Decisioni algoritmiche e diritto dei dati

Erica Palmerini[1]

1. Il GDPR e la disciplina delle decisioni algoritmiche

Molte decisioni possono essere raggiunte oggi impiegando modelli algoritmici: per individuare i contribuenti da sottoporre ad un controllo fiscale o gli stranieri a cui concedere un visto; per destinare a una certa sede gli insegnanti; per stabilire chi debba essere sottoposto a un controllo di sicurezza più accurato in aeroporto o quale imputato possa essere rilasciato su cauzione. Ancora, per determinare a quale cliente concedere un prestito, e l'eventuale tasso di interesse, o quale candidato invitare ad un colloquio per una posizione lavorativa. Nel commercio elettronico algoritmi servono a fissare il prezzo di un certo bene o servizio, a capire quali messaggi pubblicitari possono più facilmente indurci a un acquisto; nelle interazioni *on line* intercettano i nostri gusti musicali e di lettura, gli interessi rispetto ai temi di attualità, gli orientamenti politici. Queste applicazioni, guidate da trattamenti automatizzati di dati personali, classificano le persone associandole a categorie predeterminate che intendono riflettere certe caratteristiche: la capacità di reddito, la pericolosità sociale, l'affidabilità come debitore, la corrispondenza a un profilo ideale di lavoratore. L'inserzione in uno dei gruppi artificiali così creati permette di dare una risposta algoritmica agli interrogativi che ho brevemente tratteggiato negli esempi iniziali.

Non è un caso, pertanto, che la regolazione giuridica delle applicazioni dell'intelligenza artificiale abbia preso corpo finora nel contesto

[1] Professore associato di Diritto privato – Scuola Superiore Sant'Anna, Pisa. Coordinatrice del Progetto RoboLaw, finanziato dalla Commissione europea nell'ambito del Settimo Programma Quadro (2012-2014), per il quale le è stato conferito il World Technology Award 2013 nella categoria Law.

normativo della protezione dei dati personali. In attesa della definitiva approvazione dell'Artificial Intelligence Act (AI Act),[2] il Regolamento 679/2016 (GDPR) è l'ambito a cui riferirsi per la disciplina delle tecniche di analisi dei dati basate sul calcolo algoritmico e delle decisioni automatizzate. A condizione che il processo che conduce alla decisione algoritmica sia alimentato da dati personali, il GDPR troverà applicazione secondo una duplice traiettoria. Da una parte, un simile trattamento sarà soggetto ai principi e alle regole generali, che vanno dai presupposti di liceità del trattamento, alle caratteristiche del consenso quando è richiesto, al principio di finalità, di minimizzazione dei dati e così via. Da un'altra parte, vi sono previsioni specifiche, congegnate precisamente per affrontare il caso delle decisioni automatizzate, che si collocano principalmente nell'art. 22.

2. Principi di liceità del trattamento e big data analytics

Per quanto riguarda il primo versante, sono state sottolineate alcune criticità nell'applicazione di regole chiaramente non pensate per tecniche di estrazione e analisi di grandi masse di dati attraverso strumenti algoritmici.[3] Ad esempio, la previsione che impone di trattare soltanto i dati necessari a raggiungere le finalità del processo (art. 5, comma 1, lett. c) sarebbe ossimorica rispetto al concetto di *big data*, il cui valore predittivo deriva proprio dall'accumulo, dalle grandi dimensioni e dall'aggregazione, in antitesi con l'istanza di contenimento sottesa al principio di minimizzazione. Con la conseguenza che un simile vincolo, ove inteso in termini molto rigidi, potrebbe tradursi in una perdita di efficienza del processo a scapito dello stesso soggetto interessato dal trattamento. Ad esempio, una società concede piccoli prestiti, a breve termine, a persone che hanno un merito creditizio modesto che li escluderebbe dall'accesso ai canali tradizionali di finanziamento. L'ampliamento della platea dei destinatari del credito dipende dal fatto che l'impresa, anziché affidarsi ai meccanismi consueti per determinare l'affidabilità del debitore, che tengono conto di elementi limitati come precedenti ritardi di pagamento o episodi di insolvenza, dispone di una tecnologia che impiega un numero elevato di variabili, anche

[2] Si tratta della proposta della Commissione europea *for a Regulation laying down harmonised rules on artificial intelligence and amending certain Union legislative acts*, COM 2021, 206 final. Resa pubblica il 21 aprile 2021, si trova ora in una fase avanzata di negoziazione. Il testo cui faremo riferimento è quello che risulta dagli emendamenti apportati dal Consiglio, cd. Compromise Text (14954/22, 25 novembre 2022).

[3] O. Tene – J. Polonetsky, *Judged by the Tin Man: Individual Rights in the Age of Big Data*, 11 *J. on Telecomm. & High Tech. Law* 2013, 362; T.Z. Zarsky, *Incompatible: The GDPR in the Age of Big Data*, 47 *Seton Hall L. Rev.* 2017, 1009 ss. Parla di una possibile "disfunzionalità" della relazione tra GDPR e caratteristiche operative dei sistemi di IA anche R. Messinetti, *La* Privacy *e il controllo dell'identità algoritmica*, in *Contratto e Impresa*, 2021, 130 ss.

apparentemente inconferenti, che però hanno dimostrato in concreto una grande capacità predittiva.[4]
Proprio la complessità e la numerosità delle variabili che un sistema basato sui *big data* può osservare consentono di affinare i processi di decisione e di renderli più granulari, ad evitare quegli esiti discriminatori che dipendono da classificazioni più grossolane ottenute con i processi di stampo tradizionale: ad esempio l'inserimento dell'interessato in una categoria "a rischio", si tratti delle probabilità di recupero di un credito o della sicurezza di un viaggio aereo.
Qualche problema può presentare anche il principio di finalità, che richiede di identificare in anticipo gli scopi del trattamento, portarli a conoscenza dell'interessato e attenervisi nelle successive operazioni (art. 5, comma 1, lett. b). La raccolta delle informazioni, infatti, spesso avviene prima che queste siano destinate a uno specifico uso; l'ipotesi di lavoro talvolta emerge proprio durante l'opera di elaborazione dei dati, non precede l'analisi come nel paradigma tradizionale. In altre parole, sono proprio i processi algoritmici applicati a una materia grezza e capaci di rilevare *patterns* altrimenti non distinguibili che servono a "scoprire" l'utilità dei dati e del loro trattamento.[5]
Altri principi, viceversa, mantengono rilevanza anche nell'ambiente digitale grazie alla loro intrinseca flessibilità. Così, la correttezza può svolgere una funzione di garanzia tutte le volte cui non trova applicazione la disciplina specifica prevista per le decisioni interamente automatizzate. Questo ruolo è cruciale,[6] poiché, come spiegheremo, la fattispecie riguardante le decisioni automatizzate è costruita in termini restrittivi e avrà necessariamente un orizzonte operativo limitato. Tuttavia, tutte le forme di trattamento, anche quelle che non conducono a decisioni (si pensi al caso del marketing personalizzato) ovvero che includono passaggi non automatici, saranno soggette a tale principio che, unitamente a quello di *accountability*, dovrebbe informare il comportamento degli operatori a criteri di lealtà. Altrettanto vale per il requisito dell'esattezza dei dati (art. 5, comma 1, lett. d),[7] che ne impone il continuo aggiornamento, e si pone in corrispondenza ideale con la logica di funzionamento degli algoritmi,

[4] V. Mayer-Schönberger, K. Cukier, *Big Data. The essential guide to work, life and learning in the age of insight*, London, 2013, 46 s., con riferimento all'esperienza della società ZestFinance.
[5] M. Hildebrandt, *Smart Technologies and the End(s) of Law*, 25; A. Rouvroy, *"Of Data and Men". Fundamental rights and freedoms in a world of Big data*, Report for the Council of Europe, Directorate General of Human Rights and Rule of Law, 11 January 2016, 11.
[6] L.A. Bygrave, *Machine Learning, Cognitive Sovereignty and Data Protection Rights with Respect to Automated Decisions*, in *The Cambridge Handbook of Information Technology, Life Sciences and Human Rights*, Part II, *Information and Communication Technologies and Human Rights*, edited by M. Ienca, O. Pollicino, L. Liguori, E. Stefanini, R. Andorno, Cambridge, 2022, 186.
[7] Lo sottolinea A. Moretti, *Intelligenza artificiale: data protection per una governance condivisa*, in Giustizia Civile.com, 1/2020, 6 s.

che non possono produrre esiti validi se non a partire da dati in entrata veri.

