Intelligenza artificiale e dati sintetici: quando la tecnologia può diventare strumento a supporto della tutela dei diritti

Edoardo C. Raffiotta*

ARTIFICIAL INTELLIGENCE AND SYNTHETIC DATA: WHEN TECHNOLOGY CAN EMPOWER THE PROTECTION OF RIGHTS

ABSTRACT: Synthetic data is an innovative technology in the field of personal data protection and the development of artificial intelligence systems. Starting from a brief overview of the main technical elements of synthetic data, the article delves into its role within the new European digital regulatory framework. Although it is already possible to identify some of the undeniable advantages derived from this technology and the increasing consideration by legislators, there is still a need for further practical developments, and continuous dialogue between relevant stakeholders is encouraged to ensure a fair balance between innovation and the protection of fundamental rights.

KEYWORDS: Synthetic data; artificial intelligence; healthcare; personal data protection; anonymization

SOMMARIO: 1. Le problematiche etiche e giuridiche emergenti nell'economia dei dati e delle nuove tecnologie – 1.1. La descrizione dei dati sintetici e le tecniche di sintetizzazione – 1.2. *Use case* e possibili modelli virtuosi di utilizzo dei dati sintetici – 2. Analisi del contesto giuridico esistente alla luce delle disposizioni europee e nazionali: un'introduzione alla nozione giuridica di dato sintetico – 2.1. L'analisi della strategia europea sui dati – 2.2. L'intelligenza artificiale e i dati sintetici nello scenario della nuova regolazione europea – 3. Un caso emblematico: i dati sintetici e il settore sanitario – 3.1. Il nuovo Spazio europeo dei dati sanitari: l'*European Health Data Space* – 3.2. La nuova riforma del codice della privacy – 3.3. Il disegno di legge sull'intelligenza artificiale e le disposizioni in materia sanitaria – 4. Conclusioni finali e valutazione prospettiche sotto un profilo giuridico ed etico.

1. Le problematiche etiche e giuridiche emergenti nell'economia dei dati e delle nuove tecnologie

Un report della IDC del 2018 ha stimato che entro il 2025 il volume di dati generati a livello globale sarà di 175 *zettabyte*, rispetto ai 33 del 2018[2]: l'incremento esponenziale sarebbe dovuto, da un lato,

* *Professore associato di Diritto Pubblico e di Diritto dell'Innovazione e dell'Intelligenza Artificiale presso l'Università degli Studi di Milano-Bicocca. Mail: edoardo.raffiotta@unimib.it.* *La ricerca è stata condotta dall'Autore nell'ambito del PRIN (Progetto di Rilevante Interesse Nazionale) 2022 – "L'intelligenza artificiale da causa a strumento di contrasto delle discriminazioni di genere" (AiGeDi). Contributo sottoposto a doppio referaggio anonimo.*

[2] D. REINSEL, J. GANTZ, J. RYDNING, *The Digitization of the World from Edge to Core*, IDC, 2018.

all'aumento di strumenti e servizi digitali che producono dati e, dall'altro, dalla crescente capacità dei sistemi di intelligenza artificiale (IA) di analizzare i dati a disposizione e produrre una moltitudine di *output* diversi[3]. In effetti, nell'era digitale, la capacità di raccogliere, conservare e analizzare quantità massicce di dati provenienti da fonti eterogenee, come dispositivi IoT (acronimo di "*Internet of Things"),* social media, transazioni online, ha aperto nuove opportunità per l'innovazione ed il progresso tecnologico.

Le tecniche di *data analytics*, rese possibili da metodi di calcolo sempre più sofisticati, offrono notevoli vantaggi in una vasta gamma di settori, consentendo di sfruttare appieno il potenziale di enormi quantità di informazioni al fine di migliorare e ottimizzare una lunga serie di processi sia nel settore pubblico che in quello privato[4]. Si pensi, ad esempio, all'ambito sanitario, nel quale il ricorso ad algoritmi di apprendimento automatico per l'analisi di dati clinici e alle informazioni provenienti da dispositivi indossabili consente di migliorare la diagnosi e il monitoraggio delle patologie, oltre a supportare la ricerca e lo sviluppo di nuove terapie[5]. Ma le potenzialità dei *big data* si estendono, altresì, ad ulteriori e diversi rami delle politiche pubbliche consentendo di migliorare i procedimenti amministrativi e garantire una gestione più efficiente delle limitate risorse a disposizione[6]. Anche nel settore privato, ovviamente, le potenzialità dei dati sono innumerevoli: nel campo dei servizi finanziari, le tecniche di analisi dei dati permettono di rilevare frodi, ottimizzare la gestione dei rischi e offrire prodotti su misura rispetto alle esigenze dei clienti, mentre consentono alle funzioni di marketing di individuare con miglior precisione le preferenze dei consumatori e di personalizzare le strategie di comunicazione[7].

---

[3] T. COUGHLIN, *175 Zettabytes By 2025*, in *Forbes*, 2018, disponibile al seguente link: https://www.forbes.com/sites/tomcoughlin/2018/11/27/175-zettabytes-by-2025/ (ultima consultazione 08/05/2024).

[4] O. TENE, J. POLONETSKY, *Big Data for All: Privacy and User Control in the Age of Analytics*, in *Northwestern Journal of Technology and Intellectual Property*, 11, 2012, 239. Si veda anche F. BRUSCHI, V. RANA, A. PAGANI, D. SCIUTO, *Acknowledging Value of Personal Information: a Privacy Aware Data Market for Health and Social Research*, in *Proceedings of the 3rd Distributed Ledger Technology Workshop Co-located with ITASEC 2020*, Ancona, 2020.

[5] La letteratura tecnico-scientifica volta ad esaminare l'impatto dei *big data* nelle applicazioni medico-sanitarie è sterminata ed in costante aumento. Tra i molti, si rimanda a Y. WANG, L. KUNG, T.A. BYRD, *Big data analytics: Understanding its capabilities and potential benefits for healthcare organizations*, in *Technological Forecasting and Social Change*, 126, 2018, 3; N. MEHTA, A. PANDIT, *Concurrence of big data analytics and healthcare: A systematic review*, in *International Journal of Medical Informatics*, 114, 2018, 57.

[6] Oggi più che mai si avverte l'esigenza di trasformare ed ottimizzare i processi della Pubblica Amministrazione. Per una panoramica sull'impatto delle nuove tecnologie e l'impiego dei dati nel settore pubblico, si rimanda, senza pretesa di completezza, a: E.C. RAFFIOTTA, *L'erompere dell'intelligenza artificiale per lo sviluppo della Pubblica Amministrazione e dei servizi al cittadino*, in G. CERRINA FERONI, C. FONTANA, E.C. RAFFIOTTA (a cura di), *AI Anthology. Profili giuridici, economici e sociali dell'intelligenza artificiale*, Bologna, 2022, 191; P. FORTE, *Diritto amministrativo e data science. Appunti di Intelligenza Amministrativa Artificiale (AAI)*, in *P.A. Persona e Amministrazione*, 1, 2020, 247.; D.U. GALETTA, J.G. CORVALÀN, *Intelligenza artificiale per una Pubblica Amministrazione 4.0? Potenzialità, rischi e sfide della rivoluzione tecnologica in atto*, in *federalismi.it*, 3, 2019, 2; E. CHITI, B. MARCHETTI, N. RANGONE, *L'impiego di sistemi di nelle pubbliche amministrazioni italiane: prove generali*, in *BioLaw Journal – Rivista di Biodiritto*, 2, 2022, 489.

[7] M. HASAN, J. POPP, J. OLÁH, *Current landscape and influence of big data on finance*, in *Journal of Big Data*, 7(1), 2020; M. SEPE, *Innovazione tecnologica, algoritmi e intelligenza artificiale nella prestazione dei servizi finanziari*, in *Rivista Trimestrale di Diritto dell'Economia*, 3, 2019, 186. Cfr. anche STANFORD UNIVERSITY HAI, *AI and financial*

Dunque, lo sviluppo di applicazioni di IA e altre tecnologie c.d. *data-driven* hanno catalizzato l'attenzione sulla necessità di agevolare l'accesso a grossi volumi di dati anche mediante la definizione di un quadro normativo chiaro ed organico che disciplini lo scambio e la circolazione dei dati per aumentare la competitività e le opportunità di benessere in favore della collettività. Del resto, la Commissione Europea ha riconosciuto che «[i] dati ridefiniranno il nostro modo di produrre, consumare e vivere, generando benefici percepibili in ogni singolo aspetto della nostra vita: da un consumo energetico più consapevole alla tracciabilità dei prodotti, dei materiali e degli alimenti, da una vita più sana a una migliore assistenza sanitaria» e che l'Unione Europea è chiamata a «divenire un modello di riferimento per una società che, grazie ai dati, dispone di strumenti per adottare decisioni migliori, a livello sia di imprese sia di settore pubblico»[8].

Tuttavia, le nuove tecnologie, insieme ad incontestabili vantaggi, stanno ponendo alcune nuove sfide etiche e giuridiche[9], con particolare riguardo alla tutela delle libertà fondamentali dell'individuo[10] e alla tenuta di alcuni principi costituzionali[11]. Ad esempio, la capacità dei modelli e degli algoritmi di intelligenza artificiale di raccogliere, elaborare ingenti volumi di dati, di identificare *pattern* complessi e di prendere decisioni automatizzate solleva delicate questioni in ordine all'attuazione del principio di uguaglianza in un'ampia varietà di settori e attività e al corrispettivo divieto di discriminazioni[12].

Inoltre, le nuove e sofisticate tecniche di manipolazione del dato (comprese quelle che alimentano i sistemi di apprendimento automatico) presentano criticità non irrilevanti con particolare riferimento alla protezione dei dati personali[13]. Sia i soggetti pubblici che quelli privati si trovano, oggi, a gestire

services, 2021; OECD, *Artificial Intelligence, Machine Learning and Big Data in Finance: Opportunities, Challenges, and Implications for Policy Makers*, 2020.

[8] Commissione Europea, Comunicazione della Commissione al Parlamento Europeo, al Consiglio, al Comitato Economico e Sociale Europeo e al Comitato delle Regioni. Una strategia europea per i dati, COM (2020) 66 *final*.

[9] È bene precisare che non tutti i sistemi di intelligenza artificiale devono ritenersi automaticamente incompatibili con il principio di trasparenza. Vi sono, infatti, metodologie che tengono in considerazione, sin dalla fase di progettazione, le implicazioni giuridiche, sociali ed etiche nello sviluppo di nuove tecnologie, seguendo un approccio di c.d. etica anticipatoria (o "*ethics-by-design*"). Si veda a tal proposito S. UMBRELLO, M.J. BERNSTEIN, P.E. VERMAAS, A. RESSEGUIER, G. GONZALEZ, A. PORCARI, A. GRINBAUM, L. ADOMAITIS, *From speculation to reality: Enhancing anticipatory ethics for emerging technologies (ATE) in practice*, in *Technology in Society*, 74, 2023.

[10] A. SIMONCINI, *L'algoritmo incostituzionale: intelligenza artificiale e il futuro delle libertà*, in *BioLaw Journal – Rivista di BioDiritto*, 1, 2019, 63; A. D'ALOIA, *Ripensare il diritto al tempo dell'intelligenza artificiale*, in G. CERRINA FERONI, C. FONTANA, E. C. RAFFIOTTA (a cura di), *AI Anthology. Profili giuridici, economici e sociali dell'intelligenza artificiale*, Bologna, 99; C. COLAPIETRO, A. MORETTI, *L'intelligenza artificiale nel dettato costituzionale: opportunità, incertezze e tutela dei dati personali*, in *BioLaw Journal – Rivista di Biodiritto* , 3, 2020, 359. Sul tema si consiglia, altresì, la lettura del saggio di V. MAYER-SCHÖNBERGER, K. CUKIER, *Big data: una rivoluzione che trasformerà il nostro modo di vivere e già minaccia la nostra libertà*, Milano, 2013.

[11] Per una breve ma completa panoramica sull'argomento si veda C. CASONATO, *Costituzione e intelligenza artificiale: un'agenda per il prossimo futuro*, in *BioLaw Journal – Rivista di Biodiritto* , 2, 2019, 711.

[12] Si pensi ad esempio alle possibili discriminazioni sul lavoro, su cui: M. BARBERA, *Discriminazioni algoritmiche e forme di discriminazione*, in *Labour & Law Issues*, 1, 2021, 2; si rinvia altresì a M. BORZAGA, M. MAZZETTI, *Discriminazioni algoritmiche e tutela dei lavoratori: riflessioni a partire dall'Ordinanza del Tribunale di Bologna del 31 dicembre 2020*, in *BioLaw Journal – Rivista di Biodiritto* , 1, 2022, 225.

[13] Sull'inquadramento del diritto alla protezione dei dati personali nel quadro costituzionale si veda G. CERRINA FERONI, *I dati personali come oggetto di un diritto fondamentale*, in P. STANZIONE (a cura di), *I «poteri privati»*

un'enorme mole di dati. La natura stessa dei *big data*, provenienti da sorgenti eterogenee e spesso contenenti informazioni sensibili o personali, rende difficile applicare le tradizionali norme e procedure di gestione dei dati. Il ricorso a nuovi sofisticati mezzi che consentono la raccolta, l'archiviazione e, soprattutto, l'elaborazione e l'utilizzo di tali immensi *dataset* solleva interrogativi cruciali riguardo alla tutela della riservatezza e delle persone fisiche e impone un'attenta valutazione delle strategie di *governance* più adeguate al fine di usare la tecnologia in maniera responsabile e nel rispetto dei diritti fondamentali dell'individuo[14].

