Data Intermediaries Alliance

ISTITUTO ITALIANO PER LA PRIVACY E LA VALORIZZAZIONE DEI DATI

IL FUTURO REGOLATORIO DEI DATI SINTETICI

La sintetizzazione dei dati come risorsa per ricerca scientifica, innovazione e politiche pubbliche nel panorama giuridico europeo

15 luglio 2024

DOI: **10.5281/zenodo.12751884**

Autori:

Giulia Finocchiaro[i], **Antonio Landi**[ii], **Gianluca Polifrone**[iii], **Davide Ruffo**[iv], **Francesco Torlontano**[v]

Un paper promosso da **Istituto Italiano per la Privacy e la Valorizzazione dei Dati** e ***Data Intermediaries Alliance***

[i] G.Finocchiaro@istitutoprivacy.eu
[ii] A.Landi@istitutoprivacy.eu
[iii] Gianluca.Polifrone@biotecnopolo.it
[iv] Davide.Ruffo@aindo.com
[v] F.Torlontano@istitutoprivacy.eu

Indice

1. Introduzione

L'evoluzione delle tecnologie di trattamento dei dati e la loro crescente incidenza nella vita quotidiana hanno posto nuove e complesse sfide in materia di protezione dei dati personali. In questo contesto, si inserisce il dibattito sui dati sintetici, una tecnologia emergente che offre nuove prospettive per la gestione dei dati. Al riguardo, risulta essenziale considerare i dati sintetici non solo come una tecnica per garantire la privacy e la sicurezza dei dati, ma anche come un mezzo per promuovere l'innovazione e il progresso tecnologico, senza trascurare le necessarie garanzie per tutelare i diritti e le libertà degli individui. Pertanto, l'obiettivo di questo *paper* è quello di analizzare il ruolo dei dati sintetici come strumento di anonimizzazione, esplorando le implicazioni sia dal punto di vista tecnico, sia da quello giuridico. Nel primo capitolo del presente documento viene fornito un inquadramento tecnico/giuridico della nozione di dato sintetico, analizzando la differenza tra anonimizzazione e pseudonimizzazione in base a quanto prescritto dalla normativa in materia di protezione dei dati personali e ai provvedimenti e agli orientamenti delle varie autorità intervenute sul tema. Inoltre, sono state illustrate le varie tecniche di sintetizzazione in base allo stato dell'arte, con un focus sulla sintetizzazione come tecnica di anonimizzazione. Nel secondo capitolo viene analizzato il *framework* normativo attuale, nonché quello "futuro". Questa analisi include l'interpretazione delle normative esistenti e la previsione delle tendenze future in materia di protezione dei dati, ponendo particolare attenzione al ruolo dei dati sintetici nel contesto della privacy e della sicurezza dei dati. In particolare, sono state esaminate le disposizioni attualmente contenute nel Regolamento (UE) 2016/679 (di seguito anche "**GDPR**")[6], nel Data Governance Act[7], e nel Data Act[8], valutando come la sintetizzazione dei dati possa essere applicata o interpretata in base alle indicazioni contenute in questi regolamenti. Inoltre, sono state esplorate le prospettive future, come l'inclusione del dato sintetico nell'AI Act[9], l'impatto dello European Health Data Space[10], nonché l'opportunità di utilizzo della sintetizzazione nella sanità digitale rispetto al panorama italiano. L'ultimo capitolo del paper si focalizza sull'importanza crescente della sintetizzazione dei dati in contesti ulteriori rispetto a quello di ricerca scientifica, evidenziando la necessità di aggiornamenti normativi e interpretazioni evolutive per supportare l'innovazione tecnologica nel rispetto dei diritti e delle libertà degli individui. Al riguardo, è stato evidenziato come il Parere 05/2014 sulle tecniche di anonimizzazione (WP 216)[11] necessiti di un approccio rinnovato da parte delle autorità europee sul tema, considerando la sintetizzazione come rientrante all'interno delle tecniche di anonimizzazione, ove realizzata attraverso determinati parametri[12].

[6] Regolamento (UE) 2016/679 del Parlamento Europeo e del Consiglio del 27 aprile 2016 relativo alla protezione delle persone fisiche con riguardo al trattamento dei dati personali, nonché alla libera circolazione di tali dati e che abroga la direttiva 95/46/CE (regolamento generale sulla protezione dei dati).
[7] 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).
[8] 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).
[9] Proposta di Regolamento del Parlamento europeo e del Consiglio che stabilisce regole armonizzate sull'intelligenza artificiale (legge sull'intelligenza artificiale) e modifica alcuni atti legislativi dell'Unione (del 21 aprile 2021).
[10] Proposta di Regolamento del Parlamento europeo e del Consiglio sullo spazio europeo dei dati sanitari (del 3 maggio 2022).
[11] Parere 05/2014 del Gruppo di lavoro per la tutela dei dati (WP 216) sulle tecniche di anonimizzazione, adottato il 10 aprile 2014, https://ec.europa.eu/justice/article-29/documentation/opinion-recommendation/files/2014/wp216_en.pdf
[12] BEDUSCHI, A. (2024). *"Synthetic data protection: Towards a paradigm change in data regulation?"*. Big Data & Society, 11. https://doi.org/10.1177/20539517241231277.

2. Inquadramento tecnico/giuridico dei dati sintetici

2.1 Definizione di dato sintetico e possibile utilizzo

La ricerca e l'innovazione richiedono, per loro natura, l'accesso a vasti quantitativi di informazioni che, qualora riguardino persone fisiche, si qualificano come dati personali. Questo implica la necessità di un loro trattamento in conformità a normative stringenti, volte a salvaguardare la libertà e la dignità umana. Un approccio comunemente adottato per bilanciare l'avanzamento tecnologico con il rispetto per la privacy consiste nell'utilizzo di dati anonimizzati, che non rientrano nell'ambito di applicazione della normativa in materia di protezione dei dati personali.

Tuttavia, questa strategia non ha sempre prodotto i risultati auspicati, principalmente a causa del *«divario concettuale tra il pensiero legale e matematico sulla privacy dei dati»*, come opportunamente osservato da Aloni Cohen e Kobbi Nissim[13]. Recentemente, si è assistito all'emergere di un'innovativa applicazione: i **dati sintetici**, che riuscirebbe a contemperare da un lato le esigenze di tutela della privacy degli interessati, dall'altro rappresenta un sostegno all'innovazione, superando gli "ostacoli" del passato.

Per quanto concerne la definizione di dato sintetico, una possibile classificazione viene proposta dal tech champion Robert Riemann, il quale afferma che "*I dati sintetici sono dati artificiali generati da dati originali e da un modello che viene addestrato a riprodurre le caratteristiche e la struttura dei dati originali";* e *"ciò significa che i dati sintetici e i dati originali dovrebbero fornire risultati molto simili quando vengono sottoposti alla stessa analisi statistica*"[14]. Pertanto, tali dati rappresentano il risultato di processi di sintetizzazione che sfruttano algoritmi di apprendimento automatico e intelligenza artificiale per creare insiemi di dati nuovi che "imitano" le caratteristiche e le relazioni statistiche di set di dati reali, senza però contenere alcun elemento identificativo diretto. In altre parole, i dati sintetici, se correttamente generati, non solo riflettono fedelmente la distribuzione, le tendenze e i pattern dei dati originali, ma lo fanno in maniera tale da annullare qualsiasi rischio di ricondurre le informazioni a singoli individui. Questo processo non necessita della conservazione di dati "primari" legati a persone specifiche e permette di generare dati utili per analisi statistiche e di machine learning. Di conseguenza, se da un lato *"si possono creare dataset per il training di algoritmi di machine learning, così consentendo a qualsiasi modello di apprendere su una base informativa più ampia e rappresentativa (evitando quindi anche i rischi di overfitting) e simulando situazioni nuove o complesse raramente riscontrabili (o impossibili da riscontrare), dall'altra si garantisce la piena tutela della riservatezza degli interessati cui sono riferiti i dati iniziali, nel rispetto del principio di minimizzazione"*[15].

In base all'impostazione appena descritta, al dato sintetico – se reso sufficientemente anonimo da impedire o da non consentire più l'identificazione dell'interessato (utilizzando la formulazione contenuta nel Considerando 26 del GDPR) – non dovremmo riconoscere

---

[13] COHEN, A., & NISSIM, K., (14 aprile 2020). Towards formalizing the GDPR's notion of singling out, *Proceedings of the National Academy of Sciences of the United States of America*, 117 (15). 8344-8352.

[14] ROBERT RIEMANN, definizione proposta nel blog TechSonar del Comitato Europeo per la Protezione dei Dati (EDPB) https://edps.europa.eu/press-publications/publications/techsonar/synthetic-data_en.

[15] LORENZO CRISTOFARO e GABRIELE FRANCO, *"Dati sintetici, la giusta sintesi tra innovazione e privacy: stato dell'arte e scenari"*, Agenda Digitale, 26/07/2021.

garanzie in materia di protezione dei dati personali. In particolare, quanto appena affermato *"mira "a valle", e quindi all'output del processo di sintetizzazione: se il dato sintetico rispetta – ormai trasformato – i criteri di un'adeguata anonimizzazione a monte, allo stato dell'arte e della tecnica e degli elementi contestuali disponibili, esso si può sottrarre alle tutele e agli adempimenti in materia di protezione dei dati personali, per insussistenza di re-identificabilità e di personalità. In sostanza, qui la sfida sarebbe solo quella di assicurare la perdurante anonimità del dato, per evitare in toto l'applicazione del GDPR"*[16].

Per illustrare il meccanismo alla base di un algoritmo di generazione di dati sintetici, è opportuno fornire un esempio pratico. Immaginiamo di avere un annuario scolastico, contenente fotografie, dati identificativi e citazioni (relative alle ambizioni future) degli studenti dell'ultimo anno, oltre a informazioni come la media dei voti, attività extracurriculari e iscrizioni future al college. Da questo annuario, si potrebbero estrarre dati statistici significativi, come la percentuale di abbandoni scolastici o l'effetto delle attività extracurriculari sul rendimento accademico, filtrati per genere o etnia.

Ora, ipotizziamo di alimentare un algoritmo di sintetizzazione con i dati contenuti nel suddetto annuario. Otterremmo un nuovo annuario, di un'altra scuola inesistente, con studenti, fotografie e dati completamente diversi e non riconducibili a quelli originali. Nonostante ciò, interrogando questo nuovo database artificiale con le stesse domande poste a quello reale, otterremmo risposte statisticamente identiche. Questo perché il processo di sintetizzazione, se generato correttamente, crea un insieme di dati completamente nuovi, ma statisticamente equivalenti al dataset originale, senza necessità di utilizzare direttamente dati personali.

Pertanto, l'utilizzo dei dati sintetici assume un ruolo cruciale sotto molteplici aspetti. Dal punto di vista della privacy, offre una soluzione efficace per mantenere l'anonimato degli interessati, superando i limiti dell'anonimizzazione tradizionale, la quale, in alcuni contesti, può essere vulnerabile rispetto a tecniche di re-identificazione a fronte dell'avanzamento delle capacità computazionali e analitiche. I dati sintetici, quindi, permettono di lavorare con set di dati ricchi e informativi, riducendo al minimo i rischi connessi agli adempimenti prescritti dalla normativa in materia di protezione dei dati personali.

Dal punto di vista dell'innovazione e della ricerca, i dati sintetici aprono scenari inediti e promettenti[17]. Ad esempio, nel settore sanitario, consentono la condivisione di informazioni per studi e analisi senza esporre "dati sensibili" dei pazienti, accelerando la ricerca scientifica e lo sviluppo di nuove cure. Nell'ambito dell'intelligenza artificiale, facilitano l'addestramento di algoritmi su set di dati vasti e vari, migliorando la qualità dei modelli predittivi e delle simulazioni senza incorrere in questioni etiche o legali legate all'uso di dati personali[18].

Un altro aspetto significativo dei dati sintetici è la loro capacità di mitigare i *bias* presenti nei dati originali. Difatti, attraverso tecniche di generazione controllata, è possibile produrre set di dati equilibrati e rappresentativi, contribuendo a ridurre le disparità e migliorare l'equità dei sistemi di intelligenza artificiale. Questo aspetto è di fondamentale importanza al fine di

---

[16] LUCA BOLOGNINI, *"Un focus su dati sintetici e privacy, con proposte evolutive di soft regulation"*, Giuffrè Francis Lefebvre.
[17] VAN BREUGEL, B., & VAN DER SCHAAR, M. (2023). *"Beyond privacy: Navigating the opportunities and challenges of synthetic data"*. arXiv. https://arxiv.org/abs/2304.03722
[18] SAVAGE, N. (2023). *"Synthetic data could be better than real data"*. Nature. Advance online publication. https://doi.org/10.1038/d41586-023-01445-8.

garantire che le tecnologie emergenti siano inclusive e non perpetuino discriminazioni preesistenti.

In conclusione, i dati sintetici rappresenterebbero una soluzione avanzata che armonizza le esigenze di privacy, sicurezza dei dati e innovazione. Infatti, come indicato da Bellovin "*I dati sintetici rappresentano un progresso. Anche se non è una pallottola d'argento, il metodo ci permette di porre fine alla corsa agli armamenti della deidentificazione-reidentificazione e di concentrarci su ciò che conta: dati utili e privati. Per questo motivo, raccomandiamo alla comunità della privacy di accettare i dati sintetici come un valido passo avanti nel problema della privacy dei database*"[19].In generale, i dati sintetici forniscono un equilibrio tra la necessità di proteggere l'identità degli individui e la volontà di sfruttare appieno il potenziale dei dati per la ricerca e lo sviluppo tecnologico, configurandosi come strumento essenziale nell'era digitale, con applicazioni che spaziano dal settore sanitario a quello tecnologico, dalla ricerca accademica all'industria.

2.2 Anonimizzazione vs Pseudonimizzazione

L'analisi dei concetti di anonimizzazione e pseudonimizzazione, risulta essere determinante al fine di inquadrare il perimetro di applicabilità degli adempimenti prescritti dalla normativa in materia di protezione dei dati personali, considerando che (alla luce di quanto illustrato nel precedente paragrafo) la sintetizzazione rientrerebbe nella prima categoria (i.e. l'anonimizzazione).
Scendendo nel dettaglio, l'anonimizzazione e la pseudonimizzazione sono due misure tecniche fondamentali nella protezione dei dati personali, le quali si inseriscono nel contesto normativo delineato dal GDPR. Entrambe le pratiche mirano a ridurre i rischi associati al trattamento dei dati personali, ma si distinguono per metodologia e finalità.
In particolare, per quanto concerne la **nozione di anonimizzazione**, tale tecnica consiste nel processo volto a rendere anonimi i dati personali. Prendendo in considerazione il riferimento contenuto nel Considerando 26 del GDPR, i dati anonimi sono qualificati come le *"informazioni che non si riferiscono a una persona fisica identificata o identificabile o a dati personali resi sufficientemente anonimi da impedire o da non consentire più l'identificazione dell'interessato"*. Pertanto, il concetto di "anonimato" è strettamente legato all'identificabilità dell'individuo. Il suddetto considerando 26 del GDPR specifica quindi che i dati possono essere considerati anonimi o sufficientemente anonimi se, rispettivamente:

- non si riferiscono a una persona fisica identificata o identificabile, oppure
- impediscono o non consentono più l'identificazione degli interessati.