3. L'art. 22 e le decisioni automatizzate

A lungo un "angolo dormiente" del diritto della protezione dei dati,[8] la prescrizione dell'art. 22 ha assunto una rilevanza speciale[9] nella discussione al riguardo dei trattamenti algoritmici, poiché ne costituisce la prima, per quanto incompiuta, forma di regolazione.

Il contenuto dell'art. 22 può essere così sintetizzato: esso pone un divieto di trattamenti completamente automatizzati quando la decisione cui condurrebbero produce effetti giuridici o ha comunque un'importanza significativa per l'interessato (comma 1). Al divieto, di per sé già ristretto, si può fare eccezione in tre casi, in presenza cioè del consenso dell'interessato, qualora il trattamento sia necessario per concludere o eseguire un contratto ovvero se una apposita previsione lo autorizzi (comma 2). In tali ipotesi devono essere allora applicati alcuni meccanismi di garanzia per l'interessato, che consistono nel diritto di richiedere l'intervento umano, di esprimere il proprio punto di vista e di contestare la decisione (comma 3).

Il raggio di tutela espresso da queste previsioni non è particolarmente esteso, e ciò per molteplici ragioni. Anzitutto esse riguardano soltanto le *decisioni* che abbiano un impatto giuridico o altrimenti significativo sulla persona. Questa previsione lascia fuori tutte le situazioni in cui strumenti di intelligenza artificiale non servono ad assumere una decisione, bensì ad operare una rappresentazione o una classificazione della realtà. Molte volte gli algoritmi vengono impiegati per questa finalità: ad esempio, nella funzione di autocompletamento di un motore di ricerca consentono di associare al nome proprio una qualità negativa, come "terrorista",[10] oppure, a suggerirne l'uso, quello di una particolare sostanza stupefacente.[11] Di fronte a queste pratiche non potrà essere attivata la tutela di tipo preventivo assicurata dall'art. 22, bensì si farà valere il diritto all'oblio, se ve ne sono i presupposti, oppure si potrà ricorrere al rimedio risarcitorio.

Altre volte ancora la rappresentazione offensiva o comunque distorta prodotta dall'IA non riguarda un individuo determinato, bensì un gruppo etnico-religioso, una classe di persone più o meno estesa, l'intero genere

[8] R. Binns, M. Veale, *Is that your final decision? Multi-stage profiling, selective effects, and Article 22 of the GDPR*, in *International Data Privacy Law*, 2021, 11(4), 319.

[9] M. Brkan, *Do algorithms rule the world? Algorithmic decision-making and data protection in the framework of the GDPR and beyond*, in *International Journal of Law and Information Technology*, 2019, 27(2), 95 s.

[10] Garante per la protezione dei dati personali, 31 marzo 2016 n. 152, in *Corr. giur.*, 2016, 1072 ss.

[11] Trib. Roma, 13 marzo 2017.

femminile. Prima ancora della decisione vera e propria da contraddire, qui manca un interessato cui possano riferirsi mezzi di tutela di tipo individuale. La distorsione dell'informazione realizzata dal motore di ricerca che, in risposta alla digitazione del termine "*chief executive officer*", restituiva quasi esclusivamente immagini di uomini e, tra le poche immagini femminili, per prima quella della bambola "CEO Barbie"[12] non può essere corretta attraverso previsioni del tipo esaminato. Altrettanto si può dire dell'ordine dei risultati che corrisponde alla parola-chiave "jew", dove al primo posto compariva un sito con contenuti antisemiti come "Jew Watch".[13]

In buona sostanza, la rappresentazione errata, distorta o persino discriminatoria che deriva dal propagarsi e perpetuarsi di stereotipi o dalla denigrazione di aspetti della cultura e dell'identità di certi gruppi sociali difficilmente costituisce un effetto significativo: sia in sé considerata sia perché ricade non su individui, bensì su gruppi,[14] che talvolta non sono neppure costituiti intorno a una caratteristica protetta, ma sono creati dallo stesso processo di organizzazione algoritmico.[15]

È possibile pensare anche ad altre figure di collocazione incerta come la pubblicità personalizzata. Grazie allo sfruttamento di informazioni attingibili nel contesto digitale sulle propensioni e le preferenze di acquisto, e finanche sulle vulnerabilità contingenti dei destinatari, può essere molto insidiosa e spingere i consumatori ad un consumo compulsivo, riuscendo ad essere assai più precisa ed efficace delle tecniche basate sulle risultanze di studi di economia comportamentale. Ugualmente distorsivo può essere l'invio di comunicazioni personalizzate in occasione di competizioni elettorali. Eppure, queste pratiche non costituiscono decisioni, bensì processi automatizzati che hanno come destinatario un individuo, ma non si estrinsecano in un atto di volontà. Esse non sarebbero comunque tali da produrre effetti giuridici o determinare conseguenze significative nel senso richiesto dalla norma. Il loro impatto è difficile da determinare in astratto: forme anche aggressive di marketing lasciano indifferenti la maggior parte di noi[16] e condizionano invece pesantemente altri, ma il mero potenziale non basta a giustificare

[12] A. Butterly, *Google Image search for CEO has Barbie as first female result*, BBC News, 16 April 2105, <https://www.bbc.com/news/newsbeat-32332603>; S. Weber, Google, Facebook And Beyond: Why Algorithms Discriminate, Wordcrunch, 10.2.2016, <https://worldcrunch.com/tech-science/google-facebook-and-beyond-why-algorithms-discriminate>.

[13] J. Brandon, *Dropping the bomb on Google*, Wired, 11 May 2004, <https://www.wired.com/2004/05/dropping-the-bomb-on-google/>.

[14] R. Binns, M. Veale, *op. cit.*, 325 s.

[15] Si pensi alla pubblicità di un resort di lusso o di un centro sportivo esclusivo inviati solamente a persone selezionate in base al censo o alla classe sociale.

[16] M. Veale, L. Edwards, *Clarity, surprises, and further questions in the Article 29 Working Party draft guidance on automated decision-making and profiling*, in *Computer Law and Security Review*, 2018, 34, 401.

l'attivazione delle tutele. A regolare questo tipo di condotte potranno essere fonti come la disciplina delle pratiche commerciali sleali, qualora tenute in relazione a rapporti di consumo.

