LIBER AMICORUM
PER
PASQUALE COSTANZO

PAOLO ZUDDAS

INTELLIGENZA ARTIFICIALE E DISCRIMINAZIONI

16 MARZO 2020

CONSULTA ONLINE
PERIODICO TELEMATICO ISSN 1971-9892

LIBER AMICORUM
PER
PASQUALE COSTANZO

Paolo Zuddas
Intelligenza artificiale e discriminazioni

SOMMARIO: 1. Introduzione. – 2. Le principali fasi della decisione algoritmica in cui si pongono le premesse per le *AI-driven discriminations*. – 2.1. La programmazione: le discriminazioni algoritmiche dirette e indirette e i problemi connessi all'individuazione dei caratteri *proxy*. – 2.2. La configurazione del set di dati: i rischi derivanti dai dati "inquinati" e dai dati incompleti. – 2.3. (*Segue*) i margini di autonomia dell'algoritmo. – 3. Le soluzioni: il ruolo del diritto e i suoi limiti. – 3.1. I profili "interni" della decisione algoritmica: il divieto di produrre effetti discriminatori tra dimensione tecnica e dimensione etica. – 3.2. I profili "esterni": la significatività della decisione algoritmica ed i margini dell'intervento umano. – 4. In conclusione.

1. *Introduzione*

Prima di affrontare il tema posto ad oggetto dell'indagine, si ritiene opportuno offrire alcune definizioni relative a strumenti dei quali si discuterà nelle pagine seguenti, accennando sinteticamente al loro funzionamento. Tali strumenti sono essenzialmente tre: algoritmo, intelligenza artificiale, *machine learning*.

L'algoritmo può essere definito come la descrizione astratta e formalizzata di una procedura computazionale[1]: in particolare, la trascrizione di un algoritmo mediante un linguaggio di programmazione dà luogo al "programma" che consente al computer di svolgere specifiche operazioni.

Si definisce intelligenza artificiale quel ramo dell'informatica che concerne la progettazione di sistemi sia *hardware* che *software* che consentono di dotare le macchine di alcune caratteristiche tipicamente umane, quali le percezioni visive, spazio-temporali o decisionali[2].

Il *machine learning*, invece, è un tipo specifico di intelligenza artificiale che consiste in un processo automatico di individuazione di correlazioni tra variabili all'interno di un set di dati, allo scopo di compiere previsioni o stime di certi effetti[3]. In sostanza, ai sistemi di *machine learning* – o di apprendimento automatico – viene assegnato un obiettivo e fornita una vasta mole di dati da utilizzare come esempi del modo in cui l'obiettivo può essere raggiunto o dai quali far derivare modelli di decisione. Il sistema, analizzando i dati forniti, "impara" come meglio realizzare l'obiettivo richiesto[4].

---

[1] Cfr. T.Z. ZARSKY, *An analytic challenge: discrimination theory in the age of predictive analytics*, in 14 *I/S: A Journal of Law and Policy for the Information Society*, 2017, 12.

[2] In tema cfr., per tutti, M. SOMALVICO, *L'intelligenza artificiale*, Milano, 1987.

[3] Cfr. D. LEHR-P. OHM, *Playing with the data: What legal scholars should learn about machine learning*, in 51 *University of California, Davis, Law Review*, 2017, 671.

[4] Cfr. ROYAL SOCIETY (UK), *Machine learning: the power and promise of computers that learn by example* (April 2017), 19.

Il *machine learning*, dunque, basandosi principalmente su dati statistici, utilizza il metodo induttivo in alternativa al metodo logico-deduttivo della programmazione attraverso regole. Questo sistema non soltanto contraddice il metodo tradizionale seguito nella programmazione informatica ma, come è stato più volte osservato, sconta un duplice errore logico: per un verso, la pretesa di trarre forme di causazione da correlazioni statistiche (come ricorda, da ultimo, A. CELOTTO, *Come regolare gli algoritmi. Il difficile bilanciamento tra scienza, etica e diritto*, in *Analisi giuridica dell'Economia*, 1/2019, 48); per altro verso, e con specifico riguardo all'impiego di tali strumenti ai fini di produzione giuridica, la pretesa di far derivare dall'essere il dover essere (come osservano A. SIMONCINI e S. SUWEIS, *Il cambio di paradigma nell'intelligenza artificiale e il suo impatto sul diritto costituzionale*, in *Rivista di filosofia del diritto*, 1/2019, 102, rievocando la critica mossa da David Hume al giusnaturalismo). Questo non significa che nei sistemi di *machine learning* non esistano regole di partenza che precedono la "nutrizione" dell'algoritmo con i dati da utilizzare. Il vero problema è che, come vedremo, la macchina è predisposta in modo da "apprendere autonomamente", a partire dalle informazioni fornite, ai fini della risoluzione del problema posto; ciò comporta peraltro che non risulti sempre evidente e comprensibile la logica seguita nella selezione dei caratteri ricorrenti e il meccanismo di correlazione fra di loro (cfr. sul punto A. SIMONCINI-S. SUWEIS, *op. cit.*, 92). Anche sotto questo profilo, pertanto, vengono meno elementi

Correlata al concreto funzionamento dei sistemi di *machine learning* è la disponibilità e l'utilizzo dei cosiddetti *big data*, cioè di un insieme di dati estremamente cospicuo, che viene impiegato da algoritmi noti come "reti neurali" per rendere possibile l'apprendimento automatico[5]. Tanto più ampia, infatti, sarà la mole di dati messa a disposizione per alimentare il sistema di *machine learning*, tanto maggiore risulterà la capacità di apprendimento del sistema, e tanto più precisa ed efficace sarà la risposta data dall'algoritmo al problema posto.

In proposito occorre sin d'ora precisare che, se la sede più è rilevante di reperimento dei *big data* è internet, ed è stata proprio la Rete – e l'immensa quantità di informazioni in essa custodita – che ha consentito lo sviluppo dei sistemi di apprendimento automatico[6], i *big data* utilizzati dai sistemi di *machine learning* non provengono sempre dal *web*[7]. Ad oggi sono infatti disponibili numerose banche dati informatiche non necessariamente collegate alla rete internet, che possono essere in grado di offrire un adeguato *training* ai sistemi di *machine learning*. Si consideri, in proposito, che alcuni tra i più noti esperimenti che hanno visto impegnati sistemi di apprendimento automatico in attività creative particolarmente evolute si sono realizzati a seguito di *training* basati su banche dati specializzate, contenenti informazioni solo in parte disponibili in Rete: così è avvenuto nel caso delle intelligenze artificiali "alimentate" con i quadri di Rembrandt[8] o con gli spartiti di Bach[9], o che hanno imparato a dipingere elaborando le immagini di capolavori dell'arte figurativa tratte dagli archivi museali[10].

---

essenziali che caratterizzano la norma: da un lato, l'assenza di una *ratio*, cioè di una ragione giustificativa della regola (che contraddice un elemento qualificante del diritto per cui esso esprime sempre una ragione per agire in un certo modo) compromette la forza persuasiva su cui si fonda la sua obbligatorietà (e la stessa normatività); dall'altro, si contraddice un "dogma" del diritto contemporaneo, e cioè il "nesso di causalità tra i fenomeni umani", per cui la causa precede l'effetto e mai il contrario (anche sul tema spunti in A. SIMONCINI-S. SUWEIS, *op. cit.*, 93 ss.): nei sistemi di *machine learning*, infatti, le "cause" (le correlazioni poste alla base dello schema decisionale proposto dall'algoritmo) si fanno derivare dagli "effetti" (il set di dati che incorpora i modelli di decisione che emergono dalle scelte compiute in passato).

[5] Cfr. A. SIMONCINI-S. SUWEIS, *op. cit.*, 87 ss.

[6] Cfr. A. SIMONCINI-S. SUWEIS, *op. cit.*, 87 ss. Come è noto, diversi tentativi di sviluppare sistemi di *machine learning* furono operati a partire dalla fine degli anni '50 (la data di nascita ufficiale è di norma indicata nel 1956, anno in cui si organizzò il primo seminario così denominato presso il *Dartmouth College* di Hanover nel New Hampshire), ma la loro successiva implementazione fu ostacolata principalmente dalla impossibilità di reperire una mole di dati sufficiente a consentirne un corretto funzionamento (per un'accurata ricostruzione della storia dell'intelligenza artificiale, arricchita da utili spunti relativi agli scenari futuri, cfr. H. HAENLEIN-A. KAPLAN, *A brief history of artificial intelligence: On the past, present, and future of artificial intelligence*, in *California Management Review*, 2019).

[7] Sul punto si vedano ad esempio D.M. BOYD-K. CRAWFORD, *Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon* in 15 *Information, Communication & Society*, 2012, 662 ss. che definiscono i *big data* semplicemente come un insieme di dati estremamente cospicuo.

[8] Ci si riferisce al progetto di ricerca chiamato "The Next Rembrandt" e promosso negli Stati Uniti dalla *Smithsonian Institution* e condotto in collaborazione con le aziende *Ing* e *Microsoft*, con la consulenza del Politecnico olandese di Delft (TuDelft) e del museo *Rembrandthui*, che, utilizzando sistemi di apprendimento automatico, ha prodotto un ritratto nel puro stile di Rembrandt. Nell'arco di un anno e mezzo ingegneri, storici dell'arte e analisti dei dati hanno studiato 346 opere di Rembrandt, dopodiché hanno "addestrato" un sistema informatico a dipingere riproducendo esattamente lo stile dell'artista. In particolare, il sistema di intelligenza artificiale ha studiato oltre 168.000 particolari dei dipinti del pittore olandese, memorizzando le proporzioni del viso, la distanza fra gli occhi, la pennellata. Sulla base di questi dato ha prodotto un disegno, che poi una stampante 3D ha trasformato in un dipinto, sovrapponendo gli strati di colore in modo da riprodurre il tratto di Rembrandt (cfr. S. YANISKY-RAVID-S. MOORHEAD, *Generating Rembrandt: Artificial Intelligence, Accountability and Copyright – The Human-like Workers are Already Here – A New Model*, in 9 *Michigan State Law Review*, 2017).

[9] Il 21 marzo 2019 *Google* ha lanciato il suo primo *doodle* basato sull'intelligenza artificiale, che consente, attraverso un sistema di apprendimento automatico, di creare, a partire da una qualsiasi melodia caricata sul sistema, un'armonia musicale nello stesso stile barocco di Johann Sebastian Bach. Il progetto, nato in collaborazione con i team di *Google Magenta* e *PAIR*, è stato reso possibile a seguito della "nutrizione" dell'AI con 306 corali del compositore tedesco (cfr. in argomento M. KUROSU (a cura di), *Human-Computer Interaction: Perspectives on Design; Thematic Area, HCI 2019, Held As Part of the 21st HCI International Conference, HCII 2019 Orlando, USA, July 26–31, 2019, Proceedings, Part I*, Springer, 2019, 291 ss.).

