https://doi.org/10.6092/issn.1825-1927/21064 ISSN 1825-1927

Intelligenza Artificiale, ricerca scientifica e adeguamento etico-giuridico

Esigenze attuali e prospettive nell'ottica di maggior tutela degli individui

Francesco Di Tano [1]

[1] University of Bologna, Italy

Abstract: Il contributo esplora il crescente ruolo dell'Intelligenza Artificiale (IA) nella ricerca scientifica e le sfide etico-giuridiche emergenti, su protezione dei dati, trasparenza e libertà di scelta, con impatti sull'affidabilità etica dei sistemi. Premessa un'analisi degli aspetti giuridicamente rilevanti, in special modo derivanti da AI Act e GDPR, si pongono in evidenza le principali criticità che attualmente affrontano i ricercatori nel corso delle attività di raccolta ed elaborazione dei dati ai fini dello sviluppo di sistemi di IA, nonché la necessità di approcci innovativi per garantire comprensibilità e affidabilità di tali sistemi, con modelli di Explanatory AI (YAI) e nuove metodologie di valutazione d'impatto etico-giuridico, in una prospettiva di *ethics by design*.

Keywords: intelligenza artificiale, ricerca scientifica, ethics by design, decisioni automatizzate, spiegabilità

1 Introduzione

Sono trascorsi più di venti anni da quando Luciano Floridi ha (ri)definito l'*infosfera* come l'intero sistema di servizi e documenti codificati in qualsiasi supporto semiotico e fisico, contenenti qualsiasi tipo di dato, informazione e conoscenza, senza limiti dimensionali, di tipologia o struttura logica[1].

In essa vanno a integrarsi, connettersi e intrecciarsi gli elementi intelligenti e reattivi tanto della vita online quanto di quella offline. Le stesse identità degli individui non possono più prescindere dalle attività condotte in rete e dagli eventuali profili personali costruiti sui principali social media. L'online sfuma nell'offline e viceversa, in una dimensione globale che viene definita, sempre dallo stesso Floridi, "*onlife*", ove le relazioni, le comunicazioni, le attività lavorative, economiche e sociali in generale sono il frutto di una continua interazione tra la realtà materiale e quella virtuale.

Com'è ormai noto, Internet e le nuove scoperte tecnologiche hanno offerto un nuovo e dirompente modo di comunicare e interagire, rappresentando sempre più una realtà complessa indistinguibile dal mondo fisico. Oltre al piano essenzialmente materiale, dove router, server, cavi e dispositivi hardware sono percepibili dai sensi umani, è cambiata la percezione della materialità, dello spazio e delle informazioni, con conseguente diretta o indiretta influenza sul modo di comprensione della realtà circostante.

francesco.ditano@unibo.it (Francesco Di Tano);
(Francesco Di Tano);

1. Floridi, Luciano. *Philosophy and Computing: An introduction*. Routledge, 1999.

Le tecnologie digitali esprimono oramai tutta la propria forza plasmando e trasformando la quotidianità degli individui, non solo rendendosi essenziali come preziosi strumenti di supporto, ma anche sottoponendoli, in maniera più o meno consapevole, a innumerevoli processi automatizzati, spesso poco trasparenti.

L'Intelligenza Artificiale e i *big data* hanno un impatto significativo sulla società odierna, poiché molti aspetti della nostra vita sono diventati soggetti all'elaborazione (e rielaborazione) dei dati, in ciò che viene definito "*datafication*"[2].

Negli ultimi anni, si è verificata un'accelerazione rapida e dirompente dei progressi nel campo dell'Intelligenza Artificiale, guidata a sua volta da una particolare evoluzione di sempre maggiori disponibilità di dati, potenza di calcolo e capacità di apprendimento automatico dei sistemi[3]. Notevoli passi avanti sono stati compiuti nello sviluppo di modelli di base, addestrati su grandi volumi di dati, che ha poi originato la cosiddetta "IA generativa", solitamente basata su istruzioni (*prompt*) fornite dall'utente[4].

La ricerca scientifica non è immune da queste dinamiche[5]. Anzi, al contrario, specialmente nell'ultimo decennio, si è dedicata con particolare attenzione allo sviluppo e all'applicazione di sistemi di intelligenza artificiale in tutti gli ambiti sociali ed economici[6]. Sempre più progetti contemplano, nella metodologia o proprio come obiettivi di ricerca, lo sviluppo o l'applicazione di sistemi di Intelligenza Artificiale impattanti sui soggetti partecipanti[7].

Utilizzando un'ampia raccolta di dati, sia personali che generali, l'intelligenza artificiale è in grado di generare risultati preziosi per scopi scientifici e statistici. In ambito medico, supporta la diagnosi e la previsione di patologie; nelle scienze sociali, facilita l'analisi dei comportamenti sociali, economici e politici; nel settore commerciale, consente di identificare, classificare e anticipare le preferenze e le tendenze dei consumatori.

I risultati generati da queste elaborazioni possono spesso avere una natura generale, basandosi su analisi aggregate e anonime, e quindi non configurarsi come dati personali secondo la definizione del GDPR[8]. Tuttavia, in alcuni contesti, come la diagnostica medica, l'applicazione può (e oramai tende a) prevedere un approccio personalizzato, focalizzato sul paziente.

In ogni caso, l'elaborazione statistica e scientifica può coinvolgere direttamente le persone, esponendo i loro dati personali a potenziali rischi di sicurezza e abusi. Anche quando non si applica direttamente al singolo individuo, l'intelligenza artificiale può essere sviluppata e addestrata utilizzando *big data* o set di dati contenenti informazioni relative a persone fisiche identificate o identificabili[9].

2. Mai, Jens-Erik. "Big data privacy: The datafication of personal information." *The Information Society* 32.3 (2016): 192-199; van Dijck, José. "Datafication, dataism and dataveillance: Big Data between scientific paradigm and ideology." *Surveillance & Society* 12.2 (2014): 197-208; Mayer-Schönberger, Viktor e Kenneth Cukier. *Big Data: A Revolution That Will Transform How We Live, Work and Think*. John Murray Publishers, 2013.

3. Xue, Lan, e Zhenjing Pang. "Ethical governance of artificial intelligence: An integrated analytical framework." *Journal of Digital Economy*, 1.1 (2022): 44-52.

4. Resnik, David B., e Mohammad Hosseini. "The ethics of using artificial intelligence in scientific research: new guidance needed for a new tool." *AI and Ethics* (2024).

5. Casiraghi, Simone, e Niels van Dijk. "Ethics reviews in the European Union. Implications for the governance of scientific research in times of data science and Artificial Intelligence." *Law, Innovation and Technology* 16.1. (2024): 101-122; Metcalf, Jacob, e Kate Crawford. "Where are human subjects in Big Data research? The emerging ethics divide." *Big Data & Society* 3.1 (2016).

6. González-Esteban, Elsa, e Patrici Calvo. "Ethically governing artificial intelligence in the field of scientific research and innovation." *Heliyon*. 8.2 (2022).

7. Resnik, David B., e Mohammad Hosseini. *op. cit.*; Krenn, Mario, et al. "On scientific understanding with artificial intelligence." *Nature Reviews Physics* 4 (2022): 761-769; Wang, Hanchen, et al. "Scientific discovery in the age of artificial intelligence." *Nature* 620 (2023): 47-60.

8. Per GDPR si intende il Regolamento (UE) 2016/679 del Parlamento europeo e del Consiglio del 27 aprile 2016 relativo alla protezione delle persone fisiche con riguardo al trattamento dei dati personali, nonché alla libera circolazione di tali dati e che abroga la direttiva 95/46/CE (Regolamento generale sulla protezione dei dati, o GDPR).

9. Sartor, Giovanni e Francesca Lagioia. "Le decisioni algoritmiche tra etica e diritto." *Intelligenza artificiale - il diritto, i diritti, l'etica*. Giuffrè Francis Lefebvre, Milano, 2020.

Inoltre, persino i risultati di tipo puramente statistico possono avere un impatto indiretto sugli individui, poiché forniscono indicazioni sui gruppi di appartenenza, che possono influire sulle decisioni automatizzate prese nei confronti del singolo o determinare conseguenze significative[10].

