LavoroDirittiEuropa
Rivista nuova di Diritto del Lavoro

Numero 2 / 2025

Giuseppe PRIMIERO

Giustizia predittiva: riflessioni da un'analisi formale

Giustizia predittiva: riflessioni da un'analisi formale

Giuseppe PRIMIERO

Ordinario di Logica e Filosofia della Scienza, Università degli Studi di Milano

Introduzione

La cosiddetta *giustizia predittiva* è spesso evocata come un mito tecnologico o come uno scenario distopico, plasmato da suggestioni letterarie e cinematografiche (su tutte, *Minority Report* di Philip K. Dick tradotto per il grande schermo da Steven Spielberg nel 2002), giuridiche (si veda ad esempio il dibattito sul *recidivism risk assessment* nei tribunali statunitensi, sollecitato dal caso del software COMPAS - Correctional Offender Management Profiling for Alternative Sanctions) ed etiche (tra tutti, citiamo i lavori di Virginia Eubanks e Cathy O'Neil). Oggi, tuttavia, alla luce delle prime applicazioni reali di sistemi automatizzati di supporto alle decisioni giuridiche, oggetto di critiche di opacità e presunti bias razziali (Angwin et al., *ProPublica*, 2016) — l'idea di un ausilio algoritmico alle pratiche giudiziarie non è più solo una futuribile astrazione. Una recente dettagliata analisi del rapporto tra Intelligenza Artificiale e processo decisionale nel contesto giudiziario è offerta in (Galli, Sartor 2023).

Un angolo d'analisi meno esplorato ma di grande rilevanza teorica è quello della **verifica formale**, disciplina cardine della logica applicata all'informatica teorica. Essa si occupa della dimostrazione rigorosa che un sistema computazionale soddisfi certe proprietà desiderate, espresse in un linguaggio formale. In altre parole, se si intende accertare che un algoritmo si comporti sempre in conformità a una specifica funzione — ad esempio, che non discrimini sulla base di attributi irrilevanti — si costruisce un **modello formale** dell'algoritmo e lo si confronta con una **specifica** espressa tramite formule logiche.

Nell'ambito della giustizia predittiva, un aspetto centrale è ovviamente quello del tempo. E nel contesto della verifica formale, questo elemento è trattato tramite logiche temporali applicate a sistemi modellati tramite transizioni di stato. Qui il processo di verifica consiste nel determinare se, a partire da ogni possibile configurazione iniziale, una determinata proprietà — quale, ad esempio, l'equità decisionale — sia mantenuta in almeno uno oppure in tutti gli stati *futuri* del sistema. Qualora tale condizione non sia soddisfatta, è possibile individuare le configurazioni iniziali che conducono alla violazione della specifica, evidenziando così l'insorgere di comportamenti indesiderati o inaccettabili. La verifica formale per l'analisi di sistemi che evolvono deterministicamente nel tempo (Kroening et al 2018, Baier & Katoen, 2008), si è progressivamente estesa a contesti più complessi, comprendendo sia s**istemi non deterministici**, in cui esistono molteplici transizioni possibili da uno stato all'altro; che s**istemi probabilistici**, in cui le evoluzioni del sistema sono associate a distribuzioni di probabilità (es. Markov Decision Processes, Discrete-Time Markov Chain, probabilistic model checking con strumenti come PRISM).