[10] Ci si riferisce ad *AICAN*, un programma creato da *Al Rutgers' Art & AI Lab*, in grado di riconoscere stili pittorici differenti e di generare immagini innovative proprie. In particolare, l'algoritmo è stato alimentato da 80.000 immagini

Infine, va ricordato che uno degli impieghi più controversi dell'intelligenza artificiale, in relazione al quale emergono con maggiore evidenza le potenzialità discriminatorie delle decisioni algoritmiche, concerne la cosiddetta profilazione: essa consiste essenzialmente nel ricorso a sistemi automatici di elaborazione di dati per sviluppare profili che possono essere usati per prendere decisioni riguardanti persone[11].

Ciò premesso, veniamo al punto: costituisce un'evidenza ormai difficilmente eludibile il progressivo coinvolgimento dell'intelligenza artificiale nella formulazione di decisioni che producono un impatto significativo sulla vita individuale e collettiva e che sempre più interessano funzioni pubbliche fondamentali come l'amministrazione della giustizia; interessi pubblici primari come la sicurezza, la prevenzione dei reati, il contrasto al terrorismo, o il controllo dei flussi migratori; servizi pubblici essenziali come il sistema d'istruzione o l'accesso ad altri servizi sociali[12]. In tutti questi ambiti, peraltro, l'intelligenza artificiale si è dimostrata in grado di produrre risultati estremamente proficui – soprattutto in termini di maggiore efficienza e rapidità dei processi decisionali – tra i quali cui può ricomprendersi a pieno titolo anche una riduzione delle diseguaglianze [13].

Tuttavia, è vero anche il contrario. I sistemi di intelligenza artificiale, infatti, non soltanto possono commettere errori[14], ma possono anche assumere – in assenza di adeguati interventi preventivi (e talvolta, come vedremo, anche *malgrado* l'intervento preventivo) – decisioni fortemente discriminatorie[15]. L'evenienza più frequente è quella dei c.d. *bias* (associati di norma ai sistemi di *machine learning* note come reti neurali, prima accennate), cioè delle distorsioni che interessano quei sistemi informatici che "discriminano sistematicamente e ingiustamente certi individui o gruppi di individui a favore di altri", negando opportunità o generando risultati indesiderati per motivi irragionevoli o inappropriati[16].

Non a caso, tutti i principali atti di *soft law* che disciplinano a vario titolo l'utilizzo dell'intelligenza artificiale mettono in guardia rispetto al rischio di discriminazione e di *bias*.

---

che rappresentano il canone dell'arte occidentale nei cinque secoli precedenti (cfr. B. SALEH-A. ELGAMMAL, *Large-scale Classification of Fine-Art Paintings: Learning The Right Metric on The Right Feature*, in *Digital Art History*, 2/2016, 71 ss.).

[11] In argomento cfr., *ex multis*, M. HILDEBRANDT, *Defining profiling: a new type of knowledge?* in M. HILDEBRANDT-S. GUTWIRTH (a cura di), *Profiling the European citizen*, Springer, Berlin, 2008. Per dettagli ulteriori v. il Considerando 71 del regolamento (UE) 2016/679 del Parlamento europeo e del Consiglio, del 27 aprile 2016 (GDPR), citato *infra*.

[12] Cfr. sul punto, *ex multis*, G. RESTA, *Governare l'innovazione tecnologica: decisioni algoritmiche, diritti digitali e principio di uguaglianza*, in *Politica del diritto*, 2/2019, 211-212.

[13] Sul punto v. ancora G. RESTA, *op. cit.*, 218, che richiama in proposito tra l'altro: l'allocazione mirata di prestazioni sociali; lo stimolo allo sviluppo di processi partecipativi; il contrasto a frodi e all'evasione fiscale.

[14] Ciò accade quando ad esempio un algoritmo non riesce a identificare correttamente un soggetto o elabora previsioni fallaci, pur senza penalizzare sistematicamente alcuna categoria in particolare.

Si ricordi in proposito che l'esigenza di minimizzare il rischio di errori è presa in considerazione, da ultimo, dal GDPR, che nel Considerando 71, ribadisce l'opportunità che il titolare del trattamento di dati personali, "al fine di garantire un trattamento corretto e trasparente", metta in atto "misure tecniche e organizzative adeguate al fine, in particolare, che siano rettificati i fattori che comportano inesattezze dei dati e sia minimizzato il rischio di errori".

[15] Ciò accade – come sarà meglio precisato *infra* – quando un algoritmo formula previsioni scorrette discriminando regolarmente specifiche classi di soggetti. Cfr. Sul punto M. GALEOTTI, *Discriminazioni e algoritmi. Incontri e scontri tra diverse idee di* fairness, in *The Lab's Quarterly*, 4/2018, 73 ss.

[16] Così B. FRIEDMAN-H. NISSENBAUM, *Bias in Computer Systems*, in 14 *ACM Transactions on Information Systems*, 1996, 332: "*Accordingly, we use the term bias to refer to computer systems that systematically and unfairly discriminate against certain individuals or groups of individuals in favor of others. A system discriminates unfairly if it denies an opportunity or a good or if it assigns an undesirable outcome to an individual or group of individuals on grounds that are unreasonable or inappropriate*" (cfr. sul punto anche A. SIMONCINI-S. SUWEIS, *op. cit.*, 96). Posto che la discriminazione è di per sé un trattamento differenziato irrazionale, irragionevole, improprio e ingiusto, il termine più rilevante nella definizione richiamata è "sistematicamente": ciò significa, nel caso più evidente (ma, come vedremo, si verificano anche situazioni molto diverse, nelle quali l'effetto discriminatorio si realizza in modo indiretto e meno vistoso) che l'appartenenza alla categoria discriminata viene individuata dall'algoritmo come carattere rilevante per il conseguimento di un risultato indesiderato, come ad esempio l'esclusione da un beneficio.

Volendo limitare il richiamo ai documenti più rilevanti adottati negli ultimi anni in ambito europeo e nazionale, può essere utile ricordare che il principio di non discriminazione viene evocato: nella risoluzione concernente norme di diritto civile sulla robotica approvata dal parlamento europeo nel febbraio 2017 [17]; nella risoluzione su una politica industriale europea globale in materia di robotica e intelligenza artificiale approvata dal Parlamento europeo nel febbraio 2019[18]; negli orientamenti etici per un'intelligenza artificiale affidabile adottati dal gruppo di esperti istituito dalla Commissione europea nel giugno 2018[19]; nella carta etica europea sull'utilizzo dell'intelligenza artificiale nei sistemi giudiziari, adottata dalla commissione europea per l'efficacia nella giustizia del Consiglio d'Europa nel dicembre 2018[20]; nel libro bianco sull'intelligenza artificiale al servizio del cittadino adottato dall'Agenzia per l'Italia digitale nel marzo 2018[21]; nel documento che illustra la strategia nazionale per l'intelligenza artificiale predisposta dal Ministero dello sviluppo economico nel luglio 2019[22].

In che modo l'intelligenza artificiale può discriminare? All'interno della ormai amplissima letteratura disponibile in argomento, soprattutto a livello internazionale, compaiono numerosi schemi esplicativi, alcuni dei quali estremamente articolati, che individuano le possibili cause o sorgenti dei *bias* e, più in generale, delle cosiddette "*AI-driven discriminations*", cioè delle discriminazioni indotte dall'intelligenza artificiale[23]. Tuttavia, possano individuarsi essenzialmente tre ambiti – o tre diverse fasi della decisone algoritmica – nei quali possono porsi le premesse per il prodursi di un effetto discriminatorio legato all'uso di sistemi di intelligenza artificiale.

---

[17] Cfr. *Risoluzione del Parlamento europeo del 16 febbraio 2017 recante raccomandazioni alla Commissione concernenti norme di diritto civile sulla robotica*, 2015/2103(INL), *Principi etici*, nn. 10 e 13.

[18] Cfr. *Risoluzione del Parlamento europeo del 12 febbraio 2019 su una politica industriale europea globale in materia di robotica e intelligenza artificiale*, 2018/2088(INI), *1.2. Utilizzo doloso dell'intelligenza artificiale e diritti fondamentali*, n. 12; *2.4. Condizioni di sostegno: connettività, accessibilità dei dati, calcolo ad alte prestazioni e infrastruttura cloud*, n. 39; *5.2. Valori incorporati nella tecnologia – "ethical-by-design"*, n. 147; *5.4. Trasparenza, distorsioni e spiegabilità degli algoritmi*, nn. 157, 177 e 180; *6.1. Coordinamento a livello di Unione*, n. 187.

[19] Cfr. *Orientamenti etici per un'IA affidabile*, predisposta dal Gruppo di esperti ad alto livello sull'intelligenza artificiale istituito dalla Commissione europea nel giugno 2018, nel testo reso pubblico l'8 aprile 2019; in particolare, il documento indica il Principio di equità come uno dei quattro principi etici fondamentali per garantire un'IA affidabile (cfr. *Principio di equità*, n. 52); ulteriori richiami al divieto di discriminazioni nell'utilizzo dell'intelligenza artificiale si trovano in: *Legalità dell'IA*, n. 23; *2.1 I diritti fondamentali come base per un'IA affidabile*, n. 44; *Requisiti per un'IA affidabile*, n. 58-5; *Diversità, non discriminazione ed equità*, n. 80.

[20] Cfr. *Carta etica europea sull'utilizzo dell'intelligenza artificiale nei sistemi giudiziari e negli ambiti connessi*, adottata dalla CEPEJ (Commissione europea per l'efficienza della giustizia) nel corso della sua 31ª Riunione plenaria, Strasburgo, 3-4 dicembre 2018 (CEPEJ(2018)14), pp. 8 ss.

[21] Cfr. *Libro Bianco sull'Intelligenza Artificiale al servizio del cittadino*, a cura della Task force sull'Intelligenza Artificiale dell'Agenzia per l'Italia Digitale, Versione 1.0 Marzo 2018; in particolare, il documento indica la prevenzione delle diseguaglianze come una delle nove "sfide dell'IA al servizio del cittadino" (*Sfida 7: Prevenire le diseguaglianze*, pp. 62-65) e individua, tra le raccomandazioni finali, la promozione di una piattaforma nazionale dedicata allo sviluppo di soluzioni di intelligenza artificiale al fine, tra l'altro, di "organizzare e veicolare in maniera aperta i test prima del rilascio dei sistemi di IA utilizzati nella PA al fine di valutarne il comportamento e limitare le anomalie e l'amplificazione dei *bias*" (p. 73).

[22] Cfr. *Strategia Nazionale per l'Intelligenza Artificiale*, bozza per la consultazione pubblicata dal Ministero per lo Sviluppo Economico in data 31 luglio 2019; in particolare, nel documento si ribadisce che, nell'ambito della strategia del Governo italiano per l'IA, viene posto un forte accento sulla necessità di un utilizzo etico dell'intelligenza artificiale e sull'esigenza di assicurarne l'affidabilità tecnica sin dalla progettazione, nella consapevolezza che tale approccio "contribuirà a contrastare i rischi di esacerbazione delle discriminazioni e di inasprimento degli squilibri sociali e territoriali potenzialmente derivanti da un uso inconsapevole dell'IA" (p. 5).

[23] In tema cfr., *ex multis*, F.Z. BORGESIUS, *Discrimination, Artificial Intelligence and Algorithmic Decision-Making*, Council of Europe, Strasbourg, 2018, 10 ss. e B. FRIEDMAN-H. NISSENBAUM, *op. cit.*, 330 ss.