Indipendentemente dalle potenziali applicazioni future di questi sistemi, su una scala più o meno ampia di individui sulla base del contesto di riferimento, è la stessa attualità a porre rilevanti questioni etico-giuridiche riguardanti i diritti e le libertà delle persone, con particolare riferimento alla protezione dei dati personali, alla sicurezza e alla libertà di scelta[11].

Alla luce di ciò, il presente contributo si propone di analizzare il ruolo dell'Intelligenza Artificiale nella ricerca scientifica, mettendo in luce vantaggi e criticità legate al suo sviluppo e utilizzo, attraverso un approfondimento delle relative implicazioni etiche e giuridiche. L'obiettivo è quello di esaminare, attraverso un approccio qualitativo e interdisciplinare, l'attuale (nuova) disciplina in materia di Intelligenza Artificiale e di individuare possibili strumenti e adeguamenti normativi che possano garantirne un utilizzo etico nel contesto della ricerca in ambito europeo.

Attraverso l'analisi della letteratura scientifica più recente e della regolamentazione vigente, anche di *soft-law*, avente un impatto sulla produzione e sulla validazione della conoscenza scientifica, sono affrontati i dubbi interpretativi e i problemi applicativi che, in quanto tali, sono potenzialmente forieri di pericolose limitazioni dei diritti e delle libertà degli individui coinvolti. Da ultimo, si identificano possibili strumenti che, in ottica di *ethics by design*[12], riescano a guidare efficacemente l'utilizzo dei sistemi di Intelligenza Artificiale nelle attività di ricerca scientifica verso l'adesione ai principi etici coinvolti.

2 L'adeguamento etico e giuridico dei sistemi di Intelligenza Artificiale

2.1 La regolamentazione dell'Intelligenza Artificiale: dalla *trustworthy AI* all'AI Act

Le maggiori istituzioni mondiali, consapevoli del tema e pur nelle diversità culturali che le contraddistingue, hanno iniziato a porre le basi, con la medesima vocazione etica, per un governo etico-giuridico dell'Intelligenza Artificiale affidabile ("*trustworthy*"), rispettosa della dignità umana e indirizzata verso il benessere dell'uomo[13].

Proprio su questi aspetti, prim'ancora del recente AI Act[14], il Gruppo Indipendente di Esperti di Alto Livello

10. Si pensi, ad esempio, alla pubblicità online, ritagliata su misura o sulla base di categorizzazioni per gruppi, nonché alle decisioni automatizzate di compagnie assicurative o istituti bancari finalizzate alla promozione di offerte e premi o alla stipulazione di contratti, anch'esse basate sulla profilazione del cliente.

11. Weinbaum, Cortney, et al. *Ethics in Scientific Research: An Examination of Ethical Principles and Emerging Topics*. Rand Corporation Publication, 2019. 63-65.

12. L'*ethics by design* nella ricerca scientifica rappresenta un approccio secondo cui i valori etici e i diritti fondamentali (quali la dignità umana, la non discriminazione, l'autonomia, la trasparenza e la responsabilità) devono essere incorporati fin dalle prime fasi della ricerca e, in particolar modo, dello sviluppo di sistemi tecnologici e di Intelligenza Artificiale, garantendo che essi operino in modo conforme a principi etici condivisi e alle vigenti norme giuridiche. Si vedano, su tale concetto: AI4People Institute. *Towards an Ethics by Design Appoach for AI*, 2024, https://ai4people.org/wp-content/uploads/2024/06/Towards-an-Ethics-by-Design-Approach-for-AI.pdf (ultima consultazione, 20/06/2025); Brey, Philip e Brandt Dainow. "Ethics by design for artificial intelligence." *AI and Ethics* 4 (2024): 1265-1277; d'Aquin, Mathieu, Troullinou, Pinelopi, O'Connor, Noel, Cullen, Aindrias, Faller, Gráinne e Louise Holden. "Towards an 'Ethics by Design' Methodology for AI Research Projects." *AIES '18: Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society* (2018): 54-59; Dignum, Virginia, Baldoni, Matteo, Baroglio, Cristina et al. "Ethics by design: Necessity or curse?" *AIES '18: Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society* (2018): 60-66.

13. China Governance Principles for a new Generation of AI: Develop Responsible AI (2019); OECD Principles on AI (2019); Hambach Declaration on AI (2019) (Germany); ICO Guidance on AI and Data Protection (2020) (UK); White House OMB Draft Memo on the Regulation of AI (2020) (USA); Rome Call for AI Ethics (2020); Risoluzione del Parlamento europeo del 20 ottobre 2020 recante raccomandazioni alla Commissione concernenti il quadro relativo agli aspetti etici dell'intelligenza artificiale, della robotica e delle tecnologie correlate; Commissione Europea, *Libro bianco sull'intelligenza artificiale. Un approccio europeo all'eccellenza e alla fiducia* (19 febbraio 2020); Risoluzione del Parlamento europeo del 16 febbraio 2017 recante raccomandazioni alla Commissione concernenti norme di diritto civile sulla robotica.

14. Regolamento (UE) 2024/1689 del Parlamento europeo e del Consiglio, del 13 giugno 2024, che stabilisce regole armonizzate sull'intelligenza artificiale e modifica i regolamenti (CE) n, 300/2008, (UE) n, 167/2013, (UE) n, 168/2013, (UE) 2018/858,

sull'Intelligenza Artificiale, nominato dalla Commissione Europea nel giugno 2018 nell'ambito della propria strategia politica sull'intelligenza artificiale, ha elaborato gli Orientamenti Etici per un'Intelligenza Artificiale affidabile[15].

Tali linee guida, presto divenute un modello di riferimento anche nell'ambito dei programmi quadro europei per la ricerca e l'innovazione (da ultimo, Horizon Europe)[16], promuovono la necessità, per ciascun sistema di Intelligenza Artificiale, di soddisfare quattro principi fondamentali (rispetto dell'autonomia umana, prevenzione dei danni, equità ed esplicabilità) e sette requisiti fondamentali al fine di essere considerato eticamente affidabile: supervisione umana; robustezza e sicurezza; privacy e *data governance*; trasparenza; *diversity*, correttezza, assenza di discriminazione; benessere sociale e ambientale; responsabilità.

La continua espansione dei sistemi di intelligenza artificiale ha dunque indotto l'Unione europea a introdurre un quadro giuridico unificato con l'obiettivo di stabilire regole armonizzate, per l'appunto l'AI Act.

Tale corpus normativo, orientato a favorire un'adozione dell'intelligenza artificiale affidabile, centrata sull'essere umano e che garantisca alti standard di tutela per la salute, la sicurezza e i diritti fondamentali, migliorando al contempo il funzionamento del mercato interno, adotta un approccio basato sul rischio, con categorizzazione di quattro distinti livelli e corrispondenti requisiti: rischio inaccettabile, rischio elevato, rischio limitato e rischio minimo.

I sistemi classificati come "a rischio inaccettabile", che rappresentano una minaccia evidente per la sicurezza, i mezzi di sussistenza o i diritti delle persone, sono del tutto vietati. Tra questi rientrano il punteggio sociale da parte dei governi e l'identificazione biometrica in tempo reale in spazi pubblici. I sistemi definiti "ad alto rischio" sono sottoposti a severi requisiti normativi, e comprendono applicazioni in settori come infrastrutture critiche, istruzione, occupazione, servizi essenziali, sanità, forze dell'ordine, migrazione e controllo delle frontiere. Tali sistemi devono rispettare rigide procedure di test, documentazione e conformità, oltre a misure di gestione dei rischi e della qualità, per garantire trasparenza, sicurezza ed equità. Le applicazioni di IA a "rischio limitato" sono soggette a specifici obblighi di trasparenza, come l'informazione agli utenti sul fatto che stanno interagendo con un sistema di intelligenza artificiale, in modo da permettere loro di prendere decisioni consapevoli. Infine, i sistemi classificati come "a rischio minimo" prevedono requisiti normativi molto limitati, includendo ad esempio i filtri antispam e i videogiochi dotati di intelligenza artificiale.