L'impiego della **verifica formale** nell'analisi di sistemi intelligenti applicati al contesto giudiziario apre ovviamente scenari di grande rilievo teorico e pratico. La dimensione temporale dell'evoluzione del sistema assume particolare rilevanza, offrendo un quadro concettuale utile per analizzare e, potenzialmente, *controllare* il funzionamento e le implicazioni etiche di sistemi predittivi. In questo contributo intendiamo offrire brevemente due riflessioni critiche su questa

possibilità: **la complessità computazionale** e **l'equità algoritmica**.
Da un lato, ci interessa sollevare la questione della difficoltà intrinseca nel determinare, con un margine di affidabilità sufficientemente robusto, la **probabilità che un certo evento si verifichi in uno stato futuro** di un sistema. Questo implica interrogarsi sulla possibilità effettiva di stabilire se un sistema computazionale progettato per supportare decisioni giuridiche soddisfi criteri di **correttezza formale**.
Dall'altro lato, proponiamo di mettere in relazione tali riflessioni tecniche con la questione, più ampia ma strettamente connessa, dell'**equità algoritmica**: ovvero la capacità di un sistema decisionale automatizzato di operare senza introdurre o amplificare discriminazioni nei confronti di gruppi o individui, nel rispetto dei principi fondamentali dello stato di diritto.

Un esempio

Consideriamo una struttura (nel nostro caso una DTMC - Discrete-Time Markov Chain) modellata per rappresentare uno scenario semplificato di giustizia predittiva, come ad esempio l'uso di un algoritmo per valutare il rischio di recidiva di un imputato. Notoriamente, un tale sistema è stato ad es. COMPAS, il quale fornisce un rischio predetto di recidiva per ciascun individuo. L'algoritmo si aggiorna a ogni nuova decisione giudiziaria (rilascio, detenzione) e ogni nuova osservazione del comportamento dell'individuo (commette o non commette reato). Vogliamo modellare l'evoluzione dello stato giudiziario dell'individuo nel tempo per come inteso dall'algoritmo.
Una DTMC è tipicamente composta da: un insieme S di stati possibili, un insieme T di transizioni tra stati, una funzione P che assegna ad ogni transizione una probabilità, e un insieme degli stati considerati iniziali. Gli stati rappresentano le condizioni dell'individuo combinate con la valutazione del sistema predittivo. Per esempio
$S = \{$*LowRisk_Free, LowRisk_Recid, HighRisk_Free, HighRisk_Recid, Jail*$\}$
dove:
- *LowRisk_Free*: L'individuo è stato etichettato come a basso rischio e rilasciato.
- *LowRisk_Recid*: Basso rischio, ma ha commesso un nuovo reato.
- *HighRisk_Free*: Valutato come alto rischio, ma rilasciato (per errore o sovrascrittura della valutazione).
- *HighRisk_Recid*: Alto rischio, e ha commesso recidiva.
- *Jail*: L'individuo è detenuto (transizione terminale o transitoria).

Le transizioni nell'insieme T sono probabilistiche e dipendono esclusivamente dallo stato corrente (la cosidetta proprietà di Markov). La funzione $P: S \times S \rightarrow [0,1]$ specifica per ogni coppia di stati la probabilità di transizione.
Su questa base, potremmo per esempio calcolare la probabilità di recidiva entro n passi (transizioni da stato a stato associate a istanti o fasi temporali), data una valutazione iniziale. Procediamo quindi con un **esempio concreto** di **probabilità di raggiungimento (detta *reachability*)** in una **DTMC per giustizia predittiva**, simulando una semplice dinamica, e **calcoliamo la probabilità** che un individuo valutato *a basso rischio* arrivi a commettere recidiva entro 2 passi. Prendiamo un sottoinsieme rilevante di stati dell'insieme S di sopra:

- *s0 = LowRisk_Free (inizio)*
- *s1 = LowRisk_Recid*
- *s2 = Jail*

- *s3 = LowRisk_Free (rimane in libertà)*

e le seguenti transizioni:

Da → A	**Probabilià**
s0 → s1 (recidiva)	0.1
s0 → s2 (carcere)	0.05
s0 → s3 (rimane free)	0.85
s3 → s1	0.1
s3 → s2	0.05
s3 → s3	0.85