2. *Le principali fasi della decisione algoritmica in cui si pongono le premesse per le* AI-driven discriminations

La (apparente) neutralità dell'intelligenza artificiale[24] è messa a rischio, nel caso in cui una decisione sia elaborata da un sistema di *machine learning*, sotto almeno tre punti di vista[25]: la decisione algoritmica, in primo luogo, può risentire dei pregiudizi del progettista (riflettendo i preconcetti propri del programmatore o generati dai valori di riferimento dell'organizzazione in cui il programmatore opera), che possono condizionare l'impostazione dell'algoritmo attraverso, ad esempio, l'inclusione o l'esclusione di caratteri che identificano o rinviano ad una categoria protetta, a rischio di discriminazione. In secondo luogo, gli effetti discriminatori possono derivare dai contenuti del set di dati utilizzati per alimentare il sistema di *machine learning*, risultando essi stessi già "inquinati" in quanto formati da soggetti che vi hanno trasfuso i propri pregiudizi – fino al caso estremo dai dati estratti dal *web* in assenza di filtri specifici, che vede riflessi i pregiudizi diffusi nella società globale – o perché raccolti in modo tale da presentare una visione distorta della realtà. In terzo luogo, gli effetti discriminatori possono derivare dalla circostanza, tutt'altro che infrequente, che il sistema di *machine learning* individui alcune caratteristiche che rinviano indirettamente a categorie protette – associando ai loro detentori un trattamento deteriore – in modo sostanzialmente autonomo; in assenza, cioè, di specifiche indicazioni in tal senso fornite in sede di programmazione o comunque operate in sede di selezione dei dati (e talvolta, come vedremo, anche in presenza di interventi finalizzati a scongiurare tali esiti). Nei primi due casi la decisione algoritmica risulta condizionata da quelli che potrebbero definirsi pregiudizi digitali "derivati"; nel terzo caso può invece parlarsi – pur con qualche cautela, come si preciserà *infra* – di pregiudizi digitali "autonomi"[26].

2.1. *La programmazione: le discriminazioni algoritmiche dirette e indirette e i problemi connessi all'individuazione dei caratteri* proxy

Il primo e fondamentale stadio temporale nel quale è possibile intervenire allo scopo di impedire il prodursi di una decisione algoritmica discriminatoria è quello della programmazione: quella fase cioè nella quale si progetta l'algoritmo, impostando il "metodo di lavoro" che dovrà seguire il sistema di apprendimento automatico, consistente – come prima accennato – nella individuazione di correlazioni all'interno di un set di dati volta a identificare modelli di decisione. In sede di programmazione, infatti, è possibile "ordinare" al sistema di *machine learning* di escludere (cioè di considerare irrilevanti) alcuni caratteri ricorrenti nel set di dati che potrebbero condurre ad una decisione discriminatoria.

Tali caratteri possono essere essenzialmente di due tipi: anzitutto, al sistema può essere chiesto di non prendere in considerazione caratteristiche legate tradizionalmente a scelte discriminatorie, con riferimento alle discriminazioni storicamente più note: razza, colore della pelle, sesso, ecc.[27]. Gli interventi così concepiti mirano ad evitare le discriminazioni algoritmiche "dirette", cioè basate sulla esclusione di caratteri che identificano in via immediata la categoria protetta.

---

[24] In tema cfr., per tutti, M. AIROLDI, D. GAMBETTA, *Sul mito della neutralità algoritmica*, in *The Lab's Quarterly*, 2018, n. 4, 25 ss.

[25] Spunti in proposito, tra gli altri, in A. VENANZONI, *La valle del perturbante: il costituzionalismo alla prova delle intelligenze artificiali e della robotica*, in *Politica del diritto*, 2/2019, 237-238.

[26] Cfr. in argomento A. VENANZONI, *op. cit.*, 253.

[27] Si tratta delle c.d. "categorizzazioni vietate", su cui v. D.U. GALETTA-J.G. CORVALÁN, *Intelligenza Artificiale per una Pubblica Amministrazione 4.0? Potenzialità, rischi e sfide della rivoluzione tecnologica in atto*, in Federalismi.it, n. 3/2019, 21-22, definite anche come "categorie algoritmiche sospette". Nella procedura richiamata, secondo gli Autori citati, troverebbe applicazione il principio di non discriminazione algoritmica, in base al quale si richiede di operare in fase di progettazione e implementazione di algoritmi intelligenti al fine di prevenire il rischio che i sistemi di intelligenza artificiale elaborino informazioni distinguendo in base ai caratteri dianzi richiamati.

Un ulteriore intervento sull'algoritmo può essere finalizzato a rendere irrilevanti caratteri formalmente "neutri", ma che di norma ricorrono con particolare frequenza nelle categorie protette [28]: è questo il caso delle azioni volte ad evitare le c.d. *proxy discriminations* (letteralmente "discriminazioni per procura") ed a scongiurare di conseguenza le discriminazioni algoritmiche "indirette", basate cioè sulla esclusione di caratteri che rimandano indirettamente alla categoria protetta o che rinviano a caratteri distintivi non consentiti.

Vale ribadire che le *proxy discriminations* rappresentano una pratica risalente[29], cui si è fatto ricorso anche prima della nascita dei sistemi di *machine learning* per occultare pratiche discriminatorie[30]. La novità rappresentata dall'avvento dell'intelligenza artificiale è legata alla circostanza che in questo caso il carattere *proxy* potrebbe essere individuato autonomamente dal sistema di apprendimento automatico, senza che l'utilizzatore ne sia consapevole[31]: così, ad esempio, se un sistema informatico generasse autonomamente un modello che impone premi più elevati per l'assicurazione sanitaria a richiedenti che risultino iscritti ad un gruppo *Facebook* costituto per sollecitare un aumento della disponibilità del test genetico per le mutazioni dei geni BRCA correlate ad un alto rischio di sviluppare alcuni tumori, opererebbe una *proxy discrimination* sulla base dei dati genetici[32].

Un tale intervento di esclusione e di "accecamento" dell'algoritmo rispetto alle *proxies*, tuttavia, si trova a fronteggiare almeno due problemi pratici.

Il primo riguarda l'individuazione preventiva delle *proxies* potenzialmente discriminatorie. Una predeterminazione di tale caratteri (non certo completa, ma che aspiri almeno ad un livello soddisfacente di compiutezza) appare ad oggi sostanzialmente impossibile per almeno due motivi: in primo luogo, le *proxies* discriminatorie emergono di norma a seguito della concreta applicazione del sistema di apprendimento automatico (e dunque oggi possiamo basarci solo sulla limitata esperienza di *proxies* emerse in passato); in secondo luogo, i sistemi di *machine learning* operano con logiche e ruoli differenti nei diversi settori di applicazione[33]. Peraltro, la nozione giuridica di *proxy* non risulta al momento compiutamente formalizzata, oltre a scontare una certa genericità che ne rende complicata l'applicazione ai casi concreti.

Il secondo problema applicativo riguarda lo "slittamento" – operato dal sistema di *machine learning* a seguito del divieto di utilizzo di caratteri direttamente identificativi di categorie protette o delle *proxies* più "prossime" a tali categorie – verso *proxies* più "distanti", che rinviano comunque a quelle categorie, pur essendo più imprecise (e, peraltro, meno facilmente individuabili). Questo fenomeno è inevitabile quando ci si trova di fronte ad un set di dati "inquinato", in cui le decisioni assunte in passato (e sulle cui orme il sistema informatico è costretto a procedere) hanno sistematicamente discriminato alcune categorie protette[34].

È appena il caso di ricordare, in conclusione, che può anche accadere che chi programma l'algoritmo ordini al sistema informatico di selezionare a priori gli appartenenti a categorie protette

[28] Il discorso potrebbe estendersi anche ai profili attinenti al principio di eguaglianza sostanziale, considerando i caratteri che, più in generale, risultano strettamente collegati alle condizioni di soggetti deboli sul piano economico e sociale: in questa sede, tuttavia, si ragionerà unicamente dell'impatto generato delle decisioni algoritmiche sul principio di eguaglianza formale.

[29] Per una panoramica sul tema cfr. L. ALEXANDER-K. COLE, *Discrimination by Proxy*, in 14 *Constitutional Commentary*, 1997, 453 ss.

[30] V. in proposito F.Z. BORGESIUS, *op. cit.*, 13-14, che richiamano uno degli esempi più noti di *proxy discrimination*, rappresentato dall'utilizzo del codice postale quale dato indicativo dell'appartenenza ad una specifica comunità discriminata o per ragioni razziali o per ragioni socio-economiche.

[31] Sul tema cfr., ampiamente, A. PRINCE-D.B. SCHWARCZ, *Proxy Discrimination in the Age of Artificial Intelligence and Big Data*, in *Iowa Law Review*, 5 agosto 2019.

[32] Così A. PRINCE-D.B. SCHWARCZ, *op. cit.*, 5. A quest'ultimo proposito F.Z. BORGESIUS (*op. cit.*, 28-29) suggerisce che quando una organizzazione avvia un progetto che prevede il ricorso all'intelligenza artificiale dovrebbe effettuare una procedura di *risk assessment* o *risk mitigation*, impegnandosi a non prendere in considerazione caratteri potenzialmente (o indirettamente) discriminatori.

[33] Cfr. F.Z. BORGESIUS, *op. cit.*, 5.

[34] Cfr. A. PRINCE-D.B. SCHWARCZ, *op. cit.*, 29 ss.

al fine, ad esempio, di escluderle da un beneficio, o indicando espressamente il carattere immediatamente indicativo della classe[35] o – come accade più di frequente, come prima ricordato – occultando la discriminazione dietro l'indicazione un carattere *proxy*, che rinvia indirettamente alla categoria da discriminare. Il vero problema, tuttavia – è bene ribadirlo – è rappresentato dalle *proxy discriminations* di origine algoritmica: le *proxies* di origine umana sono più facilmente individuabili e sono assoggettabili ad un controllo preventivo; quelle "inintenzionali", che originano dagli algoritmi, risultano invece di norma più difficilmente individuabili e tendono di conseguenza a sfuggire alle successive verifiche[36].

2.2. *La configurazione del set di dati: i rischi derivanti dai dati "inquinati" e dai dati incompleti*

Il secondo "momento" che caratterizza l'impostazione della decisione algoritmica consiste nella configurazione del set di dati da utilizzare nel *training* del sistema di apprendimento automatico[37]. La configurazione del *data set* implica sempre una scelta che, anche in questo frangente, dovrebbe fondarsi sulla consapevolezza dei rischi legati alla possibilità di dare origine a decisioni algoritmiche discriminatorie. Tale scelta può muovere anzitutto da una opzione di base, che consiste nell'alternativa tra l'impiego di un set di dati "chiuso" ovvero il ricorso ad un set di dati tendenzialmente "aperto". Come dianzi accennato, la necessaria presenza di *big data* per alimentare il sistema di apprendimento automatico non implica necessariamente che questi dati vengano acquisiti dalla rete Internet, ma tale mole di dati può anche provenire da archivi digitali *offline* (come potrebbero essere gli archivi giudiziari o di polizia) o da banche dati formate *ad hoc* da parte degli utenti. All'interno del set di dati prescelto, si operano poi di norma ulteriori interventi di selezione. Questa selezione può implicare diversi livelli di controllo, che, occorre sottolinearlo, non sono una volta per tutte legati alla scelta di base tra sistema aperto e sistema chiuso. Anche se la "navigazione" del sistema di apprendimento automatico nel *mare magnum* del web sembrerebbe a prima vista esporre il sistema informatico ad un flusso di dati tendenzialmente incontrollabile rispetto alla – in apparenza – più rassicurante "nutrizione" dell'intelligenza artificiale con i dati provenienti da una banca dati *offline*, occorre sottolineare che non necessariamente l'acquisizione di dati dalla Rete è meno selettiva dell'acquisizione di informazioni da banche dati "chiuse". La ricerca sulla Rete può infatti riguardare il contenuto di un numero molto circoscritto di siti o di piattaforme, o svolgersi in base a parole-chiave molto precise; per contro, la banca dati *offline* può essere estremamente cospicua e la raccolta delle informazioni in essa contenute potrebbe avvenire senza alcun filtro.