Tra i rilevanti adempimenti introdotti, l'articolo 27 dell'AI Act prevede che, prima di utilizzare un sistema di IA ad alto rischio, i *deployer*[17] che siano organismi di diritto pubblico o enti privati fornitori di servizi pubblici e i *deployer* di sistemi di IA ad alto rischio debbano effettuare una valutazione d'impatto sui diritti fondamentali che l'uso di tale sistema può produrre (anche denominata *Fundamental Rights Impact Assessment*, o FRIA)[18].

L'articolo 2, paragrafo 6 dell'AI Act esclude specificamente dal proprio ambito di applicazione i sistemi di intelligenza artificiale sviluppati o utilizzati esclusivamente per la ricerca scientifica[19]. Questa esclusione,

(UE) 2018/1139 e (UE) 2019/2144 e le direttive 2014/90/UE, (UE) 2016/797 e (UE) 2020/1828, conosciuto come Regolamento sull'Intelligenza Artificiale o, per l'appunto, AI Act.

15. Gruppo Indipendente di Esperti di Alto Livello sull'intelligenza Artificiale, *Orientamenti Etici per un'Intelligenza Artificiale affidabile* (8 aprile 2019), ec.europa.eu/newsroom/dae/document.cfm?doc_id=60430 (ultima consultazione, 20/06/2025).

16. Resseguier, Anaïs, e Fabienne Ufert. "AI research ethics is in its infancy: the EU's AI Act can make it a grown-up." *Research Ethics* 20.2 (2024): 146; Casiraghi, Simone, e Niels van Dijk. *op. cit*.

17. Ai sensi dell'art. 3 dell'AI Act, il «*deployer*» è una persona fisica o giuridica, un'autorità pubblica, un'agenzia o un altro organismo che utilizza un sistema di IA sotto la propria autorità, tranne nel caso in cui il sistema di IA sia utilizzato nel corso di un'attività personale non professionale.

18. La FRIA deve contenere: a) una descrizione dei processi del *deployer* in cui il sistema di IA ad alto rischio sarà utilizzato in linea con la sua finalità prevista; b) una descrizione del periodo di tempo entro il quale ciascun sistema di IA ad alto rischio è destinato a essere utilizzato e con che frequenza; c) le categorie di persone fisiche e gruppi verosimilmente interessati dal suo uso nel contesto specifico; d) i rischi specifici di danno che possono incidere sulle categorie di persone fisiche o sui gruppi di persone individuati a norma della lettera c); e) una descrizione dell'attuazione delle misure di sorveglianza umana, secondo le istruzioni per l'uso; f) le misure da adottare qualora tali rischi si concretizzino, comprese le disposizioni relative alla governance interna e ai meccanismi di reclamo.

19. L'art. 2, par. 6 dell'AI Act stabilisce che "Il presente regolamento non si applica ai sistemi di IA o modelli di IA, ivi compresi i loro output, specificamente sviluppati e messi in servizio al solo scopo di ricerca e sviluppo scientifici".

volta a non ostacolare la libertà scientifica e l'innovazione, evidentemente riconosce la natura unica della ricerca scientifica, che spesso comporta sperimentazione, test di ipotesi e attività innovative in condizioni che potrebbero non essere in linea con i quadri normativi tradizionali, o comunque troppo stringenti. Tuttavia, solleva una serie di questioni critiche.

Innanzitutto, l'interpretazione della clausola "al solo scopo", connesso alla ricerca e allo sviluppo scientifici: la locuzione suggerisce un intento di delimitazione rigorosa, ma rimane aperta alla soggettività delle dichiarazioni d'intento e alla variabilità dei contesti applicativi. In concreto, sarà spesso difficile distinguere nettamente tra ricerca scientifica "pura" e ricerca finalizzata a sviluppi commerciali o applicazioni operative, specie nel contesto delle collaborazioni pubblico-private o delle attività di trasferimento tecnologico. La norma sembra concedere, dunque, solo un'apparente flessibilità ai ricercatori, se sol si considerano le fisiologiche implementazioni commerciali, o comunque applicative (basti pensare ai contesti clinici), di strumenti di IA sviluppati nell'ambito di progetti di ricerca[20].

Al contempo, l'assenza di criteri oggettivi per verificare tale finalità esclusiva può aprire la strada a fenomeni di *forum shopping* e di elusione delle tutele previste per i sistemi di IA ad alto rischio, giustificandone lo sviluppo iniziale sotto l'etichetta della ricerca scientifica.

Un altro nodo problematico riguarda il ruolo dell'etica e della governance nella fase di ricerca. L'AI Act, incentrandosi sulle applicazioni pratiche, rinuncia a disciplinare la fase in cui si costruiscono le fondamenta tecnologiche che daranno forma ai sistemi di IA. Ciò comporta il rischio di posticipare le tutele a un momento in cui gli effetti dei sistemi sono già difficilmente reversibili e, soprattutto, il loro *assessment* non sia più possibile in virtù del già avvenuto e consolidato sviluppo e addestramento.

In questo senso, come meglio si dirà successivamente, sarebbe auspicabile prevedere comunque meccanismi minimi di *accountability* anche per la ricerca scientifica, alla stregua delle linee guida sulla *trustworthy AI*, come ad esempio l'obbligo di una sorta di valutazione d'impatto etico.

Infine, l'esclusione totale della ricerca scientifica dalla disciplina dell'AI Act potrebbe produrre una pericolosa frammentazione a livello europeo. La ricerca scientifica, pur se teoricamente confinata in ambienti controllati, può produrre impatti concreti e immediati: si pensi alla generazione di output altamente realistici, alla manipolazione di dati biometrici o all'addestramento di modelli con dataset sensibili. L'assenza di un quadro giuridico vincolante rischia, dunque, di aprire uno spazio di irresponsabilità in cui gli effetti dannosi possono emergere prima ancora della commercializzazione di un sistema. In mancanza di norme uniformi, gli Stati membri saranno liberi di adottare approcci divergenti, con evidenti ricadute sul piano della certezza del diritto, della concorrenza tra centri di ricerca e della protezione dei diritti fondamentali. La mancanza di armonizzazione si scontra, inoltre, con la natura transnazionale della scienza e delle tecnologie digitali, rendendo inefficaci gli strumenti nazionali nel prevenire o correggere abusi.

2.2 Intelligenza Artificiale e protezione dei dati personali

Sotto il profilo privacy e *data protection*, anche le norme del GDPR sulla profilazione e sul processo decisionale automatizzato (articoli 13, 14 e 22) impattano in maniera cruciale quando le attività di ricerca, o il loro risultato (ad esempio, un prodotto), prevedono un trattamento di dati personali. Ad esempio, gli strumenti medici basati su Intelligenza Artificiale possono comportare l'assunzione di decisioni (interamente o parzialmente) automatizzate sulla salute dei pazienti reclutati per uno studio, come un software di imaging medico che determini quali immagini contengano evidenze tumorali. O, prim'ancora, assumono rilievo le fasi di training dei sistemi, basate su grandi quantità di dati (anche) personali.

Nel 2017, l'Information Commissioner's Office (ICO), l'autorità britannica per la tutela dei dati personali, quando ancora il Regno Unito rientrava tra gli Stati membri di quell'Unione Europea che di lì a poco avrebbe applicato GDPR, aveva individuato i principali aspetti distintivi dei trattamenti di dati personali su larga scala legati all'intelligenza artificiale e alla cosiddetta "*big data analytics*" rispetto alle forme più tradizionali di elaborazione. Questi aspetti, tutti potenzialmente rilevanti per la privacy e la protezione dei dati, includono

20. Resseguier, Anaïs, e Fabienne Ufert, *op. cit.*, 149.

l'impiego di algoritmi complessi, la scarsa trasparenza dei processi, la tendenza ad acquisire un volume elevatissimo di informazioni, il riutilizzo dei dati raccolti e l'adozione di nuove tipologie di dati, spesso provenienti da dispositivi piuttosto che direttamente dagli utenti[21].