Il considerando menzionato chiarisce che, per determinare se una persona fisica sia identificabile, è essenziale valutare tutte le risorse, come i metodi di riconoscimento, che il titolare del trattamento dati o una terza parte potrebbero logicamente utilizzare per riconoscere una persona fisica sia direttamente sia indirettamente. Nel valutare la plausibile possibilità di impiego di tali risorse per l'identificazione di un soggetto, lo stesso considerando 26 del GDPR prosegue precisando che è necessario considerare una varietà di fattori concreti, inclusi i costi e il tempo richiesto per identificare l'individuo, tenendo conto non solo delle tecnologie esistenti al momento del trattamento dei dati, ma anche dell'evoluzione tecnologica.

---

[19] STEVEN M. BELLOVIN, PREETAM K. DUTTA, and NATHAN REITINGER– *"Privacy and synthetic data sets"* (https://law.stanford.edu/wp-content/uploads/2019/01/Bellovin_20190129.pdf).

Alla luce delle indicazioni contenute nel GDPR, è possibile concludere che l'anonimizzazione consiste nella trasformazione dei dati personali in modo tale che l'interessato non sia più identificabile. Ciò implica un processo irreversibile per cui, anche avvalendosi di tutte le informazioni o tecniche addizionali ragionevolmente utilizzabili, non sarebbe possibile risalire all'identità dell'interessato. Pertanto, i dati anonimizzati non sono considerati dati personali e, di conseguenza, il GDPR non trova applicazione, poiché non applicabile al trattamento di informazioni anonime. L'anonimizzazione, quindi, offrirebbe garanzie forti in termini di protezione della privacy delle persone fisiche, ma implica anche la perdita di specifiche informazioni utili per determinate analisi o elaborazioni.

Invece, per quanto riguarda il **concetto di pseudonimizzazione**[20], tale tecnica viene espressamente definita dall'articolo 4, paragrafo 5, del GDPR 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"* (come, ad esempio, la separazione logica e l'accesso limitato). Pertanto, a differenza dei dati anonimizzati, i dati pseudonimizzati possono essere attribuiti all'individuo a cui si riferiscono, ma per farlo è necessario l'uso di informazioni aggiuntive. La pseudonimizzazione può quindi essere considerata come una misura utile a garantire la sicurezza nel trattamento dei dati personali, che i titolari del trattamento dovrebbero adottare in conformità agli obblighi derivanti dall'articolo 32 del GDPR; tuttavia, essa non esclude la "natura personale" dei dati e, pertanto, non preclude l'applicazione della normativa in materia di protezione dei dati personali.

A questo proposito, i *dataset* che includono dati personali possono contenere identificatori diretti e indiretti, che permettono di individuare o rendere identificabile una persona fisica. Un identificatore diretto è un'informazione specifica che fa riferimento a un individuo, come il nome o un numero di identificazione. Un identificatore indiretto (anche chiamato quasi-identificatore) è qualsiasi pezzo di informazione (ad esempio, una posizione geografica in un determinato momento o un'opinione su un certo argomento) che potrebbe essere utilizzato da qualcuno che ha informazioni su quella persona fisica al fine di re-identificare quella persona fisica nel *dataset*. La probabilità di re-identificazione è quindi la probabilità, all'interno di un determinato *dataset*, di re-identificare una persona fisica, trasformando i dati pseudonimizzati nuovamente in dati personali attraverso l'uso di tecniche di abbinamento dei dati o tecniche simili. L'utilità di un *dataset* può essere valutata in base alla capacità che le informazioni in esso contenute hanno rispetto allo scopo che si intende perseguire.

Pertanto, la pseudonimizzazione rappresenta un processo mediante il quale i dati personali vengono elaborati in modo tale da non poter essere attribuiti a un interessato specifico, senza l'utilizzo di informazioni aggiuntive. Queste informazioni aggiuntive, per poter essere considerato perfezionato il processo di pseudonimizzazione, devono essere conservate separatamente e soggette a misure tecniche e organizzative adeguate per garantire che i dati personali non siano attribuiti a una persona fisica identificata o identificabile. A differenza

[20] LUCA BOLOGNINI, CAMILLA BISTOLFI, *"Pseudonymization and impacts of Big (personal/anonymous) Data processing in the transition from the Directive 95/46/EC to the new EU General Data Protection Regulation"*, Computer Law & Security Review, Volume 33, Issue 2, 2017, Pages 171-181, ISSN 0267-3649, https://doi.org/10.1016/j.clsr.2016.11.002 . (https://www.sciencedirect.com/science/article/pii/S0267364916302151).

dell'anonimizzazione, la pseudonimizzazione è reversibile e, di conseguenza, richiede comunque il rispetto degli adempimenti previsti dalla normativa privacy.
Tuttavia, essa è concepita per ridurre i rischi per i diritti e le libertà degli interessati e aiutare i titolari del trattamento a rispettare i loro obblighi di protezione dei dati, mantenendo al contempo la possibilità di effettuare analisi sui dati pseudonimizzati.
In conclusione, entrambe le tecniche presentano vantaggi e limitazioni. L'anonimizzazione offre un livello elevato di protezione della riservatezza, rimuovendo permanentemente la possibilità di identificazione dell'interessato, ma limita l'utilità dei dati per ulteriori elaborazioni. La pseudonimizzazione, d'altro canto, preserva una certa utilità dei dati permettendo analisi più approfondite, pur richiedendo l'adozione di misure di sicurezza adeguate a prevenire la re-identificazione dell'interessato. La scelta tra anonimizzazione e pseudonimizzazione dipenderà dagli obiettivi specifici del trattamento dei dati che si vuole porre in essere e dai requisiti di protezione dei dati personali applicabili.

2.2 I 10 fraintendimenti da evitare quando si parla di anonimizzazione e pseudonimizzazione (secondo AEPD ed EDPS)

Come analizzato nel precedente paragrafo, l'anonimizzazione e la pseudonimizzazione sono due concetti fondamentali nel contesto della protezione dei dati personali, tuttavia, la complessità e la specificità tecnica che caratterizzano questi processi hanno spesso generato una serie di fraintendimenti comuni.
Nell'aprile del 2021, l'Autorità di Controllo spagnola (*Agencia Española de Protección de Datos* – "**AEPD**") e il Garante Europeo per la Protezione dei Dati (*European Data Protection Supervisor* – "**EDPS**") hanno pubblicato congiuntamente un documento che analizza i dieci fraintendimenti più diffusi relativi all'anonimizzazione[21]. Questo documento è fondamentale per comprendere le complessità e i comuni equivoci riguardanti i processi di anonimizzazione e pseudonimizzazione, specialmente alla luce dell'uso crescente che possono assumere i dati sintetici come strumento di anonimizzazione.
Di seguito, si analizzano i dieci fraintendimenti più diffusi e si forniscono spiegazioni per ciascuno di essi, analizzando le potenziali ripercussioni rispetto alla sintetizzazione, alla luce delle indicazioni fornite dalle suddette autorità.

I. **Fraintendimento:** ***"La pseudonimizzazione è la stessa cosa dell'anonimizzazione"***. La pseudonimizzazione non deve essere equiparata o confusa con l'anonimizzazione per le ragioni illustrate nella precedente sezione. Difatti, come già analizzato, il GDPR definisce la "pseudonimizzazione" come quel trattamento dei dati personali effettuato attraverso modalità tali che questi ultimi (i.e. i dati personali) non possano più essere attribuiti a un interessato senza l'uso di informazioni aggiuntive, mantenute separatamente e protette da misure tecniche e organizzative per evitare l'attribuzione a persone fisiche identificate o identificabili. Pertanto, i dati pseudonimizzati rimangono dati personali perché l'identificazione è ancora possibile attraverso l'impiego di informazioni aggiuntive. I dati anonimi, invece, non possono

[21] Agencia Española Protecciòn Datos & European Data Protection Supervisor: *"10 Misunderstandings Related to Anonymisation"*, accessible here: https://edps.europa.eu/data-protection/our-work/publications/papers/aepd-edps-joint-paper-10-misunderstandings-related_en.

essere associati a individui specifici e, una volta resi effettivamente anonimi, non rientrano nell'ambito di applicazione del GDPR.

Questo fraintendimento sottolinea l'importanza della distinzione tra pseudonimizzazione e anonimizzazione, specialmente quando si parla di sintetizzazione. Difatti, i dati sintetici, possono raggiungere un livello di anonimizzazione tale da assicurare che i soggetti, i cui dati sono stati sintetizzati, non siano identificabili, superando i limiti della pseudonimizzazione.

II. **Fraintendimento: *"La cifratura è anonimizzazione"***. La cifratura non è una tecnica di anonimizzazione, ma può essere un potente strumento di pseudonimizzazione. Il processo di cifratura impiega chiavi confidenziali per trasformare le informazioni in modo da ridurre il rischio di uso improprio, garantendo la confidenzialità per un periodo definito. In considerazione della necessità di mantenere accessibili le informazioni originarie, le trasformazioni operate dagli algoritmi di cifratura sono progettate per essere reversibili, attraverso il processo di decifratura. A tal riguardo, si evidenzia che le chiavi segrete utilizzate per la decifratura, rappresentano le "informazioni aggiuntive" (come specificato nel primo fraintendimento), che rendono i dati personali leggibili e abilitano l'identificazione. In teoria, si potrebbe considerare l'eliminazione della chiave di cifratura come un mezzo per rendere anonimi i dati cifrati; tuttavia, questa assunzione non è corretta. La cancellazione o lo stato sconosciuto della chiave di decifratura non garantisce che i dati cifrati non possano essere decifrati. Numerosi fattori influenzano la confidenzialità dei dati cifrati, in particolare a lungo termine. Questi fattori includono la robustezza dell'algoritmo e della chiave di cifratura, potenziali perdite di informazioni, complessità dell'implementazione, il volume dei dati cifrati e gli avanzamenti tecnologici (ad esempio, il calcolo quantistico).

La precisazione che la cifratura non è una tecnica di anonimizzazione, ma piuttosto una forma di pseudonimizzazione, è particolarmente rilevante anche quando si tratta di dati sintetici. Difatti, la creazione di dati sintetici non si basa su trasformazioni reversibili come la crittografia, ma mira a generare interamente nuovi set di dati che mantengono le proprietà statistiche dei dati originali, senza abilitare la re-identificazione.

III. **Fraintendimento: *"L'anonimizzazione dei dati è sempre possibile"***. In realtà non è sempre possibile ridurre il rischio di re-identificazione al di sotto di una soglia precedentemente definita mantenendo al contempo un dataset utile per uno specifico trattamento. L'anonimizzazione rappresenta un processo che cerca di trovare un equilibrio adeguato tra la riduzione del rischio di re-identificazione e la preservazione dell'utilità di un dataset per la finalità (o scopi) che intende perseguire. Tuttavia, in determinati contesti o con determinate tipologie di dati, i rischi di re-identificazione non possono essere sufficientemente mitigati. Questo scenario può sorgere quando il numero totale di potenziali individui nel dataset (definito come "universo dei soggetti") è troppo limitato, come nel caso di un dataset anonimo contenente solo 705 membri del Parlamento Europeo. Inoltre, i rischi possono persistere quando le categorie di dati sono altamente distintive tra gli individui, permettendo il loro isolamento (ad esempio, le impronte digitali dei dispositivi che accedono a un sito web specifico). Un altro caso

potenziale si verifica quando i dataset comprendono un numero sostanziale di attributi demografici[22] o dati di localizzazione[23].

Il terzo fraintendimento evidenzia quindi le sfide del processo di anonimizzazione, che andrebbero affrontate anche quando si procede con la generazione di dati sintetici. In particolare, è necessario considerare attentamente l'utilità del dataset rispetto al rischio di re-identificazione, un equilibrio che i dati sintetici cercano di raggiungere creando dati funzionali alle finalità da perseguire, ma non identificabili.

IV. **Fraintendimento: *"L'anonimizzazione è per sempre".*** Vi sono dei casi in cui l'anonimizzazione non può essere considerata permanente, poiché esiste il rischio che alcuni processi di anonimizzazione possano essere invertiti in futuro. Le circostanze potrebbero cambiare nel tempo e nuovi sviluppi tecnici e la disponibilità di informazioni aggiuntive potrebbero compromettere i precedenti processi di anonimizzazione. In particolare, le risorse informatiche e le tecnologie emergenti, o applicazioni innovative di tecnologie esistenti, disponibili per potenziali attaccanti che mirano a re-identificare un *dataset* anonimo, evolvono nel tempo. Ad esempio, attualmente, il *cloud computing* offre capacità di calcolo efficaci a livelli e prezzi precedentemente inimmaginabili. Guardando al futuro, l'avvento dei computer quantistici potrebbe ridefinire il concetto di ciò che attualmente è considerato "mezzi ragionevoli". Inoltre, la divulgazione di dati aggiuntivi nel tempo, come durante una violazione di dati personali, ha il potenziale per stabilire collegamenti tra dati precedentemente anonimi e individui identificati. La pubblicazione di registri che coprono molti decenni e contenenti informazioni altamente sensibili, come i casellari giudiziari, potrebbe ancora arrecare un danno significativo agli individui o ai loro parenti[24].

Di conseguenza, la natura dinamica degli avanzamenti tecnologici, comporta che i dati anonimizzati di oggi potrebbero diventare identificabili in futuro, un rischio che, anche se probabilmente in misura assai minore, potrebbe coinvolgere anche i dati sintetici. Questo fraintendimento enfatizza la necessità di valutazioni del rischio continue, anche dopo che i dati sintetici sono stati generati.

V. **Fraintendimento: *"L'anonimizzazione riduce sempre la probabilità di re-identificazione di un dataset a zero".*** Il processo di anonimizzazione e il modo in cui viene implementato avranno un'influenza diretta sulla probabilità dei rischi di re-identificazione. Un processo di anonimizzazione robusto cerca di ridurre il rischio di re-identificazione a una soglia predefinita. Questa soglia dipende da vari fattori, inclusa la presenza di controlli di mitigazione esistenti (assenti nel contesto della divulgazione pubblica), l'impatto potenziale sulla privacy degli individui in caso di re-identificazione, le ragioni e le capacità di un attaccante di re-identificare i dati[25].

[22] ROCHER, L., HENDRICKX, J. M., & DE MONTJOYE, Y. A. (2019). *"Estimating the success of re-identifications in incomplete datasets using generative models"*. Nature communications, 10(1), 1-9, https://www.nature.com/articles/s41467-019-10933-3.

[23] XU, F., TU, Z., LI, Y., ZHANG, P., FU, X., & JIN, D. (2017, April). *"Trajectory recovery from ash: User privacy is not preserved in aggregated mobility data"*. In Proceedings of the 26th international conference on world wide web (pp. 1241-1250), https://dl.acm.org/doi/abs/10.1145/3038912.3052620.

[24] GRAHAM, C. (2012). *"Anonymisation: managing data protection risk code of practice"*. Information Commissioner's Office https://ico.org.uk/media/1061/anonymisation-code.pdf.

[25] Vedasi al riguardo: *"External guidance on the implementation of the European Medicines Agency policy on the publication of clinical data for medicinal products for human use"* (2016), https://www.ema.europa.eu/en/documents/regulatory-procedural-guideline/external-guidance-implementation-european-medicines-agency-policy-publication-clinical-data_en-0.pdf.