La configurazione del set di dati e, in particolare, la scelta del tipo di banca dati e gli interventi di selezione del loro contenuto, è essenziale per evitare i pregiudizi digitali (che, anche in questo frangente, sono pregiudizi "derivati" dall'intervento umano nella progettazione e nella implementazione del sistema informatico). In questa fase opera infatti il principio del *garbage in-garbage out*: se i dati acquisiti dall'intelligenza artificiale sono incongrui, inesatti o non affidabili,

[35] Come è accaduto, ed esempio, nei casi di pubblicità *online* richieste a *social network* in modo da escludere ispanici o omosessuali, ricordati da G. RESTA, *op. cit.*, 218.

[36] Anche sul punto cfr. A. PRINCE-D. SCHWARCZ, *op. loc. ult. cit.*

[37] In proposito vale ribadire sin d'ora che la distinzione tra la prima e la seconda fase (ma analoghe considerazioni possono formularsi anche con riferimento alla terza) intende unicamente porre l'attenzione su tre diversi ordini di problemi concernenti il concreto operare dei sistemi di *machine learning*; nella consapevolezza che anche quelli che vengono indicati come "momenti" successivi alla programmazione iniziale individuano fasi successive del funzionamento del sistema informatico ordinate in base a una sequenza cronologica (la costruzione dell'algoritmo che individua il "problema" che il sistema dovrà risolvere – l'acquisizione dei dati nella fase di *training* – l'elaborazione dei dati immessi), ma sono comunque riconducibili alla "programmazione" iniziale (prima fase), nella quale, insieme all'assegnazione dell'obiettivo all'intelligenza artificiale, possono definirsi anche le regole da seguire nella selezione dei dati da utilizzare per l'apprendimento (seconda fase) e si scontano i limiti dell'intervento umano dai quali scaturiscono i margini di autonomia del sistema informatico nella elaborazione delle informazioni (terza fase).

condurranno inevitabilmente a decisioni inaffidabili[38]. Il rilievo di questa scelta, non a caso, è ribadito da molti degli strumenti di *soft law* intervenuti in materia che si sono dianzi richiamati: in particolare, è molto chiara sul punto la risoluzione del Parlamento europeo, prima ricordata, sulla politica industriale europea in materia di robotica del 2019, nella quale, con riferimento al *deep learning*[39], si sottolinea l'importanza della qualità dei dati utilizzati, ribadendosi che "l'utilizzo di dati di scarsa qualità, obsoleti, incompleti o inesatti può portare a previsioni inadeguate e, di conseguenza, a discriminazioni e pregiudizi"[40].

In particolare, in presenza di un set di dati "chiuso", si può procedere essenzialmente in due modi: il *data set* può essere assunto senza operare alcun filtro o, al contrario, i dati possono essere variamente selezionati. Nel primo caso sono numerose le modalità attraverso le quali può verificarsi una distorsione del risultato finale, con effetti potenzialmente discriminatori; nel caso invece in cui si operino interventi di filtraggio, un'adeguata selezione dei dati operata a monte può contribuire ad eliminare molte distorsioni.

Si consideri anzitutto l'assenza di filtri: l'algoritmo viene "alimentato" con una statistica storica delle decisioni compiute in passato, per far sì che sia in grado di assumere le medesime decisioni in futuro. Il rischio maggiore in questo caso è che il sistema si nutra di *biased data*, cioè che si alimenti di dati "inquinati", incorporando – e quindi replicando – le discriminazioni compiute nel passato.

Gli esempi a riguardo sono numerosi[41]: una delle vicende più note è quella legata all'utilizzo di COMPAS[42], un *software* alimentato in prevalenza da una banca dati costituita da precedenti giudiziari [43], in grado di predire il rischio di recidiva e di pericolosità sociale degli imputati e impiegato per decidere sull'entità e sulle modalità di esecuzione delle sanzioni penali (soprattutto in ordine alla possibilità di accedere a misure alternative alla detenzione); la successiva verifica sulla correttezza delle previsioni dimostrò che i precedenti giudiziari utilizzati, sfavorevoli ai condannati di colore, avevano indotto il sistema a sottostimare la probabilità di recidiva dei condannati bianchi ed a sovrastimare quella dei condannati neri[44]. Analoghi timori si legano all'impiego del PSA[45], un *software* utilizzato dalle Corti di diversi Stati americani per calcolare l'ammontare della cauzione o del periodo di custodia cautelare degli imputati e collegato ad una banca dati contenente casi simili, precedenti giudiziari e dati anagrafici degli imputati, alle cui decisioni i giudici tendono ad

[38] Cfr. sul punto G. RESTA, *op. cit.*, 208.

[39] Il *deep learning* è un modello di apprendimento automatico che consente a modelli computazionali composti da più livelli di elaborazione di apprendere da rappresentazioni di dati con più livelli di astrazione. Il suo impiego riguarda settori quali il riconoscimento vocale e visivo, il rilevamento di oggetti e la genomica. In argomento cfr., per tutti, Y. LECUN-Y. BENGIO-G. HINTON, *Deep learning*, in 521 *Nature***,** 2015, 436 ss.

[40] Cfr., *Risoluzione del Parlamento europeo del 12 febbraio 2019 su una politica industriale europea globale in materia di robotica e intelligenza artificiale*, cit., n. 39.

[41] Uno dei casi più risalenti è quello verificatosi negli anni '80 del secolo scorso in una Scuola di medicina Regno Unito, che ricorse ad un sistema di intelligenza artificiale per la selezione del personale: il sistema, basandosi sulle assunzioni operate in passato, penalizzò i candidati immigrati e quelli di sesso femminile (cfr. in argomento S. LOWRY-G. MACPHERSON, *A blot on the profession*, in 296 *British Medical Journal*, 1988, 657 ss.).

[42] L'acronimo corrisponde a "*Correctional Offender Management Profiling for Alternative Sanctions*".

[43] In particolare, i dati da prendere in esame erano stabiliti dalla società (privata) che forniva il software: anzitutto i precedenti giudiziari, diversi altri dati statistici, un questionario somministrato all'imputato, cui si aggiungevano ulteriori variabili non rivelati dalla società produttrice in quanto coperti dalla proprietà intellettuale (cfr. A. SIMONCINI-S. SUWEIS, *op. cit.*, 95 ss.).

Il caso è citatissimo e per molti aspetti emblematico: lo richiamano, *ex multis*, A. SIMONCINI-S. SUWEIS, *op. cit.*, *passim*; A. CELOTTO, *op. cit.*, 47; G. RESTA, *op. cit.*, 215-216 (che ricorda come in questo caso il sistema, operando su *biased data*, abbia effettuato anche una *proxy discrimination*, basata sull'acquisizione di informazioni quali residenza, istruzione, consumo di stupefacenti, collegati indirettamente all'origine razziale).

[44] Un'indagine operata successivamente all'applicazione del *software* riscontrò infatti che l'errore nella predizione della recidiva si aggirava intorno al 30%; ma, in particolare, la percentuale di falsi positivi era pari a circa al doppio per gli imputati di colore rispetto agli imputati bianchi e, per converso, la percentuale di falsi negativi era pari a quasi il doppio per gli imputati bianchi rispetto agli imputati di colore (cfr. A. SIMONCINI-S. SUWEIS, *op. cit.*, 97-98).

[45] L'acronimo corrisponde a *Public Safety Assessment*.

uniformarsi in modo pedissequo, adottando pronunce "integralmente basate sugli *scores* stabiliti" dal sistema informatico[46].

Gli esempi riportati dimostrano chiaramente che le decisioni fondate su statistiche storiche tendono inevitabilmente ad "incorporare" le distorsioni (e in particolare i pregiudizi dei decisori precedenti) e dunque, in certo senso, a cristallizzarli[47]. Da questo punto di vista il risultato appare addirittura peggiorativo della situazione esistente, in quanto la cristallizzazione comporta una perpetuazione delle distorsioni (e in questo senso un aggravamento)[48].

Un ulteriore rischio di *bias* (e dunque di possibili esiti discriminatori) può derivare - oltre che dall'utilizzo di dati "inquinati" da pregiudizi (come nel caso dell'impiego *sic et simpliciter* di statistiche storiche acquisite integralmente) - anche dall'impiego di dati incompleti, che forniscano una rappresentazione parziale della realtà[49].

Tale incompletezza genera una distorsione quando la quota di dati raccolta risulta più ampia per un certo gruppo e più esigua per un altro (o meglio, quando la quota di dati riguardante un determinato gruppo risulta troppo ampia o troppo esigua rispetto alle dimensioni e al rilievo effettivo del gruppo all'interno della popolazione considerata); e comporta un effetto discriminatorio quando il gruppo sovra o sotto-rappresentato risulta sfavorito dalla falsa rappresentazione della realtà. In particolare, la distorsione può derivare sia dalle *modalità* di raccolta dei dati, sia dallo *strumento* utilizzato per acquisirli. Un'ipotesi di errata modalità di raccolta dei dati potrebbe verificarsi nel caso in cui, ad esempio, si alimentasse un sistema di apprendimento automatico, destinato a prevedere in quali quartieri si commetteranno più reati, con le informazioni contenute in uno schedario della polizia in cui la percentuale di schedati immigrati fosse più elevata rispetto alla loro presenza nella popolazione: in tale circostanza il sistema "imparerebbe" che gli immigrati sono più inclini a commettere reati[50]. Un esempio delle distorsioni prodotte dall'utilizzo di uno strumento di raccolta dei dati inidoneo è invece offerto da un esperimento condotto nel 2011 nella città di Boston (denominato *Street Bump*), che impiegò un sistema di intelligenza artificiale per raccogliere informazioni sulla presenza di gravi irregolarità nel

[46] Così A. VENANZONI, *op. cit.*, 249.

