Riflessioni sull'evoluzione e sulla rilevanza del concetto di trasparenza algoritmica, dal GDPR all'Artificial Intelligence Act*

Alessandro Oldani

Sommario

1. La trasparenza algoritmica nel Regolamento generale sulla protezione dei dati (UE) 2016/679

Nel contesto del regolamento (UE) 2016/679 (GDPR), la trasparenza algoritmica può essere intesa come il diritto, riconosciuto agli interessati, di ottenere informazioni chiare e comprensibili sia sulle logiche alla base dei trattamenti automatizzati finalizzati a supportare o automatizzare decisioni, sia sulle motivazioni che giustificano l'esito finale di tali decisioni. Specularmente, ciò comporta anche l'obbligo - gravante sui soggetti chiamati a prendere le decisioni - di fornire ai diretti interessati (*data subjects),* le suddette informazioni[1]. La base giuridica di tale diritto è l'art. 22 GDPR, applicabile nei casi in cui l'individuo sia sottoposto a una decisione basata

* L'articolo è stato sottoposto, in conformità al regolamento della Rivista, a referaggio

unicamente sul trattamento automatizzato, compresa la profilazione, che «produca effetti giuridici che lo riguardano o che incida in modo analogo significativamente sulla sua persona»[2]. Il concetto di trasparenza algoritmica integra la tutela accordata dal GDPR al soggetto interessato, specificando – oltre al menzionato divieto di sottoporre individui a decisioni interamente automatizzate – il diritto riconosciuto a quest'ultimo ad essere informato, in termini comprensibili, circa le modalità logico/operative dell'algoritmo utilizzato, nonché la misura in cui l'*output* dello stesso abbia inciso sulla decisione finale. Al par. 2 del medesimo articolo sono altresì previste alcune eccezioni[3]. Di particolare interesse per le tematiche trattate in questa sede è la possibilità di instaurare procedimenti decisionali interamente automatizzati previo consenso dell'interessato, in cui emerge un'applicazione pratica del concetto di trasparenza algoritmica. La capacità del consenso – inteso come base giuridica legittimante il trattamento dei dati personali - di garantire un'adeguata protezione dell'individuo nei procedimenti decisionali automatizzati, infatti, è subordinata alla corretta formazione dello stesso, possibile soltanto – nel caso di specie – attraverso opportune informazioni trasmesse in termini chiari e comprensibili circa le modalità logico/operative dell'algoritmo.

Sul piano teorico, la proibizione generale del trattamento automatizzato senza intervento umano riflette la concezione giuridica europea delle nuove tecnologie e delle relative applicazioni, secondo cui l'intelligenza artificiale deve servire da strumento per l'attività umana, piuttosto che sostituirsi integralmente ad essa. Tale visione è stata definitivamente formalizzata nel c.d. AI Act, regolamento europeo sull'intelligenza artificiale, la cui proposta è stata avanzata dalla Commissione nel 2021 e il cui testo definitivo è stato approvato dal Parlamento europeo il 14 marzo 2024[4], del quale si tratterà più diffusamente in seguito, con particolare attenzione ai principi di proporzionalità, trasparenza e responsabilità che ne costituiscono l'impianto assiologico e operativo. Un accenno, seppur indiretto, al concetto di trasparenza algoritmica è infatti riscontrabile nell'art. 15, par. 1, lett. h) GDPR, in cui è sancito il diritto dell'interessato ad ottenere dal titolare del trattamento informazioni relative all'esistenza di un processo decisionale automatizzato, unitamente ad informazioni significative in merito allo stesso procedimento. La natura di tali informazioni significative è meglio

anonimo.

[1] P. Zuddas, *Brevi note sulla trasparenza algoritmica*, in *Amministrazione in Cammino*, 2, 2020, 1 ss.

[2] Regolamento (UE) 2016/679 (GDPR), art. 22.

[3] S. Borghi, *Decisioni automatizzate e profilazione: le indicazioni dei Garanti europei*, in *privacy.it*, 2 novembre 2017.

[4] Parlamento europeo, Risoluzione legislativa del 13 marzo 2024 sulla proposta di regolamento sull'intelligenza artificiale (COM(2021)0206 – C9-0146/2021 – 2021/0106(COD)).

specificata negli artt. 13, par. 2, lett. f) e 14, par. 2, lett. g), GDPR[5], nei quali - oltre all'obbligo in capo al titolare del trattamento di rendere sempre edotto il soggetto interessato qualora vi fosse un trattamento decisionale automatizzato nei suoi confronti – viene specificato che tali informazioni vertono «sulla logica utilizzata, nonché l'importanza e le conseguenze previste di tale trattamento per l'interessato».

2. Big Data, trasparenza algoritmica e i rischi per l'autodeterminazione dell'individuo

Il tema della trasparenza algoritmica ha acquisito particolare rilevanza nell'utilizzo dei Big Data in ambito politico-elettorale[6]. L'utilizzo di algoritmi sempre più "opachi", in grado di "modellare" la personalità e le preferenze di un individuo – e, attraverso le c.d. *echo chambers*[7], indirizzarne le opinioni politiche – si pone in evidente contrasto con il diritto alla spiegabilità e alla conoscenza dei processi automatizzati.
Il considerando 63 GDPR è chiaro nel sottolineare il diritto dell'individuo di comprendere le modalità con le quali vengono prese le decisioni basate sui dati personali e sulla profilazione, ed in particolar modo a "comprenderne la logica". Questo non significa divulgare lo pseudocodice[8] o

[5] S. Stefanelli, *La trasparenza dell'algoritmo base della società del futuro: ecco perché è così importante*, in *agendadigitale.eu*, 3 maggio 2023.

[6] Per una più chiara comprensione della rilevanza economica, strategica e politica connessa all'acquisizione dei dati, cfr. G. De Ruvo, *Raccolta dati, intelligenza artificiale e sicurezza nazionale: l'uso geopolitico degli strumenti giuridici americani come freno alla data governance globale. Il caso TikTok come paradigma*, in *Rivista Italiana di Informatica e Diritto. Periodico Internazionale del CNR – IGSG*, 1, 2022 e S. Piva, *Forze politiche e garanzie giuridiche: spunti di riflessione sulla pronuncia del Garante dei dati personali in tema di piattaforma Rousseau*, in *dirittifondamentali.it*, 3, 2021, 447 ss.

[7] L'essere umano tende fisiologicamente a sviluppare *bias* cognitivi di conferma, ossia la propensione a ricercare informazioni che avvalorino convinzioni già maturate, evitando invece opinioni o dati contrari suscettibili di confutarle. Una volta profilato, l'algoritmo sotteso ai social network può proporre all'individuo contenuti coerenti con le preferenze politiche e valoriali manifestate in precedenza, riducendo così lo spazio per interpretazioni alternative della realtà. Tale dinamica, combinata ai preesistenti *bias* cognitivi della mente umana, rischia di favorire processi di radicalizzazione o, come emblematicamente evidenziato dal caso *Cambridge Analytica*, di consentire a determinati attori politici di orientare le tendenze individuali e di modellare conseguentemente la linea politica. Sul punto, M. Petrocelli, *Incoscienza digitale. La risposta alla rivoluzione digitale tra innovazione, sorveglianza e postdemocrazia*, Roma, 2022.

[8] "Pseudocode is a description of what a computer does, expressed in natural language instead of programming language». E. Ekohariadi - Y. Anistyasari - R. Putra - I. Kurniawan, Assessing Computational Thinking Using Pseudocode Programming Instrument, in Proceedings of the 4th National Seminar on Applied Technology, Science, and Arts (APTEKINDO 2018), 2018, 134 ss.

la *flowchart*[9] alla base degli algoritmi – i quali risulterebbero comunque incomprensibili per la maggior parte degli utenti - bensì mettere l'utente medio nelle condizioni di svolgere le proprie attività online avendo il più possibile contezza del valore commerciale e politico derivante dalla sua profilazione, affinché possa prestarvi o meno un autonomo ed indipendente consenso. La Cassazione si è recentemente espressa circa la corretta formazione del consenso dell'utente nel caso sia profilato o sottoposto a procedimento decisionale automatizzato. Ad integrare i presupposti del libero e specifico consenso, infatti:

«è richiesto che l'aspirante associato sia in grado di conoscere l'algoritmo, inteso come procedimento affidabile per ottenere un certo risultato o risolvere un certo problema, che venga descritto all'utente in modo non ambiguo ed in maniera dettagliata, come capace di condurre al risultato in un tempo finito. Che, poi, il procedimento, come spiegato con i termini della lingua comune, sia altresì idoneo ad essere tradotto in linguaggio matematico è tanto necessario e certo, quanto irrilevante: ed invero, non è richiesto né che tale linguaggio matematico sia esteso agli utenti, né, tanto meno, che essi lo comprendano»[10].