3.1 Segue. Decisioni interamente automatizzate e automation bias

Una seconda ragione che induce a dubitare della capacità della norma di garantirci contro i rischi dei trattamenti algoritmici si lega al carattere interamente automatizzato della decisione necessario per innescare le tutele dell'art. 22: esso, infatti, si applica alla "decisione basata unicamente sul trattamento automatizzato". Benché l'intervento umano tale da escludere l'applicabilità della norma debba essere sostanziale e non meramente formale,[17] il caso più frequente è proprio l'uso dell'IA in funzione di supporto a una decisione umana, che sarà poi attribuita al suo autore. Senonché è riconosciuto come il responsabile umano sia portato ad affidarsi all'indicazione suggerita dal calcolo matematico e raramente sarà incline a discostarsi dall'esito proposto. Ciò per motivi piscologici e anche pratici, tanto più se non è in grado di controllare il percorso razionale che ha condotto al risultato o se è costretto da vincoli, di tempo e di risorse, che non incoraggiano a sviluppare una diversa convinzione. La stessa organizzazione delle attività che ricorrono ad un supporto algoritmico per rendere più rapidi ed efficienti i processi non incentiva certamente a adottare decisioni più ponderate che incidono negativamente sulla produttività individuale e complessiva. È dunque improbabile una difformità di giudizio (e nella risposta applicata) tra l'uomo e la macchina, e più verosimile l'aderenza all'esito di un processo che appare oggettivo e affidabile.

Nel noto caso dell'impiego da parte di alcune corti nordamericane di un algoritmo per determinare il rischio di recidiva degli imputati, si è potuto apprezzare come un'inferenza arbitraria preluda a esiti discriminatori sia quando il risultato prodotto dal calcolo è applicato meccanicamente sia quando il giudice si limiti a tenerlo in considerazione come elemento di una valutazione più complessa. Tuttavia, proprio questa circostanza – essere soltanto uno tra i fattori oggetto di scrutinio, non dotato in sé di un peso determinante – avrebbe reso il software compatibile con i principi del giusto processo.[18]

La difficoltà di determinare l'orizzonte applicativo dell'art. 22 dipende altresì dalla pluralità di situazioni in cui un algoritmo può intervenire nel processo decisionale: dall'offrire informazioni di partenza, operare classificazioni e provvedere una valutazione in termini probabilistici, che

[17] ART. 29 Data Protection Working Party, *Guidelines on Automated individual decisionmaking and Profiling for the purposes of Regulation 2016/679*, 6 February 2018, 20 s.
[18] Supreme Court of Wisconsin, State of Wisconsin v. Eric L. Loomis, July 13, 2016.

poi una persona soppesa e valuta autonomamente, alla selezione dei casi (*triaging*) che richiedono un apprezzamento nel merito. Se così si procede, una serie di ipotesi non arriveranno al vaglio successivo perché scartate in precedenza come non problematiche. In altri casi l'intervento automatico può essere situato alla fine del percorso, e non all'inizio, e dunque l'art. 22 potrebbe risultare applicabile unicamente per questa collocazione del processo automatizzato, che tuttavia ha previsto l'intervento di un decisore umano in uno stadio precedente.[19]
In particolare, quando l'algoritmo interviene all'inizio del processo decisionale, l'operatore umano assume il risultato prodotto (un valore, un punteggio, una classificazione che esprimono un certo significato), vi aggiunge altre informazioni e/o rivaluta quelle già sottoposte al processo automatizzato e infine decide. Ipotesi paradigmatica di decisione esclusivamente automatizzata – e come tale raffigurata nel considerando 71 del GDPR – è quella sulla concessione di un prestito assunta con strumenti automatici. Tuttavia, è frequente che l'algoritmo sia usato piuttosto per la fase di *triage*: le richieste vengono valutate e automaticamente approvate oppure deferite a un revisore umano, che emette la decisione finale. Il processo non è interamente automatizzato, e quindi non scatta la tutela dell'art. 22; la fase precedente, a sua volta, rimane fuori dalla portata degli strumenti offerti dalla *data protection*, a meno di segmentare il processo e considerarla essa stessa una decisione automatizzata che produce effetti significativi sulla persona. Ciò risulterebbe probabilmente costoso e controletterale; nondimeno, l'affidamento nella bontà del calcolo algoritmico avrà un peso sostanziale, cosicché ammettere l'intervento dell'interessato parrebbe opportuno.
Tutta questa complessità esclude che possano darsi risposte nette alla domanda sul carattere esclusivamente automatizzato della decisione; ma ci espone anche al dubbio, in caso di risposta negativa, se le tutele assicurate dall'art. 22 non siano viceversa comunque necessarie.

3.2 Segue. Gli effetti giuridici o altrimenti significativi della decisione

Alla domanda su quale sia il grado di intervento umano sufficiente a impedire l'applicazione dell'art. 22, si intreccia la questione sull'idoneità della decisione a produrre effetti giuridici sull'interessato o "che incida(no) in modo analogo significativamente sulla sua persona".
Nel caso deciso da una Corte olandese, l'algoritmo impiegato da Uber disattiva automaticamente l'account dei *drivers* quando rileva irregolarità o frodi: si tratta di una decisione automatica, ma non ha un impatto significativo, poiché ha un carattere temporaneo e l'utente, mettendosi in contatto con la società, può ottenere la riattivazione. D'altra parte, la

[19] R. Binns, M. Veale, *op. cit.*, 321 ss.

chiusura definitiva dell'account – e quindi la terminazione del rapporto di lavoro – avverrebbe dopo un attento scrutinio delle rilevazioni effettuate dall'algoritmo da parte di un operatore, e ciò la rende una decisione non basata unicamente su un trattamento automatizzato.[20]
Un altro esempio problematico si riferisce al sistema automatizzato che serve a identificare frodi nell'uso di carte di credito o carte di pagamento:[21] comportamenti anomali fanno scattare il blocco della carta o del conto, fino all'intervento di un funzionario che verifica il carattere lecito oppure illecito dei comportamenti che hanno dato luogo alla decisione. Questo effetto – anche se provvisorio – è sufficientemente significativo e dunque innesca il coinvolgimento del titolare del conto a cui chiedere il consenso al trattamento, che potrebbe, d'altronde, essere rifiutato? Sarebbe probabilmente molto oneroso per le società emittenti trattare i sistemi che routinariamente scandagliano le transazioni finanziarie alla ricerca di frodi come soggette all'art. 22, e dunque ricercare per esse una base giuridica costante come il consenso.
Il sistema (Secondary Security Screening Selection della Transportation Security Administration statunitense) in uso presso gli aeroporti americani per identificare persone sospettate di porre un rischio alla sicurezza e, quindi, inviarle a uno screening supplementare prima dell'imbarco è una decisione interamente automatizzata.[22] Ma il ritardo ed eventualmente la perdita del volo che causa sono effetti giuridici o comunque significativi?
Il problema della portata del divieto si è posto in maniera emblematica in un caso riguardante il sistema usato in Germania per l'accertamento della solvibilità di potenziali debitori. Schufa (*Schutzgemeinschaft für allgemeine Kreditsicherung*) è un'agenzia privata che calcola il merito creditizio degli individui in vista dell'ottenimento di un prestito, di un mutuo o addirittura di un contratto di locazione abitativa e, sulla base di alcune caratteristiche analizzate con metodi matematico-statistici, rilascia un certificato con il relativo punteggio. Nella vicenda che dà origine alla questione pregiudiziale sollevata presso la Corte di Giustizia, un'istituzione creditizia, ricevuto tale risultato, rifiuta di concedere un prestito. L'interessato si attiva presso l'agenzia per ottenere informazioni sui dati utilizzati e sulla logica sottostante al processo di calcolo, nonché per ottenere la cancellazione di alcuni dati che reputa errati. Schufa fornisce tuttavia soltanto alcune indicazioni basilari sul suo funzionamento e il punteggio numerico finale, senza spiegare quali particolari elementi abbia impiegato né il loro peso relativo. Adduce, nel fare ciò, che la sua attività non consiste nell'assumere decisioni, bensì nel fornire

[20] Amsterdam District Court, case C/13/692003/HA RK 20-302, (Applicants) v. UBER, 11.3.2021.
[21] R. Binns, M. Veale, *op. cit.*, 327.
[22] Anche per questo esempio cfr. R. Binns, M. Veale, *op. cit.*, 322.