[47] L'utilizzo dell'intelligenza artificiale nello svolgimento della funzione giurisdizionale, che accomuna gli esempi proposti, inoltre, induce ad interrogarsi sul mito del "giudice bot" come giudice perfetto: la specializzazione cognitiva dell'algoritmo, infatti – che comporta l'assenza di stati d'animo e di emozioni, posto che l'obiettivo del programmatore non è quello di riprodurre esattamente il funzionamento della mente umana – rende solo in apparenza l'algoritmo un giudice perfetto in quanto massimamente oggettivo (sul punto si rinvia alle considerazioni sull'aspirazione all'oggettività e la spersonalizzazione della funzione giurisdizionale di M. LUCIANI, *La decisione robotica*, in *Rivista AIC*, 3/2018, 873 ss.), perché se i dati – come i precedenti giudiziari – sono "soggettivamente inquinati" da pregiudizi, tali pregiudizi appaiono destinati a cristallizzarsi (cfr. A. VENANZONI, *op. cit.*, 253, in particolare nota 47).

Ma non solo: l'utilizzo dell'intelligenza artificiale nelle decisioni giudiziarie pone una molteplicità di problemi ulteriori. Volendone richiamare sinteticamente alcuni (rinviando, per una più ampia trattazione del tema, ad A. VENANZONI, *op. cit.*, 254 ss.): la possibilità che il sistema incontri difficoltà di ordine logico che conducano alla paralisi decisionale; Il fatto che le emozioni e le sensazioni dei giudici si connettano funzionalmente ai precedenti giudiziari; la circostanza che le opinioni del giudice siano il riflesso del dibattito dottrinale; l'insuperabile difficoltà per un algoritmo di apprezzare il *fatto* nelle sue complesse e uniche caratteristiche, evenienza possibile solo a condizione di immettere nel sistema informatico un complesso di dati pressoché infinito.

[48] Come osserva sul punto G. RESTA (*op. cit.*, 214), riportando il tema alla sua dimensione più generale, "qualora le tecniche predittive si appuntino su stati dell'uomo e sui processi sociali […] uno dei pericoli più evidenti è che le condizioni di disparità sociale esistenti in un dato momento storico si riflettono sul giudizio prognostico tramite la costruzione di profili individuali o più spesso di gruppo, composti per inferenza da fattori come la propensione al consumo, la capacità di spesa, il luogo di residenza, i trascorsi familiari, il grado di istruzione, la storia giudiziaria, ecc. Se non adeguatamente monitorate e rese neutre rispetto ai rischi di *bias* già inseriti nella selezione dei dati rilevanti, le decisioni algoritmiche che si basano su tali fattori sono atte a produrre effetti discriminatori e ad *aggravare il peso delle disuguaglianze* invece che contribuire a ridurle come pure la tecnologia potrebbe fare" (corsivi aggiunti).

[49] La differenza è ricondotta dalla letteratura in argomento alla distinzione tra *biased data* e *biased samples* (cfr. sul tema F.Z. BORGESIUS, *op. cit.*, 11-12).

[50] Cfr. F.Z. BORGESIUS, *op. cit.*, 11. Una tale evenienza peraltro rischierebbe di determinare una reazione a catena per cui si invierebbero più poliziotti nei quartieri cui vivono gli immigrati e la presenza di un maggior numero di poliziotti in quei quartieri registrerebbe un numero più elevato di reati.

manto stradale della rete viaria cittadina attraverso gli *smartphone* degli automobilisti: nei quartieri più degradati – dove le strade versavano generalmente in condizioni peggiori ma gli abitanti erano in larga parte sprovvisti di *smarthone* – le segnalazioni inviate risultarono numericamente inferiori a quelle registrate nei quartieri abitati dalle classi più abbienti; di conseguenza il sistema privilegiò gli interventi di manutenzione nei quartieri nei quali la qualità delle strade risultava generalmente migliore[51].

Al fine di scongiurare i cennati fenomeni di "cristallizzazione" dei pregiudizi generati da *biased data*, si rende necessario operare opportuni interventi di "filtraggio", selezionando accuratamente i dati statistici immessi allo scopo di evitare disallineamenti socio-culturali. Tale selezione deve anzitutto tenere conto della connessione tra dati apparentemente neutri e condizioni di disparità, eliminando preventivamente le informazioni in grado di riflettere l'appartenenza a categorie protette[52]: la "depurazione" dei dati immessi – riprendendo la metafora dell'inquinamento – comporta cioè l'eliminazione non solo dei caratteri direttamente identificativi di categorie protette, ma altresì – e anche in questo caso – l'eliminazione preventiva delle *proxies* responsabili di discriminazioni indirette; ma, anche in questo frangente, nella consapevolezza dei limiti dell'operazione, legati ai due fenomeni prima rilevati della impossibilità di predeterminare compiutamente le *proxies* discriminatorie e dello slittamento verso *proxies* più distanti. Parimenti, in presenza di *biased samples* legati alla incompletezza dei dati raccolti, è possibile intervenire migliorando le modalità e gli strumenti di raccolta dei dati.

Un cenno, in conclusione, al set di dati "aperto", che il sistema di apprendimento automatico acquisisce quando viene collegato alla rete Internet: un *data set* estremamente "appetibile", in quanto ingentissimo nello stock totale e in crescita esponenziale quanto a velocità di sviluppo, che lo rendono la sede ideale dei *big data*[53].

Il rischio più rilevante legato a tale procedura – per quanto sia possibile circoscrivere la ricerca dei dati delimitandola, come prima ricordato, soltanto a certi siti o piattaforme, o indirizzandola attraverso specifiche parole chiave con l'ausilio di motori di ricerca[54] – è legato alla circostanza che il sistema di apprendimento automatico incorpori i pregiudizi presenti nel web; o, per certi versi, considerata la copertura dell'anonimato, una versione esasperata dei pregiudizi: anche di quelli inconfessabili, che di norma non trovano espressione attraverso i canali più trasparenti e non coperti dall'anonimato[55]. Gli esempi di questo fenomeno sono numerosi e significativi. Per limitarsi ad alcuni tra i più noti, può qui ricordarsi l'esperimento condotto da *Microsoft* nel 2016, che diede vita a TAY, un'intelligenza artificiale che venne collegata ai principali *social network* per apprendere e replicare il linguaggio degli adolescenti: la macchina venne spenta dopo pochi giorni di attività, in quanto Tay si era rivelato un bot razzista, misogino e di aperte simpatie naziste[56]. O, ancora, una ricerca condotta sempre nel 2016 attraverso *Google images*, alla richiesta di individuare e riconoscere immagini ritraenti "tre ragazzi bianchi" restituì in prevalenza foto di *teenager* bianchi colti in atteggiamenti gioviali, mentre alla richiesta di immagini di "tre ragazzi neri" restituì in prevalenza foto segnaletiche della polizia[57].

---

[51] La vicenda è richiamata da F.Z. BORGESIUS, *op. cit.*, 12.

[52] Cfr. in argomento, con particolare riguardo agli effetti discriminatori legati all'impiego dei giudice-robot, M. TEGMARK, *Vita 3.0. Essere umani nell'era delle intelligenze artificiali*, Milano, 2018, 143.

[53] Cfr. A. SIMONCINI-S. SUWEIS, *op. cit.*, 89.

[54] In particolare, in quest'ultimo caso il ricorso all'intelligenza artificiale è duplice: si coinvolge l'AI del motore di ricerca nel compiere la selezione e la raccolta dei dati, su cui interverrà l'AI che li utilizzerà per il proprio *training*. Ciò rafforza, naturalmente, il rischio di distorsioni, considerata peraltro la capacità dei motori di ricerca non solo di profilare l'utenza, ma anche di orientarne le scelte.

[55] È noto infatti, ad esempio, che la Rete esaspera gli stereotipi razziali e di genere (cfr. in tema F.Z. BORGESIUS, *op. cit.*, 17).

[56] Cfr. sul punto A. VENANZONI, *op. cit.*, 238.

[57] Come ricordato da F.Z. BORGESIUS, *op. cit.*, 16.

2.3. (Segue) *i margini di autonomia dell'algoritmo*

L'individuazione delle correlazioni all'interno del set di dati, in vista della decisione da assumere, attraverso l'individuazione dei caratteri rilevanti per la soluzione del problema, rappresenta l'essenza dell'apprendimento automatico. Tuttavia – come prima rilevato – essa può essere guidata solo in parte: per il resto il meccanismo di individuazione di correlazioni resta in parte ignoto agli stessi programmatori[58].

L'ultima fase della decisione algoritmica, dunque, nella quale il sistema di apprendimento automatico analizza il set di dati ed elabora il risultato, è insieme quella decisiva ma anche la meno governabile. In particolare, è in questa fase che nasce la possibilità che l'algoritmo individui *proxies* discriminatorie non volute, soprattutto in conseguenza di quel fenomeno prima rilevato di slittamento verso *proxies* più distanti; ma è soprattutto in questa fase che sorge il rischio che il sistema generi delle nuove classi di soggetti sottoposti ad un trattamento deteriore in base a caratteristiche ricorrenti individuate dall'algoritmo in totale autonomia[59].

Occorrerà allora interrogarsi sugli strumenti offerti dal diritto – e segnatamente dalla legislazione antidiscriminatoria – per fronteggiare i pericoli richiamati, individuando gli ambiti di intervento, le modalità della disciplina e le sfide che rimangono ancora aperte.

3. *Le soluzioni: il ruolo del diritto e i suoi limiti*

Volendo interrogarsi sul modo in cui il legislatore può affrontare i rischi legati all'uso dell'intelligenza artificiale che si sono sinteticamente evocati, vale osservare, anzitutto, che tale intervento va modulato sulla base di almeno tre elementi: la finalità, l'oggetto specifico della regolazione, la modalità della disciplina del fenomeno.

La finalità è evidente: la disciplina giuridica deve riuscire a impedire (e a prevenire) le *AI-driven discriminations*, cioè le discriminazioni indotte dall'utilizzo dell'intelligenza artificiale per prendere decisioni. Perché questa finalità possa aspirare a realizzarsi occorre ampliare la prospettiva rispetto a quanto si è detto finora: se il problema è l'effetto (discriminatorio) dell'uso dell'intelligenza artificiale per assumere decisioni (con una gravità particolare nel caso in cui si prendano decisioni pubbliche), esso va affrontato sotto un profilo più generale, che induce ad interrogarsi su come regolare i *processi decisionali* in cui sono coinvolti algoritmi (avendo riguardo sia all'iter decisionale sia agli effetti sociali)[60]. Questo implica un'estensione dell'oggetto della disciplina giuridica, che deve rivolgersi essenzialmente a due diversi profili della decisione algoritmica: un profilo, per così dire, "interno", che concerne il funzionamento dell'intelligenza artificiale, nelle diverse fasi rilevanti che abbiamo esaminato, rispetto al quale occorre dettare regole volte ad evitare che, sia in sede di programmazione, sia in sede di configurazione del set di dati, si predisponga il sistema a generare modelli decisionali discriminatori; e un profilo "esterno" al funzionamento dell'intelligenza artificiale, che investe il ruolo dell'algoritmo nell'ambito della decisione finale. Su questo versante, occorre affrontare una questione cruciale, legata al ruolo assegnato all'intelligenza artificiale nella decisione umana: si tratta in sostanza del problema della *significatività* dell'algoritmo, e della connessa possibilità di un intervento umano in funzione di controllo, teso a mitigare degli effetti discriminatori della decisione algoritmica.