Riportando nel contesto della ricerca scientifica quanto chiarito dal Garante per la protezione dei dati personali in relazione alle applicazioni di IA in sanità[22], è innanzitutto evidente che le informazioni fornite agli interessati ai sensi degli articoli 13 e 14 del GDPR non possano ritenersi sufficienti.

Il titolare del trattamento ha l'obbligo di informare l'interessato sulla conduzione di attività di tal genere, con la fondamentale necessità di specificare se l'uso dei dati personali avvenga sin dalla fase di apprendimento dell'algoritmo oppure solo nella successiva fase applicativa. Al contempo, devono essere rappresentate le logiche e le caratteristiche di elaborazione dei dati e quali siano le conseguenze (ad esempio, gli eventuali vantaggi, in termini diagnostici e terapeutici, derivanti dall'utilizzo di tali nuove tecnologie), garantendo che le motivazioni alla base delle decisioni siano comprensibili.

Nel caso di processi decisionali automatizzati basati sul consenso, il titolare deve implementare misure adeguate a proteggere i diritti, le libertà e gli interessi legittimi degli interessati. Tra queste, il Gruppo di Lavoro *ex* art. 29, facendo riferimento al Considerando 71 del GDPR[23], ha sottolineato l'importanza di garantire almeno tre diritti fondamentali: richiedere l'intervento umano, esprimere il proprio punto di vista e contestare la decisione automatizzata.

La stessa letteratura ha delineato distintamente l'esistenza di un principio di conoscibilità dei dati, accompagnato dal diritto alla comprensibilità dei processi decisionali automatizzati e dei sistemi di intelligenza artificiale in generale[24]. Solo un individuo adeguatamente informato, in modo chiaro e accessibile, può esercitare pienamente il proprio diritto di accesso, esprimere un'opinione informata sulla decisione, opporsi se necessario, individuare eventuali errori e richiedere correzioni. Queste misure non solo tutelano l'interessato, ma contribuiscono anche a migliorare l'affidabilità e la correttezza dei processi decisionali automatizzati, favorendone un'applicazione più responsabile ed efficace.

---

21. Information Commissioner's Office, Big data, artificial intelligence, machine learning and data protection, https://ico.org.uk/media2/migrated/2013559/big-data-ai-ml-and-data-protection.pdf (ultima consultazione, 20/06/2025).

22. Garante per la Protezione dei Dati Personali, *Decalogo per la realizzazione di servizi sanitari nazionali attraverso sistemi di Intelligenza Artificiale*, https://www.garanteprivacy.it/documents/10160/0/Decalogo+per+la+realizzazione+di+servizi+sanitari+nazionali+attraverso+sistemi+di+Intelligenza+Artificiale.pdf/a5c4a24d-4823-e014-93bf-1543f1331670?version=2.0 (ultima consultazione, 20/06/2025).

23. Il Considerando 71 del GDPR specifica, inequivocabilmente, che «tale trattamento dovrebbe essere subordinato a garanzie adeguate, che dovrebbero comprendere la specifica informazione all'interessato e il diritto di ottenere l'intervento umano, di esprimere la propria opinione, di ottenere una spiegazione della decisione conseguita dopo tale valutazione e di contestare la decisione. Tale misura non dovrebbe riguardare un minore. Al fine di garantire un trattamento corretto e trasparente nel rispetto dell'interessato, tenendo in considerazione le circostanze e il contesto specifici in cui i dati personali sono trattati, è opportuno che il titolare del trattamento utilizzi procedure matematiche o statistiche appropriate per la profilazione, metta in atto misure tecniche e organizzative adeguate al fine di garantire, in particolare, che siano rettificati i fattori che comportano inesattezze dei dati e sia minimizzato il rischio di errori e al fine di garantire la sicurezza dei dati personali secondo una modalità che tenga conto dei potenziali rischi esistenti per gli interessi e i diritti dell'interessato e che impedisca tra l'altro effetti discriminatori nei confronti di persone fisiche sulla base della razza o dell'origine etnica, delle opinioni politiche, della religione o delle convinzioni personali, dell'appartenenza sindacale, dello status genetico, dello stato di salute o dell'orientamento sessuale, ovvero che comportano misure aventi tali effetti. Il processo decisionale automatizzato e la profilazione basati su categorie particolari di dati personali dovrebbero essere consentiti solo a determinate condizioni».

24. Palmirani, Monica. "Big Data e conoscenza." *Rivista di filosofia del diritto* IX.1 (2020): 85-88, che riconosce «accanto al diritto alla *spiegabilità* dell'algoritmo e della decisione automatica finale (ossia dell'esito) anche il principio della *conoscibilità* dei dati non tanto e non solo quelli che sono stati contribuiti o osservati dall'utente, ma anche quelli che hanno contribuito al processo decisionale quindi quelli inferiti, derivati, collettivi, statistici, anche se anonimi»; Pagallo, Ugo. "Algoritmi e conoscibilità." *Rivista di filosofia del diritto* IX.1 (2020); Malgieri, Gianclaudio. "Automated decision-making in the EU Member States: The right to explanation and other"suitable safeguards" in the national legislations." *Computer Law & Security Review* 35.5 (2019): 3-5. Si veda anche Malgieri, Gianclaudio. " 'Just' Algorithms: Justification (Beyond Explanation) of Automated Decisions Under the General Data Protection Regulation." *Law and Business* 1.1 (2021): 19-20, secondo il quale in processi decisionali automatizzati più complessi, basati sull'intelligenza artificiale, potrebbe essere difficile raggiungere un adeguato livello di spiegabilità, affrontando le cause, i fattori determinanti e i controfattuali. Una spiegazione né causale né contestuale è ritenuta inadeguata a mostrare all'interessato possibili motivi di impugnazione della decisione e quindi non è idonea ai sensi dell'articolo 22, paragrafo 3 del GDPR. Per superare questo limite della spiegazione dell'algoritmo, l'Autore propone la giustificazione della decisione automatizzata, attraverso cui spiegare non solo la logica sottostante, ma anche perché sia legalmente accettabile e conforme al GDPR.

Un ulteriore elemento di rilievo riguarda il riconoscimento e la tutela dei diritti dei soggetti interessati, considerate come partecipanti umani nei processi di ricerca, soprattutto quando l'intelligenza artificiale è utilizzata come base per decisioni automatizzate[25], inclusa la profilazione[26]. Il GDPR stesso evidenzia i potenziali rischi significativi per i diritti e le libertà individuali, derivanti dalla natura spesso poco trasparente dei processi, degli algoritmi e dei sistemi automatizzati. Questa opacità può rendere difficile per gli interessati comprendere i meccanismi alla base delle decisioni e, di conseguenza, intervenire per tutelare i propri diritti, se necessario[27].

In particolare, l'articolo 22, paragrafo 1 del GDPR[28] stabilisce che "l'interessato ha il diritto di non essere sottoposto a una decisione basata unicamente sul trattamento automatizzato, inclusa la profilazione, che produca effetti giuridici rilevanti o che incida in modo significativo sulla sua persona". Lo stesso articolo, al paragrafo 2, contempla alcune deroghe a tale diritto: se la decisione automatizzata è necessaria per la conclusione o l'esecuzione di un contratto tra l'interessato e un titolare del trattamento, se essa sia autorizzata dal diritto dell'Unione o dello Stato membro del titolare (purché preveda misure adeguate per la protezione dei diritti, delle libertà e dei legittimi interessi dell'interessato) o se si basi sul consenso esplicito dell'interessato. Tuttavia, resta comunque centrale quanto previsto dal paragrafo 3, che obbliga all'adozione di "misure appropriate per tutelare i diritti, le libertà e gli interessi legittimi dell'interessato". Tali misure comprendono almeno la possibilità di ottenere l'intervento umano da parte del responsabile del trattamento, di esprimere il proprio punto di vista e di contestare la decisione automatizzata. Infine, il paragrafo 4 dello stesso articolo affronta il caso specifico delle decisioni automatizzate che coinvolgono categorie particolari di dati personali, ritenendole ammissibili solo laddove avvengano con il consenso esplicito dell'interessato oppure per motivi di interesse pubblico rilevante, e siano necessariamente accompagnate da "misure adeguate a salvaguardare i diritti, le libertà e gli interessi legittimi dell'interessato".