Assumiamo che lo stato iniziale *P(s0) = 1*, ovvero che ci troviamo con certezza (probabilità 1) di fronte ad un individuo a basso rischio che rimane libero senza reati. Potremmo chiederci quale sia la **probabilità di raggiungere lo stato di recidiva (*s1*) entro 2 passi**, partendo da *s0*. Le possibilità di un passo di transizione da s0 sono

- *P1(s1) = 0.1 (diretto a recidiva)*
- *P1(s2) = 0.05 (va in carcere)*
- *P1(s3) = 0.85 (rimane libero)*

Da s3, le probabilità di transizione sono:

- *s3 → s1 = 0.1*
- *s3 → s2 = 0.05*
- *s3 → s3 = 0.85*

Quindi, la **probabilità di arrivare a *s1* al secondo passo** è:
*P(s1) = P(s0 → s3) * P(s3 → s1) = 0.85 * 0.1 = 0.085*

Cioè si moltiplica la probabilità che l'individuo resti libero con quella che poi raggiunga lo stato di recidiva. A questa va aggiunta la probabilità che vada direttamente in stato di recidiva:

P(recidiva entro 2 passi) = P1(s1) + P2(s1) = 0.1 + 0.085 = 0.185

Ovvero, la probabilità totale che il nostro individuo raggiunga lo stato di recidiva in al massimo due tranisizioni è del 18.5%. Questo calcolo si basa sul fatto che **la recidiva (*s1*) è uno stato finale** o che comunque ci fermiamo alla prima visita.

Questioni

Da questo piccolo esempio possiamo formulare alcune questioni fondamentali, non strettamente

formali: **quali sono le fonti e i criteri epistemici** attraverso cui attribuiamo una determinata probabilità al fatto che un individuo raggiunga lo stato di recidiva entro un numero limitato di passi del sistema? Quale giustificazione offriamo per assegnare una certa probabilità *P* a una specifica transizione tra stati? Secondo quali criteri viene fissata la distribuzione di probabilità iniziale del sistema? E soprattutto: **come possiamo assicurarci che tali scelte siano giuste ed eque** per tutti gli individui, indipendentemente dalle loro caratteristiche demografiche o socioeconomiche?
L'attuale paradigma dell'apprendimento automatico (machine learning) applicato a questioni sociali risponde ai dubbi di sopra basandosi su **tre assunzioni operative fondamentali**:

(1) la ***datafication***, ovvero l'idea che una quantità sufficientemente ampia di dati possa rappresentare fedelmente la realtà;
(2) la ***misurabilità dell'equità***, fondata su metriche standardizzate;
(3) il ***controllo del drift*** come garanzia della permanenza nel tempo di accuratezza e giustizia nei modelli.
La **prima assunzione** riflette la visione dominante del cosiddetto *soluzionismo algoritmico* (Morozov, 2013): se il mondo può essere osservato e descritto in termini di dati, allora problemi complessi – inclusi quelli sociali – possono essere risolti attraverso la modellazione computazionale. Questo approccio si è affermato soprattutto con il consolidarsi del paradigma dei *Big Data*, che ha rafforzato l'idea secondo cui un'elevata disponibilità di dati costituisce il fondamento imprescindibile per la classificazione accurata, la previsione attendibile e il controllo efficace dei fenomeni sociali (Kitchin, 2014).
La **seconda assunzione** concerne l'idea che **l'equità algoritmica possa essere definita, misurata e quindi ottimizzata**. Tra le principali metriche proposte in letteratura si annoverano:

- **Parità statistica (statistical parity)**: la probabilità di essere assegnati a una certa classe predetta deve essere uguale per i gruppi protetti e non protetti (Dwork et al., 2012);
- **Equalized Odds**: uguaglianza nei tassi di veri positivi e falsi positivi tra gruppi (Hardt, Price, & Srebro, 2016);
- **Equità controfattuale (counterfactual fairness)**: la previsione deve essere invariante rispetto a un cambiamento ipotetico del gruppo demografico (Kusner et al., 2017).