[58] È quella che si definisce la *black box*, secondo la nota e felice formula di Frank Pasquale, di norma evocata in questo contesto (si v. ad esempio: D.U. GALETTA-J.G. CORVALÁN, *op. cit.*, 15-16; A. VENANZONI, *op. cit.*, 237 ss.) per designare la sostanziale opacità dei software a codice chiuso, che operano senza mostrare il proprio metodo di lavoro.

[59] Sulla possibilità che i sistemi di apprendimento automatico – progettate per individuare correlazioni in base ad associazioni di elementi ricorrenti – introducano trattamenti differenziati in base a caratteristiche estranee alle tradizionali categorie protette, v. F.Z. BORGESIUS, *op. cit.*, 36 (e i diversi esempi ivi richiamati).

[60] Cfr. in argomento G. RESTA, *op. cit.*, 200. Il tema è di grande rilievo generale: se questi processi – specialmente quando riguardano decisioni pubbliche e lo svolgimento di funzioni pubbliche fondamentali – non sono democraticamente governati, rischiano di consolidare le posizioni di privilegio e le diseguaglianze (ivi, 201).

3.1. *I profili "interni" della decisione algoritmica: il divieto di produrre effetti discriminatori tra dimensione tecnica e dimensione etica*

Un modello di disciplina della decisione algoritmica in grado di fronteggiare le potenzialità discriminatorie connesse a tale procedura, che può rappresentare un utile e significativo un punto di riferimento per la legislazione in materia – perlomeno in ambito europeo – è costituito dal Considerando n. 71 del regolamento (UE) 2016/679 del Parlamento europeo e del Consiglio, del 27 aprile 2016 (GDPR); tale enunciato introduce quello che è stato opportunamente denominato il "principio di non discriminazione per via algoritmica"[61], invitando il titolare del trattamento di dati personali a "mett[ere] in atto misure tecniche e organizzative adeguate […] secondo una modalità che […] impedisca, tra l'altro, effetti discriminatori nei confronti di persone fisiche sulla base della razza o dell'origine etnica, delle opinioni politiche, della religione e delle convinzioni personali, dell'appartenenza sindacale, dello *status* genetico, dello stato di salute o dell'orientamento sessuale, ovvero un trattamento che comporti misure aventi tali effetti"[62].

Tale modello di disciplina presenta almeno due profili di interesse.

In primo luogo, offre utili indicazioni sul *modo* in cui vanno disciplinati i profili interni della decisione algoritmica, indicando cosa deve vietare una legislazione che ambisca ad essere efficace contro le discriminazioni algoritmiche: occorre ricordare, in proposito, che uno dei tratti distintivi più rilevanti della decisione algoritmica è rappresentato dalla mancata identificazione, in via preliminare, dei caratteri distintivi in base ai quali i soggetti destinatari della soluzione proposta dal sistema di apprendimento automatico saranno sottoposti a trattamenti differenziati. Muovendo dunque da questo dato, una legislazione antidiscriminatoria destinata a rivolgersi alle decisioni algoritmiche che concentrasse la propria attenzione specificamente sui *criteri* utilizzati nella *elaborazione* della decisione e non unicamente sul risultato, sull'effetto della decisione stessa, riuscirebbe verosimilmente ad impedire soltanto le discriminazioni algoritmiche "dirette", cioè quelle fondate sulla individuazione di caratteri in base ai quali operare la decisione che siano immediatamente identificativi della categoria protetta, nella sola eventualità (e nei ristretti margini entro i quali tale eventuale intervento opererebbe) che tali caratteri siano stati "indicati" in sede di programmazione al sistema di apprendimento automatico o che siano stati individuati dall'algoritmo come elementi in base ai quali operare un trattamento differenziato[63]. Ma nulla potrebbe contro le *proxies*, che formalmente si presentano come caratteri neutri; o, al limite, il divieto potrebbe essere esteso alle *proxies* più direttamente collegate alle categorie protette, ma sconterebbe comunque i limiti legati alla individuazione preventiva, che abbiamo visto.

Al contrario, le norme antidiscriminatorie che rivolgono la propria attenzione esclusivamente al risultato finale, vietando decisioni algoritmiche che sfavoriscano irrazionalmente gli appartenenti alle categorie protette[64], sono in grado di attenuare gli effetti del ricorso – volontario o anche

[61] Così A. SIMONCINI-S. SUWEIS, *op. cit.*, 101.

[62] Il testo non fa espresso riferimento al trattamento automatizzato, ma esso può ritenersi implicito in ragione sia del richiamo, presente nella parte omessa, alla opportunità che il titolare del trattamento "utilizzi procedure matematiche o statistiche appropriate per la profilazione", sia del riferimento al "trattamento automatizzato" dei dati personali e al "processo decisionale automatizzato" che rispettivamente aprono e chiudono il Considerando in commento.

[63] Trattamento differenziato che si fondi proprio su quella base, senza fondamento razionale: o perché ad esempio il programmatore lo ha indicato espressamente, o perché deriva dal set di dati "inquinato" che riflette discriminazioni operate nel passato.

[64] In questa prospettiva, sarebbe incostituzionale, in quanto discriminatorio, l'algoritmo che produce effetti discriminatori: cfr. in proposito A. SIMONCINI-S. SUWEIS, *op. cit.*, 102, che parlano di algoritmo "strutturalmente" incostituzionale, ma alludendo in particolare alle distorsioni presenti nel set di dati immesso nel sistema. Potrebbe tuttavia estendersi il concetto di algoritmo "strutturalmente" discriminatorio ai casi in cui l'algoritmo discrimina autonomamente, producendo un risultato discriminatorio a seguito di una selezione dei caratteri individuati dal sistema informatico come rilevanti, pure in presenza di dati *unbiased*.

involontario, nei termini dianzi descritti – alle *proxies*, limitando il prodursi di *proxy discriminations*: in questi casi occorrerà verificare se (anche a partire da caratteri "neutri") la decisione assunta dall'algoritmo avrà riservato un trattamento irragionevolmente deteriore alle categorie individuate dalla legislazione antidiscriminatoria.

Va rilevato tuttavia che neanche quest'ultima soluzione consente di superare del tutto le difficoltà connesse all'uso delle *proxies*: quando queste, infatti, sono molto "distanti" dalla categoria discriminata, non è sempre dimostrabile il nesso tra il carattere distintivo individuato dall'algoritmo come elemento in base al quale operare un trattamento differenziato e la categoria protetta che risulta sfavorita dalla decisione; riemergono cioè, ancora una volta, gli inconvenienti connessi alla difficoltà di individuare compiutamente le *proxies* vietate.

Un secondo profilo del modello di disciplina richiamato che merita di essere valorizzato è costituito dall'accento posto sulle misure tecniche e organizzative, alle quali viene affidato il compito di impedire il prodursi di effetti discriminatori. Si tratta, in certo senso, di un percorso obbligato, che deriva dall'attitudine esibita dell'enunciato in commento a non addentrarsi – opportunamente – nei meccanismi di formazione della decisione algoritmica (ma di occuparsi invece solo delle sue conseguenze finali), sottraendosi al duplice rischio di lasciarsi sfuggire "passaggi" decisivi del processo decisionale o di stabilire regole che il ritmo frenetico dell'evoluzione tecnologica condannerebbe ad una repentina obsolescenza.

In particolare, è prioritariamente nella dimensione tecnica che sembra doversi giocare la partita: è infatti costante la sollecitazione da parte dei documenti di *soft law* dianzi richiamati a costruire iter procedurali, sia di programmazione che di configurazione dei set di dati impiegati nel *training* dei sistemi di apprendimento automatico, idonei a scongiurare effetti discriminatori nella decisione finale[65]. A tale obiettivo si lega l'auspicio di un forte impegno educativo volto ad integrare la formazione di tutti i soggetti coinvolti nella progettazione dell'intelligenza artificiale, affinché acquisiscano la capacità di individuare i caratteri ed i dati "potenzialmente discriminatori": uno sforzo che nasce dalla convinzione che la piena consapevolezza dei rischi di *proxy discrimination* legati alla selezione dei caratteri rilevanti – e degli effetti distorsivi implicati dalla scelta dei dati da utilizzare o dalle modalità della loro raccolta – debba costituire un presupposto della programmazione algoritmica.

Tale esigenza postula la necessità di una saldatura tra dimensione tecnica e dimensione etica, alla quale può ricondursi una molteplicità di interventi in larga misura *preliminari* rispetto alla regolazione degli effetti della decisione algoritmica. Il primo intervento operabile "a monte" concerne il *set di valori* in base al quale orientare la programmazione[66]. È infatti soprattutto nella fase di progettazione dell'algoritmo – come è stato autorevolmente affermato[67] – che andrebbe anticipata la tutela dei valori costituzionali, tra i quali il principio di non discriminazione. In questa direzione si muove l'auspicio – dianzi sottolineato – di un impegno rafforzato nella educazione dei tecnici informatici affinché interiorizzino i valori in oggetto (dignità, libertà, ma anche non discriminazione) in modo che la progettazione dell'algoritmo ne sia ispirata[68]: è stato ribadito in proposito il ruolo essenziale svolto dalle "agenzie formative, ovvero dalle associazioni professionali o accademiche"[69], richiamando l'esempio dell'associazione statunitense della meccanica computazionale che ha pubblicato nel 2017 un documento sulla trasparenza e la responsabilità degli algoritmi, in cui si contempla, tra i principi da rispettare durante ogni fase di sviluppo e implementazione, la necessaria "consapevolezza" dei possibili effetti discriminatori dell'algoritmo.

[65] In particolare, sul rilievo cruciale, ai fini delineati, delle modalità di raccolta dei dati fruibili per i trattamenti algoritmici, cfr. le considerazioni di G. RESTA, *op. cit.*, 200.

[66] Cfr. sul punto G. RESTA, *op. cit.*, 214.

[67] Così A. SIMONCINI-S. SUWEIS (*op. cit.*, 103 ss.), che criticano le scelte operate da aziende come *Google* che fanno validare l'eticità degli algoritmi *dopo* la loro progettazione.

[68] A. SIMONCINI-S. SUWEIS, *op. cit.*, 103, parlano in proposito di tutela *by education* come premessa per un un'adeguata tutela *by design*, che condizioni a sua volta la tutela *by default*.

[69] A. SIMONCINI-S. SUWEIS, *op. cit.*, 104.

Il tema è ripreso anche negli Orientamenti etici definiti dall'Unione europea per un'intelligenza artificiale affidabile, prima richiamati, in cui si sottolinea come occorra formare sistematicamente una nuova generazione di esperti in etica dell'IA[70]. Questa sollecitazione chiama in causa il principio responsabilità e segnatamente la necessità che i programmatori di intelligenze artificiali avanzate siano considerati responsabili delle implicazioni morali di tali sistemi, come ribadito, da ultimo, nella dichiarazione di Asilomar del 2017[71]. Una sensibilità rispetto alle implicazioni etiche delle decisioni algoritmiche che è richiesta in realtà a tutti i soggetti che concorrono a vario titolo alla "costruzione" di tali decisioni: poiché infatti – come dianzi osservato – il progettista dell'algoritmo può risentire dei valori propri della organizzazione di appartenenza, i soggetti che devono essere consapevoli delle implicazioni etiche dell'algoritmo sono pertanto anche i proprietari delle imprese produttrici dei software impiegati nella decisione; occorre altresì ribadire la necessità di educare ai valori in oggetto sia i manager delle organizzazioni che impiegano sistemi di intelligenza artificiale – che devono essere a conoscenza dei rischi legati all'uso degli algoritmi[72] – sia più in generale gli utilizzatori di tali strumenti, chiamati ad acquisire una piena consapevolezza delle potenzialità discriminatorie connesse al ricorso a tali tecnologie, soprattutto quando si tratti di autorità pubbliche impegnate nell'esercizio di funzioni amministrative o giurisdizionali[73].