La decisione in questione deve essere derivata da un processo interamente automatizzato e avere un impatto giuridicamente rilevante o influire in modo sostanziale sulla vita di una persona[29]. Diversamente, se il processo decisionale includesse anche un contributo umano, seppur parziale e in ipotesi anche non determinante, la restrizione prevista dall'articolo 22 del GDPR non troverebbe applicazione, con potenziali conseguenze indesiderate sui soggetti interessati[30].

25. Gruppo di Lavoro ex art. 29, *Linee guida sul processo decisionale automatizzato relativo alle persone fisiche e sulla profilazione ai fini del regolamento 2016/679* (versione del 6 febbraio 2018), secondo cui il «processo decisionale esclusivamente automatizzato consiste nella capacità di prendere decisioni impiegando mezzi tecnologici senza coinvolgimento umano». Per un approfondimento sui processi decisionali automatizzati basati sull'intelligenza artificiale, si veda Araujo, Theo, et al. "In AI we trust? Perceptions about automated decision-making by artificial intelligence." *AI & Society* 35 (2020): 611-623.

26. Art. 4, punto 4 GDPR: «la profilazione è qualsiasi forma di trattamento automatizzato di dati personali consistente nell'utilizzo di tali dati personali per valutare determinati aspetti personali relativi a una persona fisica, in particolare per analizzare o prevedere aspetti riguardanti il rendimento professionale, la situazione economica, la salute, le preferenze personali, gli interessi, l'affidabilità, il comportamento, l'ubicazione o gli spostamenti di detta persona fisica».

27. Si veda, per un'estesa analisi della relazione tra il GDPR e l'intelligenza artificiale, Sartor, Giovanni, e Francesca Lagioia. *The impact of the General Data Protection Regulation (GDPR) on artificial intelligence*, Scientific Foresight Unit (STOA), European Parliamentary Research Service, 2020. https://www.europarl.europa.eu/RegData/etudes/STUD/2020/641530/EPRS_STU(2020)641530_EN.pdf (ultima consultazione, 20/06/2025).

28. Bygrave, Lee A. "Article 22. Automated individual decision-making, including profiling". Kuner, Christopher and Bygrave, Lee A. and Docksey, Christopher and Docksey, Christopher and Drechsler, Laura and Tosoni, Luca, *The EU General Data Protection Regulation: A Commentary/Update of Selected Articles* (2021): 96-100.

29. Il Gruppo di Lavoro ex art. 29, nelle già menzionate *Linee guida sul processo decisionale automatizzato*, chiarisce, comunque, che «il titolare del trattamento non può eludere le disposizioni dell'articolo 22 creando coinvolgimenti umani fittizi. Ad esempio, se qualcuno applica abitualmente profili generati automaticamente a persone fisiche senza avere alcuna influenza effettiva sul risultato, si tratterà comunque di una decisione basata unicamente sul trattamento automatico. Per aversi un coinvolgimento umano, il titolare del trattamento deve garantire che qualsiasi controllo della decisione sia significativo e non costituisca un semplice gesto simbolico. Il controllo dovrebbe essere effettuato da una persona che dispone dell'autorità e della competenza per modificare la decisione».

30. Sapienza, Salvatore. *Decisioni algoritmiche e diritto*. Giuffrè Francis Lefebvre, 2024, il quale introduce un'interessante tassonomia delle decisioni algoritmicamente mediate, individuando effettive differenze nello spettro delle possibili interazioni uomo-macchina derivanti da tendenze opposte, autonomia funzionale dell'algoritmo *versus* controllo funzionale sull'algoritmo e supporto dell'algoritmo *versus* delegazione all'algoritmo. Sulla base di tale sistema, Sapienza ha identificato le seguenti classi di processi decisionali: processo decisionale analogico, in cui non è coinvolto un sistema informatico; processo decisionale aritmetico, in cui l'algoritmo contribuisce a un più complesso processo decisionale umano; processo decisionale automatico, in cui l'essere umano ha il pieno controllo sul funzionamento del processo decisionale, ma la decisione è presa dall'algoritmo; processo decisionale assistito, in cui l'essere umano non controlla il funzionamento del processo decisionale, ma adotta la decisione finale a cui l'algoritmo ha

3 Le criticità nella ricerca scientifica

I vincoli e gli obblighi sopra identificati, essendo di natura generale, si applicano inevitabilmente anche ai ricercatori e agli enti di ricerca che intendono sviluppare o utilizzare sistemi decisionali automatizzati, specialmente tramite l'intelligenza artificiale. Sebbene tali limitazioni possano costituire un ostacolo rilevante per l'adozione dell'IA, l'impatto sulla ricerca scientifica risulta in parte mitigato, poiché l'obiettivo primario è principalmente – sebbene non esclusivamente – generare nuova conoscenza, piuttosto che adottare decisioni automatizzate che influenzino direttamente gli individui[31].

È senz'altro possibile che, nel corso di attività di ricerca e sviluppo, vengano progettati e addestrati sistemi di intelligenza artificiale in grado di effettuare decisioni automatizzate basate su dati personali, come nel caso di algoritmi che analizzano immagini radiografiche per identificare pazienti potenzialmente affetti da una specifica patologia. Tuttavia, un trattamento di ricerca o sperimentazione rientra nell'ambito di applicazione dell'articolo 22 del GDPR unicamente se la decisione assunta sia interamente automatizzata. Ciò non avviene, ad esempio, laddove un sistema si limitasse a fornire informazioni avanzate e innovative per un supporto in una decisione finale comunque deputata a un essere umano, che rimarrebbe dunque determinante.

In caso di sistema di intelligenza artificiale che effettivamente contempli un processo decisionale interamente automatizzato, l'applicazione dell'articolo 22 del GDPR comporta una serie di obblighi (informativi e operativi) a carico dell'ente di ricerca che richiedono una piena e precoce comprensione delle dinamiche che regolano il sistema decisionale, poiché le informazioni dettagliate sul funzionamento, la logica impiegata e le conseguenze per l'interessato dovrebbero essere comunicate all'atto della raccolta dei dati, o al più tardi entro 30 giorni ai sensi dell'art. 14 del GDPR.

Tuttavia, è da tenere in debita considerazione la probabile condizione di vulnerabilità psico-fisica o sociale del partecipante. Basti pensare a minori, anziani, persone affette da disabilità, minoranze e pazienti, che già in partenza potrebbe essere propenso alla partecipazione laddove identificasse la ricerca sperimentale come prospettiva vantaggiosa o – evento non raro nella ricerca medica – come ultima possibilità di cura.

La specifica vulnerabilità della persona esposta al potere decisionale dell'apparato tecnologico è la chiave per comprendere il senso precettivo dell'art. 22 GDPR, e il diritto alla spiegazione comprensibile costituisce una garanzia e, al tempo stesso, la base giuridica per poter assoggettare la persona al potere decisionale dello strumento di IA. Rilevano, in particolare, le capacità cognitive di comprensione delle informazioni sull'attività di ricerca o sperimentazione, sul trattamento dei dati personali e, soprattutto, sul funzionamento e sulla logica del sistema di Intelligenza Artificiale sviluppato o adoperato.

In tale contesto, emergono criticità su più fronti. *In primis*, nella grande maggioranza dei casi, in special modo quelli riguardanti lo sviluppo e l'addestramento di sistemi di IA, la raccolta dei dati personali utili ad alimentare i sistemi interviene, presso i partecipanti, in un momento primordiale della ricerca progettuale, che potrebbe successivamente muoversi verso differenti direzioni e prevedere modifiche agli algoritmi. In tali circostanze, il consenso dell'interessato, che deve essere sempre informato, attuale e specifico[32], oltre che liberamente revocabile, impone ai ricercatori di prestare particolare attenzione agli obblighi informativi, garantendo aggiornamenti precisi e puntuali sulle caratteristiche del sistema.