Sebbene raggiungere il 100% di anonimizzazione sia l'obiettivo ottimale dal punto di vista della protezione dei dati personali, in certi casi ciò potrebbe non essere fattibile, e un rischio residuo di re-identificazione deve essere preso in considerazione.

Pertanto, riconoscere che nessun processo di anonimizzazione, inclusa la generazione di dati sintetici, può garantire l'anonimizzazione assoluta, richiede un approccio informato, critico e adattabile della gestione del rischio e l'accettazione dei rischi residui.

In ogni caso, è opportuno precisare che la sintetizzazione dei dati offre un profilo di rischio di identificazione sensibilmente inferiore rispetto ad alcune metodologie tradizionali di pseudonimizzazione.

VI. **Fraintendimento: *"L'anonimizzazione è un concetto binario che non può essere misurato"***. In realtà è possibile analizzare e misurare il grado di anonimizzazione. L'espressione "dati anonimi" non dovrebbe essere interpretata come una "etichetta binaria", suggerendo che i dataset possono essere categorizzati come anonimi o non anonimi. Invece, ogni record all'interno di un dataset porta una probabilità di essere re-identificato, basata sulla sua suscettibilità di essere isolato. Un processo di anonimizzazione robusto valuta sistematicamente il rischio di re-identificazione, sottolineando la necessità di gestire e controllare questo rischio nel tempo.

Ad eccezione di casi particolari in cui i dati sono estesamente generalizzati (ad esempio, un dataset che dettaglia il conteggio annuale dei visitatori di un sito web per paese), il rischio di re-identificazione non è mai ridotto a zero.

Questo fraintendimento risulta cruciale anche nell'attività di creazione dei dati sintetici, poiché tale processo coinvolge la valutazione e la quantificazione del rischio di re-identificazione, il quale risulta di gran lunga inferiore sui dataset sintetici, rispetto alle classiche tecniche di pseudonimizzazione. Inoltre, rifletterebbe l'importanza di valutare il grado di anonimizzazione nei dataset sintetici, che non è "cristallizzato nel tempo" ma andrebbe periodicamente rivalutato[26].

VII. **Fraintendimento: *"L'anonimizzazione può essere completamente automatizzata"***. Gli strumenti automatizzati possono essere utilizzati durante il processo di anonimizzazione, tuttavia, data l'importanza del contesto nella valutazione complessiva del processo, è necessario l'intervento di esperti umani. Difatti, è necessario analizzare il dataset originale, gli obiettivi che lo stesso intende perseguire, le tecniche da applicare e il rischio di re-identificazione associato ai dati risultanti.

L'identificazione e l'eliminazione di identificatori diretti, nota anche come "mascheramento", costituiscono un aspetto cruciale del processo di anonimizzazione. Tuttavia, è necessario seguire questo passaggio con la dovuta accortezza, prendendo in considerazione le altre fonti di identificazione (indiretta), tipicamente attraverso i quasi-identificatori. Mentre gli identificatori diretti sono relativamente semplici da identificare, gli identificatori indiretti potrebbero non essere sempre evidenti, e il mancato rilevamento di questi può portare a un'inversione del processo, risultando in una re-identificazione con implicazioni per la privacy degli interessati.

L'automazione può svolgere un ruolo cruciale in alcune fasi del processo di anonimizzazione, come l'eliminazione di identificatori diretti o l'applicazione

[26] Boudewijn, A. T. P., Ferraris, A. F., Panfilo, D., Cocca, V., Zinutti, S., De Schepper, K., & Chauvenet, C. R. (2023, November). *"Privacy Measurements in Tabular Synthetic Data: State of the Art and Future Research Directions"*. In NeurIPS 2023 Workshop on Synthetic Data Generation with Generative AI.

consistente di una procedura di generalizzazione su una variabile. Tuttavia, realizzare un processo completamente automatizzato che possa identificare i quasi-identificatori in vari contesti o determinare come ottimizzare l'utilità dei dati applicando tecniche specifiche a variabili specifiche appare improbabile.

Tale fraintendimento impatterebbe anche la generazione di dati sintetici, attività che comporta l'impiego di algoritmi complessi con il coinvolgimento di matematici ed esperti, specialmente per determinare e valutare l'adeguatezza del processo di anonimizzazione e l'utilità dei dati generati.

VIII. **Fraintendimento: *"L'anonimizzazione rende i dati inutili"***. Un corretto processo di anonimizzazione mantiene i dati funzionali per uno scopo specifico (ad esempio per attività di analisi). L'obiettivo primario dell'anonimizzazione è prevenire l'identificazione delle persone fisiche all'interno di un dataset. Tuttavia, l'applicazione delle tecniche di anonimizzazione impone inevitabilmente restrizioni sugli usi potenziali del dataset risultante. Ad esempio, raggruppare le date di nascita in intervalli annuali può diminuire il rischio di re-identificazione, ma può al contempo ridurre l'utilità del dataset in determinati scenari. Ciò non implica che i dati anonimi diventino inutili; piuttosto, la loro utilità dipende dallo scopo prefissato e dal rischio di re-identificazione accettabile.

Al contrario, i dati personali non possono essere conservati permanentemente oltre il loro scopo originale, in attesa di una potenziale utilità per altre finalità. L'anonimizzazione funge da soluzione per alcuni titolari del trattamento, consentendo loro di distaccare e scartare i dati personali dal dataset mantenendo al contempo un'utilità significativa nel dataset rimanente (ad esempio, anonimizzando i *log* di accesso di un sito web mantenendo solo la data di accesso e la pagina visitata, escludendo le informazioni su chi vi ha acceduto).

Il principio di "minimizzazione dei dati" richiede al titolare del trattamento di valutare se il trattamento dei dati personali sia necessario per uno scopo specifico o se lo stesso scopo possa essere raggiunto con dati anonimi. In alcuni casi, questa valutazione può portare alla conclusione che rendere anonimi i dati non sia in linea con lo scopo inteso. In tali casi, il titolare del trattamento deve decidere tra il trattamento di dati personali (impiegando tecniche come la pseudonimizzazione) e l'applicazione del GDPR, o non trattare affatto i dati.

La creazione di dati sintetici sfida questo fraintendimento fornendo dataset che mantengono le proprietà statistiche dei dati originali per l'analisi o l'addestramento di algoritmi assicurando al contempo la privacy e, in generale, la conformità alla normativa in materia di protezione dei dati personali.

IX. **Fraintendimento: *"Seguire un processo di anonimizzazione che altri hanno utilizzato con successo porterà la nostra organizzazione a risultati equivalenti"***. I processi di anonimizzazione devono essere adattati alla natura, all'ambito, al contesto e alle finalità del trattamento, così come ai rischi di varia probabilità e gravità per i diritti e le libertà delle persone fisiche. L'anonimizzazione non può essere applicata universalmente come una ricetta standardizzata, poiché il contesto (inclusi la natura, l'ambito e gli scopi del trattamento dei dati) è probabile che vari da una situazione od organizzazione all'altra. L'efficacia di un processo di anonimizzazione, misurata dal rischio di re-identificazione, può differire significativamente in base a

fattori come il numero limitato di destinatari rispetto alla messa a disposizione dei dati al pubblico generale.

I dataset, esistenti in contesti diversi, possono presentare sfide distinte. L'incrocio di questi dataset con dati anonimi può incidere sul rischio di re-identificazione. Ad esempio, in Svezia, i dati personali dei contribuenti sono accessibili pubblicamente, mentre in Spagna no. Di conseguenza, anche se i dataset contenenti informazioni su cittadini spagnoli e svedesi subiscono un processo di anonimizzazione utilizzando la stessa procedura, i rischi di re-identificazione associati possono variare.

Questo fraintendimento sottolinea come l'effettiva anonimizzazione dipenda strettamente dal contesto in cui venga realizzata; principio applicabile, *mutatis mutandis*, anche alla generazione di dati sintetici. Ciò comporta che l'efficacia delle tecniche di anonimizzazione, inclusa la generazione di dati sintetici, dipende dalle caratteristiche specifiche dei dati e dall'uso che si intende perseguire.

X. **Fraintendimento: *"Non c'è rischio né interesse nel provare a individuare a chi si riferiscono questi dati"***. I dati personali hanno un valore di per sé, sia per gli individui stessi sia per le terze parti. La re-identificazione di un individuo potrebbe avere un impatto serio sui suoi diritti e libertà. Gli attacchi contro l'anonimizzazione possono: i) manifestarsi come tentativi deliberati di re-identificazione, ii) essere sforzi non intenzionali di re-identificazione, violazioni di dati o la divulgazione non autorizzata di dati al pubblico[27]. La prima categoria coinvolge sforzi intenzionali per re-identificare gli individui, mentre la seconda comprende scenari in cui la re-identificazione può verificarsi involontariamente. La possibilità che qualcuno re-identifichi almeno una persona in un dataset, spinto dalla curiosità, dal caso o da interessi specifici (come la ricerca scientifica, il giornalismo o attività criminali) non può essere trascurata[28].

Valutare l'impatto della re-identificazione sulla vita privata di un individuo può essere difficile poiché dipende inevitabilmente dal contesto e dalle informazioni correlate. Ad esempio, re-identificare un interessato basandosi su preferenze cinematografiche apparentemente innocue, può portare a inferenze sulle inclinazioni politiche o sull'orientamento sessuale di quella persona. Tuttavia, tali dati personali appartenenti a categorie particolari, ricevono una "protezione speciale" (ai sensi dell'art. 9 del GDPR).

Tale fraintendimento è particolarmente rilevante anche per i dati sintetici, poiché sottolinea il potenziale valore e i rischi associati alla re-identificazione (sebbene questi ultimi, come indicato in precedenza, sono notevolmente inferiori per i dati sintetizzati rispetto alle tecniche tradizionali di pseudonimizzazione). Esso enfatizza l'importanza di considerare i motivi e le conseguenze della potenziale re-identificazione, anche in dataset considerati anonimi o sintetici.

In conclusione, alla luce dell'analisi condotta nel presente paragrafo, risulta di fondamentale importanza evitare di generare confusione tra anonimizzazione e pseudonimizzazione, soprattutto in ottica di classificazione della sintetizzazione come tecnica di anonimizzazione, al fine di garantire una protezione efficace dei dati personali nel rispetto delle normative vigenti.

---

[27] KHALED EL EMAM and LUK ARBUCKLE, *"Anonymizing Health Data"* (p. 29-33).

[28] KHALED EL EMAM, ELIZABETH JONKER, LUK ARBUCKLE, BRADLEY MALIN, *"A Systematic Review of Re-Identification Attacks on Health Data"*, 11 December 2011.

Difatti la sintetizzazione, se realizzata in modo rigoroso, permetterebbe di superare i limiti imposti dalla normativa sulla protezione dei dati, offrendo nuove opportunità di utilizzo dei dati in sicurezza.

2.3 Le linee guida dell'ICO su anonimizzazione e pseudonimizzazione e "Privacy-enhancing technologies" ("PETs")

Anche l'autorità di controllo de Regno Unito: *Information Commissioner's Office* ("**ICO**") ha pubblicato, nel 2021, delle linee guida al fine di sensibilizzare il mercato riguardo ai temi dell'anonimizzazione e della pseudonimizzazione. In particolare, nel maggio 2021, ha pubblicato la prima parte (*"Introduction to anonymisation"*[29]) e, ad ottobre dello stesso anno, la seconda parte *("How do we ensure anonymisation is effective?"*[30]) di un progetto, sottoposto a consultazione pubblica, intitolato: *"Anonymisation, pseudonymisation and privacy enhancing technologies guidance"*. In questo lavoro, l'ICO ha sottolineato che i dati rappresentano la linfa vitale dell'economia digitale e la loro condivisione è fondamentale per aprire nuove opportunità. L'autorità britannica ha sottolineato l'importanza dei benefici che la condivisione dei dati può portare alle organizzazioni, agli individui e alla società nel suo complesso, ma ha individuato anche i rischi ad essa associati.

Le linee guida menzionate completano il "code of practice"[31] sull'anonimizzazione dell'ICO, il quale fornisce consigli pratici su come condividere i dati personali in conformità agli adempimenti prescritti dalla normativa in materia di *data protection*. Focalizzandosi sull'anonimizzazione, l'ICO ha indicato che essa rappresenta un approccio alternativo all'utilizzo o alla condivisione dei dati, garantendo che le persone fisiche rimangano non identificabili. Inoltre, viene precisato che le tecniche di anonimizzazione efficaci forniscono un'alternativa *"privacy-friendly"* alla condivisione dei dati personali. Tuttavia, è essenziale avere un ragionevole grado di certezza affinché la divulgazione o la condivisione di informazioni apparentemente anonime, non comporti una diffusione inappropriata di dati personali, ad esempio, attraverso la "re-identificazione".

L'autorità del Regno Unito chiarisce, altresì, che ogniqualvolta si applicano tecniche di anonimizzazione ai dati personali, questo processo rientra nella definizione di "trattamento", secondo quanto indicato dalla normativa in materia di protezione dei dati personali, implicando che tali operazioni modificano sostanzialmente le informazioni personali rendendole anonime. Questa trasformazione comporta vari passaggi, incluse l'aggregazione e l'alterazione dei dati, attività che rientrano appunto nell'alveo della nozione di trattamento. Di conseguenza, è necessario adempiere agli obblighi prescritti dal GDPR per la realizzazione di questo processo, in particolare assicurando la presenza di un'idonea base giuridica del trattamento e definendo chiaramente le finalità dell'anonimizzazione. Pertanto, sebbene l'anonimizzazione sia generalmente considerata una pratica lecita, è essenziale stabilire con precisione il perimetro della stessa e illustrare dettagliatamente le misure tecniche e organizzative adottate per conseguirla.

---

[29] Information Commissioner's Office, *"Introduction to anonymisation",* https://ico.org.uk/media/about-the-ico/consultations/2619862/anonymisation-intro-and-first-chapter.pdf.

[30] Information Commissioner's Office, *"How do we ensure anonymisation is effective?"*, https://ico.org.uk/media/about-the-ico/documents/4018606/chapter-2-anonymisation-draft.pdf.

[31] Information Commissioner's Office, *"Anonymisation: managing data protection risk code of practice"* Anonymisation: managing data protection risk code of practice (ico.org.uk).

Secondo l'ICO, determinare lo stato delle informazioni in varie circostanze è una sfida chiave. Ad esempio, si possono possedere informazioni che sono chiaramente dati personali, ma il loro *status*, quando vengono trattati da un'altra organizzazione o dal pubblico in generale, può essere poco chiaro.
Pertanto, similmente all'approccio tenuto dal EDPS e dall'AEPD, analizzato nel paragrafo precedente, l'ICO sottolinea la necessità di valutare ogni caso di anonimizzazione individualmente (come scenario a sé stante). Di conseguenza, sebbene l'anonimizzazione "assoluta" (100%) sia desiderabile, non è sempre praticabile, specialmente considerando la rapida evoluzione della tecnologia. Tuttavia, la potenziale persistenza del rischio di re-identificazione, non determina l'inefficacia della tecnica di anonimizzazione; difatti, l'autorità britannica evidenzia che la normativa in materia di protezione dei dati, non richiede che l'anonimizzazione sia completamente priva di rischi. In particolare, il rischio specifico di re-identificazione deve essere mitigato in modo tale che la possibilità che si verifichi l'evento sia sufficientemente remota.