Quanto alla dimensione organizzativa – anch'essa rilevante come presupposto della programmazione da "modulare" in chiave antidiscriminatoria – merita di essere valorizzato uno spunto che compare negli Orientamenti etici per un'intelligenza artificiale affidabile, che suggerisce la costituzione di team di progettisti provenienti da contesti, culture e discipline diverse[74].

Tuttavia, permangono ancora significative zone d'ombra in relazione ai margini di funzionamento autonomo dei sistemi di apprendimento automatico. Come prima ricordato, infatti, gli algoritmi che governano i sistemi di *machine learning* individuano autonomamente criteri distintivi suscettibili di generare nuove classi di soggetti discriminati, non sempre riconducibili alle categorie protette. Su queste nuove classi di individui la legislazione antidiscriminatoria attuale sembra poter fare poco[75]: le norme antidiscriminatorie, infatti, riescono ad impedire le discriminazioni, anche indirette, causate dall'intelligenza artificiale, ma aventi ad oggetto categorie identificate in base alle caratteristiche protette "tradizionalmente" indicate dalla legislazione in materia[76]. Trattandosi peraltro di una fase in cui il sistema informatico lavora in modo autonomo, risulta estremamente complicato stabilire rimedi preventivi. L'unica soluzione, in questo caso, andrebbe ricercata nel ricorso al principio di precauzione, che, nella sua versione più radicale, imporrebbe di non ricorrere all'intelligenza artificiale quando non sia quantificabile il rischio di distorsioni potenzialmente generabili; anche perché l'incremento dell'accuratezza della previsione aumenta il rischio di disparità di risultati tra gruppi, e dunque di effetti discriminatori[77]. Ma l'abbandono della tecnologia dovrebbe prospettarsi unicamente nella evenienza – estrema – in cui

[70] Cfr. *Orientamenti etici per un'IA affidabile*, cit., p. 3.

[71] Cfr. *Asilomar AI Principles*, adottati a conclusione della *Asilomar Conference on Beneficial AI* tenutasi in California dal 5 all'8 gennaio 2017, principio n. 9.

[72] Come suggerito, da ultimo, da F.Z. BORGESIUS, *op. cit.*, 28.

[73] Con riguardo a quest'ultimo aspetto, va ricordato che in Italia è previsto solo un obbligo di formazione del personale delle amministrazioni pubbliche all'uso delle ITC (cfr. l'art. 13 del d.lgs. 7 marzo 2005, n. 82, *Codice dell'amministrazione digitale*), peraltro non sostenuto da adeguati stanziamenti economici (come lamentano D.U. GALETTA-J.G. CORVALÁN, *op. cit.*, 14): alla luce di quanto osservato, sembra opportuno auspicare che tale formazione ricomprenda anche i rischi di *AI-driven discrimination*.

[74] Cfr. *Orientamenti etici per un'IA affidabile*, cit., *5. Diversità, non discriminazione ed equità*, n. 80, in cui si afferma che va incoraggiata l'assunzione di personale, destinato a programmare ed implementare l'AI, "proveniente da contesti, culture e discipline diverse".

[75] Cfr. F.Z. BORGESIUS, *op. cit.*, 5.

[76] Cfr. Sul punto F.Z. BORGESIUS, *op. cit.*, 20.

[77] Cfr. A. SIMONCINI-S. SUWEIS, *op. cit.*, 97. La discriminazione sembrerebbe quindi una sorta di "effetto collaterale" della precisione nella capacità predittiva.

questa "non riesc[a] più a controllare sé stessa"[78], ad esempio generando risultati del tutto incompatibili con i principi etici che tale programmazione hanno ispirato[79]; in alternativa dovrebbero individuarsi – sempre in coerenza con il principio di precauzione, inteso tuttavia in senso "debole", e colto dunque nella sua attitudine a sollecitare l'adozione di procedure che inglobino l'analisi del rischio nella implementazione del sistema informatico – soluzioni tecnologiche che consentano di migliorare l'affidabilità dei *software*, riducendo le distorsioni[80]: spunti in proposito si trovano nella Risoluzione del parlamento europeo del 2017 sul diritto civile robotico, nel quale il principio di precauzione compare tra i principi qualificanti la ricerca nel campo della robotica[81].

In questa prospettiva, per fronteggiare adeguatamente le situazioni di incertezza rispetto agli effetti discriminatori, può farsi utilmente ricorso ad alcune misure, adottabili in sede programmazione, che si trovano indicate in alcuni degli atti di *soft law* intervenuti in materia, prima ricordati. Per limitarsi, conclusivamente, a un solo esempio, andrebbe valorizzato lo spunto che compare nella Carta etica europea sull'utilizzo dell'intelligenza artificiale nei sistemi giudiziari – rivolta anzitutto a progettisti, sviluppatori e sperimentatori – che suggerisce l'effettuazione di uno *stress-test* sull'impiego di algoritmi nell'amministrazione della giustizia teso a valutare il loro impatto sulla protezione dei dati personali e sui diritti previsti dalla CEDU (incluso il rispetto del principio di eguaglianza)[82].

3.2. *I profili "esterni": la significatività della decisione algoritmica ed i margini dell'intervento umano*

I possibili effetti discriminatori prodotti dalle decisioni algoritmiche possono essere mitigati anche attraverso una disciplina che operi non soltanto, per così dire, "a monte" (nei termini finora illustrati) rispetto a tali decisioni, ma anche "a valle", richiedendo un intervento umano di controllo che consenta di individuare e correggere le eventuali distorsioni generate dall'intelligenza artificiale.

Anche questa esigenza trova riscontro anzitutto nella normativa europea: il richiamo va, anche in questo caso, al GDPR, che all'art. 22 (rubricato "processo decisionale automatizzato relativo alle persone fisiche, compresa la profilazione"), primo comma, stabilisce che "l'interessato ha il diritto di non essere sottoposto a una decisione basata unicamente sul trattamento automatizzato, compresa la profilazione, che produca effetti giuridici che lo riguardano o che incida in modo analogo

[78] Si mutua la formula da D. AMIRANTE, *Il principio precauzionale fra scienza e diritto. Profili introduttivi*, in *Diritto e gestione dell'ambiente*, 2/2002 (numero speciale su *Il principio precauzionale fra scienza e diritto*), 20.

[79] In proposito, nella *check-list* che conclude gli *Orientamenti etici per un'IA affidabile*, cit., p. 31, si suggerisce di prevedere in sede di programmazione la presenza di un "pulsante di arresto" al quale potrebbe ricorrersi anche nell'evenienza richiamata.

[80] Un peculiare rilievo rivestono, in questo ambito, le ricerche nel campo della c.d. *algorithmic bias detection*: in tema v., per tutti, R. COURTLAND, *Bias Detectives. The Researchers Striving to Make Algorithms Fair*, in 558 *Nature*, 2018, 357 ss.

[81] Cfr. *Risoluzione del Parlamento europeo del 16 febbraio 2017 recante raccomandazioni alla Commissione concernenti norme di diritto civile sulla robotica*, cit., *Principi generali riguardanti lo sviluppo della robotica e dell'intelligenza artificiale per uso civile,* nn. 7 e 23; il principio di precauzione viene richiamato anche in due documenti allegati alla risoluzione e segnatamente nel *Codice etico-deontologico degli ingegneri robotici* (v. par. *Massimizzare i vantaggi e ridurre al minimo il danno*) e nel *Codice per i comitati etici di ricerca (CER)*, che inserisce il principio di precauzione tra i principi etici di riferimento per i ricercatori nel settore della robotica.

[82] Cfr. *Carta etica europea sull'utilizzo dell'intelligenza artificiale nei sistemi giudiziari e negli ambiti connessi*, cit., *Introduzione*, n. 8, *7.3. Le sfide della "predizione" in materia penale*, n. 138, e *passim*. L'attenzione che il documento riserva al principio di non discriminazione è poi ulteriormente ribadita laddove si raccomanda l'imparzialità non solo nel garantire l'eguale accesso al sistema informatico, ma anche il diritto ad un "eguale trattamento da parte dell'intelligenza artificiale".

significativamente sulla sua persona"[83]. La disposizione stabilisce un confine minimo, operante in certo senso come "punto di partenza" delle garanzie in questo ambito: il divieto di automatizzazione integrale, che postula il necessario intervento umano nelle decisioni incidenti sulle situazioni giuridiche soggettive[84].

Si tratta di un punto di partenza anche perché la norma presenta una portata molto circoscritta e soffre di rilevanti eccezioni. In particolare, quanto ai limiti della disciplina, possono ricordarsi sinteticamente: l'applicazione ai soli soggetti sottoposti a un trattamento di dati personali[85] e la limitazione del divieto di automatizzazione totale alle sole "decisioni"[86] che siamo produttive di "effetti giuridici" sul singolo "interessato"[87].

Sono tuttavia le eccezioni che rivelano i più pericolosi punti deboli della disciplina in commento. La norma evocata, infatti, non trova applicazione non soltanto nell'evenienza in cui la decisione sia autorizzata dal diritto dell'Unione o dello Stato membro cui è soggetto il titolare del trattamento – il che dovrebbe avvenire nel caso di trattamenti giustificati da finalità di interesse pubblico, da operarsi peraltro nel pieno rispetto dei principi di legalità e di proporzionalità[88] – ma anche nella circostanza in cui essa "sia necessaria per la conclusione o l'esecuzione di un contratto tra l'interessato e un titolare del trattamento" (par. 2, lett. *a*) o "si basi sul consenso esplicito dell'interessato" (par. 2, lett. *c*). Ora, è vero che in questi ultimi due casi la decisione interamente automatizzata è possibile a condizione tuttavia che il titolare del trattamento attui "misure appropriate per tutelare i diritti di libertà e i legittimi interessi dell'interessato", garantendo "almeno il diritto di ottenere l'intervento umano da parte del titolare trattamento, di esprimere la propria opinione e di contestare la decisione" (par. 3)[89]; tuttavia, è risaputo quanto sia facile ottenere un

[83] In particolare, il Considerando 71 del GDPR specifica in cosa consista la profilazione, definendola come "una forma di trattamento automatizzato dei dati personali che valuta aspetti personali concernenti una persona fisica, in particolare al fine di analizzare o prevedere aspetti riguardanti il rendimento professionale, la situazione economica, la salute, le preferenze o gli interessi personali, l'affidabilità o il comportamento, l'ubicazione o gli spostamenti dell'interessato".

[84] Un principio che rappresenta, a ben vedere, una garanzia ineliminabile della dignità umana in senso kantiano, posto che stabilisce che una persona non può essere "oggetto passivo di decisioni assunte in forma deumanizzata" (così G. RESTA, *op. cit.*, 222).