Inoltre, le moderne tecnologie di analisi dei *big data*, in particolare quelle basate sull'intelligenza artificiale, hanno una potenza di calcolo senza precedenti e sono in grado di processare enormi volumi di dati, individuando correlazioni complesse che superano e (forse) trascendono le possibilità di comprensione umana[15]. Questi sistemi sono in grado di produrre *output* (anche imprevedibili) a partire da una quantità di informazioni che sarebbe, altrimenti, impossibile da gestire da un essere umano. Se tale circostanza può rappresentare un vantaggio, è altrettanto vero che, perlomeno, alcune complesse architetture dei sistemi di intelligenza artificiale appaiono idonee a mettere in crisi, in particolare, il principio di trasparenza, elemento fondante dell'ordinamento giuridico. Quest'ultimo, codificato nella legge n. 240 del 1990 con riferimento ai procedimenti amministrativi, estende la sua portata anche nel settore privato e si configura, in generale, quale mezzo per rintracciare la causa di un evento e, dunque, ricostruirne, ove necessario, la responsabilità. Tuttavia, la natura, talvolta, opaca dei processi di apprendimento automatico rischia di rendere difficile la piena comprensione della logica sottostante le conclusioni elaborate da taluni di questi sistemi, sollevando, in definitiva, rilevanti preoccupazioni riguardo alla possibilità di individuare un soggetto responsabile di fronte ad una decisione di cui non si conosce l'origine[16]. Ebbene, sotto questo profilo, sarà cruciale monitorare attentamente l'evoluzione dell'approccio adottato a livello comunitario: infatti, le istituzioni europee, consapevoli delle sfide appena menzionate, sono intervenute attivamente nel tentativo di trovare un bilanciamento tra l'innovazione tecnologica e la tutela dei diritti. Nel settembre 2022, la Commissione Europea ha elaborato una proposta di direttiva volta a stabilire un quadro normativo armonizzato sulla responsabilità extra-contrattuale per i danni causa da sistemi di intelligenza artificiale, introducendo, in particolare, meccanismi idonei ad agevolare l'accesso alle informazioni relative ai sistemi di intelligenza artificiale e ad alleggerire gli oneri probatori in capo agli attori che allegano un danno causato da uno di tali sistemi.

Le problematiche, sopra brevemente accennate, impongono un'attenta analisi dei vantaggi e dei rischi in gioco e richiedono un ruolo attivo di tutti gli *stakeholder* coinvolti nell'elaborazione di un quadro normativo che assicuri uno sviluppo e uno sfruttamento della tecnologia affidabile e rispettoso

delle piattaforme e le nuove frontiere della privacy, Torino, 2022, 59; S. RODOTÀ, *Tecnologia e diritti*, Il Mulino, Bologna, 2021.

[14] In generale, F. PIZZETTI, *Intelligenza artificiale, protezione dei dati personali e regolazione*, Torino, 2018; U. DE SIERVO, *Tutela dei dati personali e riservatezza*, in AA. VV., *Diritti, nuove tecnologie, trasformazioni sociali: scritti in memoria di Paolo Barile*, Padova, 2003; K.A. BAMBERGER, D.K. MULLIGAN, *Privacy on the books and on the ground*, in *Stanford Law Review*, 63, 2011, 247.

[15] A. ZWITTER, *Big Data ethics*, in *Big Data & Society*, 2, 2014, 1; T.E. FROSINI, *Il costituzionalismo nella società tecnologica*, in *Liber Amicorum per Pasquale Costanzo*, Genova, 2020.

[16] A. SIMONCINI, S. SUWEIS, *Il cambio di paradigma nell'intelligenza artificiale e il suo impatto sul diritto costituzionale*, in *Rivista di Filosofia del Diritto*, 1, 2019, 87.

dei diritti e delle libertà fondamentali. Peraltro, in un contesto in rapida evoluzione come quello in esame, è talvolta la stessa tecnologia che si tenta di regolamentare a fornire strumenti e soluzioni innovative per la protezione dei diritti costituzionali: è, per esempio, il caso delle tecniche di anonimizzazione e pseudonimizzazione nonché delle c.d. *Privacy Enhancing Technologies* (PETs), tra cui i dati sintetici che, come si avrà modo di analizzare nel presente contributo, rappresentano uno strumento di mitigazione di diversi rischi posti dalle tecnologie *data-driven* alla tutela dei dati personali e delle altre libertà fondamentali.

In questo scenario, il quadro normativo non solo deve definire regole e vincoli per un utilizzo responsabile dell'IA, ma deve anche promuovere e incentivare lo sviluppo di tecnologie abilitanti che possano contribuire attivamente alla tutela dei diritti costituzionali. Risulta fondamentale, dunque, un approccio olistico, che integri gli aspetti normativi e tecnici per garantire un ecosistema tecnologico equo, trasparente e rispettoso dei diritti umani.

1.1. La descrizione dei dati sintetici e le tecniche di sintetizzazione

In un contesto caratterizzato dall'ampia valorizzazione dell'economia dei dati e delle tecnologie ad essa correlate, si sta affermando una tendenza di crescente importanza che pare assegnare ai dati sintetici un ruolo di particolare rilievo. Tale fenomeno rappresenta un'innovativa declinazione delle molteplici potenzialità offerte dall'elaborazione e dallo sfruttamento dei dati, andando ad affiancarsi e, talvolta, a sostituirsi ai tradizionali metodi di osservazione ed elaborazione della realtà fenomenica. L'avvento dei dati sintetici costituisce un'interessante evoluzione nel panorama della *data economy* che, come sempre accade, insieme ai numerosi vantaggi, comporta nuove sfide di natura giuridica e socioeconomica.

I dati sintetici sono generati mediante particolari tecniche computazionali che permettono di emulare le caratteristiche dei dati reali che derivano da fenomeni tangibili (fisici, biologici, sociali etc.). In altri termini, i dati sintetici mantengono le stesse caratteristiche e proprietà statistiche del set di dati originari dal quale sono generati in modo da consentire a qualsiasi soggetto che analizzasse i primi di elaborare le medesime conclusioni che avrebbe tratto dallo studio del secondo[17].

Le metodologie e le tecniche di creazione e generazione di detti dati sintetici sono tra le più diverse e variegate[18]. Un primo metodo sfrutta le potenzialità delle tecniche di *regressione*, che consentono, partendo da uno specifico *dataset*, di elaborare un modello sulle relazioni tra più variabili e, nel contesto di un processo più ampio, di generare dati che sono approssimativamente simili a quelli di partenza.

Altro metodo è quello della *massima verosimiglianza* o *Maximum Likehood Extimation*, il quale restituisce i valori dei parametri di un modello di distribuzione di probabilità, basandosi su un criterio di

[17] *European Data Protection Supervisor*, *TechSonar: 2021 2022 report*, Luxembourg, 2021.

[18] Si veda G. D'ACQUISTO, *Dati sintetici: cosa sono, le applicazioni e i rischi da gestire*, in *Agenda Digitale*, 2024, disponibile al seguente link: https://www.agendadigitale.eu/sicurezza/privacy/dati-sintetici-cosa-sono-le-applicazioni-e-i-rischi-da-gestire/ (ultima consultazione 08/05/2024).

somiglianza con il *dataset* osservato e massimizzando le *chance* di verosimiglianza tra i dati sintetici generati e il *dataset* originario[19].

Infine, alcuni metodi più recenti sono basati sulle tecniche di *deep-learning* (quelli, cioè, che coinvolgono architetture di reti neurali artificiali) che includono i *variational autoencoders* (VAE)[20] e le reti generative avversarie o *Generative Adversarial Networks* (GAN). Quest'ultimi, nello specifico, risultano particolarmente utili e versatili in un ampio spettro di applicazioni, dalla generazione di contenuti (testuali, musicali e grafici) e il riconoscimento vocale fino alla generazione di dati tabellari.[21] In questo tipo di modelli, una rete neurale generativa ed una rete neurale discriminativa competono tra loro: il generatore produce dati falsi per ingannare il discriminatore che cerca di distinguere i dati reali da quelli falsificati. Entrambe le reti neurali vengono addestrate ripetutamente fino a quando il discriminatore non è in grado di identificare con precisione i dati reali e, dunque, riprodurne le caratteristiche (come da immagine qui riportata)[22].

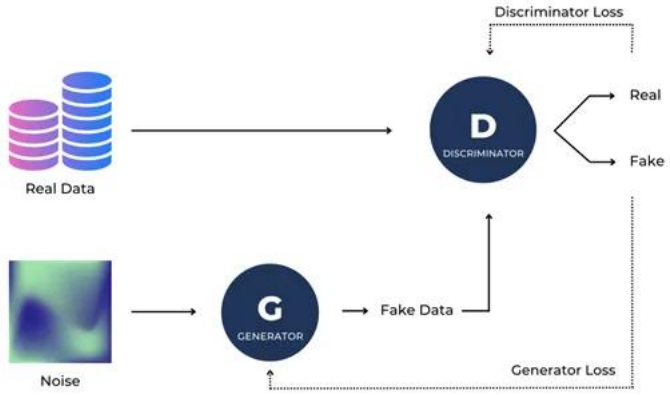

Figura 1 - Synthetic data or how to share sensitive data while staying GDPR compliant, Heka.ai, 2 giugno 2023.

Al di là delle diverse metodologie, preferibili a seconda degli ambiti di impiego, l'utilizzo di tali tecnologie pare rappresentare una sintesi tra interessi apparentemente contrapposti: da un lato, la libera circolazione e valorizzazione dei dati, in osservanza dei principi, anche di rilievo costituzionale, di libertà economica, di impresa e di circolazione di servizi e capitali e, dall'altro, l'esigenza di protezione dei dati (personali e non), sia con riferimento alla salvaguardia della vita privata, dei diritti e delle li-

[19] Si veda in generale S.R. COLE, H. CHU, S. GREENLAND, *Maximum Likelihood, Profile Likelihood, and Penalized Likelihood: A Primer*, in *American Journal of Epidemiology*, 179 (2), 2014, 252. È importante notare che l'MLE è una tecnica statistica utilizzata per stimare i parametri di un modello probabilistico: sebbene non abbia come funzione diretta la generazione di dati sintetici, una volta stimati i parametri, quest'ultimi possono essere utilizzati per generare nuovi dati sintetici che seguono la distribuzione risultante dal modello.

[20] H. DENG, *Exploring Synthetic Data for Artificial Intelligence and Autonomous Systems*, Ginevra, 2023; D.P. KINGMA, M. WELLING, *An Introduction to Variational Autoencoders*, in *Foundations and Trends® in Machine Learning*, 12, 4, 2019, 307.

[21] Nel merito, pare opportuno precisare che i VAE costituiscono la soluzione che, allo stato dell'arte, offre la migliore accuratezza nella creazione di dati tabellare e relazionali.

[22] Heka.ai, *Synthetic data or how to share sensitive data while staying GDPR compliant*, *Medium*, 2023, disponibile al seguente link: https://heka-ai.medium.com/synthetic-data-or-how-to-share-sensitive-data-while-staying-gdpr-compliant-78febdf86656 (ultima consultazione 06/05/2024).

bertà delle persone fisiche coinvolte nelle operazioni di trattamento che, in generale, della riservatezza e della tutela delle informazioni[23]. In questo senso si è soliti ricadere in una dicotomia difficilmente risolvibile: prediligere l'una a discapito dell'altra comporta, quasi inevitabilmente, dei rischi e delle conseguenti responsabilità, non solo e certamente sotto un profilo di *enforcement* delle varie *authorities* coinvolte nella supervisione di questi fenomeni (Garanti per la protezione dei dati personali, Autorità antitrust, etc.), ma anche di perdita di opportunità e benessere sociale.

1.2. *Use case* e possibili modelli virtuosi di utilizzo dei dati sintetici

Prima di addentrarsi nelle principali questioni giuridiche concernenti l'impiego dei dati sintetici alla luce del contesto normativo europeo e nazionale, pare opportuno offrire una panoramica dei principali casi studio che hanno ad oggetto il ricorso ai dati sintetici da parte di soggetti pubblici e privati.
Alcune delle più recenti applicazioni dei dati sintetici in larga scala riguardano il campo del riconoscimento facciale. I dati sintetici, infatti, rappresentano un'efficace tecnica per addestrare e sviluppare modelli destinati al riconoscimento biometrico, specie ove il set di dati utilizzato comprende sia dati generati artificialmente che dati reali[24]. Lo sviluppo delle reti neurali (*deep neural network*) e la crescente disponibilità di *dataset* di immagini facciali etichettate hanno influenzato notevolmente il panorama delle applicazioni che ricorrono a tecniche di riconoscimento facciale[25]. L'impiego di tali tecnologie è particolarmente diffuso sia tra soggetti istituzionali e delle autorità pubbliche (che, negli ultimi anni, ha fatto un crescente utilizzo di sistemi di riconoscimento facciale per fini di sicurezza e ordine pubblico) che tra i singoli individui (si pensi molto semplicemente ai meccanismi di sblocco degli smartphone mediante il riconoscimento facciale).
L'Unione Europea ha avviato il progetto Smart Borders[26], con l'obiettivo di migliorare il controllo delle frontiere dell'Unione attraverso l'implementazione di tecnologie avanzate, inclusi sistemi di riconoscimento facciale, presso i punti di ingresso dell'UE come gli aeroporti.
Un'applicazione concreta adottata nell'ambito di questo progetto è il sistema elettronico di controllo delle frontiere *eGate*, che utilizza la tecnologia di riconoscimento facciale per verificare l'identità dei passeggeri in modo automatizzato durante il processo di imbarco. I passeggeri possono registrare i propri dati biometrici, come l'immagine del volto, nei database governativi e successivamente attraversare i controlli di sicurezza in modo rapido ed efficiente[27].