3. L'avvento dell'intelligenza artificiale generativa, l'evoluzione delle modalità di trattamento dei dati personali e la risposta della Commissione europea

Nel febbraio 2020 la Commissione europea ha rilasciato una comunicazione inerente alla propria Data Strategy[11],congiuntamente al White Paper on Artificial Intelligence[12]. Si edificavano così le fondamenta per una regolazione comunitaria dei servizi digitali e delle nuove tecnologie. In seguito a tali iniziative, nel 2022 è stato adottato il Digital Services Act (DSA)[13],

[9] Un *flowchart* è definito come una metodologia grafica per presentare algoritmi, utilizzando simboli standardizzati per rappresentare operazioni, decisioni e flussi di controllo. Viene utilizzato per visualizzare la sequenza logica di un algoritmo, facilitando la comprensione e la comunicazione tra programmatori e analisti. Per un approfondimento, si veda N. Ensmenger, *The Multiple Meanings of a Flowchart*, in *Information & Culture*, 3, 2016, 321 ss.

[10] Cass. civ., sez. I, 10 ottobre 2023, n. 28358, Rv. 669166-01.

[11] Commissione europea, Comunicazione *A European strategy for data*, COM(2020) 66 final, 2020.

[12] Commissione europea, *White Paper on Artificial Intelligence – A European approach to excellence and trust*, COM(2020) 65 final, 19 febbraio 2020.

[13] Regolamento (UE) 2022/2065 del Parlamento europeo e del Consiglio del 19 ottobre 2022 relativo a un mercato unico per i servizi digitali (Digital Services Act), in *GUUE*, L 277, 27 ottobre 2022.

il quale stabilisce regole specifiche per i servizi digitali e le piattaforme online, con l'obiettivo di regolamentare la responsabilità degli intermediari online, la trasparenza e la moderazione dei contenuti, nonché la protezione dei consumatori e la concorrenza equa nel mercato digitale. Nel medesimo periodo, il 21 aprile 2021, la Commissione europea ha presentato la proposta dell'Artificial Intelligence Act ("AI Act", già menzionato nel par. 1, che mira a regolamentare l'uso e lo sviluppo dell'IA nell'Unione europea. L'AI Act stabilisce norme specifiche per garantire la sicurezza, la trasparenza e l'etica nell'uso dell'IA, con la finalità di creare un quadro normativo completo per il mercato digitale e l'uso responsabile delle tecnologie digitali nell'Unione[14]. In risposta a queste sfide emergenti, l'AI Act rappresenta un passo fondamentale per anticipare e indirizzare le implicazioni legali e sociali di tali tecnologie, cercando di colmare il divario tra progresso tecnologico e regolamentazione attuale. L'evoluzione tecnologica, infatti, spesso supera la capacità del diritto di tenerne il passo, come dimostra l'avvento dell'intelligenza artificiale generativa e l'evoluzione del *machine learning*. Le reti neurali generative sono un tipo di reti neurali artificiali utilizzate per generare nuovi dati realistici, come immagini, suoni e testi: si tratta quindi di una tecnologia che ha applicazioni nell'arte generativa, nella grafica computerizzata, nella sintesi vocale e nella creazione di contenuti multimediali. La precisione di questi modelli è ormai migliorata al punto tale da costituire – nel caso di uso scorretto e in mala fede – un serio pericolo per la libertà, la riservatezza e la dignità dell'individuo[15].

---

[14] Come si evince dal considerando 1 del Regolamento, «*The purpose of this Regulation is to improve the functioning of the internal market by laying down a uniform legal framework, in particular for the development, the placing on the market, the putting into service and the use of artificial intelligence systems (AI systems) in the Union, in accordance with Union values, to promote the uptake of human-centric and trustworthy artificial intelligence (AI), while ensuring a high level of protection of health, safety and fundamental rights, as enshrined in the Charter of Fundamental Rights of the European Union («the Charter»), including democracy, the rule of law and environmental protection, against the harmful effects of AI systems in the Union, and to support innovation. This Regulation ensures the free cross-border movement of AI-based goods and services, thus preventing Member States from imposing restrictions on the development, marketing and use of AI systems, unless explicitly authorised by this Regulation*»; regolamento (UE) 2024/1689 (AI Act), considerando 1.

[15] Il caso probabilmente più noto è rappresentato dai *deepfakes*, contenuti multimediali manipolati in modo intelligente per apparire autentici, pur essendo generati artificialmente. Tali tecniche, basate su reti neurali generative (*Generative Adversarial Networks* – GANs), consentono di creare immagini, video o tracce audio in cui un soggetto sembra dire o fare qualcosa che in realtà non ha mai detto o fatto. La tecnologia dei *deepfakes* presenta applicazioni potenzialmente legittime, ad esempio nei settori cinematografico e pubblicitario, ma comporta altresì gravi rischi di manipolazione, frode, diffamazione e disinformazione, incidendo sulla fiducia pubblica nelle fonti digitali e sull'integrità dell'informazione. Per un approfondimento sul fenomeno dei *deepfakes* e sulle sue implicazioni etiche, economiche e regolatorie, si veda J. Kietzmann - L. W. Lee - I. P. McCarthy - T. C. Kietzmann, *Deepfakes: Trick or Treat?*, in *Business Horizons*, 2, 2020, 135 ss. e Furizal -

Le potenzialità e gli ambiti applicativi sono stati compresi solo a posteriori, e ciò ha reso difficile per il legislatore europeo prevedere le conseguenze di tali tecnologie sulle piattaforme online. In effetti, l'applicabilità del Digital Services Act stesso a determinate fattispecie è messa in discussione dalle potenzialità di queste tecnologie[16].

Tralasciando la tempistica e la relativa adeguatezza della produzione normativa nella regolamentazione di tali fenomeni in costante e rapida evoluzione, nell'AI Act il legislatore europeo ha adottato un approccio "basato sul rischio", il medesimo già utilizzato per il GDPR, seppur con le dovute peculiarità e differenze. Il GDPR mostra una particolare responsabilizzazione del titolare del trattamento, il quale – considerando anche le importanti sanzioni previste - è sensibilizzato e indotto a considerare i reali rischi nel trattare dati personali. Il principio della *privacy by design* - consistente nell'approccio alla progettazione di sistemi e processi che incorpora la privacy e la protezione dei dati personali fin dalla progettazione - implica infatti una certa proattività da parte del titolare nel trattamento nella predisposizione di misure tecniche ed organizzative che contemperino le dovute garanzie e nel rispetto dei principi fondamentali nel trattamento dei dati personali di cui all'art. 5 GDPR[17]. Ciò implica un processo continuo di valutazione e monitoraggio del rischio per ogni attività di trattamento; solo quando ricorrono le condizioni che comportano un rischio elevato per i diritti e le libertà degli interessati si attiva la procedura prevista dall'art. 35 GDPR, che richiede la redazione di una valutazione d'impatto sulla protezione dei dati personali (DPIA)[18]. Il fulcro del principio di responsabiliz-

---

A. Ma'arif - H. Maghfiroh - I. Suwarno - D. Prayogi - Kariyamin - S. Lonang - A.-N. Sharkawy, *Social, legal, and ethical implications of AI-Generated deepfake pornography on digital platforms: A systematic literature review*, in *Social Sciences & Humanities Open*, 2025, 101882.
Un esempio di come un tale fenomeno possa avere impatti diretti e riflessi significativi sui diritti fondamentali e sul funzionamento della giustizia penale è offerto dallo studio sistematico condotto da M. Sandoval - M. De Almeida Vau - J. Solaas - L. Rodrigues, *Threat of Deepfakes to the Criminal Justice System: A Systematic Review*, in *Crime Science*, 2024. Gli autori evidenziano come i *deepfakes* possano compromettere la genuinità delle prove digitali, minare l'affidabilità dei processi decisionali e incidere sulla corretta amministrazione della giustizia, sollevando questioni cruciali in materia di veridicità probatoria e tutela delle garanzie processuali.

[16] M. Bassini, *Intelligenza artificiale generativa: alcune questioni problematiche*, in questa *Rivista*, 2, 2023, 391 ss.

[17] Si fa riferimento ai seguenti principi: liceità, correttezza e trasparenza; limitazione della finalità; minimizzazione dei dati; esattezza; limitazione della conservazione; integrità e riservatezza; regolamento (UE) 2016/679 *(General Data Protection Regulation),* art. 5.

[18] La valutazione d'impatto riveste un ruolo cruciale con riferimento ai *Big Data*: nel merito, cfr. S. Gobbato, *«Big Data» e «tutele convergenti» tra concorrenza, GDPR (General Data Protection Regulation – Regolamento generale sulla protezione dei dati) e Codice del consumo*, in questa *Rivista*, 3, 2019, 148 ss., e C. Coniglione, *L'utilizzo dei «Big Data» in ambito politico-elettorale e il loro impatto*

zazione risiede infatti nell'obbligo del titolare della corretta identificazione del rischio e delle conseguenti azioni per "mitigare" gli effetti compromettenti la *privacy* dell'individuo. La responsabilizzazione salvaguarda di fatto la produzione normativa da una precoce obsolescenza, in quanto pone a carico del titolare del trattamento l'attenta valutazione di ogni ipotetica situazione di rischio, la quale non può prescindere da un'analisi includente le modalità tecniche di processamento e conservazione dei dati, con annesse misure di sicurezza.
È quindi agevole rintracciare questo impianto giuridico nell'Artificial Intelligence Act, che attribuisce la responsabilità in prima battuta al fornitore dei sistemi di intelligenza artificiale. Il Regolamento prevede infatti una specifica classificazione dei sistemi in relazione al livello di rischio. Ciò significa che, per i c.d. "Sistemi ad Alto Rischio", il fornitore è tenuto all'adempimento di specifici obblighi, tra i quali la redazione di una documentazione di tipo tecnico completa ed informata, in grado di dimostrare il rispetto della normativa, unita ad implementazioni di sistemi che consentano una realistica valutazione del rischio derivanti dall'impiego dell'intelligenza artificiale. Si evince da ciò una prima e vera "istituzionalizzazione" del principio di trasparenza algoritmica, sulla cui traduzione legislativa ci si concentrerà nelle pagine che seguono.