[85] Il destinatario della disciplina è infatti unicamente "l'interessato" nel caso di "trattamento" dei suoi dati personali. Si v. sul punto G. RESTA, *op. cit.*, 223, il quale ricorda che la disciplina della decisione automatizzata nel GDPR – come conferma anche il precedente art. 15, comma 1, lett. *h)*, che prescrive la necessaria informazione sulla logica utilizzata nel trattamento – si muove esclusivamente "a partire dai dati personali".

[86] Il che escluderebbe dall'orizzonte applicativo della disposizione interventi diversi, quali ad esempio il *microtargeting* (come ricorda, da ultimo, G. RESTA, *op. cit.*, 225); ma includerebbe comunque decisioni molto rilevanti: valgano in proposito gli esempi proposti dal Considerando 71 del GDPR, primo paragrafo, che anticipa nella sostanza il contenuto della disposizione in commento: "il rifiuto automatico di una domanda di credito online" o "pratiche di assunzione elettronica senza interventi umani".

[87] Quest'ultima limitazione, in particolare, varrebbe ad escludere dall'ambito di applicazione della norma decisioni che non incidano in via diretta sulla sfera giuridica del singolo. Non sarebbe di conseguenza sottoponibile a questa disciplina, ad esempio, l'invio di *fake news* rivolte a specifiche categorie di soggetti. Resta quindi scoperta la "somma di microviolazioni individuali [suscettibile di] produrre un effetto lesivo discriminatorio per l'intero gruppo di riferimento": così G. RESTA, *op. cit.*, 226, nel ribadire la "prevalente logica individualistica" del GDPR. Portata individualistica che però contrasta con la portata prevalentemente collettiva delle decisioni algoritmiche (cfr. in argomento F.Z. BORGESIUS, *op. cit.*, 5).

[88] Come ricordano D.U. GALETTA-J.G. CORVALÁN, *op. cit.*, 1 ss. e spec. 16-17. A ciò si aggiunga che il par. 2 dell'articolo 22, lett. *b)*, stabilisce inoltre che la disciplina europea o nazionale limitativa è tenuta altresì a precisare "misure adeguate a tutela dei diritti, delle libertà e dei legittimi interessi dell'interessato".

[89] Con particolare riferimento al diritto di ottenere, da parte del titolare del trattamento, l'intervento umano, vale precisare che tale intervento "dovrà ovviamente avere carattere sostanziale e non meramente formale: infatti, il soggetto che interverrà in tale fase dovrà avere il potere di esaminare i dati su cui si è fondata la decisione, prendere in considerazione i dati aggiuntivi eventualmente messi a disposizione dall'interessato e le sue ragioni e, in ultima istanza, modificare la decisione automatizzata. Sotto diverso profilo, eventuali errori o imprecisioni nei dati esaminati potrebbero comportare classificazioni non corrette con conseguente assunzione di valutazioni fondate su presupposti errati e possibili impatti negativi sugli individui. Quindi, il titolare dovrebbe effettuare frequenti verifiche sui dati a propria disposizione attraverso *audit* sugli algoritmi utilizzati e sull'accuratezza dei processi di profilazione e

consenso in cambio dell'accesso a servizi fondati su decisioni algoritmiche[90]. Rischio peraltro aggravato dalla circostanza che il consenso esplicito dell'interessato consenta decisioni fondate sul trattamento integralmente automatizzato anche di dati sensibili[91].

Infine, non va sottaciuta un'ulteriore possibile difficoltà applicativa: se infatti va riconosciuto che l'automatizzazione integrale si realizza di fatto in un numero molto limitato di casi e sostanzialmente mai quando decide una pubblica amministrazione[92] o un'autorità giudiziaria[93], occorre sempre tener ben presente – nel caso in cui la decisione amministrativa o giudiziaria si avvalga comunque del contributo dell'intelligenza artificiale – il rischio di "cattura" della decisione da parte del sistema informatico, che si realizza di fatto pur non manifestandosi sul piano formale (e che dunque sfugge alla disciplina in commento): laddove vi sia stato un intervento umano, infatti, dimostrare che l'intelligenza artificiale abbia totalmente influenzato la decisione è una *probatio diabolica*[94], perché anche la circostanza che il decisore si sia integralmente conformato alle indicazioni provenienti del sistema informatico non costituirebbe di per sé la prova dell'assenza di un controllo o di una riponderazione della decisione che abbia condotto a condividere consapevolmente l'esito indicato dall'algoritmo[95].

4. *In conclusione*

Il timore più forte che scaturisce dall'utilizzo dell'intelligenza artificiale per assumere decisioni che incidono sulla vita collettiva – e dai connessi rischi di realizzare *AI-driven discriminations* – è legato al potenziale affermarsi di quello che è stato efficacemente definito un "nuovo medioevo digitale"[96], che rinvia allo scenario di "una società connotata da una segmentazione per caste, ove lo *status* non è però dato dalla nascita o dall'appartenenza a classificazioni sociali tradizionali (quelle su cui vigilano le norme in materia di non-discriminazione), ma da algoritmi e dai valori di coloro che li generano. Classificazioni che sono poi impiegate per prendere decisioni che coinvolgono una

---

valutazione sino a quel momento adottati": così I. DESTRI-A.M. LOTTO, *La profilazione*, in G. Cassano, V. Colarocco, G.B. Gallus, F.P. Micozzi (a cura di), *Il processo di adeguamento al GDPR*, Milano, 2018, 144.

[90] Cfr. A. SIMONCINI-S. SUWEIS, *op. cit.*, 99.

[91] Cfr. il par. 4 dell'art. 22 GDPR, che contempla, insieme alla condizione evocata, anche i motivi di interesse pubblico che giustificano il trattamento (di cui all'art. 9 GDPR, par. 2, lett. *g*) e la vigenza di "misure adeguate a tutela dei diritti delle libertà e dei legittimi interessi dell'interessato".

[92] Così G. RESTA, *op. cit.*, 225. In particolare, nel campo dell'attività amministrativa, vi sono almeno due elementi che impediscono di affidare integralmente la decisione all'intelligenza artificiale: in primo luogo, il ruolo del responsabile del procedimento nella fase istruttoria, che deve intervenire, guidare e controllare lo svolgimento dell'istruttoria procedimentale (e che non può essere sostituito dal sistema informatico); in secondo luogo, la responsabilità giuridica del provvedimento: l'errore va sempre imputato all'essere umano, titolare dell'organo competente. Si tratta, di tutta evidenza, di schemi di comportamento pensati per l'essere umano (cfr. sul punto D.U. GALETTA-J.G. CORVALÁN, *op. cit.*, 22).

[93] Con riguardo alla situazione italiana, va ricordato che la Corte costituzionale ha imposto ai giudici il divieto, nel momento in cui decidono sui diritti, di ricorrere a schemi automatici che influenzino la loro decisione, che deve essere sempre "personalizzata" (richiamando la formula che compare in A. SIMONCINI-S. SUWEIS, *op. cit.*, 101).

In materia, peraltro, è intervenuta anche la *Carta etica europea sull'utilizzo dell'intelligenza artificiale nei sistemi giudiziari e negli ambiti connessi*, cit., che ha formulato una specifica raccomandazione in ordine alla significatività della decisione algoritmica, richiamando la garanzia dell'*under user control* (oltre a ribadire che la decisione finale dev'essere assunta da un giudice umano e che devono essere assicurate sia la spiegazione dei *pattern* seguiti dall'intelligenza artificiale, sia la possibilità di proporre ricorso): si v. in particolare il *Principio n. 5* (*Principio del "controllo da parte dell'utilizzatore"*).

[94] Così A. SIMONCINI-S. SUWEIS, *op. cit.*, 100.

[95] Ai rischi evocati si aggiunge inoltre un ulteriore limite di effettività, legato al diffuso *deficit* di poteri sanzionatori esercitabili dalle autorità nazionali preposte alla protezione dei dati personali (denunciato da F.Z. BORGESIUS, *op. cit.*, 24).

[96] Cfr. G. RESTA, *op. cit.*, 233.

pluralità di soggetti, i quali però non hanno contezza della propria posizione"[97]: decisioni, peraltro, che – nei casi più infausti – si avvalgono dei dati elaborati dagli algoritmi per operare forme esasperate di controllo sociale e politico[98].

La prospettiva diventa ancor più preoccupante se si ha riguardo alle capacità di sviluppo incontrollato dei sistemi di apprendimento automatico: la possibilità che l'algoritmo sia in grado di individuare autonomamente caratteri distintivi suscettibili di generare nuove classi di soggetti potenzialmente discriminati – senza poter operare preventivamente e senza una piena consapevolezza di come ciò possa accadere – ci proietta in uno scenario preconizzato in eminenti riflessioni sullo sviluppo tecnologico[99], che avvertono da tempo come l'incremento della tecnica e della sua inarrestabile potenza sia destinato a prendere il sopravvento su qualsiasi altra forma di manifestazione e di aspirazione umana, divenendo esso stesso lo scopo primario e fondamentale dell'umanità. Letti in questa chiave, i temi qui affrontati fanno assumere un concreto rilevo alle leggi della robotica elaborate da Asimov nel 1942; o, almeno, a due delle esigenze di fondo che le hanno ispirate[100]: l'idea di giustizia, fortemente connessa al principio di non discriminazione; la necessità del controllo umano, che riporta al tema della significatività "limitata" dell'algoritmo nella decisione. Limitazione tanto più necessaria quanto più si abbia consapevolezza, ancora una volta, che "la tecnica si è da molto tempo sottratta alla mera utilizzazione come mezzo e che, al contrario, è essa stessa a trascinarsi dietro l'uomo come suo strumento, sia che egli segua ciecamente questo strappo in avanti, sia che si sforzi in continuazione di indirizzare la tecnica, quanto ai suoi effetti, verso ciò che è propizio e utile"[101].

[97] A. MANTELERO, *La gestione del rischio nel GDPR: limiti e sfide nel contesto dei Big Data e delle applicazioni di Artificial Intelligence*, in A. Mantelero, D. Poletti (a cura di) *Regolare la tecnologia: il Regolamento UE 2016/679 e la protezione dei dati personali. Un dialogo fra Italia e Spagna*, Pisa, 2018, 302.

[98] Come nel caso arcinoto del *social credit system* introdotto in Cina a partire dal 2014: in argomento si v., *ex multis*, F. LIANG-V. DAS-N. KOSTYUK-M.M. HUSSAIN, *Constructing a Data-Driven Society: China's Social Credit System as a State Surveillance Infrastructure*, in 10 *Policy & Internet, Special Issue: Social Media and Big Data in China*, 2008, 415 ss. e D. MAC SÍTHIGH-M. SIEMS, *The Chinese Social Credit System: A Model for Other Countries?*, in 82 *Modern Law Review*, 2019, 1034 ss.

[99] Sul progressivo incontrastabile dominio della tecnica nel mondo occidentale e, in prospettiva, su scala planetaria, v. soprattutto le riflessioni di E. SEVERINO, *Il destino della tecnica,* Milano, 2009.

[100] Cfr. sul punto A. CELOTTO, *op. cit.*, 47.

[101] M. HEIDEGGER, *Conferenze di Brema e Friburgo*, trad. it. di G. Gurisatti, Milano, 2002, 88.