Queste metriche, tuttavia, **sono tra loro logicamente incompatibili** (Friedler et al., 2021), e nella pratica si tende a sceglierne una come sufficiente per soddisfare i requisiti di "giustizia" tecnologica. Tale scelta riflette un approccio riduzionista, che ignora le profonde divergenze normative tra concezioni distributive, correttive o trasformative della giustizia (Young, 1990; Sen, 2009).
La **terza assunzione** riguarda la possibilità di mantenere nel tempo equità e accuratezza attraverso il cosiddetto **aggiornamento dei dati** (*dataset refresh*). In questo modello operativo, il problema del ***data drift*** – ossia del cambiamento nelle distribuzioni dei dati in input rispetto a quelli su cui il modello è stato addestrato – viene trattato come una questione tecnica da risolvere con procedure di aggiornamento continuo (Lu et al., 2018). Il principio sottostante è che, se i dati rappresentano fedelmente il mondo, e se l'equità è stata codificata attraverso metriche adeguate, allora basterà mantenere aggiornati i dati per garantire decisioni giuste nel tempo.
Tuttavia, come ampiamente sottolineato dalla critica sociotecnica (Eubanks, 2018; O'Neil, 2016; Benjamin, 2019), **nessuna quantità di dati, nessuna metrica e nessuna procedura di**

aggiornamento è stata in grado, da sola, di affrontare le cause strutturali dell'ingiustizia sociale. Gli algoritmi non hanno aumentato l'accesso ai servizi di welfare per le fasce economicamente svantaggiate, né hanno migliorato le condizioni materiali delle comunità minoritarie, come quelle afroamericane o latinoamericane, né tantomeno hanno protetto gruppi vulnerabili da pratiche di sorveglianza fiscale o discriminazione sistemica.

In definitiva, l'adozione acritica di questi tre assunti rischia di **consolidare le disuguaglianze esistenti sotto una parvenza di neutralità computazionale**, mascherando la natura politica delle scelte insite nei modelli algoritmici e rinunciando a un vero ripensamento delle forme di giustizia nelle società digitali.

Intendiamo ora mostrare come la modellizzazione formale del semplice esempio presentato nella sezione precedente offra già spunti significativi per comprendere la centralità delle questioni di **complessità computazionale** ed **equità** nella valutazione della **correttezza algoritmica** di sistemi di apprendimento automatico applicati a contesti ad alta sensibilità sociale, come quello della giustizia predittiva.

Complessità della correttezza predittiva

La definizione degli **stati** e delle **transizioni** di un sistema, così come delle **probabilità associate** agli uni e alle altre, rappresenta un passaggio cruciale nella modellizzazione formale di un sistema computazionale volto alla verifica della sua correttezza. Un'analisi rigorosa richiede, infatti, la capacità di rispondere con elevato grado di precisione alle seguenti domande: quali sono gli stati rilevanti del sistema? Come si definiscono le transizioni possibili tra questi stati? E con quali probabilità esse avvengono?

Nel caso in cui si voglia verificare, ad esempio, se un algoritmo stimi correttamente la probabilità che un individuo classificato come *"a basso rischio"* possa raggiungere uno stato di **recidiva entro due passi**, diventa essenziale specificare con chiarezza quali **stati intermedi** si interpongano tra la condizione iniziale e quella finale. Ricordiamo infatti che per la cosidetta proprietà di Markov, la probabilità di uno stato è interamente determinata dalla probabilità di transizioni a partire dagli stati che lo precedono. Questo significa che, ad esempio, introdurre uno stato intermedio che rappresenti l'accesso dell'individuo a un **programma di reinserimento lavorativo**, rappresenta un intervento che avrà necessariamente un effetto e (potremmo ipotizzare) riduce sensibilmente la probabilità di recidiva o di incorrere in un nuovo arresto. Le probabilità da **attribuire agli stati** non dovrebbero essere dunque valori arbitrari: ma sono sufficienti i **dati empirici** raccolti per l'addestramento del modello a determinarli correttamente? La qualità, la rappresentatività e la neutralità di questi dati sono determinanti nel garantire che le probabilità così ottenute siano **affidabili e giustificabili**. Solo a queste condizioni è possibile utilizzare la modellizzazione per una verifica significativa della correttezza (e dell'equità) del sistema predittivo.