4. La tutela dei diritti fondamentali nell'AI Act. Introduzione alla classificazione in base al rischio dei sistemi di intelligenza artificiale

Nel contesto dell'AI Act, l'approccio basato sul rischio - oltre alla responsabilizzazione del fornitore – include la classificazione dei sistemi di intelligenza artificiale in relazione alla potenzialità lesiva dei diritti fondamentali, tenendo in considerazione sia l'eventualità in cui il rischio abbia origine da un malfunzionamento del sistema – come nel caso dei sistemi utilizzati come componenti di sicurezza – oppure dall'uso in mala fede degli stessi, come ad esempio il furto dell'identità digitale di un individuo mediante il campionamento e la sintesi realistica della sua voce. Diversi sono pertanto gli obblighi e le garanzie del fornitore e degli utenti all'aumentare o diminuire del livello di rischio. Il Regolamento classifica i sistemi di AI in quattro fasce di rischio — inaccettabile, alto, limitato e minimo — attribuendo obblighi progressivamente più stringenti al crescere del pericolo. In particolare, il legislatore vieta gli impieghi reputati "inaccettabili", sottopone i sistemi "ad alto rischio" a un rigoroso regime di conformità, impone semplici obblighi di trasparenza per i casi di "rischio limitato" e non prevede

sulla democrazia rappresentativa, in *Nomos*, 1, 2023, 15 ss.

vincoli per le applicazioni a "rischio minimo". Sebbene il Regolamento sia già entrato in vigore, la maggior parte delle sue disposizioni si applicherà a tappe tra il 2025 e il 2026. La Commissione ha deciso di concedere un periodo relativamente lungo per permettere alle imprese e agli altri soggetti interessati di adeguarsi ai nuovi obblighi. Le normative riguardanti i sistemi AI proibiti saranno applicate dopo 6 mesi; le regole relative alle GPAI (*General Purpose AI* o intelligenza artificiale generale)[19] entreranno in vigore dopo 12 mesi, eccetto per i modelli di GPAI già disponibili sul mercato prima di tale scadenza, per i quali il termine è di 24 mesi; le disposizioni per i sistemi AI considerati ad alto rischio saranno effettive dopo 36 mesi[20]. Infine, i sistemi AI ad alto rischio già utilizzati dalle autorità pubbliche al momento dell'entrata in vigore dell'AI Act beneficeranno di un periodo di tolleranza di 48 mesi.

Passando al merito, va innanzitutto ribadito che l'inaccettabilità di alcuni sistemi implica il divieto assoluto di commercializzazione ed implementazione nella società. Dal punto di vista giuridico-costituzionale sorge dunque un primo spunto di ragionamento, circa la precisa identificazione del discrimine – talvolta sottile e delicato – in base al quale un sistema di AI viene inserito nel rischio inaccettabile piuttosto che nei sistemi ad alto rischio. Pur essendo la questione regolata nel Titolo II del Regolamento, già nella relazione introduttiva della proposta si sottolineava che l'elenco delle pratiche vietate comprende tutti i sistemi di IA il cui uso è considerato inaccettabile in quanto contrario ai valori dell'Unione[21]. Questo approccio è finalizzato a proibire pratiche contrastanti i suddetti valori, relativi al rispetto della dignità umana, della libertà, dell'uguaglianza, della democrazia e dello Stato di diritto e dei diritti fondamentali dell'Unione, compresi il

[19] Per *GPAI* si intendono quei sistemi di intelligenza artificiale aventi finalità e scopi di tipo generale, progettati per l'esecuzione di una vasta gamma di compiti anziché essere limitati a scopi specifici. Gran parte dei modelli di intelligenza artificiale generativa usufruibili dagli utenti sul web rientrano in questa categoria. Per interessanti considerazioni sulla definizione di *General Purpose Artificial Intelligence Systems* (*GPAI*), si veda General Purpose Artificial Intelligence Systems (GPAI), si veda C. I. Gutierrez - A. Aguirre - R. Uuk et al., *A Proposal for a Definition of General Purpose Artificial Intelligence Systems*, in *Digital Society*, 2, 2023, 36.

[20] Parlamento europeo, *Il Parlamento europeo approva la legge sull'intelligenza artificiale*, comunicato stampa, 13 marzo 2024, disponibile su europarl.europa.eu.

[21] Nei considerando del testo definitivo approvato dal Parlamento europeo si trovano riferimenti espliciti e dettagliati ai sistemi di intelligenza artificiale vietati dall'Unione. In particolare, il considerando 26 recita: «Per introdurre un insieme proporzionato ed efficace di norme vincolanti per i sistemi di IA, si dovrebbe seguire un approccio chiaramente definito basato sul rischio. Tale approccio dovrebbe adattare il tipo e il contenuto di tali norme all'intensità e alla portata dei rischi che i sistemi di IA possono generare. È quindi necessario vietare alcune pratiche di IA inaccettabili, stabilendo requisiti per i sistemi di IA ad alto rischio e obblighi per i relativi operatori, e stabilire obblighi di trasparenza per alcuni sistemi di IA»; regolamento (UE) 2024/1689 (AI Act), considerando 26.

diritto alla non discriminazione[22], alla protezione dei dati[23] e della vita privata e i diritti dei minori[24].

Alla base è dunque riscontrabile un'analisi "costi-benefici", un'attenta valutazione analitica del progetto, dove il "costo" è rappresentato dal rischio delle ipotetiche situazioni lesive dei diritti fondamentali che la commercializzazione e l'uso di un determinato sistema di intelligenza artificiale può comportare. Non a caso, nella relazione della proposta, − Sezione 1.1, Motivi e obiettivi della proposta − l'intelligenza artificiale è definita come «una famiglia di tecnologie in rapida evoluzione in grado di apportare una vasta gamma di benefici economici e sociali in tutto lo spettro delle attività industriali e sociali»; immediatamente dopo segue il passaggio chiave giustificante l'approccio basato sui livelli di rischio: «[t]uttavia gli stessi elementi e le stesse tecniche che alimentano i benefici socio-economici dell'IA possono altresì comportare nuovi rischi o conseguenze negative per le persone fisiche o la società». La definizione e le considerazioni in merito all'intelligenza artificiale sono state poi approfondite nei considerando 4 e 5 del testo definitivo, approvato dal Parlamento europeo il 14 marzo del 2024. L'IA è infatti considerata come un importantissimo vantaggio competitivo per le imprese, nell'ottimizzazione dei processi e nell'allocazione delle risorse. Allo stesso tempo, a seconda delle circostan-

[22] Si pensi ai potenziali rischi derivanti da eventuali *bias* algoritmici dei sistemi di intelligenza artificiale implementati nel sistema giudiziario. Sebbene l'automatizzazione di determinati procedimenti possa agevolare lo svolgimento del processo, resta comunque fondamentale un'attenta analisi dell'impatto che tali implementazioni possono avere sul giusto processo e sui diritti delle parti. Sul tema, cfr. P. R. Borges Fortes, *Paths to Digital Justice: Judicial Robots, Algorithmic Decision-Making, and Due Process*, in *Asian Journal of Law and Society*, 3, 2020, 453 ss. e A. Lorusso, *Algoritmo, provvedimento amministrativo e autotutela*, in questa *Rivista*, 1, 2023, 326 ss. È altresì evidente — considerando le attuali tecnologie — l'impossibilità di concepire un sistema di intelligenza artificiale operante in un'ottica integralmente sostitutiva della figura del giudice. Tuttavia, le nuove tecnologie potrebbero incidere profondamente su alcuni aspetti del diritto, sia sostanziale sia processuale, tra cui la determinazione dell'elemento soggettivo del reato, della prova scientifica, degli *standard* probatori e del rischio di recidiva. Quest'ultimo aspetto è approfondito in G. Tuzet, *L'algoritmo come pastore del giudice? Diritto, tecnologie, prova scientifica*, in questa *Rivista*, 1, 2020, 45 ss.

[23] Già nel 2018 la Norwegian Data Protection Authority analizzava, nel proprio report *Artificial Intelligence and Privacy*, i rischi connessi all'applicazione dell'intelligenza artificiale al trattamento e all'acquisizione di dati personali, anticipando molte delle preoccupazioni oggi affrontate dall'Artificial Intelligence Act. Per ulteriori approfondimenti, cfr. Norwegian Data Protection Authority, *Artificial Intelligence and Privacy*, in *datatilsynet.no*, 22 giugno 2018.