In base alle informazioni fornite dall'ICO nelle proprie le linee guida, i fattori chiave per l'identificazione di un individuo includono, come indicato pure dal WP29 nell'Opinione 05/2014:

1. L'identificazione di un interessato all'interno di un *dataset* da parte del titolare del trattamento o di un'altra entità (**"*singling out*"**).
2. Collegare diversi *pieces of information* - contenuti in uno o più database - riguardo allo stesso individuo o gruppi di individui (**"*linkability*"**); a causa dell'effetto cosiddetto *"mosaic effect"*, i dati individuali non forniscono informazioni, ma quando vengono combinati con altri, rivelano un quadro significativo.
3. La possibilità di dedurre, indovinare o prevedere dettagli su qualcuno in base ad altre informazioni disponibili (**"inferenze"**).

Le tecniche di anonimizzazione efficaci mirano a ridurre la potenziale occorrenza delle condizioni di identificabilità sopra menzionate, seguendo una valutazione del rischio basata su vari fattori, incluso lo stato dell'arte della tecnologia. La rapida evoluzione della tecnologia richiede che tale valutazione venga periodicamente condotta per valutare se le misure adottate nella fase iniziale (T0) rimangano valide anche nelle fasi successive (T1, T2, ecc.), o se siano necessarie nuove o diverse misure affinché i dati rimangano anonimi.
Un altro criterio suggerito dall'ICO per valutare l'efficacia della tecnica di anonimizzazione adottata è il cosiddetto "***motivated intruder test***", per valutare se un potenziale intruso sia in grado di rendere identificabili gli interessati i cui dati sono stati anonimizzati, grazie a informazioni aggiuntive in suo possesso o altrimenti accessibili/acquisibili. Il livello di competenza del potenziale intruso, all'interno del test, dovrebbe anche essere parametrato al tipo di dati coinvolti: la presenza, ad esempio, di dati finanziari, biometrici, o altrimenti altamente confidenziali, dovrebbe spingere verso l'uso di misure di sicurezza rafforzate.
Per quanto concerne, invece, la pseudonimizzazione, l'ICO avverte del rischio che, in riferimento a un dataset specifico, il termine "de-identificato" si riferisce a dati personali che sono stati trattati in modo tale da non poter essere più attribuiti, senza ulteriori informazioni, a un soggetto specifico. In tale situazione, la convinzione errata che il GDPR (o altre normative) non si applichi – come ribadito in altre circostanze, solo i dati anonimi restano fuori dal campo di applicazione del GDPR – potrebbe avere conseguenze pregiudizievoli sia per il titolare del trattamento sia per gli interessati. Difatti, il dato "de-identificato" si allinea alla definizione di

pseudonimizzazione (*ex* art. 4(5) del GDPR), differenziandosi dall'informazione anonima per il fatto che i dati de-identificati hanno subito una pseudonimizzazione piuttosto che essere stati completamente anonimizzati.
L'ICO ha anche individuato alcuni dei vantaggi connessi alla pseudonimizzazione, come di seguito riepilogati:

a. in base a quanto indicato dal considerando 29 del GDPR, le misure di pseudonimizzazione sono incoraggiate non solo come misura di sicurezza ma anche come possibile strumento per l'analisi generale dei dati;
b. la pseudonimizzazione è uno dei fattori da considerare se un titolare del trattamento decide di continuare a trattare i dati per un nuovo scopo compatibile con quello originale;
c. la pseudonimizzazione è una misura di sicurezza chiave, sia nella fase di progettazione del trattamento che durante l'attuazione di qualsiasi progetto;
d. le tecniche di pseudonimizzazione possono ridurre il rischio di danni e pregiudizi agli interessati in caso di violazione dei dati personali e possono anche facilitare la gestione dei diritti degli interessati (alcuni potrebbero non applicarsi se il titolare del trattamento è in grado di dimostrare l'impossibilità di identificare gli interessati).

Pertanto, secondo le indicazioni dell'ICO, anonimizzare o pseudonimizzare i dati personali può portare significativi benefici a un titolare del trattamento, sia in termini economici che legali, a condizione che vengano implementate le misure conformi, in linea con le migliori pratiche di mercato e le linee guida fornite dalle autorità di protezione dei dati.
Le linee guida dell'ICO, analizzate in questa sezione, sono fondamentali anche nel contesto della sintetizzazione dei dati perché stabiliscono principi chiari per trattare i dati nel rispetto della privacy, offrendo orientamenti su come applicare correttamente le tecniche di anonimizzazione e pseudonimizzazione. Queste pratiche sono essenziali nella generazione di dati sintetici, un processo, come indicato in precedenza, che mira a creare dati non riconducibili direttamente a persone fisiche specifiche, pur mantenendone l'utilità per analisi e ricerche. Seguendo le raccomandazioni dell'ICO, le organizzazioni possono gestire meglio le sfide legali ed etiche connesse alla protezione dei dati personali, garantendo che la sintetizzazione sia effettuata in modo responsabile e conforme alle normative vigenti.
Il tema della sintetizzazione è stato affrontato dall'ICO nella propria Guida[32] dedicata alle tecnologie di miglioramento della privacy (***Privacy – enhancing technologies* - PETs**), precisando che, attraverso l'impiego delle PETs, un'organizzazione può contribuire a garantire il rispetto del principio di minimizzazione dei dati e a dimostrare la protezione dei dati per progettazione e per impostazione predefinita.
Innanzitutto, occorre analizzare la nozione di PETs, come descritta dall'ICO. Le PETs sono tecnologie che incarnano i principi fondamentali della protezione dei dati, minimizzando l'uso dei dati personali, massimizzando la sicurezza dei dati e/o potenziando gli individui. La normativa in materia di protezione dei dati personali non definisce specificamente le PETs, tuttavia tale concetto comprende diverse tecnologie e tecniche. L'Agenzia dell'Unione Europea per la Cybersecurity (ENISA) si riferisce alle PETs come: *"soluzioni software e hardware, ovvero sistemi che comprendono processi tecnici, metodi o conoscenze per raggiungere specifiche funzionalità di privacy o protezione dei dati, o per proteggere contro i rischi per la privacy di un individuo o di un gruppo di persone fisiche"*.

[32] ICO – Information Commissioner's Office, *"Privacy – enhancing technologies (PETs)"*, June 2023.

Nell'analizzare le diverse tipologie di PETs, l'ICO precisa che molte di esse possono contribuire al rispetto della conformità alla protezione dei dati, inclusa la "protezione dei dati per progettazione e per impostazione predefinita", in particolare quelle che:

- riducono l'identificabilità degli individui a cui si riferiscono i dati che vengono trattati, aiutando a soddisfare il principio di minimizzazione dei dati;
- si concentrano sulla mascheratura e protezione dei dati, garantendo l'implementazione di misure di sicurezza in linea con lo 'state of the art';
- suddividono o controllano l'accesso ai dati personali, aiutando a soddisfare il principio di minimizzazione e a garantire la sicurezza dei dati, a seconda della natura del trattamento.

Inoltre, nel descrivere le PETs che generano o rielaborano dati - riducendo o eliminando l'identificabilità degli individui, interrompendo quindi la connessione tra un individuo nei dati personali originali e i dati derivati - L'ICO menziona espressamente: la privacy differenziale e i dati sintetici.
Rispetto al rapporto fra dati sintetici e anonimizzazione, l'ICO sottolinea come l'equivalenza fra dato anonimo e dato sintetico dipenda dalla possibilità o meno di dedurre i dati di natura personale dai dati sintetici generati dai processi di sintetizzazione. È quindi essenziale condurre una valutazione dei rischi di re-identificazione e inferenza/attribuzione legati ai dati sintetici, concentrandosi, per esempio, sul grado di probabilità di re-identificazione e su quali informazioni relative alle persone fisiche verrebbero rivelate se l'identificazione avesse successo.
È stato poi dimostrato che alcuni metodi di generazione di dati sintetici sono vulnerabili, per esempio, ad attacchi di *"model inversion",* che hanno l'obiettivo di risalire all'identificazione degli individui a cui si riferiscono i dati reali, o, ancora, di *"membership inference attack"*, che permettono di rivelare se nel dataset utilizzato per addestrare il modello di AI fossero presenti dati particolari (es. dati relativi alla salute).
In generale, un dato sintetico che impedisca con un elevato grado di probabilità l'identificazione dell'interessato dovrebbe essere trattato alla stregua dei dati sottoposti a processi di anonimizzazione e non potrà soggiacere, pertanto, agli obblighi previsti dalla normativa in materia di protezione dei dati personali.
Per quanto concerne la classificazione delle tecniche di sintetizzazione (in base allo stato dell'arte), l'Office for National Statistics ("**ONS**"), il più grande produttore indipendente di statistiche ufficiali del Regno Unito, ha elaborato un documento metodologico nella serie di working paper numero 16 - "Synthetic data pilot"[33]. Questo studio pilota indaga la domanda e i requisiti per i dataset sintetici ed esplora possibili strumenti per produrre dati sintetici in base a specifiche esigenze degli utenti. Tale analisi mira a comprendere come i dati sintetici possano essere utilizzati per soddisfare le necessità di analisi e ricerca, garantendo al contempo la protezione della privacy e l'integrità dei dati.
In conclusione, secondo l'ICO l'uso dei dati sintetici rappresenterebbe una valida (e, probabilmente, migliore) alternativa ai dati anonimizzati, tuttavia, non andrebbero trascurate alcune potenziali criticità derivanti, in particolare, dai possibili "*bias*" del modello di

33 Office for National Statistics *"Methodology working paper series number 16 - Synthetic data pilot"*, https://www.ons.gov.uk/methodology/methodologicalpublications/generalmethodology/onsworkingpaperseries/onsmethodologyworkingpaperseriesnumber16syntheticdatapilot.

apprendimento – vale a dire le distorsioni che possono essere causate da un utilizzo di dati reali poco rappresentativi o di scarsa qualità – o, ancora, i rischi di re-identificazione provocati da attacchi esterni (sul punto, sempre l'ICO ha evidenziato che l'uso della privacy differenziale con i dati sintetici può proteggere un set di dati da attacchi di c.d. "*linkage*", ciò tuttavia a scapito dell'utilità dei dati ottenuti). In ogni caso, l'utilizzo dei dati sintetici, come PET, fornirebbe un equilibrio tra la protezione della privacy degli individui e la conservazione dell'utilità dei dati, sostenendo le organizzazioni nel loro impegno verso la conformità alla protezione dei dati.

2.4 Posizione del EDPS sui dati sintetici

Il Garante Europeo della Protezione dei Dati (**"EDPS"**) è un'autorità di controllo indipendente che supervisiona l'applicazione del GDPR e, in generale, della normativa in materia di protezione dei dati personali da parte delle istituzioni e degli organismi dell'Unione Europea. L'EDPS ha mostrato un vivo interesse per i dati sintetici, riconoscendone il potenziale come tecnologia per la tutela della privacy.
Anche l'EDPS ribadisce che un elemento chiave per l'identificazione dell'ambito materiale della normativa europea in materia di protezione dei dati, è la definizione di dati personali: difatti, le norme del GDPR si applicano a qualsiasi informazione riguardante una persona fisica identificata o identificabile.
Al contrario, i principi della protezione dei dati non si applicano alle informazioni anonime, ovvero informazioni che non si riferiscono a una persona fisica identificata o identificabile o a dati personali resi anonimi in modo tale che l'individuo non sia o non sia più identificabile.
Per quanto concerne il carattere "anonimo" dei dati sintetici, l'EDPS ribadisce che *"se i dati sintetici si dimostrano veramente anonimi, nel senso che le informazioni, o non si riferiscono a una persona fisica identificata o identificabile, o sono state rese anonime in modo tale che l'individuo non sia più identificabile, allora siamo al di fuori dell'ambito di applicazione del quadro di protezione dei dati personali"*[34].
Secondo l'EDPS, la disponibilità di grandi quantità di dati personali, il crescente livello di interconnessione tra i sistemi digitali, l'aumento e la riduzione dei costi della potenza di calcolo e le capacità di trarre inferenze utilizzando l'Intelligenza Artificiale (AI) sfidano il concetto stesso di dati anonimi.
Allo stesso tempo, Thomas Zerdick, l'Head of Technology and Privacy del EDPS sostiene che *"c'è una tecnologia che sta guadagnando slancio nel contesto dell'intelligenza artificiale e che promette di fornire sia una robusta protezione della privacy che, contemporaneamente, la possibilità di generare dati utili quando non sono disponibili: la sintetizzazione"*[35].
Anche l'*Organisation for Economic Co-operation and Development* (OECD) definisce i dati sintetici come: *"Un approccio alla riservatezza in cui, invece di diffondere dati reali, vengono rilasciati dati sintetici che sono stati generati da uno o più population models"*.

L'EDPS descrive il concetto di generazione di dati sintetici come quel processo consistente nel prendere una fonte di dati originale (dataset) e creare nuovi dati artificiali con proprietà statistiche simili da esso. Mantenere le proprietà statistiche significa che chiunque analizzi i

---

[34] WOJCIECH WIEWIÓROWSKI, *"Synthetic data: what use cases as privacy enhancing technology"*, OPEN Webinar on synthetic data, Synthetic data: what use cases as a privacy enhancing technology? - Wojciech Wiewiórowski | European Data Protection Supervisor (europa.eu).
[35] THOMAS ZERDICK, "*Is the future of privacy synthetic?",* 14 luglio 2021, https://www.edps.europa.eu/press-publications/press-news/blog/future-privacy-synthetic_en.

dati sintetici, ad esempio un analista dei dati, dovrebbe essere in grado di trarre le stesse conclusioni statistiche dall'analisi di un determinato dataset di dati sintetici, come farebbe se avesse a disposizione i dati reali (originali). Inoltre, viene precisato che il processo di generazione, chiamato anche sintesi, può essere eseguito utilizzando diverse tecniche, come gli alberi decisionali o algoritmi di apprendimento profondo[36]. Quindi, i dati sintetici possono essere classificati in relazione al tipo di dati originali: il primo tipo utilizza set di dati reali, il secondo utilizza invece le conoscenze raccolte dagli analisti, e il terzo tipo è una combinazione di questi due. Le reti *Generative Adversarial Networks (GANs)* sono state introdotte di recente e sono comunemente utilizzate nel campo del riconoscimento delle immagini. *"Sono generalmente composti da due reti neurali che si allenano a vicenda iterativamente. La rete generatrice produce immagini sintetiche che la rete discriminatrice cerca di identificare come tale rispetto alle immagini reali"*[37].

In particolare, lo stesso Wojciech Wiewiórowski sottolinea che nessun dato personale dovrebbe risultare dalla sintetizzazione, altrimenti, il rischio potrebbe essere che i dati sintetici possano essere un "miraggio della privacy", come alcuni accademici lo hanno definito, perché alla fine *"non ci sono guadagni per la privacy"*.

Una sfida importante è rappresentata dalla possibilità di integrare caratteristiche dai record che derivano dal dataset originale all'interno dei dati sintetici (noti nella protezione dei dati come *outliers*). In casi estremi, ciò potrebbe portare all'inferenza dell'interessato. Pertanto, e similmente a quanto fatto nell'anonimizzazione, secondo Wojciech Wiewiórowski *"il processo di sintetizzazione dovrebbe considerare una preparazione precedente del dataset e, dopo che il modello è stato generato, dovrebbe essere effettuata una privacy assurance assessment, volta a garantire che nessun dato dal dataset emerga nei dati sintetici"*[38].