informazioni a terze parti, e dunque si sottrae all'ambito applicativo degli artt. 15 e 22 GDPR; che, in ogni caso, il metodo di calcolo è un segreto commerciale. L'Autorità per la protezione dei dati, cui ci si rivolge perché ingiunga di soddisfare le richieste di accesso e di cancellazione, rigetta il ricorso. La Corte amministrativa di Wiesbaden solleva allora la questione pregiudiziale se l'attribuzione di un punteggio che può essere usato per una futura decisione ricada nell'ambito dell'art. 22.[23]

Il problema suscitato da questa vicenda investe la nozione di "decisione automatizzata" che non comprende l'esito di processi, operati tramite strumenti algoritmici, che possono solo eventualmente costituire la base per una decisione assunta da altri. I dubbi riguardano, per un verso, la nozione di decisione, che non si presta immediatamente a includere operazioni dirette a rappresentare un fatto, a esprimere una valutazione, a effettuare una predizione; per un altro verso, la distinzione, sul piano tecnico-giuridico, di due trattamenti, tendenzialmente in sequenza, portati avanti da responsabili diversi e con diversa finalità. La risposta che darà la Corte dipende dalla capacità di cogliere la stretta inerenza dei due processi, pur formalmente ben distinguibili, ad evitare che la segmentazione dei percorsi che conducono a una decisione – sia dei tempi, sia dei soggetti coinvolti – porti a rendere sempre inapplicabile la norma a quelle che sono sostanzialmente decisioni esclusivamente automatizzate.[24]

4. Le garanzie previste per l'interessato

Nei casi in cui è possibile assumere una decisione interamente automatizzata ai sensi del comma 2 dell'art. 22, l'interessato ha a disposizione alcune facoltà che gli permettono di mettere in contesto la decisione:[25] il risultato del processo automatizzato, che si basa essenzialmente su inferenze statistiche, deve essere "riproporzionato" sul caso singolo. In questo momento può essere dato ingresso nella procedura a quegli elementi di conoscenza aggiuntivi, provenienti dallo stesso interessato, idonei a controbilanciare quelli sui quali si basava, o che

[23] Case C-634/21 SCHUFA Holding and Others (Scoring) and in Joint Cases C-26/22 and C-64/22 SCHUFA Holding and Others (Discharge for remaining debts).

[24] L'opinione dell'Avvocato Generale sembra indicare questa direzione. Un resoconto si può leggere in A. Häuselmann, *The ECJ's First Landmark case on Automated Decision-Making – a Report from the Oral Hearing before the First Chamber*, in *European Law Blog*, 20 February 2023, <https://europeanlawblog.eu/2023/02/20/the-ecjs-first-landmark-case-on-automated-decision-making-a-report-from-the-oral-hearing-before-the-first-chamber/>; F. Palmiotto Ettorre, *Is credit scoring an automated decision? – The Opinion of the AG Pikamäe in the Case C-634/21*, in *The Digital Constitutionalist*, <https://digi-con.org/is-credit-scoring-an-automated-decision-the-opinion-of-the-ag-pikamae-in-the-case-c-634-21/>.

[25] A.F. Fondrieschi, *A Fragile Right: The Value of Civil Law Categories and New Forms of Protection in Algorithmic Data Processing under the GDPR*, in *Oss. dir. civ. comm.*, 2019, 462 s.: "the common denominator of all these rights can be found in the need to bring out the singularity of the concrete case or, in other words, to recontextualise".

avevano un maggior peso, nel processo automatico, ed eventualmente portare alla revisione della decisione.
Strettamente legate a queste facoltà, ed anzi ad esse propedeutiche, sono le regole sull'informativa: collocate negli artt. 13 (comma 2, lett. f), 14 (comma 2, lett. g) e 15 (comma 1, lett. h), GDPR, esse consentono che l'interessato venga a conoscenza del trattamento automatizzato che lo riguarda, nonché della logica di funzionamento che vi è sottesa. Sull'interpretazione di queste norme e sul loro rapporto con il *considerando* 71, che contiene un riferimento ben più pregnante al diritto "di ottenere una spiegazione della decisione", si sono espresse molte posizioni: dall'opinione scettica, e quasi controletterale, sull'esistenza di una simile prerogativa,[26] all'altra che si interroga sulla sua utilità,[27] fino alle declinazioni del diritto alla spiegazione in termini di mera leggibilità.[28]
Al di là dell'esegesi puntuale delle singole previsioni, può dirsi invero che una logica di trasparenza pervade l'intero Regolamento e che nessun senso avrebbero le garanzie predisposte dall'art. 22 se non ci fosse la possibilità di comprendere in base a quali elementi e secondo quale percorso logico la decisione è stata assunta. Piuttosto, le difficoltà possono venire non dalla fragilità degli appigli normativi sui quali incardinare un diritto alla spiegazione, bensì da ostacoli esterni a questo quadro, in termini di praticabilità e opportunità di una trasparenza algoritmica, con cui conviene allora misurarsi.

5. La trasparenza algoritmica presa sul serio

Alcune delle obiezioni rispetto alla possibilità di ambientare un principio di trasparenza e di informazione nel contesto dei trattamenti algoritmici muovono da una logica di efficienza economica: gli algoritmi c.d. *black-box*, che tengono in considerazione una quantità più elevata di variabili e processano grandi moli di dati, sarebbero per loro stessa natura meno interpretabili e, tuttavia, più affidabili nei risultati, dunque maggiormente efficaci. A questa stregua, l'obbligo di fornire una descrizione della logica su cui si basano i processi avrebbe un impatto non desiderabile sulle attività economiche che si avvalgono di tecniche di *data analysis*: gli operatori sarebbero infatti obbligati a scegliere algoritmi più semplici e, perciò, meno validi, con possibili esiti negativi sia per la buona conduzione dell'impresa sia per il soggetto sul quale la decisione finale,

[26] S. Watcher, B. Mittelstadt, L. Floridi, *Why a Right to Explanation of Automated Decision-Making Does Not Exist in the Data Protection Regulation*, in *Int. Data Privacy Law*, 2017(7), 76 ss.
[27] L. Edwards, M. Veale, *Slave to the Algorithm? Why a Right to an Explanation is Probably Not the Remedy You Are Looking For*, in 16 *Duke Law & Tech. Rev.* 2017, 18 ss.
[28] G. Malgieri, G. Comandé, *Why a Right to Legibility of Automated Decision-Making Exists in the General Data Protection Regulation*, in *Int. Data Privacy Law*, 2017(7), 243 ss.

ad esempio ottenere un prestito, deve ricadere. In alternativa, si troverebbero costretti a impegnare risorse e competenze esperte per chiarire in termini comprensibili il modello decisionale impiegato. La necessità di rivedere i risultati del calcolo da parte di un essere umano finirebbe per vanificare il senso del ricorso all'IA, ossia automatizzare i processi per renderli più rapidi, meno costosi e più sicuri nei risultati. Obblighi di trasparenza si tradurrebbero in oneri economici e organizzativi per le imprese senza apportare reali vantaggi ai destinatari delle decisioni, finendo per essere fini a se stessi o addirittura controproducenti.[29]

Queste considerazioni radicalmente avverse all'accoglimento di una linea di tendenziale trasparenza sono invero già superate dal dato normativo e dalla formulazione testuale delle norme. La trasparenza opera come principio regolatore sia nella fase di progettazione e messa a punto dei sistemi sia nel momento dell'attività operativa.[30] Al punto che tali vincoli potrebbero persino orientare verso l'impiego esclusivo di modelli interpretabili, almeno quando la decisione in gioco è di estrema rilevanza e l'impossibilità di fornire una spiegazione sarebbe in conflitto con diritti fondamentali.[31]