---

[23] C.A. Fontanillo Lòpez, A. Elbi, *On synthetic data: a brief introduction for data protection law dummies*, *European Law Blog*, 2022, disponibile al seguente link: https://aglardor.com/index.php?title=On_synthetic_data:_a_brief_introduction_for_data_protection_law_dummies (ultima consultazione 06/05/2024).
[24] H. Qiu, B. Yu, D. Gong, Z. Li, W. Liu, D. Tao, *SynFace: Face Recognition with Synthetic Data*, in *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, IEEE, Montreal, 2021, 10860.
[25] F. Boutros, V. Struc, J. Fierrez, N. Damer, *Synthetic Data for Face Recognition: Current State and Future Prospects*, in *Image and Vision Computing*, 135, 2023, 104688.
[26] *Smart borders: European Union entry/exit system*, *EUR-lex*, 2022, disponibile al seguente link: https://eur-lex.europa.eu/EN/legal-content/summary/smart-borders-european-union-entry-exit-system.html (ultima consultazione 22/05/2024).
[27] L. Berberi, *Linate, via al sistema di riconoscimento facciale per imbarcarsi: «Controlli più rapidi senza esibire i documenti»*, in *Corriere della Sera*, 2024, disponibile al seguente link: https://milano.corriere.it/notizie/cronaca/24_maggio_08/linate-via-al-sistema-di-riconoscimento-facciale-per-

Proseguendo nell'analisi, occorre sottolineare come uno dei principali settori nel quale i dati sintetici possono manifestare il loro potenziale e portare enormi benefici è quello medico-sanitario, sia sotto il profilo clinico-diagnostico che sotto quello della ricerca[28]. Si pensi, ad esempio, alla possibilità di sviluppare una tecnologia con finalità diagnostiche, in grado di individuare con largo anticipo una forma tumorale, così da consentire un trattamento tempestivo e, dunque, più efficiente, garantendo ai pazienti maggiori *chances* di guarigione. Ebbene, tali tecnologie si basano su sistemi di intelligenza artificiale che necessitano di visualizzare un elevato numero di radiografie e, analizzando gli specifici *pattern* in esse contenute, diventano in grado di anticipare il sorgere della malattia. A tal fine, l'impiego di dati sintetici (*i.e.* immagini generate artificialmente) consentirebbe di soddisfare più agevolmente l'esigenza relativa alla raccolta delle ingenti moli di dati necessarie per il funzionamento dei sistemi in questione. Non solo. Anche la ricerca medica necessita da sempre di enormi set di dati. Tuttavia, la circolazione di tale tipologia di dati non è sempre agevole, trattandosi di dati relativi a persona fisica e, in quanto tali, soggetti alle tutele predisposte dalla normativa in materia di protezione dei dati personali (*in primis*, il GDPR). In questo contesto, pare opportuno menzionare il progetto My Health My Data ("MHMD"),[29] un network che si propone di elaborare un *hub* di dati sanitari per agevolare e sviluppare la ricerca scientifica e clinica, fornendo, al contempo, strumenti che rafforzino la privacy delle persone fisiche. Il sistema consente, da un lato, di mantenere i dati personali originali archiviati nelle *repository* locali delle singole strutture sanitarie e, dall'altro, di agevolare lo scambio di tali informazioni con terze parti mantenendo le tutele delle persone fisiche coinvolte, mediante l'impiego di tecnologie quali la *blockchain*, tecniche di crittografia e deidentificazione multilivello nonché di generazione di dati sintetici. Un analogo progetto è stato portato avanti dal *Nederlands Kanker Instituut* (NKI), un prestigioso istituto olandese specializzato nella ricerca sul cancro, il quale, mediante il ricorso a tecniche di intelligenza artificiale e a dati sintetici ha elaborato vere e proprie librerie di dati sintetici sul cancro, in modo da potenziare sia la ricerca che la diagnosi delle malattie[30].

I dati sintetici si stanno rivelando una risorsa preziosa anche nel settore della guida autonoma, offrendo una soluzione innovativa per affrontare sfide cruciali legate alla sicurezza e all'efficacia dei veicoli autonomi. Il loro potenziale consiste nel fornire una vasta gamma di scenari di guida reali (compresi quelli rari o pericolosi) permettendo ai sistemi di guida autonoma di addestrarsi su un *dataset* più completo. Inoltre, anche in questo caso, i dati sintetici consentono di superare le limitazioni legate alla disponibilità e alla diversificazione dei dati reali, migliorando significativamente le prestazioni dei sistemi di guida autonoma e contribuendo a rendere le strade più sicure e efficienti per tutti gli utenti. La società Nvidia, leader nel mercato delle GPU, ha avviato un progetto pioneristico nel

imbarcarsi-controlli-piu-rapidi-senza-esibire-i-documenti-ed527a0c-d7f2-47c0-9b9f-950320368xlk.shtml (ultima consultazione 09/05/2024).

[28] R.J. CHEN, M.Y. LU, T.Y. CHEN, D.F.K. WILLIAMSON, F. MAHMOOD, *Synthetic data in machine learning for medicine and healthcare*, in *Nature Biomedical Engineering*, 5, 2021, 493; A. B. GONZALES, G. GURUSWAMY, S.R. SMITH, *Synthetic data in health care: A narrative review*, in *PLOS Digital Health*, 2, 2023.

[29] Si veda il sito del progetto: www.myhealthmydata.eu.

[30] Si veda il sito del progetto al seguente URL: https://www.avl.nl/en/about-the-netherlands-cancer-institute/digital-oncology/.

ISSN 2284-4503

settore dell'*automotive*, denominato "NVIDIA DRIVE Sim"[31], una piattaforma di simulazione completa e realistica per il test e lo sviluppo di veicoli autonomi. All'interno di tale ambiente virtuale, dunque, è possibile addestrare e validare algoritmi di guida autonoma, sfruttando le incredibili capacità generative (potenziate proprio dai dati sintetici) di varie condizioni stradali, ambientali e di traffico. In questo modo, i fornitori di sistemi di intelligenza artificiale che operano nel settore della guida autonoma sono messi nelle condizioni di accelerare il processo di sviluppo delle proprie tecnologie, di entrare nel mercato in tempi più rapidi e con un limitato dispendio di risorse economiche.

Il ricorso ai dati sintetici si rivela particolarmente utile, altresì, per lo sviluppo di sistemi di rilevazione delle frodi finanziarie. Uno dei principali vantaggi dell'utilizzo di dati sintetici in questo contesto è la capacità di generare una vasta gamma di scenari di frode e comportamenti anomali senza dover dipendere esclusivamente dai dati reali, che potrebbero essere limitati o non rappresentativi. In linea generale, questi modelli, analizzando un grande numero di transazioni finanziarie, sono in grado di ricostruire i *pattern* nei comportamenti fisiologici dei clienti e rilevare, così, condotte sospette che potrebbero costituire attività fraudolente. Come nei casi precedenti, le prestazioni di tali sistemi dipendono in larga parte dalla quantità e dalla qualità dei dati somministrati all'algoritmo. Come già accennato, la circolazione di grandi quantità di dati, specie se relativi a persone fisiche, sconta gli ostacoli predisposti dalla normativa in materia di protezione dei dati personali e di riservatezza delle informazioni. In questo contesto, i dati sintetici sembrano essere utilizzati per testare e migliorare i sistemi di rilevazione delle frodi senza rischiare di compromettere la privacy dei clienti o di violare le normative sulla protezione dei dati.

Dunque, la *data synthesis* sembrerebbe costituire una valida alternativa allorché non sia possibile condividere i dati originali (in quanto personali) e non siano possibili o efficaci differenti tecniche di anonimizzazione o sanitizzazione[32].

2. Analisi del contesto giuridico esistente alla luce delle disposizioni europee e nazionali: un'introduzione alla nozione giuridica di dato sintetico

Se quella sin qui descritta è la tecnologia con le conseguenti opportunità che ne potrebbero derivare, guardando il diritto, ad oggi, non esiste una definizione giuridica univoca di "dato sintetico", né una disciplina completa ed organica della materia. Per analizzare le principali questioni giuridiche sottese alla generazione e all'utilizzo dei dati sintetici, occorre esaminare un corposo insieme di atti e documenti che compongono la stratificazione regolatoria di matrice europea in materia di gestione e *governance* dei dati[33].

Secondo un recente report dell'Organizzazione per la cooperazione e lo sviluppo economico (OCSE) i dati sintetici sono parte di una più ampia categoria di tecnologie, denominate "Privacy-Enhancing Technologies" (PETs), la cui finalità principale consiste nel rafforzare la tutela dei dati personali. Tut-

[31] Per maggiori informazioni sul progetto in parola, si consulti il seguente URL: https://developer.nvidia.com/drive/simulation.

[32] R. MAYER, M. HITTMEIR, A. EKELHART, *Privacy-Preserving Anomaly Detection Using Synthetic Data,* in *Data and Applications Security and Privacy,* XXXIV, 2020, 195.

[33] Per una riflessione più generale si veda E.C. RAFFIOTTA, *Dalla self-regulation alla over-regulation in ambito digitale: come (e perché) di un necessario cambio di prospettiva*, in *Osservatorio sulle fonti*, 2, 2023, 245.

tavia, l'efficacia delle PETs sotto il profilo della riservatezza dei dati dipende dall'insieme delle ulteriori misure tecniche, organizzative e legali adottate dal titolare del trattamento nell'ambito della propria strategia di *governance* del dato[34]. In particolare, l'OCSE suddivide le PETs in quattro categorie, includendo i dati sintetici nei c.d. "Data obsuscation tools", ossia strumenti di offuscamento che aggiungono "rumore" o rimuovono alcuni elementi che consentono di associare i dati ad una specifica persona fisica.

Nel corso degli anni, diverse istituzioni e organizzazioni hanno tentato di proporre alcune definizioni e categorizzazioni delle PETs, che, tuttavia, appaiono influenzate tanto dal contesto in cui sono state sviluppate quanto dallo stato dell'arte in un determinato momento storico. Nel 2021 la Federal Reserve Bank di San Francisco ha definito le PETs come un gruppo di sistemi, processi e tecniche che consentono un trattamento per estrarre valore dai dati, minimizzando, al contempo, i rischi per la privacy e la sicurezza degli individui[35]. In un report del 2016, l'Agenzia dell'Unione Europea per la cibersicurezza (ENISA) ha definito le PETs come strumenti progettati per la tutela della privacy degli utenti che fanno leva sull'utilizzo dei dati e sui metodi per conferire agli utenti un maggiore controllo sugli stessi (ad esempio, le VPNs[36]). La prospettiva considerata dall'ENISA, in questo caso, si concentra sui servizi e prodotti concreti, finalizzati a rafforzare la privacy, escludendo dall'ambito della propria analisi le restanti tecnologie (come l'anonimizzazione, la pseudonimizzazione, la sintetizzazione dei dati, etc.).

Secondo l'autorità garante della protezione dei dati personali britannica (Information Commissioner's Office, "ICO"), le PETs sono tecnologie che assicurano il rispetto dei principi in materia di protezione dei dati personali, riducendone l'utilizzo e rafforzandone la sicurezza[37]. Il documento dell'*authority* britannica fornisce indicazioni rilevanti con riferimento specifico ai dati sintetici: sono definiti come dati prodotti da appositi algoritmi generativi, in grado di replicare i *pattern* e le proprietà statistiche dei dati reali (anche quelli personali) e, poiché consentono di elaborare enormi set di dati partendo da una quantità limitata ed evitando l'ulteriore raccolta, si rivelano uno strumento particolarmente efficiente per assicurare la concreta attuazione del principio di minimizzazione dei dati[38].

Analogamente, il Comitato Consultivo della Convenzione sulla Protezione delle Persone rispetto al Trattamento Automatizzato di Dati a Carattere Personale (c.d. Convenzione 108) aveva, già nel 2019, pubblicato delle linee-guida in materia di intelligenza artificiale e protezione dei dati, precisando co-

[34] OECD, *Emerging Privacy Enhancing Technologies*, Paris, 2023.

[35] Il documento citato afferma: «[*p*]*rivacy enhancing technologies are a group of systems, processes, and techniques that enable processing to derive value from data, while minimizing the privacy and security risk to individuals*». K. ASROW, S. SAMONAS, *Privacy Enhancing Technologies: Categories Use Cases and Considerations*, Federal Reserve Bank of San Francisco, 2021, 3.

[36] *Virtual Private Networks*, ossia strumenti che consentono di creare connessioni sicure e private per i dispositivi collegati ad una rete pubblica: utilizzando la crittografia, le VPN proteggono i dati trasmessi da e verso il dispositivo, garantendo privacy e sicurezza.

[37] Information Commissioner's Office, *Privacy-enhancing technologies (PETs)*, 2023.

[38] «*Synthetic data is 'artificial' data generated by data synthesis algorithms. It replicates patterns and the statistical properties of real data (which may be personal information). It is generated from real data using a model trained to reproduce its characteristics and structure. This means that your analysis of the synthetic data should produce very similar results to analysis carried out on the original real data*». *Ibidem*, 27.

me l'uso di dati sintetici potesse rafforzare la privacy nell'ambito dello sviluppo di applicazioni di intelligenza artificiale, in quanto rappresentava un mezzo per ridurre la quantità di dati trattati a questi fini[39].

La crescente considerazione nei confronti delle *Privacy-Enhancing Technologies* e, in particolare, dei dati sintetici da parte delle istituzioni e organizzazioni pubbliche è confermata, altresì, dall'*endorsement* da parte delle autorità competenti in materia di protezione dei dati personali dei paesi appartenenti al G7 che hanno sottolineato l'importanza dello sviluppo di nuove tecniche di deidentificazione e del potenziamento di quelle esistenti, tra cui la sintetizzazione dei dati[40].

In Italia, il fenomeno dei dati sintetici non è disciplinato in maniera diretta e specifica, né sono presenti, al momento, linee guida specifiche che possano fornire indicazioni, proposte e raccomandazioni comprensive ed esaurienti in merito a tale fenomeno. Inizia, dunque, ad emergere la necessità di colmare la lacuna normativa, al fine di delineare un quadro chiaro e strutturato in grado di identificare quali *best practice* applicare nella gestione e nell'utilizzo di tali tipologie di dati.

Del resto, il tema attrae sempre maggiore considerazione nell'ambito del dibattito sociale, politico ed accademico, consapevole dell'importanza strategica dei dati sintetici nella protezione della privacy e nella promozione dell'innovazione tecnologica. Il Garante per la Protezione dei Dati Personali ha accolto favorevolmente l'adozione di iniziative legislative come il Data Governance Act (di cui si dirà *infra*) che intendono costruire un ecosistema (forse ancora troppo frammentato e disorganico) per incoraggiare «un più ampio riutilizzo dei dati detenuti dagli enti del settore pubblico, compresi i dati personali, utilizzando però ambienti di elaborazione sicuri e tecniche di anonimizzazione, come la c.d. *differential privacy* o la creazione di dati sintetici»[41].

Dalle considerazioni sinora svolte, emerge come i dati sintetici si configurino quale strumento efficace per attuare un altro pilastro fondamentale della normativa europea sulla protezione dei dati personali, ossia il principio della *privacy-by-desi*gn, favorendo lo sviluppo di prodotti e servizi che incorporino, già nella fase della loro progettazione, misure idonee a tutelare i diritti degli interessati. In definitiva, le PETs, specie se implementate congiuntamente ad ulteriori misure tecniche ed organizzative, si rivelano un elemento potenzialmente portante all'interno di un'efficace strategia di gestione del dato nel pieno rispetto del principio di *accountability*.