[24] Si tratta di una categoria particolarmente vulnerabile sul web e sui social network. L'intelligenza artificiale rappresenta, tuttavia, una grande opportunità per automatizzare la moderazione dei contenuti sulle piattaforme online. Per approfondire l'argomento e i rischi implicati, cfr. P. Dunn, *Moderazione automatizzata e discriminazione algoritmica: il caso dell'hate speech*, in *Rivista italiana di informatica e diritto*, 1, 2022, 133 ss.

ze relative alla sua specifica applicazione, sono presi seriamente in considerazione i potenziali pericoli concernenti gli interessi pubblici e i diritti fondamentali tutelati dall'Unione, includendo danni materiali, immateriali, fisici, psicologici, sociali ed economici.
A scanso di equivoci, va subito segnalato che la stragrande maggioranza degli attuali sistemi di intelligenza artificiale è riconducibile alle categorie del rischio limitato e nel rischio moderato. Per questi ultimi è previsto solamente un regime minimo "di trasparenza": ad esempio, applicazioni come *Chatbot*[25] dovrebbero riconoscere all'utente la possibilità di prendere delle decisioni informate. Fatta questa doverosa premessa e prima di trattare nel merito il tema della trasparenza algoritmica, è necessario fornire un contesto sul fulcro della tutela normativa in oggetto, ovvero la regolamentazione dei c.d. sistemi "ad alto rischio".

4.1 Una breve panoramica sui sistemi "ad alto rischio"

A differenza dei sistemi a "rischio inaccettabile", i sistemi di intelligenza artificiale ad alto rischio possono essere immessi sul mercato o messi in servizio, purché rispettino i requisiti obbligatori specificati nel Regolamento. Il fornitore è soggetto a specifici adempimenti, inerenti ad obblighi di documentazione, tracciabilità e di trasparenza, nonché ad ulteriori criteri riguardanti l'interfaccia e la sorveglianza umana, la precisione e la "robustezza" o sicurezza del sistema[26].

[25] «Sono macchine. O meglio, *chatbot*: programmi per computer in grado di "apprendere" e sviluppare una propria intelligenza, mentre dialogano con gli esseri umani. Riuscendo, almeno per qualche minuto, a spacciarsi per gente in carne e ossa. Memorizzano, cioè, parole e frasi di ogni conversazione per farne uso nelle conversazioni successive, verificandone l'effetto e affinandone di passo in passo la correttezza. Imparando dagli errori. Sono i prototipi dei personaggi virtuali che presto ci faranno da ciceroni in musei e grandi magazzini, serviranno da alter ego alle celebrità o terranno compagnia a chi è troppo solo.» P. Accolla, *La Repubblica, Affari & Finanza*, 25 settembre 2006, 25.

[26] In un'ottica di confronto, l'art. 32 del regolamento (UE) 2016/679 (*General Data Protection Regulation – GDPR*) stabilisce gli obblighi di sicurezza che i titolari e i responsabili del trattamento devono rispettare per garantire un livello di protezione adeguato al rischio connesso ai dati personali trattati. Tali obblighi comprendono la pseudonimizzazione e la cifratura dei dati personali, la capacità di assicurare la riservatezza, l'integrità, la disponibilità e la resilienza dei sistemi e dei servizi di trattamento, la possibilità di ripristinare tempestivamente l'accesso ai dati in caso di incidente fisico o tecnico, nonché l'implementazione di un processo di verifica e valutazione periodica dell'efficacia delle misure adottate. Inoltre, il titolare e il responsabile del trattamento devono tener conto della natura, dell'ambito, del contesto e delle finalità del trattamento (regolamento (UE) 2016/679, art. 32). Sul tema, cfr. J. S. Vanegas, *La violazione dei requisiti di sicurezza informatica di cui all'art. 32 del GDPR*, in *Rivista italiana di informatica e diritto*, 2, 2020, 5 ss. Nel regolamento (UE) 2024/1689 (AI Act), l'art. 15 introduce una prospettiva complementare, definendo

Per stabilire con esattezza quando un sistema di intelligenza artificiale debba essere classificato "ad alto rischio" ai sensi dell'AI Act occorre partire dall'art. 6, par. 2, che subordina tale qualifica alla compresenza di due elementi: da un lato il sistema deve costituire o integrare un "componente di sicurezza" di un prodotto — oppure presentarsi esso stesso come prodotto — disciplinato da una delle normative di armonizzazione dell'Unione richiamate nell'Allegato I; dall'altro lo stesso prodotto, o la IA venduta come prodotto autonomo, deve superare una valutazione di conformità svolta da un organismo terzo prima della sua immissione sul mercato o messa in servizio. Ne sono esempio i sensori in grado di rilevare la presenza umana o i pulsanti di arresto d'emergenza governati da logiche di IA. La nozione di "componente di sicurezza" è definita dall'art. 2, par. 3, della direttiva 2006/42/CE come quell'elemento commercializzato separatamente, destinato a una funzione protettiva, il cui guasto può compromettere la sicurezza delle persone e che non è essenziale al funzionamento primario della macchina, potendo essere sostituito da componenti equivalenti. A questi criteri l'art. 6 affianca il rinvio all'Allegato III, che individua otto ambiti di utilizzo particolarmente sensibili — biometria e identificazione, infrastrutture critiche, istruzione e formazione, occupazione e gestione delle risorse umane, accesso a servizi essenziali, applicazione della legge, migrazione e controllo delle frontiere, giustizia e processi democratici — all'interno dei quali un sistema di IA è qualificato "ad alto rischio" a prescindere dal collegamento con un prodotto fisico, salvo prova dell'assenza di un rischio significativo per la salute, la sicurezza o i diritti fondamentali. Più nel dettaglio, rientrano tra le categorie di cui all'Allegato III i sistemi di categorizzazione e identificazione biometrica delle persone fisiche; i sistemi utilizzati come componenti di sicurezza nella gestione del traffico stradale e nella fornitura di servizi essenziali; quelli impiegati per determinare l'accesso o l'assegnazione a istituti scolastici o percorsi di formazione e per valutare le prestazioni degli studenti; le soluzioni adottate nei processi di assunzione o selezione del personale; i sistemi utilizzati da autorità pubbliche per valutare l'ammissibilità a prestazioni o servizi di assistenza; nonché quelli impiegati nella gestione della migrazione, dell'asilo e del controllo delle frontiere. Analizzando tale classificazione, molto spesso si identifica un elevato livello di rischio in sistemi aventi funzionalità valutative degli in-

requisiti di *cybersecurity* e di robustezza per i sistemi di intelligenza artificiale, in particolare quelli ad alto rischio. Il legislatore europeo sottolinea che «la robustezza dei sistemi di IA ad alto rischio può essere conseguita mediante soluzioni tecniche di ridondanza, che possono includere piani di *backup* o *fail-safe*» (regolamento (UE) 2024/1689 (AI Act), art. 15). Si delinea, dunque, un duplice livello di garanzia: da un lato, la tutela della sicurezza dei dati personali, che costituisce il fulcro dell'art. 32 GDPR; dall'altro, la sicurezza intrinseca e la resilienza operativa dei sistemi di intelligenza artificiale, vero e proprio pilastro tecnico-normativo dell'*AI Act*.

dividui[27] riguardo a temi più o meno sensibili[28]. L'oggetto della valutazione è infatti correlato ai diritti fondamentali dell'uomo – prima fra tutti la dignità umana, sancita dall'art. 1 della Carta dei Diritti Fondamentali dell'Unione europea – ed ogni restrizione imposta dalla normativa è a garanzia degli stessi. Rispetto alla proposta, il testo definitivo del Regolamento - all'art. 6, par. 3 - specifica che un sistema di IA non può essere considerato ad alto rischio se non presenta un rischio significativo di danno alla salute, alla sicurezza o ai diritti fondamentali delle persone fisiche, anche non influenzando materialmente l'esito del processo decisionale.
Approfondendo opportunamente le considerazioni di cui sopra, se il sistema di IA è progettato per svolgere compiti procedurali molto limitati[29], non viene classificato come ad alto rischio. Questo tipo di IA ha scopi ben definiti e non è destinato a eseguire valutazioni ampie o critiche. Un altro caso in cui un sistema di IA non è visto come ad alto rischio riguarda il suo uso per rilevare modelli decisionali o deviazioni da questi[30]. Tali sistemi

[27] Così come esistono i *bias* cognitivi propri dell'essere umano, per implicazione logica si pone anche il problema del *bias* algoritmico, poiché gli algoritmi vengono addestrati su dati che riflettono l'interpretazione umana della realtà. Per un maggiore approfondimento sull'elevato rischio di discriminazione, cfr. S. Tommasi, *Algoritmi e nuove forme di discriminazione: uno sguardo al diritto europeo*, in *Revista de Direito Brasileira*, 10, 2020, 112 ss.