Le opinioni scettiche sulla capacità di ambientare un principio giuridico di trasparenza nella realtà digitale non considerano, in effetti, che esso è condiviso dalla stessa scienza dei dati. Al di là degli obblighi imposti dalla disciplina in materia di protezione dei dati personali, la ricerca di un punto di equilibrio tra efficienza e leggibilità degli algoritmi è un obiettivo perseguito anche sul piano scientifico. Vi è infatti piena consapevolezza che nonostante le decisioni automatizzate siano in termini generali affidabili, oggettive e depurate dell'elemento di fallibilità proprio della valutazione umana, non sono tuttavia immuni da errori. E se i falsi dell'algoritmo sono innocui nel contesto di un'indagine sperimentale, possono avere invece esiti inaccettabili quando sono impiegati per assumere decisioni sulla vita delle persone. Inoltre, poiché l'impenetrabilità previene la scoperta degli errori, non è accettabile neppure dal punto di vista del metodo scientifico. Non a caso, una linea di ricerca innovativa – un movimento composito e variegato, in realtà, emerso in risposta alle preoccupazioni evidenziate – è proprio quella

[29] N. Wallace, D. Castro, *The Impact of the EU's New Data Protection Regulation on AI*, Center for Data Innovation, March 27, 2018, <http://www2.datainnovation.org/2018-impact-gdpr-ai.pdf>; T.L. ZARSKI, *op. cit*., 1017.

[30] R. Messinetti, *op. cit.*, 166.

[31] Questa è la proposta avanzata dagli autori di uno studio interdisciplinare: A. Bibal, M. Lognoul, A. de Streel, B. Frénay, *Legal requirements on explainability in machine learning*, in *Artificial Intelligence and Law*, 2020. Mentre R. Messinetti, *op. cit.*, 167, afferma che, se non si possono rispettare le garanzie del GDPR in termini di spiegabilità della logica usata, non vi sono "le condizioni per sospendere il divieto di sottoporre la persona al processo decisionale automatizzato posto dal primo comma dell'art. 22".

rivolta a una "explainable AI",[32] da perseguire, in ipotesi, progettando algoritmi che siano in grado essi stessi di dare una spiegazione del processo che ha condotto ad un certo risultato.[33]
La trasparenza rappresenta quindi una sfida al contempo tecnologica e giuridica, e il precetto che la accoglie non è in contraddizione con le caratteristiche strutturali dei sistemi intelligenti, bensì, eventualmente, soltanto con i limiti contingenti all'attuale livello di progresso in materia di IA.
Un secondo ordine di rilievi investe il difficile contemperamento del principio di trasparenza rispetto ai diritti di privativa intellettuale o alla tutela del segreto commerciale che il titolare del trattamento potrebbe opporre. L'investimento nell'elaborazione di algoritmi innovativi ed efficienti è infatti senz'altro da proteggere, mentre una completa *disclosure* potrebbe non soltanto avvantaggiare i concorrenti, ma anche indurre i potenziali destinatari delle decisioni automatizzate ad attuare strategie di aggiramento che ugualmente andrebbero a diminuire il rendimento della tecnologia adottata.
Una possibile opzione interpretativa rivolta a conciliare queste istanze con il diritto a una spiegazione è quella che disgiunge il concetto di trasparenza proprio della comunità di *machine learning* da quello avuto di mira dalla disciplina in materia di *privacy*.[34] Il primo si riferisce essenzialmente all'accessibilità del codice sorgente, che a sua volta consente di risalire al modello matematico utilizzato. Questo contenuto, tuttavia, offre all'interessato che non abbia competenze esperte informazioni poco utili o addirittura si rivela non esigibile sul piano giuridico perché coperto da privativa. Viceversa, il precetto di trasparenza accolto dalla normativa si riferisce all'intelligibilità dell'algoritmo, ossia ad una conoscenza che sia fruibile dal profano e gli permetta di comprendere le ragioni per le quali si è arrivati a una certa conclusione. La spiegazione cui si riferiscono gli artt. 13-15 GDPR è dunque quella leggibile per l'interessato e priva di informazioni oggetto di proprietà intellettuale[35] o comunque espresse in un linguaggio matematico, che non

[32] Cfr. il programma Explainable Artificial Intelligence (XAI) della Defense Advanced Research Projects Agency (DARPA) americana, <https://www.darpa.mil/program/explainable-artificial-intelligence>, nonché il Progetto europeo, finanziato nell'ambito di Horizon 2020, XAI – Explanation of AI Decision Making, <https://xai-project.eu/>.
[33] A.D. Selbst, S. Barocas, *The Intuitive Appeal of Explainable Machines*, 87 *Fordham Law Rev.* 2018, 1085 ss.
[34] K. Yeung, A. Weller, *How is 'transparency' understood by legal scholars and the machine learning community?* in *Being Profiled. Cogitas Ergo Sum: 10 Years of Profiling the European Citizens*, edited by E. Bayamloğliu, I. Baraliuc, L. Janssens, M. Hildebrandt, Amsterdam, 2018; A. Bibal, M Lognoul, A. de Streel, B. Frénay, *op. cit.*
[35] L. Edwards, M. Veale, *op. cit.*, 54 ss., propongono molteplici tipologie di spiegazione possibili della logica di un algoritmo compatibili con la tutela di eventuali diritti di privativa. Similmente, M. Brkan, G. Bonnet, *Legal and Technical Feasibility of the GDPR's Quest for Explanation of Algorithmic*

avrebbe senso dischiudere. La trasparenza nel senso del GDPR non coincide pertanto con la *disclosure* del codice o del modello matematico, che in sé non è necessaria e non sarebbe comunque sufficiente a dare attuazione al suo significato autentico.
In modo particolare, come esprime meglio il termine "*meaningful*" – il predicato assegnato all'informazione nel testo inglese del Regolamento – occorre misurare tale qualità rispetto al singolo interessato:[36] si tratta cioè di una nozione che ha meno il senso dell'importanza "quantitativa" e più quello della capacità di trasmettere all'individuo elementi utili alla sua comprensione e all'eventuale esercizio delle scelte che ne conseguono. In tal senso è possibile attingere anche alla "*meaningful trasparency*"[37] del diritto dei contratti dei consumatori, dove l'informazione dovuta si attesta su quella necessaria per compiere scelte consapevoli, mentre deve essere evitato sia il sovraccarico di informazioni, che ha un effetto controproducente, sia la restituzione di contenuti non pertinenti o troppo complessi, e dunque incomprensibili.
La ricostruzione delle perplessità, di ordine tecnico e di ordine giuridico, circa una trasparenza intesa in termini rigidi e assoluti rende il "diritto a una spiegazione" una posizione di interesse flessibile, adattabile e non sempre giustiziabile,[38] ma non elimina la necessità di prendere sul serio la trasparenza come obiettivo di politica del diritto e come criterio ordinante nell'organizzazione delle attività di trattamento automatizzato dei dati e di produzione delle decisioni. Ciò significa che la tracciabilità del percorso logico di decisione deve costituire, fin dove possibile, una scelta di *design* che lo sviluppatore può imprimere, in modo particolare attraverso le impostazioni di *logging*.[39] Là dove una completa trasparenza non sia attingibile allo stato dell'arte, l'algoritmo validato da un impiego consolidato che lo dimostri sicuro negli esiti può essere spiegato attraverso una semplice ipotesi di funzionamento che trovi conferma nel dato di esperienza. In determinati contesti, tuttavia, il principio di trasparenza può finire persino per significare un divieto dell'uso di algoritmi che abbiano

Decision: of Black Boxes, White Boxes and Fata Morganas, in *European Journal of Risk Regulation*, 2020, 11(1), 40 ss.