In secondo luogo, a differenza dei ricercatori, i sistemi di IA e le reti neurali non possono spiegare il proprio pensiero e, quando non è possibile comprendere la logica seguita dall'algoritmo per raggiungere il proprio

contribuito; processo decisionale automatizzato, in cui l'essere umano non controlla funzionalmente il processo decisionale, interamente condotto e finalizzato dall'algoritmo. Alla luce di ciò, si condivide la preoccupazione dell'Autore sull'inapplicabilità degli strumenti di controllo ex art. 22 del GDPR, in quanto non interamente automatizzati, ai processi decisionali assistiti da algoritmi dotati di autonomia funzionale e dunque in grado di sfuggire al controllo umano.

31. Meszaros, Janos, e Chih-Hsing Ho. "AI research and data protection: Can the same rules apply for commercial and academic research under the GDPR?" *Computer Law & Security Review* 41 (2021): 5.

32. Art. 4, n. 11) GDPR: «consenso dell'interessato»: qualsiasi manifestazione di volontà libera, specifica, informata e inequivocabile dell'interessato, con la quale lo stesso manifesta il proprio assenso, mediante dichiarazione o azione positiva inequivocabile, che i dati personali che lo riguardano siano oggetto di trattamento».

obiettivo (c.d. "*black box*"), il sistema diventa oscuro, impenetrabile dal controllo esterno, con conseguente complicazione di supervisione e revisione etica[33].

Inoltre, eventuali inesattezze o alterazioni nei dati acquisiti o diffusi, così come nel funzionamento dei sistemi automatizzati, potrebbero portare a conclusioni errate dal punto di vista scientifico e a giudizi basati su previsioni errate, con possibili impatti negativi sui soggetti coinvolti. Questo rischio è particolarmente rilevante quando il sistema viene testato nell'ambito di una sperimentazione clinica necessaria per ottenere la certificazione come dispositivo medico, prima di un suo effettivo utilizzo in ambito diagnostico. In tale fase, eventuali errori predittivi o bias nei dati potrebbero influenzare negativamente i risultati della ricerca e avere ripercussioni sui soggetti coinvolti nella sperimentazione. Per questo motivo, i ricercatori dovrebbero adottare un meccanismo di verifica periodica dei dataset utilizzati, correggendo tempestivamente eventuali anomalie e rivalutando la precisione e la rilevanza dei dati, oltre a riesaminare il design e l'implementazione del sistema.

A tutto ciò si può aggiungere la potenziale capacità dei sistemi di IA di aumentare l'identificabilità di dati apparentemente anonimi, poiché consentono di collegare dati non identificati (inclusi dati anonimizzati o pseudonimizzati) alle persone interessate[34].

La reidentificazione può essere vista come uno specifico tipo di inferenza di dati personali, attraverso cui un identificatore personale è associato a dati precedentemente non identificati, che, di conseguenza, diventano dati personali. Inoltre, i sistemi di IA possono dedurre, mediante inferenze, nuove informazioni sui soggetti interessati, applicando modelli algoritmici ai loro dati personali già conosciuti. Qualificando le informazioni ricavate come nuovi dati personali, come si desume dal Parere 4/2007 sul concetto di dati personali[35] e dalle Linee guida sul processo decisionale automatizzato relativo alle persone fisiche e sulla profilazione[36] dell'allora Gruppo di lavoro "Art. 29", le inferenze automatizzate attivano tutti gli adempimenti che il trattamento dei dati personali comporta ai sensi del GDPR (e delle eventuali normative nazionali): la necessità di una base giuridica, le condizioni per il trattamento dei dati sensibili o per un processo decisionale automatizzato, i diritti dell'interessato[37].

Di conseguenza, anche un riutilizzo massivo di dati apparentemente anonimi, già frutto di elaborazioni da precedenti attività di ricerca scientifica, può avere ricadute sull'attuazione delle norme etico-giuridiche.

Trattandosi, in tal caso, di evidente trattamento di dati personali, lo sviluppo del sistema dovrebbe aderire ai principi chiave della tutela dei dati, tra cui in particolare la *privacy by design*, che richiede l'integrazione di misure di sicurezza fin dalla progettazione, e la *privacy by default*, che assicura la protezione dei dati durante l'intero ciclo di vita del sistema. Devono essere rispettati anche i principi di legalità, correttezza e trasparenza, insieme alla minimizzazione e all'accuratezza dei dati trattati.

4 Possibili soluzioni di *assessment* etico-giuridico in ottica di *ethics by design*

La traduzione in pratica dei requisiti e delle condizioni di carattere etico-giuridico, spesso oggetto di controllo da parte delle istituzioni finanziatrici i progetti di ricerca, costituisce, specialmente in assenza di approcci

33. Fasan Marta. *Intelligenza artificiale e costituzionalismo contemporaneo: principi, diritti e modelli in prospettiva comparata*. Università degli Studi di Trento, 2024, 77-84. Scaffardi, Lucia. "La medicina alla prova dell'Intelligenza Artificiale: Medicine to the test of Artificial Intelligence." *DPCE Online*, 51.1 (2022): 352.

34. Sartor, Giovanni, e Francesca Lagioia. *The impact of the General Data Protection Regulation (GDPR) on artificial intelligence*, *op. cit*.

35. https://www.garanteprivacy.it/documents/10160/10704/ARTICOLO+29+-+WP+136.pdf/339f9753-f2bc-41ed-b720-0e12f0a56801?version=1.1 (ultima consultazione, 20/06/2025).

36. https://ec.europa.eu/newsroom/article29/items/612053 (ultima consultazione, 20/06/2025).

37. Sartor, Giovanni, e Francesca Lagioia. *The impact of the General Data Protection Regulation (GDPR) on artificial intelligence*, *op. cit*.

condivisi e validati per la creazione di sistemi di intelligenza artificiale affidabili[38], una delle attuali sfide per i ricercatori e i relativi enti di appartenenza.

Per soddisfare le complesse e costanti esigenze di trasparenza e adeguatezza dei sistemi di IA nella ricerca scientifica, sono emersi, nella recente letteratura, metodologie e modelli atti a indirizzare utilmente la progettazione, lo sviluppo e l'implementazione di tali sistemi verso il rispetto dei principi etici in questione[39].

Alcuni approcci si focalizzano sugli interventi nelle fasi iniziali dello sviluppo, promuovendo una maggiore sensibilizzazione degli sviluppatori alle problematiche etiche[40], la creazione di team eterogenei e multidisciplinari[41], l'integrazione di valori morali nei sistemi tramite una progettazione proattiva[42], nonché il controllo dei modelli decisionali e del codice su cui si basano[43]. Altri strumenti, come le valutazioni d'impatto[44], analizzano gli effetti generati dall'utilizzo dei sistemi decisionali automatizzati o si concentrano sul contesto, prevedendo la partecipazione diretta degli operatori umani[45]. Altri processi, invece, adottano un approccio fondato sull'etica, confrontando costantemente i sistemi con i principi e le regole pertinenti per monitorarne o orientarne il comportamento[46].

Sotto il profilo più strettamente legato alla trasparenza dei sistemi di IA[47], e dunque alla conoscibilità e al controllo, modelli statici di *eXplainable AI* (XAI) rischiano di non soddisfare adeguatamente le esigenze di tutela dei diritti e delle libertà degli interessati coinvolti, anche sotto un profilo organizzativo e di efficienza delle attività di ricerca. È stato difatti osservato in letteratura come tali modelli, incentrati su un approccio generalizzato (definito *one-size-fits-all*), siano incapaci di illustrare in maniera pragmatica e incentrata sull'utente il processo decisionale automatizzato, contrariamente a quanto in realtà sarebbe richiesto dal GDPR, dalle linee guida del Gruppo di Esperti di Alto Livello sull'Intelligenza Artificiale e ora dall'AI Act, ossia informazioni chiare e soprattutto comprensibili dal singolo individuo destinatario del trattamento o dell'applicazione del sistema[48]. Questi strumenti basati su XAI forniscono, in maniera generale e per qualunque utente, informazioni su perché, come e cosa, presupponendo che siano fruibili e utili a priori per chiunque[49]. Ciò, tuttavia, può non essere l'ideale per soddisfare il fondamentale requisito della spiegabilità e della comprensione da parte

38. Mittelstadt, Brent. "Principles alone cannot guarantee ethical AI." *Nature Machine Intelligence* 1.11 (2019): 501-507.