[28] Basti pensare al potenziale impatto dei sistemi di riconoscimento facciale basati su intelligenza artificiale utilizzati dalle autorità pubbliche. Sul tema si è già espressa l'Alta Corte di giustizia nel caso *R (Bridges) v. The Chief Constable of South Wales Police*. Per un'analisi ampia e approfondita del caso, cfr. A. Pin, *There Is No «Silver Bullet»: The Artificial Face Recognition's First Judicial Scrutiny*, in *DPCE Online,* 4, 2019, 3075 ss.

[29] Si pensi, ad esempio, a un sistema di intelligenza artificiale utilizzato da un museo per fornire suggerimenti personalizzati ai visitatori in base alle opere già visualizzate o ai percorsi scelti all'interno dell'edificio. In tal caso, il sistema elabora informazioni di carattere contestuale, ma non adotta decisioni autonome né produce effetti giuridici o significativi sulla persona. Tale applicazione, avente finalità puramente informative e di supporto all'esperienza utente, non rientra tra i sistemi ad alto rischio ai sensi dell'art. 6 dell'*AI Act*, poiché non incide sui diritti fondamentali né comporta rischi di discriminazione o danni materiali. È tuttavia opportuno evidenziare che la valutazione del rischio varia in funzione del contesto e dell'utilizzo concreto del sistema di intelligenza artificiale, potendo mutare al variare delle finalità, del tipo di dati trattati o del grado di autonomia decisionale. Sulla gestione e valutazione del rischio nel quadro dell'*AI Act*, si veda J. Schuett, *Risk Management in the Artificial Intelligence Act*, in *European Journal of Risk Regulation*, 2, 2023, 367 ss.

[30] Un esempio di tale sistema è uno strumento di IA che, dato un certo modello di valutazione (*grading pattern*) di un insegnante, può essere utilizzato per verificare *ex post* se l'insegnante possa aver deviato da tale modello, segnalando potenziali incongruenze o anomalie. In questo caso, il sistema agisce come un meccanismo di *flagging* che non prende la decisione finale, ma fornisce *input* per una successiva revisione umana. Un'interessante prospettiva sull'utilizzo dell'intelligenza artificiale nel contesto educativo è offerta da C. Kooli e N. Yusuf, i quali analizzano come i *Large Language Models* (*LLM*), e in particolare *ChatGPT*, possano contribuire a trasformare le modalità di valutazione scolastica, migliorandone l'efficienza, la coerenza

analizzano le decisioni passate per identificare consistenze o discrepanze, ma non sostituiscono il giudizio umano. Sono strumenti di supporto che necessitano di una revisione umana per garantire che le valutazioni siano corrette e appropriate. Questi sistemi preparano il materiale informativo necessario per decisioni importanti, agendo come un passaggio preliminare nel processo decisionale più ampio.

Rispetto alla proposta è intercorsa dunque una notevole evoluzione circa le modalità di classificazione dei sistemi: l'alto rischio non dipende soltanto dall'ambito di applicazione del sistema, bensì dalla funzione decisiva dell'intelligenza artificiale nel processo decisionale e nella sua potenziale capacità di causare danni significativi autonomamente.

L'obiettivo del Regolamento rimane comunque la responsabilizzazione dei fornitori, ponendo a loro carico precisi obblighi e procedure nello sviluppo e nella commercializzazione di intelligenze artificiali. Va tuttavia rilevato che l'individuazione di ipotesi per le quali un fornitore possa essere ritenuto responsabile degli effettivi risultati ottenuti dall'IA è operazione indubbiamente complessa; ciò in quanto il medesimo fornitore può non avere contezza dello scopo effettivo dell'utilizzo, né essere in grado di fornire un ragionevole e provato riscontro degli effetti a lungo termine al momento dell'immissione[31]. Sul tema risulta cruciale la figura del *deployer,* che verrà debitamente approfondita nel paragrafo seguente.

Esposte – seppur sommariamente – le ragioni e i criteri alla base delle suddette classificazioni, è agevole comprendere che la previsione di "requisiti obbligatori" in capo ai fornitori[32] o *providers* nella fase di progettazione e sviluppo, uniti ad una stringente valutazione di conformità - da effettuare *ex ante,* prima della commercializzazione - rappresentano la tutela sostanziale del Regolamento.

---

e la trasparenza. Gli autori mostrano come l'IA possa essere impiegata per supportare la correzione automatica e l'analisi qualitativa delle prove, mantenendo un ruolo complementare rispetto al giudizio umano e promuovendo una didattica più equa e personalizzata. Per un approfondimento, si veda C. Kooli - N. Yusuf, *Transforming Educational Assessment: Insights into the Use of ChatGPT and Large Language Models in Grading*, in *International Journal of Human-Computer Interaction*, 2024, 3388 ss.

[31] M. R. Carbone, *AI Act. Come procede il regolamento UE sull'intelligenza artificiale: il nuovo testo*, in *agendadigitale.eu*, 2 settembre 2022.

[32] Nel contesto dell'*Artificial Intelligence Act*, il «fornitore» (*provider*) è definito come la persona fisica o giuridica, l'autorità pubblica, l'agenzia o altro organismo che sviluppa un sistema di intelligenza artificiale — o ne fa sviluppare uno — e lo immette sul mercato o lo mette in servizio sotto il proprio nome o marchio, sia a titolo oneroso sia gratuito (regolamento (UE) 2024/1689 (AI Act), art. 3, punto 3). Tale definizione non comprende automaticamente l'importatore né il distributore, individuati separatamente ai punti 6 e 7 dello stesso articolo, sebbene tali soggetti possano, in determinate circostanze, assumere alcuni obblighi propri del fornitore.

5. La tutela sostanziale dell'AI Act. Il ruolo dei *providers* e dei *deployers*

Prima di approfondire la rilevanza del concetto di trasparenza algoritmica nella compliance all'Artificial Intelligence Act, è opportuno fare chiarezza sui soggetti definiti dal Regolamento. Analogamente ai soggetti definiti dal GDPR, come ad esempio il "titolare del trattamento" ed il "responsabile del trattamento", è necessario evidenziare la differenza tra il *provider*, ossia il fornitore, e il *deployer*. Per fornitore – ai sensi dell'art. 3, par. 3 del Titolo I del Regolamento – si intende quell'entità che progetta, sviluppa o produce sistemi di AI per uso diretto o per distribuzione sul mercato europeo, a titolo gratuito o a pagamento. Il *deployer* è una persona fisica o giuridica, un'autorità pubblica, un'agenzia o un altro organismo che utilizza il sistema di IA sotto la propria autorità, ad eccezione del caso in cui il sistema venga utilizzato nel corso di un'attività personale non professionale. Il regolamento pone a carico del *deployer* specifiche responsabilità, che completano e specificano quelle dei fornitori. Già il considerando 91 dell'AI Act chiarisce che, per i sistemi di intelligenza artificiale classificati "ad alto rischio", la conformità deve estendersi oltre la fase di progettazione e includere l'adozione di misure tecniche e organizzative capaci, da un lato, di assicurare l'uso del sistema in linea con le istruzioni del fornitore e, dall'altro, di consentire un monitoraggio costante delle prestazioni lungo l'intero ciclo di vita. Questa esigenza si traduce, all'art. 12, nell'obbligo di registrare in modo automatico e sistematico i c.d. *log*, vale a dire i dati di tracciamento degli eventi rilevanti, i quali devono essere accurati, affidabili e prontamente disponibili a fini di *audit* e di indagine *post-market*. La continuità di tale tracciabilità garantisce che eventuali anomalie o incidenti possano essere identificati e gestiti anche dopo l'immissione del prodotto sul mercato, inserendosi nel più ampio sistema di gestione del rischio delineato dall'art. 9.

A ciò si affianca l'art. 26, che impone ai *deployer* di assicurare che il personale incaricato della supervisione e dell'utilizzo del sistema disponga di un adeguato livello di alfabetizzazione in materia di IA, nonché della formazione e dell'autorità necessarie per intervenire ove occorra. L'efficacia di tutte queste cautele risulterebbe, tuttavia, compromessa se il fornitore non rispettasse i doveri di trasparenza sanciti dall'art. 13: istruzioni d'uso chiare, complete e comprensibili rappresentano infatti la condizione imprescindibile affinché i log siano interpretati correttamente, il monitoraggio sia significativo e le competenze degli operatori possano dispiegarsi in modo effettivo. In sintesi, il Regolamento costruisce un circuito virtuoso tra tracciabilità tecnica, controllo umano informato e trasparenza documentale, destinato a preservare la sicurezza e l'affidabilità dei sistemi di IA ad alto rischio per tutta la loro vita operativa.