Il processo di progettazione di un modello formale per la verifica di un sistema di apprendimento automatico presenta dunque un livello intrinseco di **complessità modellistica** che incide in modo determinante sull'esito della valutazione. Una variazione apparentemente marginale – ad esempio, l'attribuzione di una probabilità di recidiva del **15% invece che del 18,5%** – può tradursi in **conseguenze sostanziali**, come la differenza tra l'assegnazione alla libertà vigilata e la permanenza in uno stato di restrizione della libertà personale (Angwin et al., 2016).

Questa complessità non è solo di natura tecnica, ma riflette scelte **epistemiche e normative** lungo tutto il ciclo di sviluppo: dalla **selezione e rappresentazione dei dati di addestramento**, alla loro **categorizzazione**, fino alla definizione – spesso arbitraria o scarsamente giustificata – di **soglie decisionali** (*decision thresholds*) per la classificazione automatica (Barocas, Hardt, & Narayanan, 2019). Tali soglie, benché tecnicamente configurabili, comportano effetti normativi

impliciti, poiché stabiliscono confini discreti su continui di rischio o merito, influenzando direttamente **l'accesso a diritti o l'esercizio di libertà fondamentali** (Eubanks, 2018; Selbst & Barocas, 2018).
Nel contesto della **giustizia predittiva**, queste scelte assumono un'importanza critica: la loro opacità e il loro potenziale impatto discriminatorio richiedono non solo trasparenza tecnica, ma anche una **valutazione etica e giuridica strutturata**, orientata al principio di giustizia sostanziale, e non solo procedurale.
Infine, non possiamo non osservare che la nostra capacità di costruire modelli fedeli di comportamento di un sistema complesso ed incerto come quello di un algoritmo di apprendimento automatico è ulteriormente limitato dalla opacità di questi sistemi e dall'eventualità che l'accesso sia limitato da questioni di segreto industriale.

Equità nelle valutazioni iniziali e nelle transizioni
Un secondo elemento cruciale nell'analisi della modellizzazione di sistemi soggetti a evoluzione temporale per la verifica della raggiungibilità di proprietà riguarda l'attribuzione di valutazioni probabilistiche sia agli stati iniziali, sia alle transizioni tra stati. In un contesto computazionale tradizionale — ad esempio, nella modellazione di un sistema di gestione del traffico — tali assegnazioni sono generalmente motivate da scelte di design predeterminate. Si pensi, ad esempio, alla necessità che in un sistema semaforico la probabilità che un semaforo sia rosso quando l'altro è verde sia determinata al fine di evitare incidenti; oppure al vincolo secondo cui la transizione dal rosso al giallo debba avvenire con la stessa probabilità della transizione dal verde al giallo, per ragioni di simmetria e sicurezza.
Tuttavia, quando tali strumenti di modellizzazione vengono applicati a contesti di rilevanza sociale, la neutralità di queste scelte viene meno. La determinazione delle probabilità iniziali e delle traiettorie evolutive plausibili assume allora una valenza normativa, e può generare o amplificare situazioni di diseguaglianza. Si consideri, a titolo esemplificativo, un sistema che, basandosi su dati storici, attribuisce una probabilità di transizione dallo stato "libero" a quello "recidivo" pari a 0.1 per un individuo con determinate caratteristiche (ad esempio, bianco) e pari a 0.3 per un individuo con caratteristiche differenti (ad esempio, un individuo di colore). In tal caso, si costruisce implicitamente una struttura decisionale che codifica una maggiore probabilità di recidiva per individui appartenenti a determinati gruppi demografici, configurando una dinamica discriminatoria.
Quando queste stime probabilistiche vengono presentate come estrapolazioni "scientifiche" da dati presuntamente oggettivi, la questione si sposta inevitabilmente di nuovo sulla qualità del dato stesso. Tipologia, completezza, e coerenza dei dati raccolti concorrono infatti a costruire una rappresentazione del mondo che può riflettere — e talvolta rafforzare — pregiudizi sistemici. Ad esempio, una narrativa secondo cui è più probabile che un uomo di colore sia recidivo rispetto a un uomo bianco può emergere per ragioni fattuali da una raccolta e interpretazione dei dati strutturalmente distorte. È dunque responsabilità etica e scientifica di chi sviluppa questi modelli assicurarsi che tali distorsioni siano identificate, comprese e, ove possibile, corrette anche in una prospettiva di evoluzione temporale e con l'obiettivo del miglioramento dello *status quo* (Quaresmini & Primiero, 2024).