[36] A.D. Selbst, J. Powles, *Meaningful information and the right to explanation*, in *Int. Data Privacy Law*, 2017(7), 236. L.A. Bygrave, *op. cit.*, 182, sottolinea come il predicato esprima la necessità di un'informazione modellata sulle capacità cognitive dell'interessato.

[37] M. Durovic, *European Law on Unfair Commercial Practices and Contract Law*, Hart Publishing, Oxford-Portland, 2016, 117.

[38] Persino l'AI Act sembra avere accolto una forma debole di trasparenza, rivolta soltanto agli utenti del sistema e non ai destinatari finali, tra i requisiti di sicurezza che i sistemi algoritmici devono osservare in una logica *ex ante* (art. 13): cfr. M. Ebers *et al.*, *The European Commission's Proposal for an Artificial Intelligence Act – A Critical Assessment by Members of the Robotics and AI Society (RAILS)*, in *Multidisciplinary Scientific Journal*, 2021, 596.

[39] J. Bryson, *The Artificial Intelligence of the Ethics of Artificial Intelligence: An Introductory Overview for Law and Regulation*, in *The Oxford Handbook of Ethics of Artificial Intelligence*, edited by M. Dubber, F. Pasquale, e S. Das, Oxford, 2019.

una logica di funzionamento completamente oscura. Tale divieto può derivarsi direttamente dalla impossibilità pratica di soddisfare gli obblighi imposti dal Regolamento, che rende illegittimo un trattamento non conforme con i suoi precetti. Ma la scelta di astenersi dall'impiegare algoritmi indecifrabili può anche corrispondere a incentivi che provengono da altri terreni. Se, infatti, l'obiettivo della esplicabilità è troppo debolmente affermato nella disciplina della *data protection*, può ricevere sostegno dall'affermarsi di standard nell'ambito della responsabilità contrattuale ed extracontrattuale. Questi potrebbero cioè portare a includere, nella valutazione dell'attività che si avvale di modelli computazionali, il livello di comprensione della logica dell'algoritmo utilizzato e "premiare" il comportamento che, a questo riguardo, si dimostra più accorto e avveduto.[40]

6. I limiti delle tutele offerte dalla data protection

L'analisi puntuale della disciplina ha fatto emergere come la regolazione degli algoritmi non possa essere lasciata unicamente alla prospettiva della *data protection*,[41] benché quest'ultima sia logicamente essenziale.

Le regole che disegnano l'architrave di tale sistema normativo, infatti, rispondono a logiche non corrispondenti ai bisogni di tutela che emergono nei confronti dei trattamenti algoritmici, quando non sono addirittura incompatibili con essi.

Un parametro di liceità centrale nel primo sistema normativo è quello della esattezza e dell'aggiornamento dei dati. Si tratta di un presupposto valido anche rispetto al trattamento automatizzato, che può restituire risultati validi soltanto a partire da dati corretti; esso, tuttavia, non esaurisce il principio di qualità dei dati, che costituisce ora uno dei cardini della sicurezza algoritmica cui è improntato l'AI Act.[42] Tale caratteristica esprime infatti anche un'istanza di ricchezza e inclusività del campione di informazioni, che deve essere in grado di rappresentare tutte le variabili presenti nella popolazione che sarà oggetto della valutazione automatizzata.

Qualche esempio tratto dalla casistica di studi e sperimentazioni in essere permette di comprendere meglio questa affermazione: i software usati in

[40] Ad ambientare il principio di trasparenza nel più generale contesto della responsabilità aquiliana e contrattuale, in particolare nel settore della responsabilità medica, è lo studio di P. Hacker, R. Krestel, S. Grundmann, F. Naumann, *Explainable AI under contract and tort law: legal incentives and technical challenges*, in *Artificial Intelligence and Law*, 2020.

[41] C. Sabelli, M. Tallacchini, *From Privacy to Algorithms' Fairness*, in *Privacy and Identity Management. The Smart Revolution*, a cura di M. Hansen *et al.*, 2017; A. VEDDER, *Why data protection and transparency are not enough when facing social problems of machine learning in a big data context*, in *Being Profiled*, cit., 2018.

[42] Cfr. l'art. 10 intitolato "Data and Data Governance".

medicina, e in particolare in dermatologia, per leggere immagini della cute e rilevare le lesioni a carattere patologico si sono dimostrati molto efficaci, ma la percentuale di successo decresce esponenzialmente quando la pelle che costituisce lo sfondo della lesione è di colore scuro.[43] Ciò dipende dalla presenza insufficiente nei set di dati impiegati per il training dell'algoritmo di immagini provenienti da persone di diversa etnia rispetto a quella caucasica. In un fallimento di questo tipo incorre anche il software per il riconoscimento facciale di Amazon che, in un esperimento condotto dalla American Civil Liberties Union, scambia politici e membri del Congresso americano per soggetti accusati di un crimine, ritratti in un database di foto segnaletiche. In particolare, l'errore ricadeva più frequentemente su persone di origine afroamericana o latina.[44]

Una lacuna simile potenzialmente affligge i modelli utilizzati in ambito lavorativo per la selezione dei candidati a una determinata posizione: se il genere femminile è poco rappresentato nelle esperienze pregresse di assunzione, ciò ne determina un funzionamento imperfetto e rende meno affidabile la valutazione per il gruppo sottorappresentato.[45] L'incompletezza del *data set* può peraltro dipendere dal divario tecnologico: certi gruppi di soggetti sono meno presenti nei campioni utilizzati proprio perché sfuggono alla *datafication*, ossia la loro vita e le loro attività economiche non sono coinvolte a sufficienza nella generazione di dati che si deve alla pervasività dei dispositivi tecnologici.

È anzitutto in questa prospettiva che si apprezza la differenza con le garanzie derivanti dalla normativa sul trattamento dei dati personali: quest'ultima assicura, in funzione di tutela dell'individuo, l'esattezza e la completezza dei dati, ad evitare che informazioni erronee, parziali o non aggiornate ne riflettano un'immagine distorta o incompiuta. Per un funzionamento ottimale dell'algoritmo, la base di dati deve invece essere anche inclusiva.

Un altro presidio centrale nel contesto della protezione dei dati sono le restrizioni poste al trattamento di categorie particolari di dati, che danno luogo al divieto di raccoglierle o di processarle se non in situazioni specifiche e con le opportune garanzie. Ma questi stessi limiti, ambientati nel contesto dei trattamenti algoritmici, rischiano, per un verso, di essere

[43] Young *et al.*, *Artificial Intelligence in dermatology: a primer*, in *Journal of Investigative Dermatology*, 2020, 140, 1504 ss.; A.S. Adamson, A. Smith, *Machine Learning and Health Care Disparities in Dermatology*, in *JAMA*, 2018, 154(11), 1247 ss.

[44] N. Singer, *Amazon's Facial Recognition Wrongly Identifies 28 Lawmakers, A.C.L.U. Says*, The New York Times, July 26, 2018, <https://www.nytimes.com/2018/07/26/technology/amazon-aclu-facial-recognition-congress.html>.

[45] J. Kroll *et al.*, *Accountable Algorithms*, 165 *Univ. Penn. Law Rev.* 2017, 681. In tema cfr. A. Kelly-Lyth, *Challenging Biased Hiring Algorithms. The Adequacy and Enforcement of EU-derived Law*, in *Oxford J. of Legal Studies*, 2021, 41(4), 899 ss.

facilmente elusi; per un altro verso, possono divenire una barriera rispetto agli obiettivi di giustizia e non discriminazione. Dal primo punto di vista, dati che non appartengono a categorie specialmente protette possono costituire un indice rilevante per inferirne informazioni sullo stato di salute, l'orientamento sessuale, le opinioni politiche, l'origine etnica. Ad esempio, il codice postale dell'indirizzo di residenza rappresenta notoriamente un *proxy* per l'appartenenza a minoranze etniche, specialmente in determinati territori e insediamenti urbani;[46] le ricerche effettuate su internet per parole chiave possono essere indicative di una condizione patologica. Ciò dimostra come, specialmente nel mondo dei *big data*, le restrizioni al trattamento dei dati sono da sole insufficienti. Ma vi è di più, perché potrebbero persino ridondare in una minore accuratezza dei risultati, con un effetto controproducente.