L'art. 27 dell'AI Act attribuisce ai *deployer* di sistemi di IA ad alto rischio l'obbligo di redigere, prima della messa in uso, una Fundamental Rights Impact Assessment (FRIA) volta a valutare le ripercussioni che l'impiego del sistema può produrre sui diritti fondamentali delle persone coinvolte. Tale valutazione deve descrivere in modo puntuale: i processi operativi in cui il sistema sarà impiegato, il periodo e la frequenza di utilizzo, le categorie di individui potenzialmente interessate, i rischi specifici connessi all'applicazione concreta, le misure di supervisione umana previste nonché le azioni da intraprendere qualora tali rischi dovessero materializzarsi. L'obbligo riguarda i soggetti di diritto pubblico e le entità private che erogano servizi di pubblico interesse, nonché i *deployer* dei sistemi rientranti nei punti 5-b e 5-c dell'Allegato III, con la sola esclusione dei sistemi destinati alle infrastrutture critiche di cui al punto 2 dello stesso allegato. Il legislatore prevede che i risultati della FRIA siano notificati all'autorità di sorveglianza del mercato utilizzando il modello standardizzato che sarà messo a disposizione dall'AI Office. Inoltre, qualora una parte delle analisi sia già coperta da una valutazione d'impatto sulla protezione dei dati (DPIA) ai sensi dell'art. 35 GDPR, la FRIA deve integrarla, evitando duplicazioni e garantendo coerenza metodologica. Questa previsione, perfettamente in linea con il principio di accountability mutuato dal GDPR, responsabilizza il *deployer* sin dalle prime fasi di progetto, imponendogli di prevenire e mitigare in modo ragionevole i potenziali pregiudizi[33] per la salute, la sicurezza e le libertà fondamentali connessi all'impiego del sistema di IA

I fornitori sono infatti tenuti a redigere una documentazione di tipo tecnico "completa ed informata", in grado di dimostrare il rispetto della normativa, unita ad implementazioni di sistemi che consentano la predisposizione di una realistica valutazione del rischio derivanti dall'impiego dell'intelligenza artificiale. Sono inoltre previsti obblighi di tracciabilità delle decisioni dei sistemi ad alto rischio, provvedendo a creare sistemi di registrazione automatica dei log, cioè una sorta di "scatola nera" funzionale alla ricostruzione accurata degli eventi caratterizzanti un determinato procedimento decisionale.

6. Il primo livello di trasparenza

Tali imposizioni sono strettamente collegate al principio di trasparenza, che vincola il fornitore a predisporre informazioni chiare, precise e com-

[33] Parallelismo altresì approfondito in A. Alaimo, *Il regolamento sull'intelligenza artificiale, dalla proposta della Commissione al testo approvato dal Parlamento. Ha ancora senso il pensiero pessimistico*, in *federalismi.it*, 25, 2023, 136 ss.

prensibili[34]. Nonostante l'assonanza semantica, va precisato che si tratta qui di un principio diverso da quello sancito dal GDPR in merito alla trasparenza del trattamento dei dati personali[35] , dato che nella normativa in esame si fa riferimento ad un concetto avanzato di trasparenza algoritmica, strettamente funzionale all'autonomia e al controllo sulle modalità operative dei sistemi di AI. Ciò non significa la divulgazione del codice sorgente, bensì «la riduzione/eliminazione dell'asimmetria informativa finalizzata ad acquisire la capacità di descrivere, ispezionare e riprodurre correttamente i meccanismi attraverso cui i sistemi di intelligenza artificiale assumono decisioni ed imparano ad adattarsi ai loro ambienti»[36]. Comprendere a fondo la logica decisionale che governa l'algoritmo—nonché le modalità con cui il flusso di dati viene raccolto, preprocessato e alimentato nel modello—è condizione indispensabile per individuare l'origine di eventuali malfunzionamenti, ricostruirne la ratio causale e interrompere il rischio di reiterazione dell'errore. In questa prospettiva, l'art. 13 dell'AI Act impone al fornitore un livello di trasparenza tale da mettere il *deployer* in grado di decodificare le dinamiche interne del sistema, interpretarne correttamente gli output e applicare le necessarie misure correttive; solo così l'utilizzatore può monitorare il funzionamento dell'IA in modo informato, prevenendo derive sistemiche e garantendo che l'impiego del modello resti coerente con i requisiti di affidabilità e tutela dei diritti fondamentali che permeano l'intero impianto regolatorio. Fondamentale risulta il par. 3 di tale art., contenente un'accurata descrizione del contenuto delle istruzioni per l'uso che il fornitore deve fornire al *deployer*. In primo luogo, è necessario includere l'identità e i dati di contatto del fornitore e, se presente, del suo rappresentante autorizzato. Le istruzioni devono poi descrivere in modo accurato le caratteristiche, le capacità e i limiti delle prestazioni del sistema di IA. Il contenuto delle istruzioni deve chiarire lo scopo specifico del sistema, il livello di accuratezza atteso e i requisiti di robustezza e sicurezza informatica previsti dall'art. 15. Devono inoltre essere richiamate tutte le condizioni note in grado di incidere su tali parametri, nonché gli scenari di uso improprio ragionevolmente prevedibili che potrebbero generare rischi. Laddove pertinente, occorre spiegare la capacità del modello di fornire spiegazioni comprensibili dei propri output e precisare le prestazioni rilevate in relazione ai gruppi di destinatari. Risulta quindi necessario

[34] Si fa riferimento a un linguaggio rivolto all'utente medio che si interfaccia e utilizza il sistema di intelligenza artificiale, evitando il più possibile il linguaggio tecnico o strettamente comprensibile solo da esperti del settore.

[35] Regolamento (UE) 2016/679 del Parlamento europeo e del Consiglio, 27 aprile 2016 (GDPR), art. 12.

[36] M. Bonafè - C. Trevisi, *Intelligenza artificiale, l'algoritmo trasparente: un rebus ancora da sciogliere*, in *agendadigitale.eu*, 15 luglio 2019.

descrivere in dettaglio le caratteristiche dei dati di input nonché dei set di addestramento, validazione e test impiegati, sempre in rapporto alla finalità funzionale dell'IA.
Le istruzioni devono poi menzionare ogni modifica predeterminata delle prestazioni impostata dal fornitore in sede di valutazione di conformità, a garanzia di trasparenza e per consentire agli utenti di comprendere le basi dell'approvazione del sistema. Nel medesimo documento vanno illustrate le misure di supervisione umana previste dall'art. 14 e le soluzioni tecniche predisposte per facilitare una corretta interpretazione dei risultati. Non meno rilevante è l'indicazione delle risorse computazionali e hardware necessarie, della durata di vita prevista e degli interventi di manutenzione (inclusi gli aggiornamenti software) necessari al regolare funzionamento. Infine, qualora applicabile, occorre descrivere i meccanismi interni che consentono ai *deployer* di raccogliere, archiviare e interpretare correttamente i log ai sensi dell'art. 12, affinché tutte le operazioni siano tracciabili e sottoposte a verifiche di sicurezza e conformità.
Quanto appena descritto - per mera categorizzazione di tipo dottrinale - può intendersi come il primo livello di trasparenza, ovvero la trasparenza garantita dal *provider* affinché il *deployer* possa agire in conformità con il Regolamento, in accordo con quanto ampiamente descritto dal considerando 71[37]. Rimane ora da analizzare il secondo livello, ossia la trasparenza

[37] Il considerando 71 del regolamento (UE) 2024/1689 (AI Act) richiama l'attenzione su uno dei principi cardine della disciplina: la trasparenza, quale condizione preliminare per l'immissione sul mercato dei sistemi di intelligenza artificiale ad alto rischio. Il legislatore europeo riconosce che la complessità e l'opacità tecnologica possono ostacolare la comprensione effettiva del funzionamento di tali sistemi da parte dei loro utilizzatori — i cosiddetti *deployer* — e, di conseguenza, compromettere un uso consapevole e responsabile degli stessi. Per mitigare tale rischio, il Regolamento impone ai fornitori di progettare sistemi di intelligenza artificiale che risultino effettivamente intellegibili: chi li utilizza deve poter comprendere, in termini pratici, le modalità di elaborazione dei dati da parte dell'algoritmo, i risultati attesi e i limiti prestazionali entro cui il sistema opera. Questa esigenza di comprensione si traduce in un articolato obbligo informativo. Ogni sistema ad alto rischio deve essere accompagnato da istruzioni d'uso complete, redatte in una lingua comprensibile per il pubblico di riferimento, contenenti la descrizione delle caratteristiche tecniche, delle capacità e dei limiti del sistema, nonché degli scenari di utilizzo previsti e di quelli espressamente vietati. Devono inoltre essere indicate le condizioni — note o ragionevolmente prevedibili — in cui il comportamento del sistema può variare, in particolare quando un intervento del *deployer* possa incidere sulle prestazioni. Le istruzioni devono altresì specificare eventuali modifiche di *performance* già valutate in sede di procedura di conformità e le misure di supervisione umana necessarie per interpretare correttamente gli *output*. A fini di effettiva fruibilità, il considerando invita a integrare tali informazioni con esempi illustrativi: casi d'uso, limiti operativi e contesti nei quali il sistema di intelligenza artificiale non dovrebbe essere applicato. L'obiettivo è duplice: da un lato, agevolare il *deployer* nella selezione e nell'impiego del sistema più idoneo alle proprie esigenze e, dall'altro, assicurare un utilizzo conforme ai principi di sicurezza, tutela della salute e rispetto dei diritti fondamentali. In ultima

accordata all'utente finale interessato, ultimo beneficiario dell'operatività delle intelligenze artificiali.