Conclusioni su controllo ed equità

L'adozione sempre più ampia di sistemi di intelligenza artificiale (IA) in ambiti ad alta rilevanza sociale – come la giustizia predittiva, l'assistenza sociale o il credito – ha posto con urgenza la questione del **controllo tecnico** sui modelli e della **garanzia di equità** nei loro risultati. Questi due aspetti, spesso trattati separatamente, devono essere considerati congiuntamente: un sistema è controllabile solo se il suo comportamento è comprensibile e prevedibile; e tale controllo ha valore solo se è esercitato per garantire l'equità verso tutte le classi di individui.
Il controllo nei sistemi di IA può essere inteso come la capacità di **prevedere e limitare il comportamento del sistema** entro margini accettabili, rispetto a determinati requisiti. In contesti ad alta sensibilità etica, come quello giudiziario, questo implica la possibilità di verificare che un sistema risponda a requisiti di accuratezza, robustezza e non discriminazione. Tuttavia, molti modelli di apprendimento automatico, soprattutto quelli di tipo black-box, sfuggono a una valutazione formale della loro correttezza. Ciò rende difficile dimostrare, ad esempio, che una decisione non sia stata influenzata da **bias latenti** presenti nei dati (Barocas et al., 2019; Selbst & Barocas, 2018).
La formalizzazione dell'equità nei sistemi IA si basa spesso su metriche come la **parità statistica**, la **parità di opportunità** o la **fairness controfattuale**. Tuttavia, ciascuna di queste metriche implica una visione normativa differente della giustizia, ed è noto che esse non possono essere soddisfatte contemporaneamente in presenza di distribuzioni di base differenti (Kleinberg et al., 2016). Inoltre, l'adozione di una metrica rispetto a un'altra non è una decisione neutra: essa riflette un modello implicito di **giustizia sociale** che può favorire o penalizzare certi gruppi.
Un sistema è pienamente controllabile solo se possiamo **verificare formalmente** che esso rispetti determinate proprietà di equità. Questo richiede la modellizzazione esplicita delle transizioni tra stati, delle distribuzioni di probabilità, e dei meccanismi decisionali interni. In particolare, l'uso di **modelli probabilistici formalmente verificabili --** come ad esempio le catene di Markov a tempo discreto (DTMC) per sistemi non deterministici che evolvono nel tempo -- può costituire uno strumento utile per analizzare in modo sistematico i rischi di **disparità nei risultati** e per esplorare gli effetti controfattuali di politiche alternative.
Tuttavia, la complessità che abbiamo illustrato nel modellare i processi decisionali che sistemi automatici dovrebbero riprodurre e predirre ci impone di trovare metodi di verifica e validazione che siano appropriati e applicabili. Questa necessità richiede la combinazione di verifica *ex-ante,* dove il il controllo avviene per guidare lo sviluppo, e verifica *post-hoc* (e.g. la strategia usata in (Coraglia et al. 2023)*,* dove il processo di controllo è attuato per verificare che il comportamento di modelli già esistenti e a volte inaccessibili sia corretto secondo criteri definiti dai supervisori umani.
Controllo ed equità non possono più essere considerati elementi secondari nell'ingegneria dei sistemi intelligenti: devono essere progettati fin dall'inizio come **vincoli strutturali** del sistema stesso. Il loro intreccio è il terreno su cui si gioca la legittimità dell'IA nei processi decisionali pubblici come quello della giustizia.
L'analisi formale qui considerata si applica tipicamente al risultato prodotto da un modello computazionale, inteso come l'output finale generato da un sistema automatico di supporto alla decisione. Questo risultato è oggetto di valutazione tecnica sulla base di criteri algoritmici e di conformità al modello sottostante, coerentemente con gli approcci della logica computazionale e della verifica formale.