Il divieto di trattare informazioni sensibili, rendendo cieco l'algoritmo rispetto a caratteristiche protette, è considerato infatti una strategia deficitaria: anzitutto perché molte caratteristiche, come l'essere donna o appartenere a una minoranza etnica, sono codificate nei nostri comportamenti, dal tipo di vestiti che acquistiamo ai cibi che consumiamo, dal livello di educazione al luogo in cui abitiamo, e possono essere facilmente inferite.[47] Gli algoritmi di *machine learning* sono particolarmente efficienti a trovare variabili latenti e a scoprire caratteristiche sensibili, anche quando non ricevono un input diretto. Ma, soprattutto, tale (apparente) neutralità previene la possibilità di rilevare i pregiudizi da cui è affetto il set di dati impiegati per costruire e validare l'algoritmo e, di conseguenza, i suoi risultati. Preso atto che soltanto contemplando dati sensibili nella costruzione dei modelli può essere conseguito un risultato algoritmico equo, l'AI Act ha previsto un'esenzione al loro trattamento per i fornitori di sistemi ad alto rischio (art. 10(5)).[48]

Un'ultima osservazione, infine, può convincere dell'insufficienza di una regolazione basata esclusivamente sulla disciplina del trattamento dei dati personali. I processi di calcolo che operano su dati corretti possono dare risultati comunque indesiderabili e discriminatori, là dove in parte

[46] Non a caso, l'art. 31(1) n 3 del German Data Protection Act vieta di prendere in considerazione, almeno come parametro esclusivo, l'indirizzo di residenza a fini di *credit scoring*, perché è noto che risulta strettamente correlato all'origine etnica. Con riferimento alla variabile della razza nelle decisioni relative al prezzo del credito, cfr. T.B. Gillis, J.L. Spiess, *Big Data and Discrimination*, in 86 *Univ. Chicago L. Rev.*, 2019, 468 ss.

[47] C. Dwork, *The emerging theory of algorithmic fairness*, <https://www.microsoft.com/en-us/research/video/the-emerging-theory-of-algorithmic-fairness/>.

[48] I. Žliobaitė, B. Custers, *Using sensitive personal data may be necessary for avoiding discrimination in data-driven decision models*, in *Artificial Intelligence and Law*, 2016, 24, 183 ss.; G. Resta, *Cosa c'è di 'europeo' nella proposta di Regolamento UE sull'intelligenza artificiale?*, in *Dir. inf.*, 2022, 336 ss.

riflettono disuguaglianze effettivamente diffuse nella realtà, in parte addirittura le amplificano. Istruttiva, al riguardo, è la vicenda relativa alla Medical School del St. George's Hospital di Londra, che aveva introdotto un sistema automatico di valutazione delle domande per le posizioni di specializzazione messe a disposizione annualmente. Per "insegnare" al sistema il modo di operare, vi immette i giudizi delle commissioni di selezione degli anni precedenti. Un buon numero delle domande rigettate aveva in comune alcuni elementi: una peggiore conoscenza della lingua inglese – dimostrata da errori ortografici o sintattici nella presentazione della domanda – ma altresì aspetti come il luogo di nascita, il cognome o l'indirizzo di residenza, spesso indicativi di una certa provenienza geografica o della condizione di immigrazione. Il software generalizza questa correlazione e applica un giudizio negativo (o comunque comparativamente peggiore) alle domande che condividevano questi fattori, del tutto estrinseci rispetto alle reali qualità del candidato.[49] In effetti, anche quando la costruzione del modello si basa su decisioni pregresse etichettate come corrette, non si è al riparo dal rischio di una distorsione dei criteri di scelta e valutazione.

Nel caso, già ricordato, della compagnia ZestFinance che concede piccoli prestiti a breve termine a un tasso di interesse sensibilmente più basso di quello che offre il mercato, la tecnologia impiegata per selezionare i debitori analizza, tra l'altro, informazioni come il tempo e la cura impiegati per compilare la domanda di accesso al finanziamento, la presenza di errori ortografici, se ci si è dedicati a leggere le condizioni di contratto. Il modello predittivo utilizzato si basa sull'assunto che la fedeltà alle regole, anche grammaticali, ovvero l'interesse per il contenuto del contratto cui ci si vincola sia sintomatico di una maggiore propensione a rispettare gli impegni in termini più generali. Sennonché attribuire rilevanza generalizzata alla padronanza della lingua scritta e alla capacità di comprensione delle condizioni contrattuali può svantaggiare persone con una bassa alfabetizzazione o soggetti immigrati di recente, senza che tali condizioni abbiano una reale incidenza sulla solvibilità individuale.[50]

Secondo un lavoro pubblicato su *Science*, un algoritmo in uso presso il sistema sanitario statunitense discrimina gli afroamericani poiché usa la spesa sostenuta per le cure come un *proxy* per i bisogni effettivi di terapia; riflette, in questo modo, le disuguaglianze nella distribuzione della ricchezza nella popolazione americana.[51] Ancora una volta, da una

[49] S. Barocas, A.D. Selbst, *Big Data's Disparate Impact*, 104 *California L. Rev.* 2016, 682.
[50] C. O'Neil, *Weapons of Math Destruction. How Big Data Increases Inequality and Threatens Democracy*, New York, 2016, 128.
[51] Z. Obermeyer *at al.*, *Dissecting racial bias in an algorithm used to manage the health of populations*, in *Science*, 25 Oct 2019, vol. 366, issue 6464, 447 ss.

situazione apparentemente registrata in maniera neutrale possono essere tratte inferenze che perpetuano e aggravano le disuguaglianze.

Non è dunque sufficiente ogni cautela nel formare e processare una base di dati esatti, aggiornati e inclusivi, bensì occorre che la progettazione del sistema di calcolo avvenga secondo criteri di scientificità e validità statistica. Si tratta, in altre parole, di fissare regole che governino l'architettura degli algoritmi secondo standard asseverati dalla comunità scientifica in tutte le fasi di elaborazione del modello:[52] dalla raccolta dei dati al modo in cui si procede a "etichettarli"; dalla specificazione del problema da risolvere in termini formali alla scelta del numero di fattori da prendere in considerazione, oltre che del loro peso, per avere una risposta valida; dal riconoscimento dei "pregiudizi" impliciti nei dati o insiti nelle decisioni passate usate come esempi all'introduzione di correttivi.

7. Una strategia integrata per la regolazione degli algoritmi

La creazione di un quadro di regole di sicurezza per i sistemi algoritmici è l'obiettivo dell'AI Act. Nello spazio limitato di questo intervento non è possibile analizzare nel dettaglio il contenuto e l'efficacia delle norme che si stanno delineando per governare la tecnologia algoritmica. Possono tuttavia essere proposti alcuni spunti per riflettere sulla capacità della regolazione emergente di corrispondere alle lacune evidenziate nel sistema della protezione dei dati e integrare così il quadro delle tutele.