39. Mökander, Jakob, et al. "Ethics-Based Auditing of Automated Decision-Making Systems: Nature, Scope, e Limitations." *Science and Engineering Ethics* 27.4 (2021): 3-4.

40. Floridi, Luciano, et al. "AI4People—An ethical framework for a good AI society: Opportunities, risks, principles, e recommendations." *Minds and Machines* 28.4 (2018): 689-707.

41. Sánchez-Monedero, Javier, Lina Dencik, e Lilian Edwards. "What does it mean to 'solve' the problem of discrimination in hiring? Social, technical and legal perspectives from the UK on automated hiring systems." *FAT* '20: Proceedings of the 2020 Conference on Fairness, Accountability, e Transparency* (2020): 458-468.

42. Aizenberg, Evgeni, e Jeroen van den Hoven. "Designing for human rights in AI." *Big Data & Society* 7.2 (2020); van de Poel, Ibo. "Embedding Values in Artifcial Antelligence (AI) Systems." *Minds and Machines* 30.3 (2020): 385-409.

43. Dennis, Louise A., et al. "Practical verification of decision-making in agent-based autonomous systems." *Automated Software Engineering* 23.3 (2016): 305-359.

44. ECP, *Artifcial intelligence impact assessment*, 2018, ecp.nl/wp-content/uploads/2019/01/Artificial-Intelligence-Impact-Assessment-English.pdf (ultima consultazione, 20/06/2025).

45. Jotterand, Fabrice, e Clara Bosco. "Keeping the"Human in the Loop" in the Age of Artificial Intelligence: Accompanying Commentary for "Correcting the Brain?" by Rainey and Erden." *Science and Engineering Ethics* 26.5 (2020): 2455-2460; Rahwan, Iyad. "Society-in-the-loop: programming the algorithmic social contract." *Ethics and Information Technology* 20.1 (2018): 5-14.

46. Mökander, Jakob, et al. *op. cit.*, 9-17; Zicari, Roberto V., et al. "Z-Inspection®: A Process to Assess Trustworthy AI." *IEEE Transactions on Technology and Society* 2.2 (2021): 83-97.

47. Per una panoramica sulla spiegabilità dei sistemi IA, si veda in particolare Santosuosso, Amedeo e Giovanni Sartor. *Decidere con l'IA. Intelligenze artificiali e naturali nel diritto*. Il Mulino, 2024, 132-139.

48. Sapienza, Salvatore. *op. cit.*, il quale approfondisce altresì il modello di conoscibilità dell'algoritmo proposto dall'AI Act (ancora nella versione di proposta di regolamento), rapportandolo alla spiegabilità dell'algoritmo nel contesto del GDPR e riconoscendone una triplice portata: abilitante (a vantaggio del supervisore), attestativa o conformativa (a supporto delle istituzioni competenti per la verifica dell'adeguamento del sistema) e azionabile (a beneficio dei soggetti destinatari delle decisioni algoritmiche). Tale tripartizione, che corrisponde ad esigenze e condizioni distinte, evidenzierebbe, dunque, l'impraticabilità di un approccio *one-size-fits-all*.

49. Sovrano, Francesco, e Fabio Vitali. "Explanatory artificial intelligence (YAI): human-centered explanations of explainable AI and complex data." *Data Mining and Knowledge Discovery* 38 (2024): 3141-3168.

dei soggetti interessati. La XAI, difatti, pur ampliando la leggibilità dei modelli, può fallire nel fornire spiegazioni effettivamente utili per gli utenti finali, specialmente quando questi non possiedano competenze tecniche adeguate[50].

I modelli di *explanatorY AI* (YAI), pur basandosi sulla spiegabilità, intendono colmare tale gap e muovere oltre, ridefinendo il concetto di spiegazione in senso relazionale, dinamico e comunicativo, e dunque organizzando e articolando tutte le informazioni spiegabili in narrazioni centrate sull'utente all'interno di uno spazio esplicativo. Tutto ciò affinché sia l'utente stesso, nell'esplorare tale spazio esplicativo in modo interattivo attraverso un'apposita interfaccia, a prodursi la spiegazione più adatta alle proprie necessità[51].

Rispetto alla XAI, il paradigma della YAI sposta l'attenzione dal modello all'interazione con l'utente, configurandosi come un'intelligenza artificiale che non si limita a essere spiegabile, ma che è, in senso proprio, in grado di spiegare.

Tale riconfigurazione comporta una trasformazione strutturale e funzionale nell'architettura dei sistemi intelligenti, i quali non vengono più concepiti unicamente come strumenti soggetti a tecniche di interpretabilità *ex post*, bensì come agenti autonomi dotati della capacità di generare spiegazioni contestuali, orientate alla persona e sensibili al contesto. Lo YAI si articola, infatti, come un modello dialogico e interattivo, capace di produrre giustificazioni non solo *ex post*, ma anche in itinere rispetto al processo decisionale, adattando il contenuto esplicativo al livello cognitivo, al ruolo e alle aspettative dell'interlocutore umano. Questa caratteristica la distingue profondamente dall'approccio XAI, che, nella maggior parte delle sue implementazioni, rimane fortemente legato a tecniche descrittive e visuali, come le rappresentazioni grafiche delle salienze, l'analisi delle feature, oppure l'utilizzo di strumenti, che, pur avendo una funzione di trasparenza, risultano spesso inaccessibili a utenti non esperti o comunque esterni al dominio tecnico-scientifico.

La YAI rappresenta, pertanto, una risorsa prospettica di particolare rilevanza anche nel contesto della ricerca scientifica, in quanto consente di migliorare la tracciabilità, la riproducibilità e la verificabilità delle scelte operate da modelli computazionali all'interno del processo.

La fornitura delle informazioni all'interessato, nonché le attività di adeguamento del sistema di IA sotto il profilo etico, seguendo uno (o più) dei metodi disponibili, potrebbero però richiedere tempo e risorse anche oltre le possibilità. Ciò è tanto più vero nel contesto della ricerca scientifica, ove solitamente i cronoprogrammi – per lo meno nella ricerca competitiva finanziata – sono piuttosto tassativi.

Tuttavia, l'etica e la protezione dei dati personali nella ricerca sono divenuti aspetti imprescindibili, pretesi e verificati, come già ricordato, dalle stesse istituzioni finanziatrici. Di conseguenza, adottando un approccio di *ethics by design* analitico e interdisciplinare, i ricercatori devono mettere debitamente in conto il tempo, le risorse e la metodologia opportuna per condurre efficaci azioni di *assessment* etico-giuridico dei sistemi di IA e di *explicability by design*.

Bibliography

AI4People Institute. *Towards an Ethics by Design Appoach for AI*, 2024.

Aizenberg, Evgeni, e Jeroen van den Hoven. "Designing for human rights in AI." *Big Data & Society* 7.2 (2020).

50. Doshi-Velez, Finale, e Been Kim, "Towards A Rigorous Science of Interpretable Machine Learning." *arxiv:1702.08608* (2017).

51. La fornitura di spiegazioni orientate agli obiettivi dell'utente implica la spiegazione dei soli fatti rilevanti per l'utente, in base alle sue conoscenze, interessi e altre peculiarità che lo rendono unico, con esigenze altrettanto uniche e suscettibili di mutevolezza nel tempo. Per approfondimenti su *explanatorY AI* (YAI) e sul rapporto con *eXplainable AI* (XAI), si vedano in particolare: Palmirani, Monica. "Interpretabilità, conoscibilità, spiegabilità dei processi decisionali automatizzati." *XXVI lezioni di Diritto dell'Intelligenza Artificiale*. Giappichelli, Torino, 2021. 66-79; Sovrano, Francesco, e Fabio Vitali, *op. cit.*; Sovrano, Francesco, Fabio Vitali, e Monica Palmirani. "Making Things Explainable vs Explaining: Requirements and Challenges Under the GDPR." *AI Approaches to the Complexity of Legal Systems XI-XII. AICOL AICOL XAILA 2020 2018 2020. Lecture Notes in Computer Science()* 13048 (2021): 169-182.