Inoltre, l'EDPS individua i potenziali impatti positivi e negativi che la sintetizzazione può avere sulla protezione dei dati personali. Per quanto concerne **i risvolti positivi,** la sintetizzazione potrebbe:

- **Migliorare la privacy nelle tecnologie:** da un approccio alla protezione dei dati fin dalla progettazione, questa tecnologia potrebbe fornire, sulla base di una *privacy assurance assessment*, un valore aggiunto per la privacy degli individui, i cui dati personali non devono essere divulgati.
- **Miglioramento della correttezza**: i dati sintetici potrebbero contribuire a mitigare i pregiudizi utilizzando set di dati sintetici *corretti/esatti* per addestrare i modelli di intelligenza artificiale. Questi set di dati sono manipolati per avere una migliore "rappresentatività del mondo" (ad esempio, senza discriminazioni di genere o razziali).

Invece, per quanto concerne i **risvolti negativi:**

- **La gestione dell'output può essere complessa:** specialmente nei dataset complessi, il modo migliore per garantire che l'output sia accurato e consistente è confrontare i dati

[36] NOWOK, B., G.M. RAAB & C. DIBBEN (2016), synthpop: *"Bespoke creation of synthetic data in R."*, Journal of Statistical Software, 74:1-26; DOI:10.18637/jss.v074.i11. Retrieved from https://www.jstatsoft.org/article/view/v074i11.
[37] ROBERT RIEMANN, Tech Champion, *"Synthetic Data"*, Synthetic Data | European Data Protection Supervisor (europa.eu)
[38] WOJCIECH WIEWIÓROWSKI, *"Synthetic data: what use cases as privacy enhancing technology"*, OPEN Webinar on synthetic data, Synthetic data: what use cases as a privacy enhancing technology? - Wojciech Wiewiórowski | European Data Protection Supervisor (europa.eu).

sintetici con i dati originali o con dati annotati manualmente. Tuttavia, per questo confronto, è nuovamente richiesto l'accesso ai dati originali.

- **Difficoltà nel mappare gli outliers**: i dati sintetici possono solo imitare i dati del mondo reale; non sono una replica. Pertanto, i dati sintetici potrebbero non coprire alcuni outliers presenti nei dati originali. Tuttavia, per alcune applicazioni, gli outliers nei dati possono essere più importanti dei punti dati regolari.
- **La qualità del modello dipende dalla fonte dei dati:** la qualità dei dati sintetici è fortemente correlata alla qualità dei dati originali e al modello di generazione dei dati. I dati sintetici possono riflettere i *bias* presenti nei dati originali. Inoltre, la manipolazione dei dataset per creare dataset sintetici equi potrebbe risultare in dati inaccurati.

Questi aspetti sottolineano l'importanza di un'approfondita valutazione e di un attento sviluppo quando si generano e si utilizzano dati sintetici. È essenziale adottare metodologie robuste sia nella generazione di dati sintetici sia nel loro utilizzo, per assicurare che i dati siano non solo utili ma anche privi di *bias* non intenzionali. In aggiunta, quando si confrontano dati sintetici e originali, è fondamentale mantenere elevati standard di protezione della privacy per prevenire la rivelazione involontaria di informazioni "sensibili". Queste sfide evidenziano il bisogno di un equilibrio tra l'utilizzo innovativo dei dati sintetici e la garanzia che tale uso non comprometta la qualità dei risultati o la privacy degli individui coinvolti.
Al riguardo, poiché secondo l'EDPS l'utilizzo dei dati sintetici nel contesto della protezione dei dati potrebbe rivelarsi estremamente utile, per approfondire tale tema il 17 giugno 2021, l'EDPS ha tenuto un webinar intitolato *"Synthetic data: what use cases as a privacy enhancing technology?"*, al quale hanno partecipato circa 170 esperti in materia, da entrambe le sponde dell'Atlantico, per lo più provenienti dall'industria, ma anche accademici. Il fulcro della discussione si è basato sull'utilizzo dei dati sintetici, invece dei dati reali, come misura di privacy applicata in determinati settori (soprattutto in ambito sanitario) e in casi pratici, come progetti di intelligenza artificiale e *data science* o test di *software* e simulazioni per valutazioni tecnologiche.
Secondo l'EDPS, la sfida che potrebbe essere vinta attraverso la sintetizzazione, è quella di avere dati che siano ancora utili per gli scopi prefissati, ad esempio la ricerca medica, perché mantengono le stesse proprietà statistiche dei dati originali, ma non sono più quelli originariamente raccolti dagli individui.
Durante il webinar i vari relatori hanno condiviso il loro punto di vista su come l'uso della sintetizzazione vada valutata in relazione ai metodi di anonimizzazione "classici" dei dati (originali) e se essa offre un beneficio rilevante. Sono state condivise anche metodologie per calcolare quanto l'uso dei dati sintetici possa ridurre i rischi per la privacy e persino le soglie di rischio per la privacy comunemente accettate.
L'EDPS ha precisato che, indipendentemente dalla qualifica dei dati sintetici come dati personali, sembra ragionevole sostenere che, da un approccio di protezione dei dati *by design*, questa tecnologia fornisce un valore aggiunto per la privacy degli individui, rispetto alla divulgazione dei dati originali.
In conclusione, come anticipato in precedenza, in materia di sintetizzazione l'EDPS suggerisce di effettuare un ***privacy assurance assessment*** cruciale per verificare che i dati sintetici generati non siano dati personali reali. Questo processo valuta il rischio di identificazione delle

persone fisiche nei dati sintetici generati e quante nuove informazioni su di loro potrebbero essere rivelate in caso di identificazione riuscita[39].

2.6 Sentenza "Deloitte" del Tribunale della CGUE: i nuovi scenari

Con la sentenza del 26 aprile 2023[40] (di seguito la "**Sentenza**") il Tribunale della Corte di Giustizia dell'Unione Europea (di seguito anche "**Corte**" o "**CGUE**"), ha espresso le sue considerazioni sui concetti di "pseudonimizzazione" e "anonimizzazione" nel contesto del trattamento dei dati che vedeva principalmente coinvolte due organizzazioni: una che agiva in qualità di titolare del trattamento e mittente di un insieme di dati, mentre l'altra agiva come destinataria di tali dati.

Il Tribunale della CGUE ha fornito un'interpretazione che appare – per le ragioni che verranno spiegate di seguito – innovativa, innescando un vivace dibattito tra i professionisti del settore. Questo dibattito è indubbiamente alimentato dall'importanza cruciale che questi due concetti rivestono per il diritto alla protezione dei dati.

Come già indicato, stabilire se un dato sia qualificabile come personale o meno è fondamentale per la (non)applicazione del GDPR. Ancora, vale la pena notare che il GDPR non si applica al trattamento di informazioni anonime, mentre si applica pienamente al trattamento di dati pseudonimizzati – con tutte le conseguenti implicazioni.

2.6.1 Il Caso

Il caso esaminato dalla Corte ha visto principalmente coinvolti il **Comitato di risoluzione unico** (di seguito **"CRU"**), ossia l'autorità dell'Unione Europea per la risoluzione delle crisi bancarie, con la missione di garantire la risoluzione ordinata delle banche in difficoltà e **Deloitte**, coinvolta come valutatore indipendente incaricato dal CRU per condurre valutazioni. Le valutazioni miravano ad accertare se gli azionisti e i creditori del Banco Popular avrebbero ricevuto un trattamento migliore qualora la banca avesse subito una procedura di insolvenza regolare.

In sintesi, e ai fini della nostra analisi, il CRU, nell'ambito delle sue attività istituzionali, ha raccolto informazioni relative agli azionisti e ai creditori che hanno partecipato alla procedura relativa al diritto di essere ascoltati (di seguito denominati anche "**Parti Interessate**"). Queste informazioni includevano i loro dati identificativi (necessari ai fini dell'audit) e le loro osservazioni personali su aspetti rilevanti per le suddette valutazioni. Successivamente, il CRU ha sottoposto a pseudonimizzazione le osservazioni ricevute e le ha trasmesse a Deloitte senza fornire informazioni aggiuntive necessarie per la re-identificazione. Pertanto, il CRU sarebbe rimasto l'unico ente capace di collegare le osservazioni ai dati che permettono l'identificazione degli autori.

Inoltre, durante la raccolta dei dati, il CRU aveva informato le Parti Interessate riguardo al trattamento dei loro dati personali, ma non aveva precisato che le osservazioni sarebbero state trasmesse a Deloitte. Questa lacuna informativa, secondo cinque Parti Interessate reclamanti, avrebbe violato l'obbligo di informare gli interessati circa i destinatari dei dati personali trattati, secondo quanto prescritto dal GDPR, motivo per cui ha avuto origine l'*iter* che ha condotto alla Sentenza in commento.

---

[39] Boudewijn, A. T. P., Ferraris, A. F., Panfilo, D., Cocca, V., Zinutti, S., De Schepper, K., & Chauvenet, C. R. (2023, November). *"Privacy Measurements in Tabular Synthetic Data: State of the Art and Future Research Directions"*. In NeurIPS 2023 Workshop on Synthetic Data Generation with Generative AI.

[40] Sentenza del Tribunale (Ottava Sezione ampliata), della Corte di Giustizia dell'Unione Europea, del 26 aprile 2023, https://eur-lex.europa.eu/legal-content/IT/TXT/?uri=CELEX:62020TJ0557#t-ECR_62020TJ0557_IT_01-E0001.

2.6.2 Punti Chiave della Sentenza

Dopo aver stabilito che le osservazioni e i punti di vista delle Parti Interessate rientrano nella definizione di "dato personale", come contenuta nell'art. 4(1) del GDPR, e dato che Deloitte non aveva accesso a informazioni che permettessero l'identificazione degli autori delle osservazioni, la Corte si è interrogata se Deloitte dovesse essere considerata come "destinatario" e, quindi, se le informazioni a essa trasmesse a costituissero, a suo riguardo, "dati personali" – ovvero informazioni relative a una persona fisica identificata o identificabile. Essenzialmente, senza affrontare specificamente il caso concreto, la Corte critica la decisione del Garante Europeo per la Protezione dei Dati ("**EDPS**"), che si era pronunciato sulla questione a seguito dei reclami ricevute, perché aveva concluso che le osservazioni trasmesse a Deloitte potessero essere classificate come "dati personali" solo perché il CRU, titolare del trattamento ed entità separata da Deloitte, possedeva informazioni aggiuntive per collegare queste osservazioni ai rispettivi autori.

La posizione della Corte può essere efficacemente riassunta nei punti 96 e 97 della Sentenza, che sono parzialmente riprodotti di seguito:

- **96**. *"Certamente"* – qui il Tribunale conviene con l'EDPS, e richiama la sentenza c.d. Breyer della Corte di Giustizia dell'UE del 19 ottobre 2016, causa C-582/14[41] – *"[…] il fatto che le informazioni aggiuntive necessarie per identificare gli autori delle osservazioni ricevute […] non fossero in possesso di Deloitte, bensì del CRU, non è idoneo a escludere a priori che le informazioni trasmesse a Deloitte costituissero dati personali* ***per quest'ultima****"*.
- **97**. *"Tuttavia"* – prosegue il Tribunale facendo leva sulla citata sentenza c.d. Breyer – *"[…] per stabilire se le informazioni trasmesse a Deloitte costituissero dati personali,* ***occorre porsi dal punto di vista di quest'ultima*** *per determinare se le informazioni che le sono state trasmesse si riferiscano a «persone identificabili»"*.

In linea con il punto 96, l'assenza di possesso di informazioni aggiuntive da parte del destinatario potrebbe escludere che le informazioni trasmesse (anche se detenute da un'altra entità) costituiscano dati personali per il destinatario.

2.6.3 Le conseguenze della Sentenza

La Sentenza stabilisce quindi un principio importante nel contesto della normativa privacy: la percezione del mittente riguardo alla natura "personale" dei dati non determina automaticamente la stessa classificazione per il destinatario. Questo implica che se il destinatario non considera i dati ricevuti come personali, il titolare del trattamento non ha l'obbligo di informare gli interessati riguardo al trattamento dei loro dati da parte dell'entità ricevente. Tale interpretazione solleva questioni rilevanti che impatterebbero anche sulla trasmissione e sul trattamento dei dati sintetici, enfatizzando la necessità di valutare caso per caso la natura dei dati scambiati tra le diverse entità.

[41] Sentenza della Corte di Giustizia dell'Unione Europea (Seconda Sezione), del 19 ottobre 2016, PATRICK BREYER contro BUNDESREPUBLIK DEUTSCHLAND, EUR-Lex - 62014CJ0582 - IT - EUR-Lex (europa.eu).

Parrebbe, dunque, che la corte *"introduca un'accezione relativistica anche in senso soggettivo dei concetti di anonimizzazione e pseudonimizzazione"*[42], chiarendo che:

- è necessario considerare la prospettiva di Deloitte per determinare se le informazioni trasmesse a quest'ultima costituiscano dati personali, e inoltre
- il fatto che il CRU detenga informazioni aggiuntive non *"[…] non è idoneo a escludere a priori che le informazioni trasmesse a Deloitte costituissero dati personali per quest'ultima"*.

Quanto sopra sembra non allinearsi perfettamente con un'interpretazione letterale delle norme pertinenti, per le ragioni di seguito illustrate.

Secondo il considerando 26 del GDPR, disposizione già analizzata in precedenza, sembrerebbe che l'anonimato dei dati debba essere valutato in relazione al soggetto specifico coinvolto in una particolare fase del trattamento piuttosto che nel suo insieme. Per chiarire, se, all'interno di un'attività, i dati trattati potessero essere attribuiti da una qualsiasi delle diverse entità coinvolte nel trattamento alla persona a cui i dati si riferiscono, allora tali dati in quel contesto specifico dovrebbero essere considerati dati personali. A sostegno di questa interpretazione ci sono almeno due elementi letterali del Considerando 26 del GDPR: *"Per stabilire l'identificabilità di una persona è opportuno considerare tutti i mezzi, come l'individuazione, di cui* ***il titolare del trattamento o un terzo*** *può ragionevolmente avvalersi per identificare detta persona fisica* ***direttamente o indirettamente****."*. Pertanto, il legislatore europeo non sembra preoccupato di quale entità possa identificare la persona; sarebbe sufficiente che almeno una di esse possa farlo. Allo stesso modo, i metodi con cui viene effettuata l'identificazione, sia diretta che indiretta, non sembrano avere importanza. Difatti, l'uso del termine "indirettamente" suggerisce, ad esempio, casi in cui il titolare del trattamento è in grado di identificare la persona attraverso il responsabile del trattamento, e viceversa.

Secondo questa interpretazione, come sostenuto dall'EDPS, per l'identificabilità di una persona da un dataset pseudonimizzato, non importerebbe quale entità detenga le informazioni aggiuntive; sarebbe sufficiente che almeno un'entità le detenga.

Riguardo alla pseudonimizzazione, le regole menzionate non sembrano ammettere la possibilità che i dati pseudonimizzati in un certo contesto possano essere considerati anonimi nello stesso contesto, a seconda dell'entità che li elabora. Questo è evidente dalla definizione stessa di "pseudonimizzazione", come fornita in precedenza, che introduce, come condizione affinché un trattamento dei dati sia considerato pseudonimizzato, il fatto che *"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"*. In altre parole, la conservazione separata delle informazioni aggiuntive e le misure di sicurezza prese per proteggere tali informazioni (incluso, naturalmente, limitare l'accesso ad esse) sono elementi indispensabili affinché un dataset sia considerato pseudonimizzato. È sorprendente, quindi, che non rendere accessibili le informazioni aggiuntive al titolare del trattamento potrebbe teoricamente essere considerato come rendere tali dati anonimi, quindi non personali.