In questo contesto, non è sufficiente sviluppare una strategia regolatoria per affrontare i nuovi fenomeni digitali emergenti, ma è, altresì, necessario investire nella ricerca e fornire incentivi strategici alle imprese operanti nel settore dello sviluppo di sistemi e tecnologie destinati alla protezione della privacy[42]. Tali investimenti sono essenziali per promuovere innovazioni tecnologiche che possano ef-

[39] Comitato Consultivo della Convenzione sulla Protezione delle Persone rispetto al Trattamento Automatizzato di Dati a Carattere Personale, *Linee-guida in materia di intelligenza artificiale e protezione dei dati*, 2019.

[40] *Roundtable of G7 Data Protection and Privacy Authorities*, *Promoting Data Free Flow with Trust and Knowledge Sharing about the Prospects for International Data Spaces*, 2022. Si veda altresì il report del Forum economico mondiale: *World Economic Forum*, *Top 10 Emerging Technologies of 2024*, 2024.

[41] G. CERRINA FERONI, *Luci e ombre della Data Strategy europea*, *Agenda Digitale*, 2022, disponibile al seguente link: https://www.agendadigitale.eu/sicurezza/privacy/cerrina-feroni-garante-privacy-luci-e-ombre-della-data-strategy-europea/ (ultima consultazione 17/05/2024).

[42] C. MANGANELLI, *Intervento alla sessione su «Intranets e servizi globali»*, *Garante della Protezione dei Dati Personali*, 2020, disponibile al seguente link: https://www.garanteprivacy.it:443/home/docweb/-/docweb-display/docweb/46858 (ultima consultazione 17/05/2024).

ficacemente rispondere alle sfide della tutela dei dati personali in un ambiente che, come delineato, è destinato a diventare sempre più digitale. Le imprese, ove incentivate da adeguate politiche di sostegno, possono giocare un ruolo cruciale nell'elaborazione di soluzioni all'avanguardia, in grado di garantire non solo la conformità alle normative esistenti, ma anche di anticipare e mitigare i rischi legati alla privacy con un approccio proattivo ed olistico che unisca misure di *governance* a soluzioni tecnologiche in un'ottica di tutela dei diritti fondamentali degli individui.

2.1. L'analisi della strategia europea sui dati

Il nuovo ecosistema digitale globale e le recenti innovazioni tecnologiche hanno reso necessario una nuova strategia volta alla regolamentazione dei processi tecnologici e di *governance* dei dati. In questo senso, il legislatore europeo ha creato una fitta rete di disposizioni, spesso intersecate tra loro, che, se da un lato rischiano di sfociare in un'*over-regulation* del fenomeno digitale, dall'altro assicurano – o quantomeno mirano a farlo – un sistema digitale coerente, affidabile ed efficiente, che possa elevare l'Unione Europea a modello e standard di riferimento nella regolamentazione di questi (complessi) fenomeni tecnologici e digitali[43].

L'Unione Europea ha, dunque, iniziato ad implementare un programma normativo con lo scopo di istituire un mercato unico dei dati che miri a rafforzare la competitività degli Stati membri (soprattutto, rispetto alle grandi potenze internazionali[44]), garantendo, al contempo la tutela dei diritti e delle libertà fondamentali degli individui nonché la sovranità digitale degli Stati membri[45].

Il fulcro del quadro regolatorio cui punta il legislatore comunitario è costituito dai recenti regolamenti noti come "Data Governance Act"[46] e "Data Act"[47]. Questi atti normativi rappresentano pilastri fondamentali nella strategia europea per promuovere un ecosistema digitale trasparente, flessibile ed efficiente, volto a garantire una gestione sicura e responsabile dei dati, incentivando l'innovazione e la competitività nel mercato unico digitale[48].

Il Data Governance Act (DGA) intende fornire un quadro giuridico armonizzato per il riutilizzo e la condivisione dei dati detenuti da enti pubblici, anche grazie all'operato di soggetti appositamente istituiti a tale scopo (*i.e.* gli intermediari di dati e le organizzazioni di altruismo dei dati) e ha ad og-

[43] G. Cerrina Feroni, *Luci e ombre della Data Strategy europea*, cit.

[44] L'ultimo rapporto sull'IA del centro di ricerca dell'università di Stanford ha indicato che, nel 2023, gli investimenti privati tra Stati Uniti e Cina ammontano a più di 70 miliardi di dollari, mentre il primo paese dell'Unione Europea è la Germania con investimenti inferiori ai 2 miliardi di dollari. Cfr. *Stanford University HAI*, *Artificial Intelligence Index Report 2024*, 2024.

[45] *Commissione Europea*, Comunicazione della Commissione al Parlamento Europeo, al Consiglio, al Comitato Economico e Sociale Europeo e al Comitato delle Regioni. Competitività a lungo termine dell'UE: prospettive oltre il 2030, COM(2023) 168 *final*.

[46] Regolamento (UE) 2022/868 del Parlamento Europeo e del Consiglio del 30 maggio 2022 relativo alla governance europea dei dati e che modifica il regolamento (UE) 2018/1724 (regolamento sulla *governance* dei dati).

[47] Regolamento (UE) 2023/2854 del Parlamento europeo e del Consiglio, del 13 dicembre 2023, riguardante norme armonizzate sull'accesso equo ai dati e sul loro utilizzo e che modifica il regolamento (UE) 2017/2394 e la direttiva (UE) 2020/1828 (regolamento sui dati).

[48] Commissione Europea, Comunicazione della Commissione al Parlamento Europeo, al Consiglio, al Comitato Economico e Sociale Europeo e al Comitato delle Regioni. Una strategia europea per i dati, COM(2020) 66 *final*, 2020.

getto sia i dati personali (senza pregiudizio delle tutele di cui al GDPR) che quelli non personali[49]. In particolare, il DGA mira a rafforzare la fiducia nella condivisione dei dati da parte degli enti pubblici, imponendo agli Stati membri di implementare misure organizzative (ad esempio, accordi di riservatezza) e tecniche (tra cui la sintetizzazione dei dati) che assicurino la tutela dei dati personali e la riservatezza delle ulteriori categorie di dati non personali considerati protetti.

A tale ultimo proposito, è emblematico il considerando n. 7 del DGA il quale afferma che «[e]sistono tecniche che consentono l'analisi di banche dati contenenti dati personali, quali l'anonimizzazione, la privacy differenziale, la generalizzazione, la soppressione e la casualizzazione, l'utilizzo di dati sintetici o metodi analoghi, nonché altri metodi all'avanguardia di tutela della vita privata che potrebbero contribuire a un trattamento dei dati maggiormente rispettoso della vita privata». Secondo il legislatore europeo «l'applicazione di tali tecniche, unite a valutazioni d'impatto globali in materia di protezione dei dati e ad altre tutele può contribuire a una maggiore sicurezza nell'utilizzo e riutilizzo dei dati personali».

Gli Stati membri possono disciplinare autonomamente le condizioni di riutilizzo dei dati e attribuire agli enti pubblici la facoltà di concedere o negare l'accesso ai dati. In ogni caso, gli enti pubblici devono concedere l'accesso per il riutilizzo dei dati soltanto dopo aver garantito l'anonimizzazione dei dati personali o qualsiasi altro «metodo» per garantire la tutela delle tipologie di dati protetti[50].

Tale approccio conferma l'interpretazione, già elaborata dall'OCSE e analizzata nelle pagine precedenti, secondo cui le tecnologie PETs, ove utilizzate congiuntamente ad ulteriori misure tecniche ed organizzative (incluse, tra l'altro, policy di sicurezza, valutazioni d'impatto), garantiscono un «ambiente di trattamento sicuro, fornito o controllato dall'ente pubblico» e contribuiscono, in ultima analisi, ad effettuare un bilanciamento tra l'interesse all'accesso e alla condivisione dei dati a fini di riutilizzo e l'esigenza di riservatezza e di tutela dei diritti e alle libertà individuali.

Un ulteriore passo verso la realizzazione del mercato unico dei dati è rappresentato dall'approvazione del Data Act, entrato in vigore in data 11 gennaio 2024 e applicabile a partire dal 12 settembre 2025.

In un'ottica di complementarità con il DGA, il Data Act mira ad aumentare le possibilità di utilizzo e circolazione dei dati (anche in questo caso, sia quelli personali che quelli protetti) e, in particolare, quelli generati dai prodotti connessi e dai correlati servizi digitali. A tal fine, il regolamento in esame ha fissato diverse disposizioni intese a disciplinare la messa a disposizione dei dati, pur rimanendo sensibile alle tutele necessarie per garantire il rispetto dei diritti degli interessati o degli altri titolari dei dati.

I prodotti connessi e i servizi correlati devono essere progettati e forniti in modo da rendere, per impostazione predefinita, i dati dagli stessi generati accessibili agli utenti, senza rinunciare ad adeguati

[49] La letteratura offre alcuni spunti iniziali sulla nuova disciplina europea. Cfr. F. BRAVO, *Intermediazione di dati personali e servizi di data sharing dal GDPR al Data Governance Act*, in *Contratto e Impresa. Europa*, 1, 2021, 199.; P. KAMOCKI, K. LINDEN, A. PUKSAS, A. KELLI, *EU Data Governance Act: Outlining a Potential Role for CLARIN*, in *CLARIN Annual Conference*, 2023, 57; H. RICHTER, *The public interest dimension of the single market for data: public undertakings as a model for regulating private data sharing*, in *European Law Journal*, 29(1-2), 2023, 91; S. TRANQUILLI, *Il nuovo citoyen européen nell'epoca del Data governance act*, in *Rivista di Digital Politics*, 1-2, 2022, 179.

[50] Cfr. art. 5 Reg. UE 2022/868.

standard di sicurezza. Inoltre, la condivisione dei dati può avvenire anche tra imprese: i titolari dei dati, dietro una specifica richiesta di un utente, sono obbligati a rendere accessibili a terzi i dati di cui dispongono, fatto salvo, in questo caso, il rispetto da parte di tali terzi sia delle condizioni concordate con l'utente che della normativa in materia di protezione dei dati personali stabilita dall'Unione Europea e dal diritto degli Stati membri.

A tal proposito, le tecniche di sintetizzazione dei dati possono svolgere un ruolo cruciale. Il Data Act, infatti, riconosce l'importanza di quelle tecnologie che «consentono di applicare gli algoritmi ai dati e di ricavare informazioni preziose senza la trasmissione tra le parti o la copia non necessaria dei dati stessi, siano essi grezzi o strutturati»[51]. Parrebbe, qui, esserci un riferimento implicito, tra l'altro, ai dati sintetici, ricompresi nel novero di misure tecniche ed organizzative idonee a soddisfare i requisiti imposti dal principio della minimizzazione e da quello della *privacy-by-design*, assicurando, in ultima analisi, la tutela dei diritti fondamentali delle persone.

L'impianto normativo sin qui analizzato lascia impregiudicato il rispetto dei principi, delle garanzie e dei diritti delineati dal regolamento generale sulla protezione dei dati (meglio conosciuto con l'acronimo della sua denominazione inglese "GDPR"), caposaldo della disciplina europea che prevede specifiche tutele e limitazioni al trattamento dei dati personali.

A questo proposito una delle questioni più urgenti riguarda la natura giuridica dei dati sintetici: a seconda della loro configurazione come dato personale o no, deriveranno precise conseguenze in ordine all'applicabilità del regime giuridico previsto dal regolamento.

Il GDPR, infatti, si applica al trattamento interamente o parzialmente automatizzato di dati personali, a loro volta definiti come «qualsiasi informazione riguardante una persona fisica identificata o identificabile» (art. 4 par. 1 n. 1 GDPR). Il regolamento precisa che «si considera identificabile la persona fisica che può essere identificata, direttamente o indirettamente, con particolare riferimento a un identificativo come il nome, un numero di identificazione, dati relativi all'ubicazione, un identificativo online o a uno o più elementi caratteristici della sua identità fisica, fisiologica, genetica, psichica, economica, culturale o sociale».

Alla luce delle caratteristiche genetiche dei dati sintetici descritte nei paragrafi precedenti, la questione verte sulla possibilità di eliminare ogni collegamento tra i dati personali originali e quelli generati artificialmente, così da poter escludere ogni collegamento tra i secondi ed eventuali interessati.

Occorre, preliminarmente, constatare che lo stato dell'arte non comprende tecniche, riconosciute universalmente, che siano in grado di eliminare permanentemente, sotto il profilo meramente tecnico, il rischio di reidentificazione nei processi di anonimizzazione dei dati. In questo contesto, non pare possibile stabilire *a priori* la natura giuridica dei dati sintetici. Le difficoltà di qualificazione dipendono, infatti, dal rischio, diverso a seconda dei casi, di reidentificazione, ossia dalla concreta possibilità di risalire, mediante complesse tecniche computazionali, ai dati contenuti nel *dataset* originario. Il modo in cui tale tipologia di dati viene generata fa sì che essi possano venire inquadrati come dati anonimi quando non sia possibile individuare alcuna persona fisica cui i dati si riferiscono, mentre

[51] Cfr. considerando 8, Reg. UE 2023/2854.

debbano essere assoggettati alla normativa in materia di protezione dei dati personali qualora sia possibile riferire talune informazioni a specifici soggetti[52].

Si rende, dunque, necessaria un'analisi delle singole fattispecie che valuti il rispetto della normativa sulla protezione dei dati personali in un'ottica di *accountability*[53] e che identifichi il livello di rischio «accettabile» [54], ossia una soglia oltre la quale si ritiene ragionevolmente eluso il rischio di identificazione.