Araujo, Theo, et al. "In AI we trust? Perceptions about automated decision-making by artificial intelligence." *AI & Society* 35 (2020): 611-623.

Brey, Philip e Brandt Dainow. "Ethics by design for artificial intelligence." *AI and Ethics* 4 (2024): 1265-1277.

Bygrave, Lee A. "Article 22. Automated individual decision-making, including profiling". Kuner, Christopher and Bygrave, Lee A. and Docksey, Christopher and Docksey, Christopher and Drechsler, Laura and Tosoni, Luca, *The EU General Data Protection Regulation: A Commentary/Update of Selected Articles* (2021): 96-100.

Casiraghi, Simone, e Niels van Dijk. "Ethics reviews in the European Union. Implications for the governance of scientific research in times of data science and Artificial Intelligence." *Law, Innovation and Technology* 16.1. (2024): 101-122.

d'Aquin, Mathieu, Troullinou, Pinelopi, O'Connor, Noel, Cullen, Aindrias, Faller, Gráinne e Louise Holden. "Towards an 'Ethics by Design' Methodology for AI Research Projects." *AIES '18: Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society* (2018): 54-59.

Dennis, Louise A., et al. "Practical verification of decision-making in agent-based autonomous systems." *Automated Software Engineering* 23.3 (2016): 305-359.

Dignum, Virginia, Baldoni, Matteo, Baroglio, Cristina et al. "Ethics by design: Necessity or curse?" *AIES '18: Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society* (2018): 60-66.

Doshi-Velez, Finale, e Been Kim, "Towards A Rigorous Science of Interpretable Machine Learning." *arxiv:1702.08608* (2017).

Fasan Marta. *Intelligenza artificiale e costituzionalismo contemporaneo: principi, diritti e modelli in prospettiva comparata*. Università degli Studi di Trento, 2024, 77-84.

Floridi, Luciano. *Philosophy and Computing: An introduction*. Routledge, 1999.

Floridi, Luciano, et al. "AI4People—An ethical framework for a good AI society: Opportunities, risks, principles, e recommendations." *Minds and Machines* 28.4 (2018): 689-707.

González-Esteban, Elsa, e Patrici Calvo. "Ethically governing artificial intelligence in the field of scientific research and innovation." *Heliyon*. 8.2 (2022).

Jotterand, Fabrice, e Clara Bosco. "Keeping the"Human in the Loop" in the Age of Artificial Intelligence: Accompanying Commentary for "Correcting the Brain?" by Rainey and Erden." *Science and Engineering Ethics* 26.5 (2020): 2455-2460.

Krenn, Mario, et al. "On scientific understanding with artificial intelligence." *Nature Reviews Physics 4* (2022): 761-769.

Mai, Jens-Erik. "Big data privacy: The datafication of personal information." *The Information Society* 32.3 (2016): 192-199.

Malgieri, Gianclaudio. "Automated decision-making in the EU Member States: The right to explanation and other"suitable safeguards" in the national legislations." *Computer Law & Security Review* 35.5 (2019).

Malgieri, Gianclaudio. " 'Just' Algorithms: Justification (Beyond Explanation) of Automated Decisions Under the General Data Protection Regulation." *Law and Business* 1.1 (2021): 16-28.

Mayer-Schönberger, Viktor e Kenneth Cukier. *Big Data: A Revolution That Will Transform How We Live, Work and Think*. John Murray Publishers, 2013.

Meszaros, Janos, e Chih-Hsing Ho. "AI research and data protection: Can the same rules apply for commercial and academic research under the GDPR?" *Computer Law & Security Review* 41 (2021).

Metcalf, Jacob, e Kate Crawford. "Where are human subjects in Big Data research? The emerging ethics divide." *Big Data & Society* 3.1 (2016).

Mittelstadt, Brent. "Principles alone cannot guarantee ethical AI." *Nature Machine Intelligence* 1.11 (2019): 501-507.

Mökander, Jakob, et al. "Ethics-Based Auditing of Automated Decision-Making Systems: Nature, Scope, e Limitations." *Science and Engineering Ethics* 27.4 (2021).

Pagallo, Ugo. "Algoritmi e conoscibilità." Rivista di filosofia del diritto IX.1 (2020): 93-106.

Palmirani, Monica. "Interpretabilità, conoscibilità, spiegabilità dei processi decisionali automatizzati." *XXVI lezioni di Diritto dell'Intelligenza Artificiale*. Giappichelli, Torino, 2021. 66-79.

Palmirani, Monica. "Big Data e conoscenza." Rivista di filosofia del diritto IX.1 (2020): 73-92.

Rahwan, Iyad. "Society-in-the-loop: programming the algorithmic social contract." *Ethics and Information Technology* 20.1 (2018): 5-14.

Resnik, David B., e Mohammad Hosseini. "The ethics of using artificial intelligence in scientific research: new guidance needed for a new tool." *AI and Ethics* (2024).

Resseguier, Anaïs, e Fabienne Ufert. "AI research ethics is in its infancy: the EU's AI Act can make it a grown-up." Research Ethics 20.2 (2024).

Sánchez-Monedero, Javier, Lina Dencik, e Lilian Edwards. "What does it mean to 'solve' the problem of discrimination in hiring? Social, technical and legal perspectives from the UK on automated hiring systems." *FAT* '20: Proceedings of the 2020 Conference on Fairness, Accountability, e Transparency* (2020): 458-468.

Santosuosso, Amedeo e Giovanni Sartor. *Decidere con l'IA. Intelligenze artificiali e naturali nel diritto*. Il Mulino, 2024, 132-139.

Sapienza, Salvatore. *Decisioni algoritmiche e diritto*. Giuffrè Francis Lefebvre, 2024.

Sartor, Giovanni, e Francesca Lagioia. *The impact of the General Data Protection Regulation (GDPR) on artificial intelligence*, Scientific Foresight Unit (STOA), European Parliamentary Research Service, 2020.

Sartor, Giovanni and Francesca Lagioia. "Le decisioni algoritmiche tra etica e diritto." *Intelligenza artificiale - il diritto, i diritti, l'etica*. Giuffrè Francis Lefebvre, Milano, 2020.

Scaffardi, Lucia. "La medicina alla prova dell'Intelligenza Artificiale: Medicine to the test of Artificial Intelligence." *DPCE Online*, 51.1 (2022): 349-359.

Sovrano, Francesco, e Fabio Vitali. "Explanatory artificial intelligence (YAI): human-centered explanations of explainable AI and complex data." Data Mining and Knowledge Discovery 38 (2024): 3141-3168.

Sovrano, Francesco, Fabio Vitali, e Monica Palmirani. "Making Things Explainable vs Explaining: Requirements and Challenges Under the GDPR." *AI Approaches to the Complexity of Legal Systems XI-XII. AICOL AICOL XAILA 2020 2018 2020. Lecture Notes in Computer Science()* 13048 (2021): 169-182.

van de Poel, Ibo. "Embedding Values in Artifcial Antelligence (AI) Systems." *Minds and Machines* 30.3 (2020): 385-409.

van Dijck, José. "Datafication, dataism and dataveillance: Big Data between scientific paradigm and ideology." *Surveillance & Society* 12.2 (2014): 197-208

Wang, Hanchen, et al. "Scientific discovery in the age of artificial intelligence." *Nature* 620 (2023): 47-60.

Weinbaum, Cortney, et al. *Ethics in Scientific Research: An Examination of Ethical Principles and Emerging Topics*. Rand Corporation Publication, 2019.

Xue, Lan, e Zhenjing Pang. "Ethical governance of artificial intelligence: An integrated analytical framework." *Journal of Digital Economy*, 1.1 (2022): 44-52.

Zicari, Roberto V., et al. "Z-Inspection®: A Process to Assess Trustworthy AI." *IEEE Transactions on Technology and Society* 2.2 (2021): 83-97.