In conclusione, lungi dal voler avventurarsi in un'interpretazione affrettata o eccessivamente estensiva di alcuni passaggi della Sentenza, in ogni caso, sarà cruciale valutare con attenzione

[42] CAROLA CAPUTO e GIOVANNI FERORELLI, *"Pseudonimizzazione" o "anonimizzazione"? Sui dati, questo è il dilemma (e non solo)*, Agenda Digitale, 12 maggio 2023.

il grado di identificabilità degli interessati da parte del destinatario dei dati. Per qualificare i dati come anonimi dal punto di vista del destinatario, è essenziale una valutazione documentata, considerando anche le pattuizioni contrattuali mirate a minimizzare il rischio di re-identificazione, compreso il rischio che questi venga a conoscenza, in qualsiasi modo, delle informazioni aggiuntive necessarie appunto alla re-identificazione.
Questo tema è cruciale anche per il trattamento dei dati sintetici, dato che questi ultimi vengono generati per mantenere l'utilità dei dati originali senza compromettere l'identificabilità degli individui. La Sentenza elaborerebbe quindi un corollario, applicabile anche in materia di sintetizzazione, il quale richiederebbe di valutare attentamente se e come i dati sintetici possano essere considerati dati personali, a seconda della capacità di identificazione che si mantiene nei dati generati artificialmente e delle informazioni accessibili al destinatario.
Ancora una volta saranno il principio di "accountability" e un approccio basato sul rischio a dover guidare il titolare del trattamento in queste valutazioni, enfatizzando la necessità di procedure accurate per la gestione dei dati trasmessi.

2.7 Le tecniche di sintetizzazione in base allo stato dell'arte

Nel contesto attuale, segnato dal rapido avanzamento dell'Intelligenza Artificiale (IA) e del Machine Learning (ML), l'emergere dei dati sintetici rappresenta un cambiamento significativo, offrendo nuove opportunità per l'innovazione nel rispetto delle complesse normative e degli aspetti etici legati alla privacy[43]. Tali dati, generati tramite algoritmi avanzati costituiscono una soluzione efficace per affrontare le sfide legate alla raccolta, condivisione e gestione dei dati in conformità con le attuali normative sulla protezione dei dati personali[44]

Questi dati, favorendo lo sviluppo e il perfezionamento dei modelli di intelligenza artificiale e machine learning, eludono i pericoli associati all'uso di dati sensibili e stimolano la ricerca in ambiti critici quali il settore medico, finanziario e della sicurezza. La capacità di replicare insiemi complessi e diversificati permette agli specialisti di testare ipotesi e proporre soluzioni innovative, altrimenti impraticabili per le restrizioni sulla disponibilità di dati reali.

Al contempo, l'impiego di dati sintetici solleva questioni etiche di rilievo, sottolineando l'urgenza di adottare misure adeguate per mitigare eventuali pregiudizi o distorsioni.

La generazione di dati sintetici richiede una valutazione attenta delle disparità presenti nei dataset reali, con l'obiettivo di assicurare che lo sviluppo tecnologico supporti principi di equità e giustizia sociale[45]. Inoltre, mentre regolamenti come il GDPR forniscono un quadro normativo per la protezione dei dati personali, l'integrazione di tali standard nell'ambito dei dati sintetici solleva questioni nuove[46]. Gli sviluppatori devono trovare un equilibrio tra la protezione della privacy e l'ottimizzazione dell'utilità dei dati sintetici, cercando un compromesso che rispetti sia la privacy individuale sia le esigenze di ricerca e innovazione. Gestire tali sfide richiede un dialogo costante tra esperti informatici, legislatori, esperti di etica

[43] GANEV, G., XU, K., & DE CRISTOFARO, E. (2022)."*Understanding how differentially private generative models spend their privacy budget*". arXiv preprint arXiv:2305.10994.
[44] EL EMAM, K. (2022). *"Precaution, ethics and risk: Perspectives on regulating non-identifiable data"*.
[45] GANEV, G., XU, K., & DE CRISTOFARO, E. (2022). "*Understanding how differentially private generative models spend their privacy budget*". arXiv preprint arXiv:2305.10994.
[46] EL EMAM, K. (2022). *"Precaution, ethics and risk: Perspectives on regulating non-identifiable data"*.

e rappresentati della società civile per plasmare un futuro in cui intelligenza artificiale e apprendimento automatico possano evolversi in armonia con i diritti fondamentali dell'uomo e nel rispetto della protezione dei dati personali[47].

Tanto premesso, si procede, di seguito, all'analisi dei metodi di generazione dei dati sintetici allo stato dell'arte.

2.7.1 Generative Adversarial Networks (GANs)

Le Generative Adversarial Networks (GANs)[48], introdotte da Ian Goodfellow nel 2014, rappresentano un modello innovativo per la generazione di dati sintetici. Le reti avversarie generative sono composte da un'architettura dualistica che comprende un generatore (G) e un discriminatore (D). Il generatore ha il compito di creare dati che imitino fedelmente la distribuzione dei dati reali; il discriminatore, invece, si dedica all'analisi comparativa di tali dati con gli equivalenti reali, perseguendo l'identificazione della loro natura, autentica o simulata.

Nei primi stadi di addestramento, i dati generati risultano chiaramente distinguibili dagli originali; tuttavia, mediante un processo iterativo di affinamento, il generatore incrementa la sua capacità di creare output sempre più indiscernibili dal dato reale. Analogamente, il discriminatore affina le sue capacità nel riconoscere le sottili differenze tra i dati generati e quelli reali. Questa dinamica competitiva induce entrambe le reti a perfezionarsi reciprocamente attraverso iterazioni successive, culminando nella capacità del generatore di eludere efficacemente il discriminatore, con l'obiettivo di produrre dati sintetici che siano indistinguibili da quelli reali.

Nell'ambito artistico, le GANs hanno inaugurato nuove prospettive espressive, consentendo la sperimentazione con stili visivi distintivi. Iniziative quali "DeepArt" hanno manifestato l'abilità delle GANs di trasfigurare fotografie in creazioni artistiche[49] che emulano le tecniche di illustri pittori, ponendo in evidenza la convergenza tra tecnologia e arte.

Nel settore medico, le GANs trovano applicazione nella generazione di immagini diagnostiche per la formazione e la validazione di strumenti diagnostici basati su AI, fornendo un mezzo per produrre estese collezioni di immagini adatte all'addestramento di modelli senza compromettere la riservatezza dei dati dei pazienti[50].

L'impiego delle GANs non si esaurisce nella generazione di immagini. Modelli come il CTGAN, infatti, sono definiti specificamente per la generazione di dati tabulari, mentre il PATE-GAN estende tali capacità integrando la privacy differenziale, con l'obiettivo di garantire che i dati sintetici tutelino la privacy degli individui[51][52]. A tal riguardo, va rilevato che le applicazioni dei modelli Generative Adversarial Networks (GANs) mostrano una notevole versatilità, essendo

---

[47] GRESER, J. (2023). *"Synthetic Data and Medical AI – Where Do We Stand?"*.
[48] GOODFELLOW, I., et al. (2014). *"Generative adversarial networks"*. arXiv preprint arXiv:1406.2661.
[49] GATYS, L. A., ECKER, A., & BETHGE, M. (2017). *"Neural style transfer: A review"*. arXiv preprint arXiv:1706.08127.
[50] KISELEVA, A. (2020). *"A review of the application of GANs in medical imaging and healthcare"*. Journal of Healthcare Engineering.
[51] XU, L., SKOULARIDOU, M., CUESTA-INFANTE, A., & VEERAMACHANENI, K. (2019). *"Modeling tabular data using Conditional GAN"*. Advances in Neural Information Processing Systems (NeurIPS). Retrieved from https://arxiv.org/abs/1907.00503.
[52] JORDON, J., YOON, J., & VAN DER SCHAAR, M. (2019). *"PATE-GAN: Generating synthetic data with differential privacy guarantees"*. International Conference on Learning Representations (ICLR). Retrieved from https://openreview.net/pdf?id=S1zk9iRqF7.

impiegabili con differenti tipologie di dati e in molteplici contesti operativi. Tale caratteristica contribuisce in modo significativo alla valorizzazione della loro rilevanza nel processo di sviluppo e perfezionamento dei sistemi di intelligenza artificiale.

Sebbene le GANs aprano scenari di opportunità senza precedenti, esse comportano sfide tecniche come il fenomeno del *mode collapse*, che restringe la varietà degli output generati, e richiedono un'intensa capacità computazionale e una calibrazione meticolosa dei parametri per l'addestramento efficace. Le implicazioni relative alla privacy, emergenti dalla fedeltà dei dati sintetici ai corrispettivi reali, sollevano interrogativi di riservatezza, esponendo a rischi quali gli "attacchi di inversione".

Tanto premesso, le applicazioni di successo delle GANs sono numerose, come evidenziato da "*This Person Does Not Exist*"[53], che produce immagini di volti umani di incredibilmente realistici, e dal progresso nei sistemi di diagnosi medica assistita dall'intelligenza artificiale, che beneficiano della generazione di dataset sintetici per l'addestramento, incrementando la precisione e l'affidabilità.

2.7.2 Modelli Autoregressivi e CNN: profondità tecnica e applicazioni

I modelli autoregressivi, notevoli per la loro capacità di prevedere valori futuri a partire da dati storici, si rivelano un approccio chiave nell'analisi delle sequenze temporali. Questi modelli trovano applicazione estesa in vari settori che richiedono previsioni accurate basate su osservazioni passate, come le previsioni meteorologiche, le analisi di mercato finanziario e il monitoraggio della salute tramite dati biometrici[54]. Un esempio di tale metodologia è l'impiego degli Alberi di Classificazione e Regressione (CART) nel SynthPop di Gillian Raab, i quali, benché non si configurino come reti neurali, applicano principi autorregressivi per la generazione di dati sintetici basati su osservazioni storiche[55].

A ciò si aggiunge che, le Reti Neurali Convoluzionali (CNN) hanno migliorato l'elaborazione dei dati visivi. Utilizzando strati convoluzionali che simulano il meccanismo di percezione visiva dell'uomo, le CNN sono capaci di individuare elementi distintivi nelle immagini, diventando strumenti indispensabili in applicazioni come il riconoscimento di oggetti, l'analisi diagnostica medica e la sorveglianza automatizzata[56].

Dal punto di vista tecnico, i modelli autoregressivi si distinguono per semplicità ed efficacia nel modellare le relazioni lineari tra punti dati successivi. Tuttavia, possono trovare limitazioni in presenza di dinamiche non lineari o di dipendenze a lungo termine che sfuggono ai modelli più elementari. In queste circostanze, le varianti come le Reti Neurali Ricorrenti (RNN) e in particolare le LSTM (Long Short-Term Memory) si presentano come soluzioni a tali limiti, fornendo ai modelli una capacità di memoria a lungo termine per affrontare dipendenze temporali elaborate. Recentemente, i transformer si sono affermati come un'alternativa altamente efficace, impiegando meccanismi di self-attention per catturare dipendenze lungo

---

[53] https://www.thispersondoesnotexist.com/

[54] GANEV, G., OPRISANU, B., & DE CRISTOFARO, E. (2022). *"Robin Hood and Matthew Effects: Differential privacy has disparate impact on synthetic data"*. In Proceedings of the 39th International Conference on Machine Learning (ICML 2022).

[55] NOWOK, B., G.M. RAAB & C. DIBBEN (2016), *"synthpop: Bespoke creation of synthetic data in R."*, Journal of Statistical Software, 74:1-26; DOI:10.18637/jss.v074.i11. Retrieved from https://www.jstatsoft.org/article/view/v074i11.

[56] LECUN, Y., BOTTOU, L., BENGIO, Y., & HAFFNER, P. (1998). *"Gradient-based learning applied to document recognition"*. Proceedings of the IEEE, 86(11), 2278-2324.

sequenze estese senza le limitazioni di elaborazione sequenziale tipiche delle RNN. Tale caratteristica li rende adatti per compiti che implicano relazioni temporali di vasta portata[57].

Le CNN, attraverso l'uso di filtri convoluzionali, elaborano immagini mantenendo la relazione spaziale tra i pixel, ciò le rende particolarmente efficaci nell'identificare schemi e texture.

L'ottimizzazione dell'efficacia di queste tecnologie è oggetto di ricerca verso modelli ibridi che integrano le capacità dei modelli autoregressivi con quelle delle CNN. Per esempio, nel campo del riconoscimento vocale, l'unione di modelli autoregressivi con le CNN permette di esaminare sequenze audio nel tempo, catturando sia la dinamica temporale che le caratteristiche frequenziali del segnale[58].

2.7.3 Variational autoencoder (VAE) e modelli di diffusione

Il Variational Autoencoder (VAE) è considerata una tecnologia all'avanguardia nell'ambito del deep learning per la creazione di dati sintetici, con particolare attenzione alla modellazione di dati complessi e relazionali. Questo modello si distingue per la sua capacità di apprendere rappresentazioni latenti dense di dati ad alta dimensionalità, consentendo così la generazione di nuovi dati che mantengono le stesse caratteristiche statistiche dei dati originali. Un elemento chiave del VAE è l'utilizzo di una funzione di perdita che include un termine di regolarizzazione chiamato divergenza KL, il quale misura quanto la distribuzione appresa nello spazio latente si discosti da una distribuzione predefinita, tipicamente gaussiana. Tale approccio spinge il modello a organizzare lo spazio latente in modo strutturato e continuo, agevolando così la produzione di nuovi dati diversificati. A tal riguardo si nota che gli Autoencoder Variazionali (VAEs) non sono circoscritti alla elaborazione di dati immagine; dimostrano, invece, notevole efficacia anche nella generazione di dati sintetici per tabelle e database relazionali, inclusi quelli che comportano tipologie di dati avanzate. Pertanto, la versatilità dei VAEs li rende adatti a numerose applicazioni, che includono sia la sintesi di immagini sia la generazione di dati tabulari complessi[59].

I modelli di diffusione, un'altra avanzata metodologia per la generazione di dati sintetici, condividono diverse somiglianze con gli Autoencoder Variazionali (VAEs), in particolare nella loro capacità di modellare distribuzioni di dati complesse. I modelli di diffusione operano definendo un processo diretto che aggiunge gradualmente rumore ai dati, trasformandoli in una distribuzione di rumore. Successivamente, una rete neurale viene addestrata per approssimare il processo inverso, rimuovendo gradualmente il rumore per recuperare la distribuzione originale dei dati. Questo processo iterativo consente ai modelli di diffusione di produrre dati sintetici altamente realistici che somigliano strettamente ai dati del mondo reale[60].

---

[57] VASWANI, A., SHAZEER, N., PARMAR, N., USZKOREIT, J., JONES, L., GOMEZ, A. N., KAISER, Ł., & POLOSUKHIN, I. (2017). *"Attention is All You Need. Advances in Neural Information Processing Systems"*, 30, 5998-6008. Retrieved from https://arxiv.org/abs/1706.03762.

[58] GRAVES, A., MOHAMED, A., & HINTON, G. (2013). *"Speech recognition with deep recurrent neural networks"*. 2013 IEEE International Conference on Acoustics, Speech, and Signal Processing, 6645-6649.