Nel contesto giuridico, tuttavia, la valutazione di tale output da parte del giudice non si limita alla sua correttezza formale o tecnica. Al contrario, essa si inserisce all'interno di un processo interpretativo complesso, in cui il giudice esercita un'autonoma capacità fondata sulla conoscenza della normativa vigente, sull'analisi del caso concreto e sulla valutazione di fattori extratecnici quali i principi generali dell'ordinamento, l'equità e il contesto socioculturale della decisione.
In quest'ottica, è auspicabile che il giudice sia giuridicamente legittimato a discostarsi dall'esito fornito da un sistema automatico, anche qualora questo sia tecnicamente corretto. Tale posizione è coerente con il principio della centralità dell'essere umano nei processi decisionali automatizzati, sancito, tra gli altri, dal Regolamento (UE) 2016/679 (GDPR), art. 22, e ripreso nella proposta di AI Act della Commissione Europea, che impone meccanismi di controllo umano ("human oversight") su sistemi di intelligenza artificiale ad alto rischio.
In sintesi, la correttezza tecnica dell'output non garantisce di per sé l'accettabilità giuridica o la vincolatività della decisione, riaffermando la distinzione fondamentale tra validazione tecnica e legittimazione giuridica della decisione automatica (Floridi et al., 2018; Hildebrandt, 2015).

Bibliografia

Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (2016, May 23). *Machine bias: There's software used across the country to predict future criminals. And it's biased against Blacks*. ProPublica. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

Baier, C., & Katoen, J.-P. (2008). *Principles of model checking*. MIT Press.

Barocas, S., Hardt, M., & Narayanan, A. (2019). *Fairness and machine learning: Limitations and opportunities*. http://fairmlbook.org

Benjamin, R. (2019). *Race after technology: Abolitionist tools for the new Jim code*. Polity Press.

Coraglia, G., D'Asaro, F. A., Genco, F. A., Giannuzzi, D., Posillipo, D., Primiero, G., & Quaggio, C. (2024). BRIOxAlkemy: A bias detecting tool. In G. Boella, F. A. D'Asaro, A. Dyoub, L. Gorrieri, F. A. Lisi, C. Manganini, & G. Primiero (Eds.), *BEWARE 2023: Bias, Ethical AI, Explainability and the Role of Logic and Logic Programming* (Vol. 3615, pp. 44–60). CEUR Workshop Proceedings. https://ceur-ws.org/Vol-3615/paper4.pdf

Dwork, C., Hardt, M., Pitassi, T., Reingold, O., & Zemel, R. (2012). Fairness through awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference* (pp. 214–226). https://doi.org/10.1145/2090236.2090255

Eubanks, V. (2018). *Automating inequality: How high-tech tools profile, police, and punish the poor*. St. Martin's Press.