7. La tutela degli "interessati", il secondo livello di trasparenza e gli obblighi dei *providers* e dei *deployers* nei confronti degli utenti finali

Sebbene nell'art. 3 del Titolo I del Regolamento non vi sia alcun riferimento ai soggetti interessati, intesi come "beneficiari" dei servizi offerti con l'ausilio o il supporto di intelligenze artificiali, possiamo inquadrare delle particolari categorie di interessati analizzando i considerando del Regolamento. Il considerando 51 analizza l'uso dell'IA nei delicati processi che regolano l'accesso a prestazioni essenziali, pubbliche o private. Gli interessati sono i cittadini che ne fanno richiesta; se un algoritmo decide in merito alla concessione, riduzione o revoca di tali prestazioni, si applica l'art. 22 GDPR. L'interessato deve essere dunque informato dal soggetto erogante il servizio dell'eventuale decisione basata su procedimento decisionale automatizzato, affinché sia possibile l'esercizio di tutti i diritti riconosciuti, inclusa la richiesta di informazioni aggiuntive. Ulteriori obblighi in capo ai «fornitori» - questa volta da intendersi come l'insieme di *providers* e *deployers* - nei confronti delle persone fisiche, sono sanciti nel Titolo IV del Regolamento «*Transparency obligations for providers and deployers of certain AI Systems*» all'art. 50. Il primo paragrafo sancisce un preciso obbligo per i fornitori che sviluppano sistemi di AI destinati all'interazione diretta con persone fisiche. Tali sistemi devono essere progettati e sviluppati in modo tale che le persone siano informate dell'interazione con un sistema di AI, a meno che ciò non sia ovvio data la situazione e il contesto di utilizzo[38]. L'obbligo di maggiore rilevanza – a tutela della libertà di espressione e di informazione sancita dall'art. 11 della Carta dei diritti fondamentali dell'Unione europea– è relativo alla necessità di informazione nel caso di generazione di contenuti *deep fakes* con sistemi di intelligenza artificiale. I *deployers* che generano o manipolano contenuti visivi o audio raffiguranti individui o situazioni, avendo come risultato un prodotto difficilmente distinguibile dalla realtà, devono rivelare che il contenuto è stato generato o manipolato

analisi, la trasparenza richiesta dal considerando 71 non si riduce a un mero adempimento documentale, ma rappresenta uno strumento funzionale alla più ampia *accountability* che permea l'intero impianto regolatorio dell'*Artificial Intelligence Act* (considerando 71).

[38] A titolo di esempio, supponiamo che un'azienda stia sviluppando un assistente virtuale basato su intelligenza artificiale per l'assistenza clienti in un sito web di e-commerce. L'azienda fornitrice avrà l'obbligo di progettare e sviluppare questo sistema in modo che i clienti siano informati dell'interazione con un sistema di AI, a meno che ciò non risulti ovvio data la situazione e il contesto di utilizzo.

artificialmente. Lo stesso obbligo circa l'informazione di un contenuto generato artificialmente si applica a chi genera e pubblica testi – aventi rilevanza di pubblico interesse, come ad esempio un articolo di cronaca – a meno che il testo non sia stato sottoposto a revisione umana. Inoltre - nel caso in cui servissero ulteriori approfondimenti circa la natura artificiale o meno di un contenuto, come ad esempio in contesti giudiziari – il Regolamento pone a carico dei fornitori l'obbligo di garantire che qualsiasi contenuto generato da sistemi di intelligenza artificiale abbia un formato *machine-readable*[39] e di conseguenza identificabile inequivocabilmente come artificiale. Infine, il considerando 91 attribuisce ai *deployers* il ruolo fondamentale di informare le «persone fisiche soggette all'uso del sistema di IA ad alto rischio», rendendole opportunamente edotte circa lo scopo previsto, il tipo di decisioni prese e il diritto a ricevere tutte le informazioni necessarie ai sensi del Regolamento.

8. Interpretabilità e spiegabilità come presupposti della trasparenza algoritmica[40]

Per garantire un'efficace trasparenza algoritmica, è fondamentale distinguere e integrare due concetti chiave: l'interpretabilità e la spiegabilità. L'interpretabilità[41] si riferisce alla capacità di un esperto di comprendere il funzionamento interno di un sistema algoritmico, analizzando come i dati vengono elaborati e come le decisioni vengono generate. Questo aspetto è cruciale per i tecnici e gli sviluppatori, che devono poter «aprire la black box»[42] per verificare la correttezza logica e funzionale del sistema, ad esempio tramite l'analisi dei pesi dei modelli o la visualizzazione delle attivazioni delle reti

[39] Un formato *machine-readable* (leggibile dalle macchine) si riferisce a qualsiasi formato di dati che può essere facilmente elaborato da un computer. D. Amadi - S. Kiwuwa-Muyingo - T. Bhattacharjee - A. Taylor - A. Kiragga - M. Ochola - C. Kanjala - A. Gregory - K. Tomlin - J. Todd - J. Greenfield, *Making Metadata Machine-Readable as the First Step to Providing Findable, Accessible, Interoperable, and Reusable Population Health Data:Framework Development and Implementation Study*, in *Online Journal of Public Health Informatics*, 2024, e56237., costituisce un ottimo esempio applicativo del concetto di *machine-readable*, poiché dimostra come la strutturazione e standardizzazione dei metadati nel contesto sanitario possano rendere i dati realmente *Findable, Accessible, Interoperable and Reusable* (*FAIR*), migliorandone la qualità, l'integrazione e la riusabilità nei sistemi di salute pubblica.

[40] Si ringrazia il Dott. Giorgio Presepio per la collaborazione nella redazione e revisione del presente paragrafo.

[41] F. Doshi-Velez - B. Kim, *Towards a Rigorous Science of Interpretable Machine Learning*, in *arXiv*, 2017, 1702.08608.

[42] F. Pasquale, *The Black Box Society: The Secret Algorithms That Control Money and Information*, Cambridge (MA), 2015.

neurali per comprendere quali caratteristiche sono considerate rilevanti. La spiegabilità[43], d'altro canto, riguarda la possibilità per un utente non esperto – come un cittadino, un giudice o un consumatore – di comprendere le ragioni alla base di una specifica decisione prodotta da un algoritmo. La spiegabilità si concentra sul "perché" una decisione è stata presa, rendendola intelligibile a un pubblico non tecnico, ad esempio spiegando che "la tua richiesta di prestito è stata rifiutata perché il tuo reddito è inferiore alla soglia X e il tuo rapporto debito/reddito supera Y".

Questa distinzione evidenzia un doppio livello di comprensibilità necessario per la piena realizzazione della trasparenza. A un livello tecnico, la comprensione è rivolta a sviluppatori, responsabili della protezione dei dati (DPO) e revisori (audit), che necessitano di una visione approfondita del modello per assicurarne la correttezza, l'affidabilità e la conformità normativa. Questo si traduce nella necessità di strumenti e tecniche che permettano di analizzare la logica interna del sistema e i suoi *bias* intrinseci. A un livello sociale, la comprensione è destinata a coloro che sono direttamente influenzati dalle decisioni algoritmiche, i quali devono poter capire il perché di una determinata decisione per poter esercitare i propri diritti e per mantenere la fiducia nel sistema. Come sottolineato in letteratura, la fiducia pubblica nell'IA dipende dalla sua capacità di essere spiegata in modo significativo agli utenti finali.

Esiste una chiara connessione normativa tra questi concetti e il quadro giuridico europeo. L'art. 13 dell'AI Act, ad esempio, impone al fornitore di un sistema di intelligenza artificiale di fornire al *deployer* le istruzioni necessarie affinché quest'ultimo possa comprendere e utilizzare correttamente il sistema. Questo requisito mira a garantire un'adeguata interpretabilità a livello tecnico, permettendo al *deployer* di adempiere ai propri obblighi di supervisione e monitoraggio, inclusa una valutazione d'impatto sui diritti fondamentali (FRIA) efficace.

Parallelamente, l'art. 22 GDPR, applicabile alle decisioni basate unicamente sul trattamento automatizzato, inclusa la profilazione, che producano effetti giuridici significativi per l'individuo, impone l'obbligo di fornire «informazioni significative in merito alla logica utilizzata». Questo requisito si traduce nell'esigenza pratica della spiegabilità, poiché le informazioni devono essere presentate in termini comprensibili all'interessato.

Il quadro normativo europeo, pertanto, riconosce implicitamente l'importanza di interpretabilità e spiegabilità. Tali obblighi, già trattati nei paragrafi precedenti, trovano qui conferma nella necessità di garantire autonomia e consapevolezza all'utente. Il considerando 71 dell'AI Act rafforza

[43] M. Ribeiro - S. Singh - C. Guestrin, *Nothing Else Matters: Model-Agnostic Explanations by Identifying Prediction Invariance*, in *arXiv*, 2016, 1611.05817.

ulteriormente questa necessità, richiamando esplicitamente l'importanza di fornire una spiegazione comprensibile degli *output* dei sistemi di intelligenza artificiale agli utenti finali. In particolare, per i sistemi ad alto rischio, l'obbligo di rendere intellegibile la logica del processo decisionale è essenziale per consentire agli individui di comprendere come vengono prese le decisioni che li riguardano e di esercitare i propri diritti, come il diritto di ottenere un intervento umano o di contestare la decisione. La capacità di comprendere la "logica utilizzata" è dunque un requisito fondamentale per la validità del consenso e per l'esercizio dell'autodeterminazione, specialmente in contesti che possono influenzare diritti fondamentali. Garantire interpretabilità e spiegabilità non è solo un obbligo legale, ma anche un imperativo etico per costruire sistemi di IA che siano equi, responsabili e affidabili.