[59] PANFILO, D., BOUDEWIJN, A., SACCANI, S., COSER, A., SVARA, B., CHAUVENET, C. R., MAMI, C. A., & MEDVET, E. (2023). *"A deep learning-based pipeline for the generation of synthetic tabular data"*. IEEE Access.

[60] HO, J., JAIN, A., & ABBEEL, P. (2020). *"Denoising Diffusion Probabilistic Models"*. Advances in Neural Information Processing Systems, 33, 6840-6851. Retrieved from https://arxiv.org/abs/2006.11239.

Autoencoder Variazionali (VAEs) e modelli di diffusione offrono soluzioni robuste per la generazione di dati sintetici ad alta fedeltà. I modelli di diffusione, analogamente ai VAEs, sono efficaci nel catturare strutture di dati complesse e sono stati applicati con successo in vari domini, tra cui la sintesi di immagini e la previsione di serie temporali. Entrambi i metodi consentono la generazione di dati sintetici che preservano le proprietà statistiche dei dati originali, garantendo al contempo la privacy e la conformità alle normative sulla protezione dei dati[61].

2.8 Garanzie per la privacy

La presente analisi, che si propone di contestualizzare i dati sintetici e il loro valore aggiunto nel tessuto sociale contemporaneo, richiede un'analisi approfondita del concetto di privacy all'interno di un complesso quadro normativo e operativo relativo alla sicurezza dei dati.

Questo processo comporta il passaggio da concetti astratti di riservatezza verso la creazione di indicatori quantificabili e misurabili oggettivamente, richiedendo un'indagine approfondita sulle caratteristiche e sui meccanismi specifici.

Tradurre le garanzie di riservatezza nell'uso di dati sintetici in parametri quantificabili diventa essenziale per stabilire standard di privacy chiari e pragmatici. A tal riguardo, i meccanismi per la tutela della privacy mirano a offrire un criterio uniforme per valutare l'efficacia delle misure di protezione, cruciali per quantificare il rischio di re-identificazione o divulgazione di informazioni sensibili che potrebbero minare la privacy degli individui.

Prima di esaminare le garanzie di privacy integrate nella generazione di dati sintetici, è necessario comprendere i dati sintetici in relazione alle loro proprietà, quali fedeltà e utilità. Nel merito, l'utilità di un dato sintetico dipende dalla sua applicabilità a specifiche finalità o insiemi di compiti, mentre la fedeltà misura quanto accuratamente il dato sintetico replichi il dato reale di riferimento.

Per preservare i dati nella generazione di dati sintetici possono essere utilizzati molteplici metodi per proteggere la privacy nella generazione di dati sintetici, come applicare tecniche di k-anonimato, implementare controlli per prevenire l'overfitting e garantire la plausible deniability.

Tra questi, la Privacy Differenziale rappresenta uno degli approcci più discussi e utilizzati. Questa metodologia, basata su principi matematici, contribuisce a prevenire la de-anonimizzazione introducendo rumore[62].

Un aspetto rilevante della Privacy Differenziale, che merita particolare attenzione in questo contesto, è l'allocazione del budget di privacy, rappresentato dalla lettera greca ε (epsilon). Tale parametro limita la quantità di informazioni che possono essere rivelate su un individuo, influenzando direttamente la generazione di dati sintetici e il rischio di de-anonimizzazione. In merito, si rileva una mancanza di consenso sulla scelta ottimale del parametro (ε, δ), poiché valori più elevati sono associati a un livello inferiore di garanzia della privacy. Studi recenti suggeriscono che un singolo valore di epsilon possa offrire vari livelli di protezione

[61] MAMI, C. A., COSER, A., MEDVET, E., BOUDEWIJN, A. T. P., VOLPE, M., WHITWORTH, M., SVARA, B., SGROI, G., PANFILO, D., & EVAL., S. (2022). *"Generating Realistic Synthetic Relational Data through Graph Variational Autoencoders"*. arXiv:2211.16889v1.
[62] GU, W. (2021). *"Privacy preserving synthetic data generation"*. Freie Universität Berlin, Berlin, Germany. Retrieved from https://www.mi.fu-berlin.de/inf/groups/ag-idm/theseses/2021_Gu_BSc.pdf.

a seconda del contesto specifico, rendendo particolarmente complessa la determinazione di un valore adeguato negli scenari pratici. Tali considerazioni assumono particolare importanza per la generazione di dati sintetici, che comporta la divulgazione di informazioni più dettagliate rispetto ad altre applicazioni della privacy differenziale. Diversamente dalla privacy differenziale applicata alle statistiche aggregate - dove le linee guida attuali indicano che la DP potrebbe non fornire una protezione adeguata - o ai classificatori AI, la generazione di dati sintetici implica il rilascio di dataset completi, seppur sintetici, granulari o di modelli generativi[63]. L'insufficienza della privacy differenziale nel proteggere sia la privacy sia la fedeltà dei dati sintetici, a prescindere dal valore di epsilon, è ulteriormente dimostrata dalle valutazioni degli algoritmi di generazione di dati sintetici[64].

Di seguito, i vantaggi e gli svantaggi legati all'allocazione di un budget della privacy basso e alto al fine di contestualizzare l'applicazione della *differential privacy*.

Budget della Privacy Basso: Protezione Massima, Limiti nell'Utilità

Un budget della privacy ridotto significa imporre un rigido controllo sulla quantità di informazioni personali che possono essere rivelate[65]. Questa cautela aumenta significativamente la protezione della privacy, ma implica anche che i dati sintetici generati saranno meno fedeli agli originali, con possibili distorsioni che ne limitano l'applicabilità.

Fidelity bassa: In questo scenario, i dati sintetici differiscono sostanzialmente dai dati reali, risultando in una rappresentazione meno accurata di specifiche caratteristiche. Tale distanza mira a impedire l'identificazione diretta o indiretta degli individui nei dataset[66].

Utility limitata: La minore fedeltà incide negativamente sull'utilità dei dati per analisi dettagliate[67]. Inoltre, la privacy differenziale, specialmente con un budget ridotto, può comportare un aumento di bias nei dati generati[68].

Budget della Privacy Alto: Maggiore Utilità a Scapito della Privacy

Di contro, un budget della privacy elevato permette di generare dati sintetici che mantengono una grande somiglianza con quelli originali. Ciò favorisce analisi più dettagliate e risultati più affidabili, ma introduce, al contempo, un rischio maggiore di compromissione della privacy.

---

[63] LEE, J., & CLIFTON, C. (2011). *"How much is enough? Choosing ε for differential privacy"*. In J. Domingo-Ferrer & I. Tinnirello (Eds.), Information security (pp. 325-340). Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-24861-0_22.

[64] National Institute of Standards and Technology (NIST). (2020). Privacy Collaborative Research Cycle Archive. Retrieved from https://pages.nist.gov/privacy_collaborative_research_cycle/pages/archive.html.

[65] DWORK, C., MCSHERRY, F., NISSIM, K., & SMITH, A. (2006). *"Calibrating noise to sensitivity in private data analysis"*. In Proceedings of the Theory of Cryptography Conference (TCC), Lecture Notes in Computer Science, vol 3876. Springer.

[66] DWORK, C., & ROTH, A. (2014). *"The Algorithmic Foundations of Differential Privacy"*. Foundations and Trends® in Theoretical Computer Science, 9(3–4), 211–407.

[67] LI, N., QARDAJI, W., & SU, D. (2013). *"On sampling, anonymization, and differential privacy or, k-anonymization meets differential privacy"*. In Proceedings of the 7th ACM Symposium on Information, Computer and Communications Security (pp. 32-33). ACM.

[68] GANEV, V., LI, W., SURIYAKUMAR, V., LIN, D., MIN, M. R., & YOON, J. (2022). *"Revisiting pre-trained models for tabular data"*. In K. Chaudhuri, S. Jegelka, L. Song, C. Szepesvári, G. Niu, & S. Sabato (Eds.), Proceedings of the 39th International Conference on Machine Learning (ICML 2022), 162, 6888-6902. Retrieved from https://proceedings.mlr.press/v162/ganev22a.html.

Alta Fidelity: Con un budget più generoso, i dati sintetici rispecchiano da vicino i dati reali, conservando una maggiore quantità di dettagli individuali. Questo incrementa la precisione delle analisi ma solleva questioni sulla potenziale re-identificazione degli individui[69].

Utility potenziata: L'accuratezza migliorata dei dati sintetici espande le loro applicazioni, permettendo agli analisti di trarre conclusioni più precise e di sfruttare i dati per una vasta gamma di ricerche[70].

Nonostante la scelta di un budget della privacy sia fondamentale per determinare il livello di protezione e l'utilità dei dati sintetici, è importante riconoscere che non esiste una soluzione univoca.

Il bilanciamento tra privacy e utilità richiede che la scelta relativa all'allocazione del budget di privacy sia guidata dal contesto specifico e dagli obiettivi prefissati per l'analisi.

La scelta sull'allocazione del budget, in altri termini, costituisce il limite intrinseco della *differential privacy* come *privacy preserving guarantees* in quanto la scelta sul posizionamento di ε (epsilon) incide direttamente sulle proprietà di fidelity e utility del dato generato[71].

Nel processo di generazione dei dati sintetici, l'adozione di metriche per la tutela della privacy si rivela fondamentale per misurare e assicurare la sicurezza dei dati elaborati. Questi strumenti forniscono criteri oggettivi per valutare l'efficienza delle strategie adottate per proteggere le informazioni sensibili, promuovendo l'uso responsabile dei dati sintetici in ambiti come ricerca e sviluppo, senza mettere a rischio la riservatezza degli individui nei dataset originali[72].

Equilibrare l'utilità dei dati con la loro protezione si presenta come un obiettivo complesso. I dati sintetici devono riflettere le proprietà statistiche dei dati originali per supportare analisi predittive e modelli efficaci. Allo stesso tempo, tuttavia, occorre impedire qualsiasi divulgazione di dati personali.

Premesso quanto sopra, l'utilizzo di metriche come strumenti di prevenzione della privacy si rivela uno strumento efficace per affrontare tali sfide assicurando al contempo utilità e fedeltà elevate dei dati sintetici rispetto ai dati reali. Questa metodologia guida con precisione lo sviluppo e l'impiego di dataset sintetici in diversi ambiti, soddisfacendo equamente le necessità di tutela della privacy e le aspettative analitiche.

La capacità dei dati sintetici di conservare le caratteristiche statistiche e le correlazioni del dataset originale, mantenendo la privacy, definisce la loro utilità.

La fedeltà indica la precisione con cui i dati sintetici replicano le distribuzioni dei dati originali. Una alta fedeltà garantisce che i dati sintetici possano sostituire in modo affidabile gli originali

---

[69] ABADI, M., CHU, A., GOODFELLOW, I., MCMAHAN, H. B., MIRONOV, I., TALWAR, K., & ZHANG, L. (2016). *"Deep Learning with Differential Privacy"*. In Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security. ACM.
[70] MCSHERRY, F. (2009). *"Privacy Integrated Queries: An Extensible Platform for Privacy-Preserving Data Analysis"*. In Proceedings of the 2009 ACM SIGMOD International Conference on Management of Data. ACM.
[71] LEE, J., & CLIFTON, C. (2011). *"How much is enough? Choosing ε for differential privacy"*. In J. Domingo-Ferrer & I. Tinnirello (Eds.), Information security (pp. 325-340). Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-24861-0_22.
[72] SOSNOVCHYK, D. (2021). *"Evaluating the privacy of synthetic data using the nearest neighbor distance"*. (Bachelor's thesis, Freie Universität Berlin). Retrieved from https://www.mi.fu-berlin.de/inf/groups/ag-idm/theseses/2021-Sosnovchyk-BSc.pdf.

nelle analisi, senza alterare i risultati a causa di distorsioni. L'impiego di metriche specifiche applicate a metodologie diverse offre una modalità per valutare la fedeltà del dato.

I dati sintetici dovrebbero differire dai dati originali abbastanza da prevenire la riconoscibilità degli individui, pur mantenendo una somiglianza che permetta analisi valide. Questo bilanciamento garantisce che i dati sintetici rimangano utili e sicuri per le applicazioni desiderate[73]. I dati sintetici, analogamente a qualsiasi altro tipo di dati, possono incorporare o manifestare *bias*. Questo può avvenire durante la fase di generazione dei dati sintetici, in particolare se i modelli o gli algoritmi utilizzati per la loro creazione apprendono da set di dati reali che già contengono *bias*. Ad esempio, se un set di dati reali utilizzato per addestrare un modello generativo presenta una rappresentazione squilibrata di determinati gruppi demografici, è probabile che anche i dati sintetici prodotti riflettano questo squilibrio. A tal riguardo, la presenza di *bias* solleva preoccupazioni significative, specialmente quando tali dati vengono impiegati per addestrare o testare sistemi di intelligenza artificiale (IA) che saranno poi utilizzati in applicazioni reali. Tanto premesso, i *bias* possono perpetuare o addirittura amplificare le disuguaglianze esistenti, portando a decisioni ingiuste o discriminatorie.

Le tecniche discusse riflettono un impegno significativo verso la creazione etica e responsabile di dati sintetici, dimostrando come l'innovazione tecnologica possa essere guidata da principi di privacy e sicurezza. Guardando al futuro, il necessario equilibrio tra l'espansione delle capacità AI e le crescenti esigenze di protezione dei dati personali invita a una riflessione approfondita e a un impegno continuo nella ricerca.[74].

2.8.1 Metodologie di sintetizzazione avanzate

A seguito dell'analisi dei diversi metodi generativi dei dati sintetici, l'attenzione si concentra sulle metodologie sviluppate, di recente, da startup innovative.

La formulazione di una metodologia per la generazione di dati sintetici comprende sia l'applicazione di tecniche generative sia l'impiego degli strumenti adeguati per la valutazione dell'affidabilità e dell'utilità dei dati sintetici, nonché per la tutela della riservatezza. Conseguentemente, l'analisi della metodologia di generazione di dati sintetici deve necessariamente prendere in considerazione i fattori sopra menzionati. Sul tema, nell'ambito della valutazione dei metodi di generazione dei dati sintetici, ricercatori e professionisti del settore esplorano nuove metodologie al fine di garantire la protezione dei dati personali. Alcune tecniche impiegano algoritmi di apprendimento per catturare le strutture intrinseche dei dati originali in modo probabilistico, preservando informazioni statistiche fondamentali. Successivamente, un modello generativo, addestrato sulla base della distribuzione del dato, procede alla generazione del dato sintetico.

Per quanto concerne, invece, la sicurezza e la protezione dei dati personali, possono essere impiegate metriche basate sulla distanza, utilizzate come garanzie di preservazione della privacy, che valutano la vicinanza tra i dati sintetici e quelli reali.

---

[73] EL EMAM, K., MOSQUERA, L., & HOPTROFF, R. (n.d.). *"Evaluating Synthetic Data Utility"*. In Practical Synthetic Data Generation. O'Reilly Media. Retrieved from https://www.oreilly.com/library/view/practical-synthetic-data/9781492072737/ch04.html.

[74] CASTELNOVO, A., CRUPI, R., INVERARDI, N., REGOLI, D., & COSENTINI, A. (2022). *"Investigating Bias with a Synthetic Data Generator: Empirical Evidence and Philosophical Interpretation"*. arXiv:2209.05889v1 https://arxiv.org/abs/2209.05889.