Peraltro, l'analisi delle metriche in base alle quali accertare il rischio di reidentificazione dovrà necessariamente tenere conto di diversi criteri stabiliti direttamente dal GDPR: il considerando 26 GDPR, infatti, afferma che «[p]er accertare la ragionevole probabilità di utilizzo dei mezzi per identificare la persona fisica, si dovrebbe prendere in considerazione l'insieme dei fattori obiettivi, tra cui i costi e il tempo necessario per l'identificazione, tenendo conto sia delle tecnologie disponibili al momento del trattamento, sia degli sviluppi tecnologici». Il quadro normativo elaborato dal legislatore europeo sembrerebbe lasciare un discreto margine di manovra, ancorando il concetto di identificabilità al tasso di probabilità con cui un soggetto potrebbe avere accesso a mezzi e tecnologie che consentano la reidentificazione, anche alla luce delle risorse economiche e di tempo necessarie; in altri termini, si esclude una nozione assoluta di anonimizzazione che imponga l'impossibilità, sotto il profilo puramente tecnico, di individuare l'interessato cui si riferisce il dato anonimizzato[55].

In questo contesto, i dati sintetici rappresentano una promettente soluzione nell'ambito della protezione dei dati personali, offrendo un approccio innovativo per mitigare il rischio di reidentificazione degli individui.

Tuttavia, trattare aprioristicamente i dati sintetici come dati non personali costituirebbe una «semplificazione eccessiva»: è, infatti, opportuno sottolineare che l'efficacia dei dati sintetici nel prevenire la reidentificazione può dipendere da una serie di variabili, tra cui la sofisticatezza dei modelli di generazione, il contesto e le finalità di utilizzo, la qualità e la tipologia dei dataset di riferimento.

In ogni caso, anche laddove i dati sintetici non fossero in grado di escludere, entro la soglia della ragionevolezza, il rischio di reidentificazione, questi potrebbero venire qualificati come dati pseudonimizzati. L'art. 4 n. 5 GDPR definisce la pseudonimizzazione come «il trattamento dei dati personali in modo tale che i dati personali non possano più essere attribuiti a un interessato specifico senza l'utilizzo di informazioni aggiuntive, a condizione che tali informazioni aggiuntive siano conservate separatamente e soggette a misure tecniche e organizzative intese a garantire che tali dati personali non siano attribuiti a una persona fisica identificata o identificabile».

Peraltro, pare opportuno sottolineare che anche la pseudonimizzazione costituisce uno strumento adeguato a garantire il rispetto del principio *privacy-by-design* e *by-default* e può rappresentare, spe-

[52] A tal proposito, in dottrina è stato suggerito un nuovo approccio in favore del superamento di un dualismo tra dati personali e non personali. Cfr. A. BEDUSCHI, *Synthetic data protection: Towards a paradigm change in data regulation?*, in *Big Data & Society*, 11(1), 2024.

[53] F. BROZZETTI, *I dati sintetici: panacea della privacy?*, *TopLegal*, 2024 https://www.toplegal.it/art/i-dati-sintetici-panacea-della-privacy/ (ultima consultazione 06/05/2024). Sul punto si veda anche il parere del Gruppo di Lavoro Articolo 29, *Parere 05/2014 sulle tecniche di anonimizzazione*, 2014.

[54] K. EL EMAM, C. ALVAREZ, *A critical appraisal of the Article 29 Working Party Opinion 05/2014 on data anonymization techniques*, in *International Data Privacy Law*, 5(1), 2015, 73.

[55] C.A.F. LÓPEZ, A. ELBI, *On the legal nature of synthetic data*, in *NeurIPS 2022 Workshop on Synthetic Data for Empowering ML Research*, 2022.

cie se implementato congiuntamente a misure tecniche e organizzative complementari, un mezzo efficace per assicurare la conformità delle attività dei titolari e dei responsabili del trattamento ai requisiti stabiliti dal GDPR in un'ottica di *accountability*[56].

L'interpretazione appena delineata appare, in ultimo luogo, compatibile le potenzialità *privacy-enhancing* dei dati sintetici: questi ultimi, di fatto, potrebbero apparire idonei ad impedire, secondo un criterio di ragionevolezza, il rischio di reidentificazione e anche laddove, a seguito di puntuali verifiche, non potessero ritenersi tali, rappresenterebbero, comunque, uno degli strumenti più efficaci per soddisfare gli standard di sicurezza e protezione dei dati personali enucleati nel GDPR e per ridurre i rischi per le libertà e i diritti fondamentali degli interessati.

2.2. L'intelligenza artificiale e i dati sintetici nello scenario nella nuova regolazione europea

I dati sintetici sono profondamente legati ai sistemi di intelligenza artificiale, sia perché costituiscono un fattore determinante per il loro sviluppo, sia perché sono da quest'ultimi generati[57]. Come ampiamente descritto sopra, uno dei principali vantaggi pratici offerti dai dati sintetici, da cui le aziende possono trarre rilevanti benefici, consiste nella maggiore disponibilità di set di dati di grandi dimensioni e nella, conseguente, possibilità di superare le difficoltà inerenti la raccolta e l'elaborazione di banche dati reali: gli sviluppatori potranno, da un lato, creare i propri modelli con maggiore efficienza e con un minor dispendio di risorse e, dall'altro, estendere la gamma di scenari e contesti di addestramento, ottenendo risultati fondati su set di dati più rappresentativi e, dunque, più precisi ed accurati[58].

In questo contesto, si inserisce la recente normativa europea che mira a fornire una regolamentazione trasversale dell'intelligenza artificiale. Il primo agosto 2024 è entrato in vigore il regolamento europeo n. 2024/1689 per l'adozione di norme armonizzate sull'intelligenza artificiale (meglio noto come "*Artificial Intelligence Act*" o "*AI Act*"), il cui scopo è quello di stabilire una cornice normativa comprensiva e uniforme per lo sviluppo e l'uso di sistemi di intelligenza artificiale incentrati sull'uomo[59].

L'*AI Act* mira a garantire che i sistemi di IA immessi sul mercato unico europeo siano sicuri e rispettosi dei diritti fondamentali, promuovendo al contempo lo sviluppo e l'implementazione di tecnologie all'avanguardia. Il nuovo regolamento adotta un approccio basato sul rischio, classificando i sistemi di

[56] Sul tema della centralità del principio della *privacy-by-design*, si suggerisce G. CERRINA FERONI, *Intelligenza artificiale e protezione dei dati personali: percorsi di analisi*, in G. CERRINA FERONI, C. FONTANA, E.C. RAFFIOTTA (a cura di), *AI Anthology. Profili giuridici, economici e sociali dell'intelligenza artificiale*, Bologna, 2022. L'autrice, qui, sostiene che il principio della *privacy-by-design* di cui all'art. 25 GDPR sembra essere, in particolare, lo strumento più idoneo a soddisfare le esigenze di tutela dei diritti fondamentali nello sviluppo tecnologico e, pertanto, possa essere elevato a principio di rango «quasi» costituzionale.

[57] Per una panoramica sul ruolo dei dati sintetici nello sviluppo dei sistemi di intelligenza artificiale nonché sulle tecniche di generazione, anche tramite, GANs si rinvia a S.I. NIKOLENKO, *Synthetic Data for Deep Learning*, CLXXIV, Cham, 2021; Á. FIGUEIRA, B. VAZ, *Survey on Synthetic Data Generation, Evaluation Methods and GANs*, in *Mathematics*, 10(15), 2022, 2733.

[58] Cfr. K.D.V. CHAITANYA, M.K. YOGI, *Role of Synthetic Data for Improved AI Accuracy*, in *Journal of Artificial Intelligence and Capsule Networks*, 5(3), 2023, 330.

[59] Sulla necessità di un approccio antropocentrico dell'intelligenza artificiale si legga L. VIOLANTE, A. PAJNO, *Diritto e etica dell'intelligenza artificiale. Presentazione*, in *BioLaw Journal – Rivista di Biodiritto* , 3, 2019, 179.

IA in quattro livelli di rischio (inaccettabile, alto, limitato e minimo), cui corrispondono diversi fasci di oneri e obbligazioni in capo agli operatori economici coinvolti nella catena di produzione e utilizzo di tali sistemi.

Pur non essendo questa la sede più opportuna per una completa ricostruzione del testo normativo, vale la pena menzionare alcuni dei principi fondanti del nuovo testo normativo che guidano lo sviluppo dei sistemi di intelligenza artificiale lungo l'intero ciclo di vita.

L'approccio europeo appare improntato a dare priorità alla tutela dei diritti e delle libertà costituzionalmente garantite e a promuovere la diffusione di un'intelligenza artificiale antropocentrica e affidabile[60]. A tal fine, riveste un ruolo fondamentale le misure di protezione dei dati personali, considerato l'uso generale massiccio di dati (anche quelli personali) per lo sviluppo di sistemi di IA.[61] Inoltre, occorre rilevare come il nuovo regolamento stabilisca requisiti specifici volti a tutelarne la riservatezza e l'integrità nelle diverse fasi del ciclo di vita dei sistemi di IA e contenga numerosi richiami al regolamento generale sulla protezione dei dati, lasciando impregiudicate le garanzie da questo predisposte[62].

A tale proposito, l'*AI Act* riconosce esplicitamente i dati sintetici come strumenti che possono giocare un ruolo fondamentale, sia per il miglioramento della qualità dei set di dati utilizzati nei sistemi di IA e la correzione di eventuali distorsioni che per la mitigazione dei rischi associati al trattamento dei dati personali.

Di fatti, nel tentativo di trovare un equilibrio tra tutela dei dati personali e affidabilità dei sistemi di IA, l'art. 10 AIA consente ai fornitori di sistemi di intelligenza artificiale ad alto rischio di trattare anche categorie particolari di dati personali, ove necessari per le correzioni delle eventuali distorsioni di

[60] Cfr. art. 1 AI Act secondo cui «[l]o scopo del presente regolamento è migliorare il funzionamento del mercato interno e promuovere la diffusione di un'intelligenza artificiale (IA) antropocentrica e affidabile, garantendo nel contempo un livello elevato di protezione della salute, della sicurezza e dei diritti fondamentali sanciti dalla Carta dei diritti fondamentali dell'Unione europea, compresi la democrazia, lo Stato di diritto e la protezione dell'ambiente, contro gli effetti nocivi dei sistemi di IA nell'Unione, e promuovendo l'innovazione».

[61] Sono numerosi le posizioni che evidenziano le sfide in materia di privacy poste dall'intelligenza artificiale. Tra i molti testi, si segnala: D.J. SOLOVE, *Artificial Intelligence and Privacy*, in *Florida Law Review*, 77, (in fase di pubblicazione), disponibile al seguente link: https://ssrn.com/abstract=4713111; J. KING, C. MEINHARDT, *Rethinking Privacy in the AI Era*, Stanford, 2024.; A. D'ALOIA, *Ripensare il diritto al tempo dell'intelligenza artificiale*, cit.; G. CERRINA FERONI, *Intelligenza artificiale e protezione dei dati personali: percorsi di analisi*, cit. Peraltro, si ricorda che il Garante della Protezione dei Dati Personali è stato la prima *authority* europea ad emettere un provvedimento nei confronti di OpenAI (la società sviluppatrice dell'ormai celebre ChatGPT) con il quale, di fatto, si è disposto un blocco temporaneo del servizio per violazione delle disposizioni del GDPR: si veda Garante della Protezione dei Dati Personali, *Provvedimento del 30 marzo 2023, n. 112*.

[62] L'*AI Act*, consapevole dell'inevitabile connessione tra i dati ed il funzionamento dei modelli di IA, conferisce particolare importanza alla normativa in materia di protezione dei dati personali. Tra le molteplici disposizioni che contengono rinvii al GDPR e norme limitrofe, si ricordano: l'articolo 1 che fa salvo, in via generale, «[i]l diritto dell'Unione in materia di protezione dei dati personali, della vita privata e della riservatezza delle comunicazioni» (con specifico riferimento al regolamento (UE) 2016/679 e (UE) 2018/1725 nonché alla direttiva 2002/58/CE e (UE) 2016/680) rispetto ai trattamenti personali effettuati nell'ambito delle attività che ricadono nell'ambito di applicazione del regolamento; l'art. 10 che disciplina le misure di gestione e *governance* dei dati utilizzati per l'addestramento, la convalida e la prova dei modelli; l'art. 27 che annuncia la necessità di coordinare la nuova valutazione d'impatto sui diritti fondamentali con quella richiesta ai sensi dell'art. 35 GDPR.

tali sistemi, a condizione che tale finalità non possa essere raggiunta mediante strumenti alternativi, compresi i dati anonimi o sintetici.

Analogamente, il ricorso ai dati sintetici in luogo di quelli personali viene favorito quale misura a tutela della protezione dei dati personali con riferimento allo sviluppo dei sistemi di intelligenza artificiale nell'ambito degli spazi di sperimentazione normativa (le c.d. *sandbox* regolatorie). L'art. 59 AIA, infatti, subordina la possibilità di effettuare un trattamento di dati personali – originariamente raccolti per altre finalità – per l'addestramento di sistemi di intelligenza artificiale nel contesto delle *sandbox* regolatorie ad una serie di condizioni, tra cui la circostanza che i requisiti indicati dalla sezione 2 del capo III AIA (tra cui spiccano quelli relativi alla qualità, pertinenza, rappresentatività e correttezza dei dati) non possano essere soddisfatti mediante il ricorso a dati anonimizzati, sintetici o altri dati non personali.

L'impianto normativo, qui brevemente descritto, parrebbe configurare un preciso onere che impone, nei contesti specificati, di ricorrere ai dati sintetici o anonimi prima di procedere al trattamento di dati personali. Tali tipologie di dati, dunque, ricoprono un ruolo fondamentale sia con riferimento allo sviluppo dei sistemi di intelligenza artificiale che sotto il profilo di tutela della privacy: la lettera degli artt. 10 e 58 AIA, infatti, affiancando i dati sintetici a quelli anonimi come alternativa primaria ed obbligatoria al dato personale, sembrerebbe implicitamente suggerire la loro classificazione come dati non personali, enfatizzandone l'importanza strategica nel bilanciamento tra innovazione tecnologica e protezione dei diritti individuali.

Infine, una delle principali novità nella disciplina dei dati sintetici introdotta dal nuovo *AI Act* è rappresentata dagli obblighi di trasparenza relativi ad alcuni sistemi di intelligenza artificiale, imposti sia ai fornitori che ai c.d. *deployer*, ossia gli utilizzatori professionali di tali sistemi.

