raggiunto il grado di maturità necessario per essere adottati in contesti d'uso reali.

D'altro canto, sul piano legale, una prima ipotesi potrebbe essere quella di generalizzare la giurisprudenza della Corte Suprema degli Stati Uniti nella causa *Loomis*, introducendo anche negli ordinamenti giuridici europei una regola probatoria secondo la quale i risultati dei sistemi A/IA possono essere valutati dal giudice, purché non costituiscano l'elemento unico e decisivo a carico dell'imputato (83). Questa soluzione, forse già in linea con l'approccio omnicomprensivo elaborato dalla Corte di Strasburgo, non sembra tuttavia del tutto soddisfacente.

Per l'imputato, infatti, contestare il merito della valutazione fornita dal sistema A/IA rimane sempre difficile, se non impossibile. Da ciò deriva, di conseguenza, anche l'impossibilità di esercitare un ricorso davvero effettivo contro la decisione fondata su tali elementi.

Difficilmente, inoltre, questa lacuna può essere colmata dalla garanzia di un intervento successivo da parte dell'autorità giudiziaria "umana". Abbiamo già illustrato le ragioni per cui spesso un essere umano non è nella condizione di poter efficacemente riesaminare una decisione automatizzata. In questo senso, la disposizione dell'art. 11 della Direttiva 2016/680 può essere letta più come l'affermazione della necessità di non delegare *in toto* il potere decisionale ai sistemi A/IA (anche ai fini di eventuali richieste di risarcimento) piuttosto che come una disposizione realmente in grado di ottenere tale effetto. Per garantire che i mezzi di ricorso a disposizione degli indagati o imputati siano veramente effettivi, sia quando la decisione è totalmente automatizzata sia quando è basata solo in parte elementi prodotti da sistemi A/IA, appare più appropriato un cambio di prospettiva.

Si potrebbe, in tal senso, introdurre nel processo penale il diritto a far riesaminare le valutazioni generate da un sistema A/IA da un altro sistema automatizzato. Ad esempio, se il giudice di primo grado emette una decisione (sulla quantificazione della pena ma, in prospettiva, anche sulla colpevolezza) basata in tutto o in parte su valutazioni effettuate dal sistema X, l'imputato dovrebbe avere il diritto, dinanzi alla corte d'appello, di far ripetere tale valutazione da un sistema Z.

Non sarebbe tuttavia sufficiente che entrambi i sistemi siano certificati e validati per l'utilizzo nel processo penale; per garantire un rimedio effettivo, questi dovrebbero anche essere progettati e sviluppati da produttori diversi. Questa prospettiva, ancora inesplorata nel settore della giustizia penale, trova da tempo applicazione in settori *safety critical*, quali ad esempio l'aviazione, dove l'adozione di tecnologie ridondanti è generalmente considerato il metodo migliore per ridurre il rischio di incidenti e aumentare l'affidabilità dei controlli (84). Essa

(81) In questo senso, si veda anche la Commissione europea (n 80), 26, punto (107), secondo cui: "Poiché non ci si può aspettare che tutti siano in grado di comprendere appieno il funzionamento e gli effetti dei sistemi di IA, occorre valutare la possibilità di ricorrere a organizzazioni che possano certificare per il pubblico generale che un sistema di IA è trasparente, responsabile ed equo53. Tali certificazioni applicherebbero norme elaborate per diversi campi di applicazione e diverse tecniche di IA, opportunamente allineate alle norme industriali e sociali dei vari contesti. La certificazione non può tuttavia mai sostituire la responsabilità e dovrebbe essere quindi integrata da quadri di accountability, tra cui clausole di esclusione della responsabilità nonché meccanismi correttivi e di riesame".

(82) Cfr. Guidotti – Monreale – Ruggieri – Turini – Giannotti - Pedreschi, *A survey of methods for explaining black box models*, in ACM *Comput. Surv.* 51, 5, 2018, 93:1–93:42.

(83) Corte EDU, *Al-Khawaja e Tahery v UK* (n 30).

(84) Dove si suggerisce addirittura suggerito di mantenere l'utilizzo di tecnologie diverse per la stessa funzione, anche quando una di esse sia nettamente migliore delle altre, cfr. Jones, *Common cause failures and ultra reliability*, in 42a Conferenza Internazionale sui Sistemi Ambientali (2012), 3602, all'indirizzo <https://arc.aiaa.org/doi/abs/10.2514/6.2012-3602>.

si basa sul principio di c.d. ridondanza (*redundancy*), secondo cui le stesse informazioni devono essere elaborate simultaneamente da un certo numero di sistemi diversi ma con le medesime funzioni. La diversità può essere ottenuta adottando approcci alternativi nello sviluppo degli algoritmi, impiegando diversi *team* di programmatori e selezionando diversi componenti *hardware* e *software* (85).

L'applicazione del *redundancy approach* alle decisioni automatizzate nel processo penale richiederebbe di rendere disponibili, presso ogni distretto giudiziario, una gamma di sistemi A/IA certificati e validati (ad esempio, tramite la creazione di un apposito albo). Questo ventaglio di opzioni dovrebbe essere sufficientemente ampio da consentire ai giudici di appello di scegliere, per la falsificazione della valutazione prodotta in primo grado (o durante le indagini preliminari), un sistema diverso da quello già applicato. Garantire tale diritto potrebbe consentire al giudice (umano) di accedere ad una seconda valutazione automatizzata, dandogli la possibilità di applicare in modo effettivo i criteri della logica (umana) nell'operazione di comparazione e, quindi, di realizzare (o provare a realizzare) un rimedio veramente effettivo nei confronti dell'imputato.

L'introduzione strutturale di tali sistemi, che rappresenterebbe, in un certo senso, una versione tecnologicamente aggiornata dell'istituto della perizia, potrebbe anche contribuire a ridurre la disparità fra imputati che hanno le risorse per far esaminare le valutazioni automatizzate (e quindi per cercare di contestarle) e coloro che invece ne sono privi. Il rischio di discriminazioni di tipo economico, in astratto riscontrabili in tutti i processi e specialmente in quelli dove la difesa può trarre giovamento dalla nomina di un consulente tecnico, appare infatti particolarmente iniquo alla luce degli attuali ambiti di applicazione dei sistemi A/IA. Come si è illustrato, invero, questi sono ad oggi esclusivamente utilizzati in riferimento a reati cd. "di strada", cioè reati in cui si concentrano buona parte degli imputati economicamente svantaggiati. In tal senso, un uso non debitamente regolato (e ragionato) dei sistemi A/IA rischia certamente di comportare un peggioramento nella possibilità di esercitare in modo efficace i propri diritti fondamentali.

Gli algoritmi e l'intelligenza artificiale, però, stanno dimostrando di avere grandi potenzialità nella trasformazione *in toto* delle dinamiche decisionali tipiche del processo penale; potenzialità che ancora non sembrano appieno comprese e sfruttate. Forse è giunto il tempo di metterle anche al servizio dei diritti della difesa.

(85) Downer, *When failure is an option: Redundancy, reliability and regulation in complex technical systems*, Discussion Paper no. 53, Centre for Analysis of Risk and Regulation, London School of Economics, 2009.