Floridi, L., Cowls, J., Beltrametti, M. *et al.* AI4People—An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations. *Minds & Machines* **28**, 689–707 (2018). https://doi.org/10.1007/s11023-018-9482-5

Friedler, S. A., Scheidegger, C. and Venkatasubramanian, S. (2021). The (Im)possibility of fairness: different value systems require different mechanisms for fair decision making. Commun. ACM 64, 4 (April 2021), 136–143. https://doi.org/10.1145/3433949

Galli, F., & Sartor, G. (2023). AI approaches to predictive justice: A critical assessment. *Human(ities) and Rights – Global Network Journal*, 5(2), 165–217.

Hardt, M., Price, E., & Srebro, N. (2016). Equality of opportunity in supervised learning. In *Advances in Neural Information Processing Systems*, 29. https://proceedings.neurips.cc/paper_files/paper/2016/hash/9d2682367c3935defcb1f9e247a97c0d-Abstract.html

Hildebrandt, M. (2015). *Smart technologies and the end(s) of law: Novel entanglements of law and technology*. Edward Elgar Publishing.

Huth, M., & Ryan, M. (2004). *Logic in computer science: Modelling and reasoning about systems* (2nd ed.). Cambridge University Press.

Kitchin, R. (2014). *The data revolution: Big data, open data, data infrastructures and their consequences*. SAGE.

Daniel Kroening, Doron Peled, Edmund M. Clarke Jr., Helmut Veith and Orna Grumberg (2018), *Model Checking*, 2nd edition, MIT Press.

Kusner, M. J., Loftus, J., Russell, C., & Silva, R. (2017). Counterfactual fairness. In *Advances in*

Neural Information Processing Systems, 30. https://proceedings.neurips.cc/paper_files/paper/2017/hash/a486cd07e4ac3d270571622f4f316ec5-Abstract.html

Kwiatkowska, M., Norman, G., & Parker, D. (2011). PRISM 4.0: Verification of probabilistic real-time systems. In Gopalakrishnan, G., & Qadeer, S. (Eds.), *Computer Aided Verification (CAV 2011)* (pp. 585–591). Springer. https://doi.org/10.1007/978-3-642-22110-1_47

Lu, J., Liu, A., Dong, F., Gu, F., Gama, J., & Zhang, G. (2018). Learning under concept drift: A review. *IEEE Transactions on Knowledge and Data Engineering*, 31(12), 2346–2363. https://doi.org/10.1109/TKDE.2018.2876857

Middlestadt, B. D., Allo, P., Taddeo, M., Wachter, S., & Floridi, L. (2016). The ethics of algorithms: Mapping the debate. *Big Data & Society*, 3(2), 1–21. https://doi.org/10.1177/2053951716679679

Morozov, E. (2013). *To save everything, click here: The folly of technological solutionism*. PublicAffairs.

O'Neil, C. (2016). *Weapons of math destruction: How big data increases inequality and threatens democracy*. Crown Publishing.

Quaresmini, C., & Primiero, G. (2024). Data quality dimensions for fair AI. In R. Calegari, V. Dignum, & B. O'Sullivan (Eds.), *Proceedings of the 2nd Workshop on Fairness and Bias in AI (AEQUITAS 2024), co-located with the 27th European Conference on Artificial Intelligence (ECAI 2024)* (Vol. 3808, pp. 1–16). CEUR-WS.org. https://ceur-ws.org/Vol-3808/paper12.pdf

Sen, A. (2009). *The idea of justice*. Harvard University Press.

Selbst, A. D., & Barocas, S. (2018). The intuitive appeal of explainable machines. *Fordham Law Review*, 87(3), 1085–1139. https://ir.lawnet.fordham.edu/cgi/viewcontent.cgi?article=5569&context=flr

Wachter, S., Mittelstadt, B., & Floridi, L. (2017). Why a right to explanation of automated decision-making does not exist in the General Data Protection Regulation. *International Data Privacy Law*, 7(2), 76–99. https://doi.org/10.1093/idpl/ipx005

Young, I. M. (1990). *Justice and the politics of difference*. Princeton University Press.