Il principio di trasparenza costituisce uno dei pilastri sul quale si regge la complessa architettura normativa delineata dall'Unione Europea in materia di *governance* dei dati e dell'intelligenza artificiale e rappresenta un elemento imprescindibile per la tutela dei diritti e delle libertà fondamentali. Esso si configura quale cardine ineludibile dell'impianto legislativo comunitario sul quale si appoggia un complesso sistema di regole e vincoli giuridici improntato alla massima apertura, chiarezza e intelligibilità dei processi decisionali e delle logiche sottese al funzionamento dei sistemi di intelligenza artificiale che vengono immessi nel mercato.

La *ratio* che si cela dietro l'affermazione di tale principio corre su un duplice binario: da un lato, la trasparenza consente di controllare l'iter logico seguito da un algoritmo di apprendimento automatico cui è stata attribuito, in qualche forma, un potere decisionale e, dunque, consente di valutarne l'accuratezza, promuovendo una maggiore affidabilità dello stesso; dall'altro, la necessità di conoscere le motivazioni che hanno determinato la decisione di un sistema di intelligenza artificiale risiede nella possibilità di sottoporla a sindacato e di rintracciarne, ove necessario, la responsabilità[63].

Tuttavia, come già accennato, uno dei principali problemi posti dall'intelligenza artificiale consiste proprio nel fatto che non è sempre possibile conoscere l'*iter* logico seguito dalla macchina[64]: è il c.d.

[63] Si veda a tale riguardo Gruppo di esperti ad alto livello sull'intelligenza artificiale, *Orientamenti etici per un'IA affidabile*, Bruxelles, 2019.

[64] Si richiama, qui, il caso di *AlphaGo*, un sistema di intelligenza artificiale sviluppato da *DeepMind*, autore dell'ormai nota "mossa 37", che, rimasta incompresa ai professionisti del gioco Go, ha consegnato la vittoria al

ISSN 2284-4503

problema della *black box*, ossia il fenomeno in base al quale, pur essendo conoscibili i dati che sono stati forniti al modello e l'*output* da quest'ultimo prodotto, non è possibile spiegare le ragioni, le logiche ed, in generale, il percorso seguito da un algoritmo nelle proprie determinazioni.

L'AI Act prevede una serie di obblighi che impongono, in primo luogo, al fornitore di un sistema di IA di garantire che il suo funzionamento sia «sufficientemente trasparente da consentire ai *deployer* di interpretare l'*output* del sistema e utilizzarlo adeguatamente»[65]. La disposizione, se considerata isolatamente, sembrerebbe stabilire un'obbligazione di risultato, senza indicare i mezzi necessari a conseguirlo. Tuttavia, è possibile reperire qualche elemento in più: ai sensi dell'art. 13 AIA, infatti, il sistema di IA deve essere accompagnato dalle relative istruzioni d'uso, le quali, a loro volta, informazioni relative alla capacità e alle caratteristiche tecniche del sistema, ove pertinenti a «spiegarne l'output». Sotto questo profilo, sarà centrale il ruolo dei soggetti privati (i principali sviluppatori della tecnologia) nella definizione di standard e *best practice* in grado di fornire maggiori indicazioni sugli elementi che consentirebbero l'interpretazione dell'*output*. A tal proposito, occorre sottolineare che l'art. 13 in esame stabilisce che le istruzioni d'uso devono contenere, se ritenuto opportuno, anche indicazioni relativamente ai dati di *input* o qualsiasi altra informazione pertinente in termini di set di dati di addestramento, convalida e prova del sistema. La norma, dunque, sembrerebbe riconoscere un certo ruolo al set di dati usato per lo sviluppo del sistema, che potrebbe rivelare informazioni utili sul suo funzionamento. Sotto questo profilo, sembra legittimo chiedersi se la disposizione appena esaminata non comporti l'obbligo in capo al fornitore di un sistema di IA di rivelare, tra l'altro, l'esistenza di dati sintetici nel *dataset* di addestramento e, se necessario, le tecniche di generazione utilizzate. L'impianto così delineato appare coerente con l'obbligo di adottare specifiche misure di *governance* volte a documentare i processi di raccolta e l'origine dei dati nonché quello di garantire che i set di addestramento siano sufficientemente rappresentativi, esenti da errori e completi.

Occorre, tuttavia, una precisazione: le informazioni concernenti i dati di addestramento possono solo essere uno degli elementi necessari per garantire la spiegabilità e la trasparenza degli *output* dei sistemi di intelligenza artificiale più complessi e sofisticati, ma non appaiono sufficienti per assicurare autonomamente la completa attuazione del principio di trasparenza.

In realtà, allo stato, una concreta attuazione del principio di trasparenza potrebbe rimanere, in talune circostanze, particolarmente ardua: ciò non solo in ragione delle difficoltà ad assoggettare a un sindacato tecnico le motivazioni sottese alle decisioni algoritmiche, bensì anche perché i tentativi di tradurre il codice del sistema in un linguaggio comprensibile agli utilizzatori rischiano di generare un elevato margine di superficialità ed approssimazione, tale da non rendere con esattezza la formula tecnica seguita dall'algoritmo[66]. Di fronte all'erompere dell'intelligenza artificiale, sembrano, dunque,

sistema durante un match disputato con il campione del mondo Lee Sedol; C. METZ, *In Two Moves, AlphaGo and Lee Sedol Redefined the Future*, *Wired*, 2016, disponibile al seguente link: https://www.wired.com/2016/03/two-moves-alphago-lee-sedol-redefined-future/ (ultima consultazione 27/05/2024).

[65] Cfr. art. 13 *AI Act.*

[66] M.R. ALLEGRI, *Il valore della trasparenza nei sistemi sociali algoritmici: questioni di ordine linguistico e costituzionale*, in *AIC. Associazione Italiana dei Costituzionalisti*, 2023, disponibile al seguente link: https://www.associazionedeicostituzionalisti.it/it/la-lettera/12-2023-liberta-di-ricerca-e-intelligenza-

indebolirsi alcuni dei principi fondamentali dell'ordinamento: se è vero che ci troviamo in presenza di un nuovo soggetto con funzioni «cognitive»[67], tale soggetto esercita quelle funzioni con modalità del tutto nuove e, in parte, sconosciute.[68] Ciò premesso, sorge un interrogativo sulla necessità di elaborare un nuovo standard di trasparenza: in altri termini, potrebbe essere opportuna una nozione del concetto di trasparenza, che, più che garantire una perfetta intellegibilità delle decisioni algoritmiche, si concentri sul funzionamento della relazione tra la macchina e l'uomo, ossia tra l'*output* del sistema e la decisione presa, in ultima istanza, dall'essere umano.[69] In sostanza, stante le difficoltà intrinseche nel rendere pienamente comprensibili sistemi automatizzati complessi, è legittimo chiedersi se non sia più adeguato alle esigenze regolatorie porre l'accento su un livello di trasparenza algoritmica qualitativamente diverso che, alla necessità di analizzare l'iter logico dell'algoritmo, sostituisca la volontà di definire i nuovi contorni della relazione tra uomo e macchina.

Tuttavia, in assenza di una prassi applicativa consolidata, una valutazione esaustiva sull'effettività dei vincoli imposti dal principio di trasparenza, così come delineato dall'*AI Act*, è ancora prematura: per ora, è possibile evidenziare come il complesso articolato di obblighi predisposto dall'*AI Act* assolva la funzione di mitigare il rischio di sviluppo e utilizzo di sistemi sottratti a qualsivoglia forma di sindacato, pur senza precludere *tout court* il ricorso alla tecnologia[70].

Il principio di trasparenza, infine, rileva, altresì, sotto un ulteriore e distinto profilo. Se fino ad ora si è discusso della trasparenza relativamente alla logica utilizzato dal sistema di intelligenza artificiale, il regolamento europeo si occupa di elaborare i margini entro i quali possono muoversi taluni sistemi di intelligenza artificiale con particolari caratteristiche. È il caso dei sistemi con finalità generali in grado di produrre output di testo o altri contenuti audio, video e grafici (i c.d. *deep fake*): in relazione a tali sistemi, l'art. 50 AIA prevede, da un lato, un obbligo in capo ai fornitori di garantire che gli output possano essere individuati come generati artificialmente, e, dall'altro, che i *deployer* di tali sistemi debbano rivelare l'eventuale origine artificiale del contenuto.

Ancora una volta, emergono dubbi sulle concrete modalità di implementazione dei suddetti obblighi di trasparenza e sulle misure (tecniche e organizzative) per assicurarne il rispetto da parte di tutti gli operatori economici. Dunque, anche in questo caso, sarà necessaria una chiara definizione di regole, standard e *best practice* che consentano di stabilire quali misure tecniche ed organizzative possano venire implementate al fine di mitigare il rischio di un uso improprio dei dati sintetici.

---

artificiale/il-valore-della-trasparenza-nei-sistemi-sociali-algoritmici-questioni-di-ordine-linguistico-e-costituzionale (ultima consultazione 30/05/2024).

[67] A. SIMONCINI, *Il linguaggio dell'intelligenza artificiale e la tutela costituzionale dei diritti*, in *Rivista AIC*, 2, 2023, 1, 8.

[68] Sul punto si legga ancora A. SIMONCINI, S. SUWEIS, *Il cambio di paradigma nell'intelligenza artificiale e il suo impatto sul diritto costituzionale*, cit., 92 che spiega come l'approccio stocastico delle moderne tecniche di intelligenza artificiale costituisca una «una forte limitazione» in quanto la macchina «non fornisce una spiegazione causale o analitica dei pattern (proprietà ricorrenti)».

[69] D.U. GALETTA, *Digitalizzazione, intelligenza artificiale e Pubbliche Amministrazioni: il nuovo codice dei contratti pubblici e le sfide che ci attendono*, in *federalismi.it*, 12, 2023, iv.

[70] C. CASONATO, B. MARCHETTI, *Prime osservazioni sulla proposta di regolamento dell'Unione Europea in materia di intelligenza artificiale*, in *BioLaw Journal – Rivista di Biodiritto* , 3, 2021, 415; C. COLAPIETRO, *Gli algoritmi tra trasparenza e protezione dei dati personali*, in *federalismi.it*, 5, 2023, 151.

3. Un caso d'uso emblematico: i dati sintetici e il settore sanitario

Come accennato brevemente nell'introduzione, le moderne tecniche di analisi dei dati, compresa l'intelligenza artificiale, nonché le tecnologie volte a rafforzare la tutela della privacy, tra cui i dati sintetici, manifestano pienamente le proprie potenzialità soprattutto nel settore sanitario[71]. In generale, i dati rivestono un'importanza cruciale per il miglioramento delle pratiche cliniche[72], per la ricerca medica[73] e l'ottimizzazione dei servizi di assistenza e rappresentano una risorsa preziosa per lo sviluppo di modelli di intelligenza artificiale in grado di alimentare sofisticati sistemi con finalità di cura[74]. Secondo un rapporto della società di consulenza McKinsey, tuttavia, una delle principali sfide del settore *life sciences*, che include il settore sanitario e *pharma*, è la mancanza di accesso a dati di qualità e di piattaforme integrate che garantiscano la messa a disposizione di dati[75]. È in questo contesto che occorre leggere ed interpretare la strategia dell'Unione Europea sui dati che, oltre all'adozione di misure volte a favorire in generale la circolazione dei dati (anche personali), prevede iniziative legislative verticali con il preciso scopo di creare degli ecosistemi di dati in specifici settori dell'ordinamento.

3.1. Il nuovo Spazio europeo dei dati sanitari: l'European Health Data Space

Nella sua ultima sessione di legislatura il Parlamento europeo ha approvato in via definitiva – con 445 voti a favore, 142 contrari e 39 astenuti – la proposta di regolamento dell'Unione Europea istitutiva dello spazio europeo dei dati sanitari ("*European Health Data Space*" o "EHDS"[76]) recependo l'accordo politico raggiunto con il Consiglio dell'Unione Europea. Lo Spazio europeo dei dati sanitari sarà un

[71] La letteratura che esamina le potenzialità dei big data e lo sviluppo delle nuove tecnologie nel campo sanitario è sterminata. Cfr. tra gli altri, *European Commission*, *Study on health data, digital health and artificial intelligence in healthcare*, 2022.; WHO, *Regulatory considerations on artificial intelligence for health*, 2023; *European Commission. Joint Research Centre*, *Artificial intelligence for healthcare and well-being during exceptional times: a recent landscape from a European perspective*, Luxembourg, 2023.; J. LI, B.J. CAIRNS, J. LI, T. ZHU, *Generating synthetic mixed-type longitudinal electronic health records for artificial intelligent applications*, in *NPJ Digital Medicine*, 6, 2021; A. R. GONÇALVES, P. RAY, B. SOPER, J. STEVENS, L. COYLE, A.P. SALES, *Generation and evaluation of synthetic patient data*, in *BMC Medical Research Methodology*, 20, 2020. Si veda anche: G. FIORIGLIO, *La protezione dei dati sanitari nella società algoritmica. Profili informatico-giuridici*, in *Journal of Ethics and Legal Technologies*, 3(2), 2021, 80; E.A. FERIOLI, *L'intelligenza artificiale nei servizi sociali e sanitari: una nuova sfida al ruolo delle istituzioni pubbliche nel welfare italiano?*, in *BioLaw Journal – Rivista di Biodiritto* , 1, 2019, 163.

[72] G.M.O. FARES, *Telemedicine and contact tracing apps in times of pandemic*, in *Ius et Salus*, 2022.

[73] Per una panoramica sull'uso dei dati per finalità di ricerca scientifica in generale, si veda P. GUARDA, *Il regime giuridico dei dati della ricerca scientifica*, Napoli, 2021; G. BINCOLETTO, P. GUARDA, *A proactive GDPR-compliant solution for fostering medical scientific research as a secondary use of personal health data*, in *Opinio Juris in Comparatione*, 1, 2021, 43.

[74] Si ricorda il recente decalogo emanato dal Garante della Protezione dei Dati Personali che passa in rassegna alcune delle principali questioni da considerare ai fini della realizzazione di servizi sanitari nazionali attraverso sistemi di intelligenza artificiale: si veda Garante della Protezione dei Dati Personali, *Decalogo per la realizzazione di servizi sanitari nazionali attraverso sistemi di intelligenza artificiale*, 2023.