Nonostante i progressi normativi, i modelli attuali di intelligenza artificiale presentano dei limiti intrinseci. Molti algoritmi, in particolare le reti neurali profonde e i modelli generativi, operano come "scatole nere" (*black-box*), rendendo difficile una spiegazione diretta e trasparente del loro funzionamento interno. Questa opacità, ampiamente discussa nella letteratura scientifica[44], deriva dalla loro complessità e dal modo in cui apprendono da grandi quantità di dati. Spesso, esiste un trade-off tra l'accuratezza di un modello e la sua interpretabilità: un sistema altamente preciso potrebbe essere intrinsecamente meno comprensibile, e viceversa. Questo *trade-off* rappresenta una delle sfide più significative nello sviluppo di sistemi XAI[45]. Questa sfida tecnica ha significative implicazioni giuridiche. La spiegabilità diventa una vera e propria condizione di legittimità per l'automazione decisionale. In assenza di una spiegazione intelligibile, diviene estremamente difficile per gli individui esercitare pienamente i loro diritti, come il dirit-

[44] Per una rassegna sistematica sullo stato dell'arte dell'*Explainable Artificial Intelligence* (*XAI*) (v. nota n. 43), con particolare attenzione ai metodi di interpretazione dei modelli *black-box*, alle sfide di trasparenza e ai risvolti etico-normativi in ambiti critici come sanità, finanza e trasporti autonomi, si veda V. Hassija - V. Chamola - A. Mahapatra - A. Singal - D. Goel - K. Huang - S. Scardapane - I. Spinelli - M. Mahmud - A. Hussain, *Interpreting Black-Box Models: A Review on Explainable Artificial Intelligence*, in *Cognitive Computation*, 1, 2023, 45 ss. Questo problema è particolarmente evidente in ambito sanitario; per una panoramica completa, si veda Z. Salahuddin-H. C. Woodruff-A. Chatterjee-P. Lambin, *Transparency of Deep Neural Networks for Medical Image Analysis: A Review of Interpretability Methods*, in *Computers in Biology and Medicine*, 2022, 105919.

[45] Un sistema *XAI* è un sistema di intelligenza artificiale capace di fornire spiegazioni sulle proprie decisioni o previsioni, offrendo agli utenti una comprensione del comportamento del modello e delle influenze che portano a un certo *output*. Si veda in merito A. Barredo Arrieta - N. Díaz-Rodríguez - J. Del Ser - A. Bennetot - S. Tabik - A. Barbado - S. Garcia - S. Gil-Lopez - D. Molina - R. Benjamins - R. Chatila - F. Herrera, *Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges Toward Responsible AI*, in *Information Fusion*, 2020, 82 ss.

to di accesso ai dati personali, il diritto di opposizione a un trattamento automatizzato o il diritto di presentare un reclamo. L'introduzione di un sistema di "scatola nera" (questa volta intesa come un sistema di registrazione automatica dei *log*) funzionale alla ricostruzione accurata degli eventi, come previsto dall'art. 12 dell'AI Act per i sistemi ad alto rischio, mira proprio a colmare questa lacuna di tracciabilità e spiegabilità.
Le prospettive future in questo campo sono strettamente legate allo sviluppo della cosiddetta "XAI" (*Explainable AI*)[46], un'area di ricerca che si propone di sviluppare tecniche e metodologie per rendere i sistemi di intelligenza artificiale più trasparenti e comprensibili. La *XAI* rappresenta non solo una sfida tecnica, ma anche giuridica, poiché le soluzioni sviluppate dovranno essere in grado di soddisfare i requisiti normativi di trasparenza e spiegabilità imposti dal GDPR e dall'AI Act. Le soluzioni *XAI* non mirano solo a spiegare il «perché» di una decisione, ma anche a fornire controfattuali (cosa sarebbe dovuto cambiare perché la decisione fosse diversa) o a identificare i fattori più influenti in una decisione, offrendo così maggiore controllo e comprensione all'utente. A tal riguardo, la standardizzazione internazionale, come quella definita dalla norma ISO/IEC TR 24028[47] (*Trustworthiness in AI*), identifica la spiegabilità come una componente chiave dell'affidabilità dei sistemi di intelligenza artificiale, riconoscendone l'importanza per la fiducia del pubblico e la conformità legale. L'evoluzione del concetto di trasparenza algoritmica, dal GDPR all'AI Act, riflette la crescente consapevolezza dell'importanza di garantire agli individui non solo la protezione dei loro dati, ma anche la capacità di comprendere e controllare le decisioni automatizzate che li riguardano, salvaguardando così la loro autodeterminazione e i diritti fondamentali.

9. Conclusioni

Il significato sostanziale della trasparenza algoritmica si è progressivamente ampliato in parallelo all'evoluzione tecnologica e alla crescente incidenza che quest'ultima esercita sull'effettività e sulla tutela dei diritti fondamentali della persona. Già nell'era della profilazione tradizionale la piena consapevolezza delle conseguenze dei processi decisionali automatizzati costituiva un requisito imprescindibile per la tutela della privacy e, soprattutto, dell'autodeterminazione individuale. Nel 2016 il giornalista Steve

[46] A. Adadi - M. Berrada, *Peeking Inside the Black Box: A Survey on Explainable Artificial Intelligence (XAI)*, in *IEEE Access*, 2018, 52138 ss.

[47] ISO/IEC, *Information technology – Artificial intelligence – Overview of trustworthiness in artificial intelligence*, ISO/IEC TR 24028:2020, 2020.

Lohr introdusse il termine «dataismo»[48] per descrivere una potenziale "dittatura digitale" in cui l'autonomia umana rischia di soccombere al peso degli algoritmi. Con i sistemi di intelligenza artificiale di nuova generazione, la posta in gioco è divenuta ancor più alta: l'ambito applicativo di tali tecnologie travalica l'analisi dei dati, interessando l'erogazione di servizi pubblici, la gestione del traffico, le componenti di sicurezza dei macchinari industriali, i veicoli a guida autonoma e, attraverso l'IA generativa, la produzione di contenuti testuali, visivi e sonori difficilmente distinguibili da quelli reali. La protezione non riguarda più soltanto la riservatezza delle informazioni e la libertà di scelta, ma si estende alla stessa integrità psico-fisica degli individui e alla tenuta dell'ecosistema informativo. In tale contesto il principio di *accountability* si dilata fino a investire direttamente i fornitori di sistemi di IA ad alto rischio, i quali devono assicurare un livello di trasparenza elevato lungo l'intero ciclo di vita del prodotto. Da un lato ciò implica mettere i *deployer* nella condizione di conoscere in modo esaustivo le caratteristiche tecniche, i limiti operativi e le condizioni di impiego sicuro, così da consentire un controllo effettivo del sistema anche dopo la commercializzazione; dall'altro, comporta fornire agli utenti finali informazioni chiare sui benefici e sui possibili pericoli derivanti da usi impropri o fraudolenti, affinché il rapporto con l'IA rimanga consapevole e bilanciato. Solo una trasparenza intesa in questa duplice accezione può consolidare la fiducia nella società dei dati e scongiurare il rischio che l'innovazione tecnologica si traduca in una nuova forma di eterodirezione digitale.

[48] S. Lohr, *Data-ism. Inside the Big Data Revolution*, Londra, 2016. Sul tema, cfr. I. Syllaidopoulos, *Dataism: The Rise of a Data-Driven World? A Guide for Data-Oriented Policy and Management*, in *International Journal of Multidisciplinary Research and Analysis*, 2023, 3580 ss.

Alessandro Oldani

Abstract

Il contributo esamina l'evoluzione del concetto di trasparenza algoritmica nel diritto europeo, ripercorrendone lo sviluppo dal GDPR fino all'AI Act. Dopo aver analizzato il fondamento giuridico nell'art. 22 GDPR, il lavoro si sofferma sulla classificazione dei sistemi di IA in base al rischio e sugli obblighi di trasparenza per fornitori e *deployer*. Vengono distinti i livelli di trasparenza tecnico-operativa e sostanziale, concludendo con una riflessione su interpretabilità e spiegabilità nel quadro della *Explainable AI* (XAI).

This article examines the evolution of algorithmic transparency in EU law, tracing its development from the GDPR to the AI Act. After analysing the legal basis under art. 22 GDPR, the paper addresses the AI Act's risk-based classification and the transparency obligations imposed on providers and deployers. Two distinct levels of transparency are identified: technical-operational and substantive, concluding with a reflection on interpretability and explainability within the framework of Explainable AI (XAI).

Keywords

algorithmic transparency – artificial intelligence – GDPR – AI Act – self-determination