Una prima osservazione si riferisce alla definizione dei sistemi ad alto rischio, gli unici per i quali sono dettati requisiti di sicurezza e obblighi di certificazione. Si tratta di una nozione ampia e costruita su un doppio binario, destinata tuttavia a non comprendere alcune delle ipotesi che abbiamo esemplificato. Gli algoritmi impiegati in contesti digitali, dai social network al commercio elettronico, non saranno verosimilmente soggetti a questa disciplina. Pratiche come la pubblicità personalizzata e la discriminazione dei prezzi potranno trovare collocazione nel divieto di pratiche commerciali scorrette, e nei rimedi, anche individuali, che tale ambito normativo mette a disposizione, ma si trovano ai margini del processo di regolazione che investe il fenomeno dell'intelligenza artificiale in sé. Lo stesso accade per la manipolazione dell'informazione attraverso i cd. *deepfakes*, per i quali sussiste solo un obbligo di trasparenza (art. 52). Altri applicativi lambiscono la sfera della salute, ma se non costituiscono dispositivi medici in senso stretto oppure non servono per amministrare servizi di emergenza (cfr. Allegato III, n. 5, lett. c), non

[52] Di "Legality by design" parla G. Resta, *Governare l'innovazione tecnologica: decisioni algoritmiche, diritti digitali e principio di uguaglianza*, in *Pol. dir.*, 2019, 218 ss.

ricadono nell'ambito di applicazione della proposta di Regolamento.[53] Che la categoria del rischio elevato sia stata definita in termini troppo restrittivi è dunque una delle critiche avanzate nei confronti del Regolamento.

Un secondo rilievo attiene al metodo della co-regolazione cui ancora una volta la Commissione si rivolge per disciplinare i prodotti europei: ossia, con la fissazione di requisiti generali di sicurezza nel testo vincolante e la delega agli organismi di standardizzazione per la determinazione delle specifiche tecniche che permettono di soddisfarli. La tendenza di tali organizzazioni a diventare "i veri 'legislatori' del settore", come accade in altri comparti industriali,[54] è resa più problematica dall'oggetto del tutto peculiare del loro intervento. La costruzione dei modelli algoritmici coinvolge infatti la dimensione dei diritti fondamentali, e il design dei dispositivi dovrebbe pertanto riflettere anche giudizi di valore e soluzioni di bilanciamento tra esigenze in conflitto. L'assenza di competenze all'interno degli organismi capaci di cogliere questa speciale natura dei rischi, nonché la loro più generale mancanza di trasparenza e di rappresentatività in ordine a scelte che incidono sui diritti della persona e delle collettività, acuiscono le preoccupazioni verso un approccio considerato tecnocratico e orientato soprattutto al mercato.[55]

A ciò si aggiunga che il rispetto dei criteri di sicurezza è quasi sempre lasciato ad una autovalutazione del fornitore del sistema di IA, e solo in casi limitati è previsto l'intervento di controllo, ai fini della certificazione, di un soggetto indipendente.[56] Ancora sul piano dell'effettività della disciplina, una lacuna è considerata la mancanza di diritti e rimedi individuali che possano accompagnare l'*enforcement* pubblicistico.[57] I meccanismi di *governance* non sembrano dunque soddisfare l'esigenza che l'analisi appena condotta ha messo in evidenza: ossia, quella di un monitoraggio del funzionamento degli algoritmi sia in fase di certificazione sia successivo alla loro messa in commercio o comunque

[53] H. van Kolfschooten, *Conspicuous by its absence: health in the European Commission's Artificial Intelligence Act*, BMJ Opinion, 2021, <https://blogs.bmj.com/bmj/2021/07/30/conspicuous-by-its-absence-health-in-the-european-commissions-artificial-intelligence-act/>; M. Ebers *et al.*, *op. cit.*, 594.
[54] G. Resta, *Cosa c'è di 'europeo'*, cit., 341 s.
[55] M. Ebers *et al.*, *op. cit.*, 594 s.; M. Veale, F. Borgesius, *Demistifying the draft EU Artificial Intelligence Act*, in *Computer Law Rev Int.*, 4/2021, 105; N. Smuha *et al.*, *How the EU can achieve Trustworthy AI: A Response to the EU Commission Proposal for an Artificial Intelligence Act*, 5 August 2021, 54.
[56] M. Ebers *et al.*, *op. cit.*, 595 s.; M. Veale, F. Borgesius, *op. cit.*, 106; particolarmente critici sul punto L. Edwards, *Regulating AI in Europe: four problems and four solutions*, Ada Lovelace Institute Expert opinion, March 2022, 23 ss.; N. Smuha *et al.*, *op. cit.*, 37 ss.
[57] M. Ebers *et al.*, *op. cit.*, 599 s.; M. Veale, F. Borgesius, *op. cit.*, 111; G. Resta, *Cosa c'è di 'europeo'*, cit., 342; L. Edwards, *op. cit.*, 10 s.; N. Smuha *et al.*, *op. cit.*, 44 ss. Questa impressione è solo in parte mitigata dall'introduzione nel Testo di Compromesso dell'art. 63(11) sul diritto di rivolgere un reclamo all'autorità di sorveglianza del mercato competente.

alla loro applicazione. Tali forme di controllo, in particolare, dovrebbero essere istituzionalizzate, e non lasciate all'iniziativa spontanea di individui o gruppi organizzati che non assicurano la sistematicità e il sapere esperto necessari.[58]
Nel merito, infine, dei requisiti di sicurezza, ancora molto limitata appare l'attenzione al carattere non discriminatorio dell'algoritmo. Costruire un algoritmo privo di pregiudizi non è semplice, e ciò per diverse ragioni: gli esiti possono essere discriminatori quando riflettono una disuguale distribuzione di qualità rilevanti per l'obiettivo di ricerca assegnato, che è propria della realtà di partenza. Cercare di eliminare o circoscrivere questi effetti, ridimensionando, ad esempio, il peso di variabili che invece a livello statistico sono molto predittive, può portare a diminuire l'accuratezza dei risultati. D'altra parte, ridurre le disuguaglianze sociali richiede altrimenti di introdurre dei correttivi che potranno agire alla stregua di "azioni positive"[59] o avere una funzione distributiva.[60]

Vi è da aggiungere che attenuare le distorsioni può essere oneroso, poiché presuppone di cambiare l'impostazione del modello ed ottenere dati aggiuntivi non sempre liberamente o facilmente accessibili. Occorre pertanto definire uno standard di qualità nella predisposizione della collezione di dati e nella progettazione dell'algoritmo che sia sostenibile per le imprese private e tuttavia improntato a un rigore che il regolatore considera adeguato al contesto di impiego. In relazione a tale profilo, anche per quanto già detto con riguardo alla definizione dei sistemi ad alto rischio, l'approccio dell'AI Act non sembra ancora del tutto compiuto.

[58] I *bias* che affliggevano il software COMPAS usato in diversi Stati americani per determinare la pericolosità sociale e quindi il rischio di recidiva degli arrestati sono stati disvelati da un'indagine dell'associazione di giornalismo investigativo ProPublica, che ha messo a confronto, per un periodo protratto per circa due anni, i risultati prodotti dall'algoritmo con l'effettivo ritorno a delinquere degli arrestati: J. Angwin *et al.*, ProPublica, *Machine Bias. There's software used across the country to predict future criminals. And it's biased against blacks*, May 23, 2016, <https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing>. Nel caso che ha coinvolto la Medical School del St. George's Hospital di Londra, soltanto su iniziativa di due docenti viene avviata un'inchiesta della Commissione per l'uguaglianza razziale del Governo britannico che condanna infine la Scuola per discriminazione nelle sue *policies* di ammissione: O. Schwartz, *Untold History of AI: Algorithmic Bias Was Born in the 1980's. A medical school thought a computer program would make the admission process fairer – but it did just the opposite*, IEEE Spectrum, 15 April 2019, <https://spectrum.ieee.org/untold-history-of-ai-the-birth-of-machine-bias>.
[59] Barocas, Selbst, *op. cit.*, 728.
[60] La correzione a fini distributivi è problematica quando l'algoritmo viene usato nei rapporti tra privati: A.F. Fondrieschi, *op. cit.*, 469.