[75] B. ALBRECHT, S. DE FREMOND, T. DEVENYNS, R.T. LI, D. TINKOFF, L. VAN DER VEKEN, *10 new trends in life sciences analytics & digital*, in *McKinsey&Company*, 2023.

[76] Proposta di regolamento del Parlamento Europeo e del Consiglio sullo spazio europeo dei dati sanitari, 2022/0140 (COD)

pilastro fondamentale dell'Unione europea della Salute ed è il primo spazio comune di dati dell'UE in un settore specifico a emergere dalla strategia europea per i dati. In sostanza è la prima *lex specialis* rispetto al *Data Governance Act* che, come *lex generalis* fissa, le regole sulla circolazione intersettoriale di dati in spazi europei comuni dei dati. L'EHDS consentirà ai cittadini-pazienti di assumere il pieno controllo dei propri dati sanitari, facilitandone lo scambio per la fornitura di assistenza sanitaria in tutta l'UE (il cosiddetto uso primario dei dati che circoleranno mediante una apposita piattaforma elettronica europea). Si promuove anche un vero e proprio mercato unico per i sistemi di cartelle cliniche elettroniche e si attua, inoltre, un sistema coerente, affidabile ed efficiente per il riutilizzo dei dati sanitari a fini di ricerca, innovazione, sviluppo di medicina e app personalizzate, addestramento di algoritmi di IA, elaborazione di politiche e attività normative: il cosiddetto uso secondario dei dati, per cui sarà attivata una seconda e apposita piattaforma elettronica.

Sebbene l'*European Health Data Space* presenti vantaggi considerevoli, la piena realizzazione di uno spazio europeo comune pone rilevanti sfide in relazione alla privacy, alla sicurezza e alla interoperabilità dei dati: l'obiettivo di favorire l'accesso ai dati sanitari per uso primario e secondario non deve compromettere la tutela dei diritti fondamentali degli individui. L'EHDS, del resto, lascia impregiudicata l'articolato impianto di garanzie elaborato dalla normativa in materia di protezione dei dati personali che, considerata la natura particolare dei dati sanitari, continuerà a trovare piena applicazione, in assenza di disposizioni speciali.

Gli organismi responsabili dell'accesso ai dati sono tenuti a condividere i dati nel rispetto del principio di minimizzazione e di limitazione della finalità, privilegiando la fornitura in forma anonimizzata qualora ciò risulti sufficiente rispetto alle finalità del trattamento del soggetto richiedente; laddove invece l'anonimizzazione non consenta di perseguire tali finalità, l'accesso dovrà essere garantito previa pseudonimizzazione dei dati, i cui codici di decodifica sono detenuti esclusivamente dall'organismo responsabile di riferimento, con divieto per l'utente di compiere operazioni di reidentificazione, pena l'applicazione di adeguate sanzioni[77]. Come evidenziato dall'EDPS, il ricorso ai dati sintetici si presenta quale strumento idoneo, insieme ad ulteriori misure tecniche e organizzative, ad agevolare la condivisione dei dati, mantenendo le garanzie di privacy e sicurezza richieste dalla normativa di settore[78].

Peraltro, la proposta di regolamento consente il c.d. uso secondario dei dati ai fini di «addestramento, prova e valutazione degli algoritmi, anche nell'ambito di dispositivi medici, sistemi di IA e applicazioni di sanità digitale»[79]. In questo modo, l'EHDS potrebbe rappresentare una porta di accesso a dati sanitari con lo scopo favorire la creazione di *dataset* sintetici che vengano sfruttati per affinare applicazioni di IA nel campo sanitario.

In ogni caso, è stato rilevato come nonostante i passi in avanti fatti con l'*European Health Data Space* per semplificare la condivisione e l'utilizzo dei dati sanitari in tutta l'UE, rimanga l'esigenza di garanti-

[77] Cfr. art. 44 EHDS.

[78] *European Data Protection Supervisor*, Opinion 3/2020 on the European strategy for data, 2020. Il punto 66 afferma: *«[t]here are a number of promising technological solutions, such as use of synthetic data, which may, inter alia, facilitate access to training data for machine learning. While there are still uncertainty and open questions related to possible feasibility and the efficacy of such solutions to mitigate data protection risks, the EDPS encourages the Commission to invest in further research and tests*».

[79] Cfr. art. 34 par. 1 lett. g) EHDS.

re una maggiore armonizzazione con riferimento alla protezione dei dati personali in ambito sanitario. Il GDPR, infatti, consente agli Stati membri di prevedere specifiche deroghe sia sotto il profilo dell'esercizio dei diritti da parte degli interessati[80] sia con riferimento alle basi giuridiche per il trattamento di tali dati, rischiando di favorire approcci differenziati tra gli Stati membri che potrebbero rappresentare un ostacolo alla raccolta e all'utilizzo dei dati all'interno dello spazio sanitario europeo[81].

A tal fine, l'*European Data Protection Supervisor* ha, tra l'altro, incoraggiato l'adozione di Codici di Condotta aventi ad oggetto specifiche linee guida per il trattamento dei dati sanitari al fine di fornire un quadro chiaro ed uniforme, auspicando misure che favoriscano la fiducia degli individui e dei pazienti nel sistema di condivisione di dati che l'UE si prefigge di realizzare[82]. Ebbene, a tal fine, l'EDPS ha sottolineato l'importanza cruciale di incorporare alcune forme di garanzia per la protezione dei dati nello sviluppo e nella *governance* dello spazio europeo dei dati sanitari: l'autorità europea ha, pertanto, incoraggiato, da un lato, il rispetto dei principi di cui all'art. 5 GDPR da parte dei soggetti che rendono disponibili i dati e, dall'altro, il ricorso a misure organizzative e tecniche a protezione dei dati personali, tra cui tecniche di anonimizzazione e aggregazione nonché, in subordine, di pseudonimizzazione[83].

Sebbene il regolamento non sia ancora entrato in vigore, è possibile evidenziare già in questa fase i vantaggi apportati dalle PETs e, in particolare, dalla sintetizzazione dei dati: esse costituiscono uno strumento efficace per rispettare gli standard richiesti dalla normativa europea e nazionale e favorire, così, la condivisione dei dati sanitari all'interno dello spazio comune europeo, con il fine ultimo di promuovere il miglioramento delle prestazioni e dei servizi sanitari nonché il progresso della ricerca medico-scientifica.

3.2. La nuova riforma del codice della privacy

Come detto, tra gli obiettivi sottesi alla creazione del nuovo spazio europeo dei dati sanitari vi è quello di incentivare il riutilizzo dei dati sanitari per finalità diverse rispetto a quelle che hanno originariamente motivato la loro raccolta: in altri termini, il legislatore europeo intende promuovere forme di impiego secondario di tali informazioni, consentendone lo sfruttamento per finalità ulteriori e distinte da quelle iniziali, ampliandone dunque le potenziali applicazioni.

[80] *European Data Protection Supervisor*, *A Preliminary Opinion on data protection and scientific research*, 2020, disponibile al seguente link: https://www.edps.europa.eu/sites/default/files/publication/20-01-06_opinion_research_en.pdf (ultima consultazione 21/05/2024).

[81] P. DE HERT, A. KISELEVA, *Creating a European Health Data Space: Obstacles in Four Key Legal Areas*, in *European Pharmaceutical Law Review*, 5, 1, 2021, 21.

[82] *European Data Protection Supervisor*, *Preliminary Opinion 8/2020 on the European Health Data Space*, 2020, punto n. 17.

[83] Ibidem, p. 8. Il punto 23 della menzionata opinione afferma: *«[a]s to the sources of that data, we are of the view that this should be a responsibility of the national Member States' contact points/permit authorities, who would assess the validity and quality of the type of data submitted by the parties sharing such data. Following the necessity and proportionality principles, the EDPS believes that the data made available to the EHDS should as a general rule be made anonymised and aggregated. If this were not possible due to the nature of the data at stake and the purpose of the processing, this should at least be pseudonymized*».

Tuttavia, se il c.d. uso secondario dei dati nel campo sanitario, da un lato, può aprire nuove prospettive per la ricerca scientifica e lo sviluppo di terapie innovative, dall'altro lato, solleva legittimi timori riguardo ai rischi di violazione della privacy e della riservatezza dei soggetti cui i dati si riferiscono. Tali preoccupazioni sono alla base dell'impianto normativo in materia di protezione dei dati personali e hanno giustificato l'adozione di principi, garanzie e vincoli all'utilizzo delle informazioni di natura sanitaria che ne hanno ostacolato la circolazione e lo sfruttamento.

Nel contesto europeo e nazionale, il consenso dell'interessato è stato tradizionalmente considerato una delle basi giuridiche fondamentali per legittimare il trattamento dei dati personali di carattere sanitario. Come noto, l'art. 9 GDPR impone un generale divieto generale di trattare categorie particolari di dati personali, inclusi quelli relativi allo stato di salute. Tuttavia, la medesima fornisce un elenco tassativo di eccezioni che legittimano il trattamento di dati sanitari. Tra queste ipotesi derogatorie, la principale è rappresentata dall'acquisizione del consenso della persona cui i dati si riferiscono. Tuttavia, quando si tratta dell'utilizzo dei dati per finalità di ricerca scientifica, soprattutto nel caso della ricerca medica e biomedica, il consenso si rivela spesso un fondamento giuridico inadeguato e insufficiente.[84] Il riferimento, in particolare, è agli studi osservazionali retrospettivi che, si basano su set di dati già disponibili e raccolti, anche molto tempo prima, per il perseguimento di finalità diverse ed ulteriori e in relazione ai quali risulta impossibile o particolarmente difficile raccogliere il consenso dell'interessato, il quale potrebbe risultare irraggiungibile. Pertanto, il regolamento generale sulla protezione dei dati e le normative nazionali di recepimento hanno individuato basi giuridiche alternative per il trattamento dei dati a fini di ricerca, come l'interesse pubblico rilevante o il perseguimento di finalità di ricerca scientifica. Nel tentativo di contemperare l'esigenza di un maggiore accesso ai dati per finalità di ricerca, l'art. 89 GDPR impone l'adozione di misure tecniche ed organizzative al fine di garantire il principio di minimizzazione dei dati, favorendo l'adozione di tecniche di pseudonimizzazione o altri trattamenti che impediscano di reidentificare l'interessato.

Peraltro, la disciplina nazionale aveva integrato il sistema comunitario introducendo, in conformità con quanto stabilito all'art. 36 GDPR, l'obbligo di consultazione preventiva dinanzi al Garante della Protezione dei Dati Personali, la cui autorizzazione era *condicio sine qua non* per il trattamento di dati sanitari per finalità di ricerca medico-scientifica. La norma aveva, di fatto, posto delle restrizioni particolarmente stringenti all'utilizzo secondario dei dati per tali finalità, generando un freno non indifferente nel campo della ricerca biomedica e delle sperimentazioni cliniche. L'art. 110 Codice Privacy *ante* riforma, infatti, prevedeva un complesso *iter* burocratico, che implicava la necessità di effettuare una valutazione d'impatto ai sensi dell'art. 35 GDPR e una consultazione preventiva dinanzi al Garante ai sensi dell'art. 36 del GDPR.

A questo proposito, occorre dare conto della recente riforma del Codice della Privacy, e in particolare dell'articolo 110, che presenta profili di interesse per quanto concerne l'utilizzo dei dati personali, con le sue implicazioni per i dati sintetici, nel contesto della ricerca scientifica. La modifica legislativa, introdotta con c.d. il "decreto PNRR bis"[85], ha, infatti, semplificato le procedure per il trattamento dei

[84] Si veda C. CASONATO, *Intelligenza artificiale e diritto costituzionale: prime considerazioni*, in *Diritto pubblico comparato ed europeo*, Fascicolo Speciale, 2019, 101.

[85] Legge 29 aprile 2024, n. 56 che ha convertito in legge, con modificazioni, il decreto-legge 2 marzo 2024, n. 19, recante ulteriori disposizioni urgenti per l'attuazione del Piano nazionale di ripresa e resilienza (PNRR).

dati a fini di ricerca, eliminando l'obbligo di consultazione preventiva del Garante della Privacy per i dati relativi alla salute trattati a fini di ricerca scientifica.
Un simile cambio di rotta appare maggiormente conforme allo spirito promosso dalle istituzioni europee con l'affermazione del principio di *accountability* e le ulteriori iniziative volte alla creazione di uno spazio comune di condivisione e circolazione dei dati sanitari tra gli Stati membri, al fine di promuovere, tra l'altro, la ricerca medica e, dunque, lo sviluppo di terapie e soluzioni innovative per il benessere dei cittadini.
Infatti, l'intervento legislativo, allentando i precedenti limiti restrittivi, favorisce un quadro giuridico più aperto alla condivisione e allo sfruttamento del potenziale informativo racchiuso nei patrimoni di dati sanitari, senza rinunciare agli adeguati livelli di tutela dei diritti fondamentali degli interessati attraverso l'adozione di misure tecniche e organizzative all'avanguardia, in linea con i principi della privacy by design e by default.
L'intervento normativo, infatti, non mira a rimuovere qualsivoglia tutela prevista a favore dei dati personali degli interessati, ma appare come un tentativo di adottare un approccio più flessibile e adattabile alle esigenze attuali della ricerca, in linea con i principi di proporzionalità e necessarietà previsti dal GDPR.[86] Rimane, infatti, la necessità da parte del titolare del trattamento di dati sanitari per fini di ricerca medica, non solo di effettuare la valutazione di impatto sul proprio programma di ricerca, ma anche di adottare misure appropriate per tutelare i diritti, le libertà e i legittimi interessi dell'interessato. Tra queste misure, le *privacy enhancement technologies*, compreso il ricorso ai dati sintetici, possono giocare un ruolo fondamentale: questi ultimi possono offrire notevoli benefici per lo sviluppo e il progresso della ricerca medico-scientifica, garantendo, al contempo, un elevato standard di protezione dei dati riservati.[87]
In conclusione, la riforma dell'articolo 110 del Codice della Privacy, con le sue implicazioni sui dati sintetici, segna un passo importante verso un utilizzo più efficace e flessibile dei dati nella ricerca, in linea con un approccio europeo sempre più aperto nei confronti dell'uso e della circolazione dei dati sanitari, e semina un terreno florido per il progresso nello sviluppo di terapie e pratiche cliniche innovative nel rispetto dei diritti fondamentali.

3.3. Il disegno di legge sull'intelligenza artificiale e le disposizioni in materia sanitaria

Nell'analisi del quadro normativo applicabile al trattamento di dati sanitari, occorre dare conto di un ulteriore sviluppo concernente la regolazione nazionale dell'intelligenza artificiale e dei processi di trattamento finalizzati al suo addestramento. Il Consiglio dei ministri italiano, infatti, ha approvato, in data 23 aprile 2024, un disegno di legge con lo scopo di elaborare norme di principio e disposizioni specifiche in materia di intelligenza artificiale (DDL IA). L'iniziativa legislativa, ora sottoposta all'esame del Parlamento, si occupa di vari aspetti relativi all'utilizzo dell'intelligenza artificiale, dalla

[86] *Proposta di riforma per la privacy e la ricerca scientifica - Tavolo Salute di State of Privacy*, *Istituto Italiano per la Privacy e la Valorizzazione dei Dati*, 2024, disponibile al seguente link: https://www.istitutoitalianoprivacy.it/2024/01/10/proposta-di-riforma-per-la-privacy-e-la-ricerca-scientifica-tavolo-salute-di-state-of-privacy/ (ultima consultazione 16/05/2024).
[87] Per una discussione più approfondita sulle potenzialità dei dati sintetici ai fini di tutela dei dati sanitari si veda il saggio di K.E. EMAM, L. MOSQUERA, J. BASS, *Evaluating Identity Disclosure Risk in Fully Synthetic Health Data: Model Development and Validation*, in *Journal of Medical Internet Research*, 22, 11, 2020.

previsione di regole e principi generali alla predisposizione di specifiche disposizioni settoriali (ad es. riferite al lavoro, alle professioni intellettuali, alla Pubblica Amministrazione, all'attività giudiziaria, *etc.*), in particolare, con riferimento al settore sanitario assume una rilevanza particolare. Il DDL IA, infatti, oltre ad alcune dichiarazioni di principio generali sulla regolazione dell'intelligenza artificiale, contiene specifiche previsioni verticali volte a definire le condizioni e i limiti per l'impiego di sistemi di IA in specifici settori, tra cui quello sanitario, con rilevanti implicazioni sulla gestione dei dati relativi alla salute. Sebbene si tratti di un'iniziativa legislativa ancora in una fase embrionale, si rende opportuna una disamina delle disposizioni contenuto nel disegno di legge al fine di delineare compiutamente la cornice regolatoria di riferimento, ponendosi in un'ottica di coordinamento e complementarità rispetto al *corpus* giuridico europeo.

In particolare, risultano di particolare interesse le previsioni di cui agli artt. 7 e 8 del disegno di legge in esame. Il primo, per vero, parrebbe limitarsi a dichiarazioni di principio relative al ruolo cruciale che l'intelligenza artificiale può ricoprire nel sistema sanitario. Viene, infatti, sottolineato che l'impiego di tecnologie è in grado di apportare un contributo significativo per lo sviluppo di metodologie innovative con finalità di prevenzione e cura delle patologie. Ciò nondimeno, tale avanzamento tecnologico deve necessariamente coniugarsi con un rigoroso rispetto dei diritti e delle libertà fondamentali dell'individuo, prestando particolare attenzione alle esigenze di tutela dei dati personali.

Proseguendo nell'analisi delle disposizioni relative al settore sanitario, l'art. 8 fornisce, non senza criticità, una disciplina del trattamento di dati per finalità di ricerca e sperimentazione, laddove stabilisce che «i trattamenti di dati, anche personali, eseguiti da soggetti pubblici e privati senza scopo di lucro per la ricerca e la sperimentazione scientifica nella realizzazione di sistemi di intelligenza artificiale per finalità di prevenzione, diagnosi e cura di malattie, sviluppo di farmaci, terapie e tecnologie riabilitative, realizzazione di apparati medicali […], di salute pubblica, incolumità della persona, salute e sicurezza sanitaria, in quanto necessari ai fini della realizzazione e dell'utilizzazione di banche dati e modelli di base» sono ritenuti di interesse pubblico, ai sensi dell'art. 9 GDPR par. 2 lett g)». L'attuale formulazione parrebbe fondare i trattamenti di dati finalizzati alla sperimentazione e allo sviluppo di modelli di intelligenza artificiale con finalità di cura e di sperimentazione sull'interesse pubblico, ai sensi del richiamo alla base giuridica di cui alla lettera g) dell'art. 9 GDPR, rendendo, dunque, superfluo il consenso degli interessati. Il comma 2 della disposizione in esame, inoltre, fornisce una base giuridica per il c.d. uso secondario dei dati personali (purché privi degli elementi identificativi diretti) per lo sviluppo di sistemi di intelligenza artificiale con finalità di cura o di sperimentazione. Tuttavia, tale trattamento è sottoposto sia all'approvazione dei comitati etici che a quella del Garante della Protezione dei Dati Personali, che potrebbe, *rec sic stantibus*, bloccare il trattamento entro 30 giorni dalla data in cui gli è stato notificato.

La norma presenta, in effetti, alcuni profili critici che meritano un approfondimento. Da un lato, ci si occupa esclusivamente dei soggetti, pubblici o privati, senza scopo di lucro, lasciando una lacuna normativa per quanto riguarda le medesime attività ove svolte da soggetti che perseguono uno scopo di lucro. Dall'altro, il nuovo regime parrebbe essere in contrasto con la disciplina emergente dalla riforma del Codice Privacy, di cui si è detto sopra. In effetti, ai sensi del novellato art. 110 Codice Privacy, è stata rimossa l'autorizzazione preventiva del Garante per l'uso secondario di dati ai fini di ricerca medica-scientifica, rimanendo esclusivamente l'obbligo di adottare misure appropriate per tu-

telare i diritti, le libertà e i legittimi interessi dell'interessato. Ebbene, il regime che verrebbe introdotto con l'attuale formulazione del DDL IA rischierebbe di generare un'antinomia difficilmente risolvibile, aprendo la strada ad incertezze giuridiche in un campo che, paradossalmente, necessita di un consolidamento dei già labili confini di manovra.

Di fronte al *vulnus* giuridico appena delineato, si rende necessaria un'ulteriore riflessione volta a considerare come la tecnologia, pur sollevando inediti e complessi interrogativi normativi, possa, al contempo, fornire nuovi potenziali strumenti risolutivi di tali criticità. Ebbene, nel contesto sinora descritto, appare evidente che il ricorso a tecniche che consentano di garantire maggiori tutele per la riservatezza dei dati sia spesso idoneo a soddisfare i requisiti posti dalla normativa: anche in questo caso, l'uso di dati sintetici potrebbe contribuire alla rimozione degli elementi identificativi che la norma sopra analizzata pone quale *condicio sine qua non* per il trattamento secondario dei dati finalizzati allo sviluppo di sistemi di intelligenza artificiale.[88] Inoltre, ove utilizzati insieme ad altre tecniche di cifratura e anonimizzazione nonché a misure organizzative a tutela della riservatezza, i dati sintetici potrebbero contribuire a raggiungere uno standard di anonimizzazione tale da escludere l'applicazione delle garanzie previste dal GDPR e, dunque, rendere più agevole la condivisione e lo sfruttamento dei dati per lo sviluppo di sistemi di intelligenza artificiale. A tale fine, le certificazioni europee potrebbero ricoprire un ruolo significativo quale ulteriore strumento per supportare la conformità dei processi e dei risultati della sintetizzazione alla normativa vigente per fornire elementi di mitigazione di responsabilità a carico dei titolari e dei responsabili dei trattamenti. Inoltre, uno schema di certificazione, fondato sui criteri di approvati dalle autorità di controllo, potrebbe, da un lato, rafforzare la fiducia da parte di tutti gli *stakeholder* coinvolti e, dall'altro, favorire il monitoraggio (e, dunque, anche il perfezionamento) delle tecniche impiegate per generare i dati sintetici anche con riferimento alle garanzie per i dati personali.

Peraltro, sempre alla luce della formulazione attuale, si prospetta un ampio potere del Garante della Protezione dei Dati Personali: investita del potere di autorizzare il trattamento finalizzato allo sviluppo di sistemi di intelligenza artificiale con finalità di cura o di sperimentazione, l'*authority* italiana potrebbe trovarsi nella posizione di esaminare l'adeguatezza delle misure adottate dal titolare del trattamento, esprimendosi anche sulla natura dei dati sintetici nell'ambito dello sviluppo di sistemi e applicazioni di IA in campo sanitario nonché sulla loro idoneità a consentire una condivisione dei dati, nel rispetto delle garanzie in materia di protezione e sicurezza dei dati personali.

Il disegno di legge, qui brevemente analizzato, prescrive, da un lato, un quadro in linea con i principi giuridici già affermati a livello comunitario, soprattutto con riferimento alla protezione dei dati personali e all'impiego di sistemi di intelligenza artificiale; dall'altro, tuttavia, presenta dei profili di criticità che richiederebbero un maggior coordinamento con le fonti europee e nazionali, così da garantire maggiore certezza giuridica agli operatori economici e promuovere, dunque, uno sviluppo tecnologico nel rispetto dei diritti fondamentali.

[88] Si veda a tal proposito Garante della Protezione dei Dati Personali, *Provvedimento del 2 agosto 2024, n. 477*.

4. Conclusioni finali e valutazioni prospettiche sotto un profilo giuridico ed etico

I dati sintetici rappresentano una tecnologia innovativa che sta assumendo un ruolo sempre crescente in vari settori, da quello sanitario allo sviluppo delle applicazioni di guida autonoma. Nell'epoca dei *big data*, il ricorso ai dati sintetici riveste un ruolo fondamentale, soprattutto, con riferimento all'addestramento e allo sviluppo di sistemi di intelligenza artificiale.

L'adozione di tali tecnologie promette significativi vantaggi sotto diversi aspetti. In primo luogo, i dati sintetici consentono di ridurre sensibilmente il dispendio di risorse economiche ed organizzative connesse alla creazione di quei dataset di grandi dimensioni, necessari per l'addestramento efficace di algoritmi di *machine learning*. Inoltre, i processi di sintetizzazione mettono a disposizione enormi volumi di elevata qualità che possono essere utilizzati, spesso, per correggere le distorsioni provocate dai dataset tradizionali (così come, peraltro, suggerito dall'art. 10 AIA, sopra analizzato).

Oltre a questi vantaggi prettamente tecnici ed economici, i dati sintetici rappresentano un'innovativa tecnologia di protezione dei dati personali (PETs), potenzialmente in grado di fornire garanzie più elevate in termini di tutela della privacy e della sicurezza delle informazioni. Infatti, i dati sintetici, essendo generati artificialmente senza utilizzare dati personali reali, possono essere impiegati per addestrare modelli di intelligenza artificiale senza esporre direttamente i dati degli individui.

Nonostante rimanga meritevole di considerazione il *trade-off* tra utilità dei dati generati e tutela della riservatezza, le moderne tecniche di sintetizzazione dei dati consentono di implementare strumenti innovativi a protezione dei dati personali, potenzialmente in grado di fornire garanzie più elevate in termini di privacy e sicurezza delle informazioni. Del resto, l'analisi della normativa esistente non sembra accogliere una nozione assoluta di anonimizzazione, bensì esclusivamente una tale da escludere, secondo un criterio di ragionevolezza e proporzionalità, il rischio di reidentificazione. I dati sintetici, eventualmente integrati dall'implementazione di ulteriori misure organizzative e tecniche - laddove, come precedentemente illustrato, se ne ravvisi la necessità - appaiono dunque idonei a garantire gli standard di sicurezza e protezione richiesti.

Il recente AI Act sembrerebbe confermare tale lettura, qualificando i dati sintetici come valida alternativa, insieme ai dati anonimi, ai dati personali nel contesto della correzione dei *bias* dei sistemi già sviluppati oppure dello sviluppo dei sistemi nel perimetro degli spazi di sperimentazione normativa previsti dal regolamento.

Si aggiunga, inoltre, che, in questo contesto, si inserisce il nuovo quadro normativo europeo, volto a promuovere la creazione di un mercato comune dei dati, preservando al contempo standard elevati in materia di privacy e sicurezza. Tale quadro normativo comprende il Data Governance Act, il Data Act, la proposta di regolamento inteso ad istituire uno spazio europeo dei dati sanitari (European Health Data Space), i quali, pur con diverse sfumature, mirano a incentivare la condivisione e il riutilizzo dei dati, attribuendo un ruolo fondamentale alle tecniche di sintetizzazione dei dati come misure idonee a garantire il rispetto dei principi di protezione dei dati personali e di *accountability*.

Dunque, i dati sintetici possono svolgere un ruolo cruciale, consentendo, nel contesto di una comprensiva strategia di gestione del dato, lo sfruttamento delle potenzialità offerte dai dati e dall'intelligenza artificiale, senza compromettere la tutela dei diritti fondamentali degli individui.

Tuttavia, è necessario sottolineare che il nuovo assetto normativo europeo si trova ancora in una fase di definizione. Pertanto, sarà fondamentale attendere l'evoluzione della prassi applicativa per

comprendere appieno le modalità di coordinamento tra le diverse disposizioni in materia di dati e IA, nonché le implicazioni concrete dell'utilizzo dei dati sintetici in un ambiente tecnico e normativo dinamico e in costante trasformazione.

Sarà, inoltre, essenziale il ruolo di organizzazioni pubbliche e private per elaborare linee-guida e *best practice* per l'impiego dei dati sintetici, al fine di garantire che tali tecnologie siano adoperate in modo responsabile e conforme ai principi etici e normativi. In particolare, sarà necessario affrontare sfide legate alla qualità, alla trasparenza e alla *governance* dei dati sintetici, assicurando che i modelli di generazione siano robusti, accurati e privi di *bias* indesiderati.

Solo attraverso un'attenta osservazione delle prassi emergenti e un costante dialogo tra legislatori, *stakeholder* e comunità scientifica, sarà possibile valutare pienamente l'impatto e le opportunità offerte da dati sintetici nell'ambito del nuovo quadro regolamentare europeo e gettare, così, le basi per un ecosistema digitale più sicuro, etico e rispettoso dei diritti fondamentali degli individui.