Se i dati sintetici risultano troppo simili ai dati reali, questi vengono modificati o rimossi per evitare il rischio di re-identificazione.

Ulteriori controlli sui rischi di re-identificazione possono essere applicati alla metodologia per valutare la possibilità di attacchi di re-identificazione.

L'adozione di tali tecniche e strumenti garantisce che i dati sintetici mantengano un elevato grado di fedeltà rispetto ai dati originali, assicurando contestualmente la protezione della privacy degli individui e minimizzando i rischi di re-identificazione. Questo approccio integrato consente di coniugare la necessità di un'analisi accurata e utile dei dati con l'irrinunciabile esigenza di tutelare i diritti e le libertà delle persone fisiche, in conformità con le normative vigenti in materia di protezione dei dati.

Un ulteriore aspetto che contribuisce a ridurre i rischi sopra individuati è rappresentato dalla possibilità di eliminare dati specifici dall'input di sintetizzazione. Si tratta di un aspetto della sintetizzazione dei dati che non riguarda il modello generativo in sé, quanto piuttosto specifiche fasi iniziali del processo. In particolare, la sintetizzazione di dati specifici può rappresentare un punto di vulnerabilità: l'esposizione di dati, come ad esempio i codici di avviamento postale di piccoli centri, può facilitare la re-identificazione degli individui. Ciò assume particolare rilevanza nelle piccole comunità, dove l'unicità del codice postale accresce significativamente l'efficacia degli algoritmi di re-identificazione. Ad esempio, in un paese con 100 residenti e un solo CAP, pubblicare dati sintetici inclusivi di tale CAP potrebbe semplificare notevolmente l'individuazione della persona corrispondente, poiché gli algoritmi di re-identificazione avrebbero pochi candidati da considerare, compromettendo l'anonimato.

Pertanto, la sintetizzazione dei dati, se correttamente gestita attraverso l'identificazione e l'eliminazione delle categorie di dati specifici, come il CAP, e l'adozione di metriche di valutazione adeguate, offre una solida protezione contro i rischi di re-identificazione, preservando al contempo l'utilità dei dati per analisi e applicazioni successive.

L'adozione di metodologie avanzate per la generazione di dati sintetici, che includono l'integrazione di garanzie per la preservazione della privacy, permette di analizzare i dati sintetizzati alla luce dei concetti di pseudonimizzazione e anonimizzazione previsti dal GDPR.

Tanto premesso, qualora i dati sintetici, pur mantenendo la capacità di fornire risultati analitici comparabili ai dati originali, non contengono né riflettono alcun elemento specifico che possa condurre all'identificazione dell'interessato sono da considerarsi dati anonimi (vedasi al riguardo l'orientamento del EDPS descritto al precedente paragrafo 2.5). In questo scenario, come evidenziato nel Considerando 26 del GDPR, i dati trattati hanno subito un processo di anonimizzazione tale per cui l'individuo non è più identificabile, escludendo di conseguenza i dati dal campo di applicazione del GDPR. Di conseguenza, è importante riconoscere che, benché la pseudonimizzazione costituisca una strategia di protezione dati notevolmente efficace, la trasformazione dei dati in una forma che garantisca l'impossibilità di riconoscere gli individui coinvolti, come illustrato precedentemente, si qualifica come una forma di anonimizzazione. Ciò conferisce ai dati sintetizzati un valore aggiunto in termini di protezione della privacy, consentendo allo stesso tempo l'utilizzo di tali dati per finalità di analisi, ricerca e sviluppo, senza incorrere nelle restrizioni imposte dal GDPR. Questo approccio si allinea alle indicazioni fornite dall'EDPS, valorizzando l'impiego di dati sintetici come strumento per coniugare innovazione e rispetto della privacy nell'era digitale.

3.Normativa attuale e futura relativamente al dato sintetico

3.1 Attuale

3.1.1 Uso dei dati sanitari per finalità di ricerca scientifica nel GDPR (artt. 9 e 89): sintetizzazione come tecnica di anonimizzazione?

La regolamentazione europea sul trattamento dei dati personali (i.e., il Regolamento UE 2016/679 o "GDPR"), pur non definendo espressamente il concetto di "ricerca scientifica", traccia un perimetro piuttosto ampio al cui interno è ricompresa una serie eterogenea di attività di ricerca che attengono, fra gli altri, allo sviluppo tecnologico, alla ricerca finanziata da privati nonché agli studi svolti nel settore della sanità pubblica (v. Considerando 159 del GDPR[75]).
Sul punto, sono intervenuti anche i Garanti europei, chiarendo come la nozione di ricerca scientifica non possa essere estesa oltre il suo significato comune, da intendersi principalmente come progetto di ricerca istituito in conformità con le pertinenti norme metodologiche e deontologiche settoriali, in linea con le buone prassi[76].

A prescindere dall'esatta connotazione da attribuire all'ambito della ricerca scientifica, il legislatore europeo, attraverso le norme del GDPR, ne ha riconosciuto il ruolo fondamentale all'interno della società, approntando un regime giuridico per certi versi meno rigoroso rispetto agli obblighi generali stabiliti per il trattamento dei dati personali per altre finalità.
Per quanto qui di interesse, il paradigma normativo a cui fare riferimento in relazione al trattamento di dati personali per finalità di ricerca scientifica è rappresentato dall'art. 89 del GDPR, rubricato "*Garanzie e deroghe relative al trattamento a fini di archiviazione nel pubblico interesse, di* ***ricerca scientifica*** *o storica o a fini statistici*", avente carattere sia programmatico che precettivo.
La formulazione "aperta" di cui al comma 2 dell'art. 89 lascia al diritto nazionale dei singoli Stati membri la possibilità di intervenire con discipline normative integrative, mediante l'introduzione di deroghe e adeguamenti ai diritti previsti dagli artt. 15 ("diritto di accesso"), 16 ("diritto di rettifica"), 18 ("diritto di limitazione di trattamento") e 21 ("diritto di opposizione"), a condizione che l'esercizio dei predetti diritti rischi di rendere impossibile o di pregiudicare gravemente il conseguimento delle finalità specifiche del trattamento e che tali limitazioni siano necessarie al raggiungimento delle finalità perseguite dal titolare del trattamento.
L'applicazione delle deroghe sopra richiamate è comunque sempre subordinata all'attuazione di specifiche e adeguate garanzie e misure di sicurezza tecniche e organizzative volte a tutelare i diritti e le libertà degli interessati.
Il regime delineato dal legislatore europeo in relazione ai trattamenti che si svolgono nel settore della ricerca scientifica contempla, inoltre, alcune eccezioni ai principi fondamentali applicabili al trattamento di dati personali - in particolare quello di limitazione delle finalità e di

[75] Il Considerando 159 del GDPR stabilisce che *"il trattamento di dati personali per finalità di ricerca scientifica dovrebbe essere interpretato in senso lato e includere ad esempio sviluppo tecnologico e dimostrazione, ricerca fondamentale, ricerca applicata e ricerca finanziata da privati, oltre a tenere conto dell'obiettivo dell'Unione di istituire uno spazio europeo della ricerca ai sensi dell'articolo 179, paragrafo 1, TFUE. Le finalità di ricerca scientifica dovrebbero altresì includere gli studi svolti nell'interesse del settore della sanità pubblica"*.
[76] Cfr. EDPB, *"Linee guida 5/2020 sul consenso ai sensi del regolamento (UE) 2016/679"* adottate il 4 maggio 2020.

conservazione dei dati[77] - nonché alcune specifiche condizioni che riguardano il trattamento dei dati rientranti nelle categorie particolari di cui all'art. 9 del GDPR.
Come noto, il primo punto relativo alle limitazioni delle finalità involge la questione – molto dibattuta e spinosa – concernente l'utilizzo dei dati personali per finalità ulteriori rispetto a quelle iniziali (c.d. *"secondary use"*), tematica, questa, peraltro già oggetto di analisi da parte dei Garanti europei nell'Opinione 3/2013 del 2 aprile 2013 (v. sez. 3.2.3 *"Further processing for historical, statistical or scientific purposes"*).
Il riuso dei dati è esplicitamente ammesso dall'art. 5, par. 1, lett. b) del GDPR, che contempla una presunzione di compatibilità dell'ulteriore trattamento, nell'ipotesi, per esempio, di trattamento per finalità di ricerca scientifica ai sensi dell'art. 89 del GDPR[78]. Sotto questo profilo, il Garante europeo della protezione dei dati ("EDPS") ha fornito un primo chiarimento[79] sul concetto di presunzione di compatibilità enucleato nella norma citata, escludendo che questa possa costituire un'autorizzazione generale a riutilizzare i dati, in qualsiasi circostanza, per finalità storiche, statistiche o scientifiche. Pur dovendosi valutare, caso per caso, le caratteristiche di ciascun trattamento, in linea di principio i dati personali raccolti in un determinato ambito possono essere ulteriormente trattati per scopi di ricerca scientifica (dall'originario o da un nuovo titolare del trattamento), a condizione che siano poste in essere le menzionate garanzie adeguate ex art. 89 del GDPR.

Il secondo ambito di deroga su cui interviene l'art. 89 del GDPR attiene all'obbligo di conservazione dei dati personali per un periodo di tempo determinato, non superiore al conseguimento delle finalità indicate dal titolare del trattamento. Detta regola generale subisce un'eccezione che permette la conservazione dei dati personali per un periodo di tempo più ampio nell'ipotesi di trattamenti effettuati per fini di ricerca scientifica o storica ovvero di archiviazione nel pubblico interesse o statistici, purché, in questo caso, siano attuate misure tecniche e organizzative adeguate a tutelare i diritti e le libertà degli interessati e il trattamento sia svolto esclusivamente per le suindicate finalità.
Infine, rispetto al trattamento di dati appartenenti alle c.d. categorie particolari, l'art. 9, par. 2, lett. j) del GDPR rende non applicabile il divieto generale di trattare tali dati se il trattamento è finalizzato alla ricerca scientifica e si basa sul diritto nazionale o dell'Unione, se è proporzionato a tale finalità, se rispetta l'essenza del diritto alla protezione dei dati e prevede misure appropriate e specifiche per tutelare i diritti fondamentali e le libertà degli interessati.
Va tuttavia specificato che rispetto al trattamento dei dati genetici, biometrici e relativi alla salute, l'ordinamento italiano, nell'adeguare il quadro normativo interno alle previsioni del GDPR, ha esercitato la facoltà di cui all'art. 9, par. 4, inserendo nel D.Lgs. 196/2003 ("Codice Privacy"), mediante il D.Lgs. 101/2018, l'art. 2-septies, che introduce ulteriori condizioni per il trattamento di tali tipi di dati.

---

[77] Art. 5, par. 1, lett. e) GDPR: *"I dati personali possono essere conservati per periodi più lunghi a condizione che siano trattati esclusivamente a fini di archiviazione nel pubblico interesse, di ricerca scientifica o storica o a fini statistici, conformemente all'articolo 89, paragrafo 1, fatta salva l'attuazione di misure tecniche e organizzative adeguate richieste dal presente regolamento a tutela dei diritti e delle libertà dell'interessato"*.
[78] Art. 5, par. 1, lett. b) GDPR: *"...un ulteriore trattamento dei dati personali ai fini di archiviazione nel pubblico interesse, di ricerca scientifica o storica o a fini statistici non è, conformemente all'articolo 89, paragrafo 1, considerato incompatibile con le finalità iniziali"*.
[79] Cfr. Garante europeo della protezione dei dati, *"A Preliminary Opinion on data protection and scientific research"* (6 gennaio 2020).

Limitandoci, in questa sede, alla disciplina contenuta nel GDPR, deve porsi l'attenzione sulle garanzie richieste dalla normativa europea a presidio dei diritti e delle libertà degli interessati. Tali garanzie hanno lo scopo precipuo di assicurare la predisposizione di adeguate misure (tecniche ed organizzative) a tutela, in particolare, del principio di minimizzazione (art. 5, par. 1, lett. c), GDPR).
Fra le garanzie espressamente indicate dall'art. 89 del GDPR, da leggere in combinato disposto con l'art. 32 del GDPR, figurano la pseudonimizzazione e l'anonimizzazione dei dati personali. In particolare, dal dettato normativo emerge un approccio cautelativo adottato dal legislatore, che vede un'applicazione graduata di tali misure. In altri termini, laddove le finalità di ricerca scientifica possano essere conseguite attraverso l'utilizzo di dati anonimi, allora dovrà privilegiarsi il ricorso a tecniche di anonimizzazione che consentano la totale e irreversibile de-identificazione del dato. Diversamente, ed in via subordinata, tali garanzie potrebbero includere la pseudonimizzazione, purché il trattamento consenta la realizzazione delle finalità suindicate, sempre nel rispetto del principio di minimizzazione.

Affinché un dato personale possa essere considerato anonimo, dovrà infatti essere sottoposto a un processo di anonimizzazione secondo le tecniche indicate dai Garanti europei – e.g., mascheramento, randomizzazione, generalizzazione ecc. – nonché in base al nuovo standard ISO/IEC 27559:2022 *"Information security, cybersecurity and privacy protection – Privacy enhancing data de-identification framework"*.
La valutazione sull'efficacia del processo di anonimizzazione di volta in volta adottato dovrà tenere conto, in ultima analisi, anche delle specifiche indicazioni fornite nei provvedimenti delle Autorità di Controllo[80], nonché dalla giurisprudenza (nazionale ed europea)[81].
Nel settore della ricerca scientifica – in particolare quella sanitaria – è nota l'importanza di possedere un insieme di dati che presenti caratteristiche di qualità e completezza, rappresentando tali elementi presupposti indefettibili per poter effettuare studi e ricerche accurati ed efficaci. Tuttavia, tali requisiti devono necessariamente contemperarsi con l'esigenza di tutelare la riservatezza degli interessati e delle informazioni che li riguardano.
Sul punto, è altrettanto risaputo che un set di dati completamente anonimo può presentare alcuni limiti che derivano, soprattutto, dalla minor accuratezza delle analisi svolte su tali dati e dalla perdita di relazioni fra i dati stessi.
Per tali ragioni, a supporto della ricerca scientifica (in particolare sanitaria), si stanno diffondendo metodi innovativi che prevedono l'utilizzo di dati c.d. sintetici creati sulla base di modelli statistici elaborati da un algoritmo di Intelligenza Artificiale, in grado di fornire una soluzione alternativa all'anonimizzazione "tradizionale".
L'uso di dati sintetici sembra quindi essere una soluzione estremamente interessante per le aziende del settore tecnologico (sanitario ma non solo) che necessitano di elaborare e analizzare grandi quantità di dati, ed in particolare per le aziende che sviluppano sistemi di intelligenza artificiale che richiedono un notevole volume di dati per l'addestramento degli algoritmi.
Un ulteriore elemento critico, sotto il profilo giuridico, riguarda invece la fase iniziale di elaborazione del dato personale "reale" con cui costruire il modello che dovrà consentire al sistema di Intelligenza Artificiale di generare dati sintetici non personali e di addestrarlo.
L'uso di dati personali per lo sviluppo di modelli di apprendimento o per l'addestramento di sistemi IA generativa rappresenta, infatti, un tipico caso di trattamento di dati per finalità

[80] Si veda *ex multis* provv. GPDP n. 226 del 1° giugno 2023 (caso "ThinS.r.l.").
[81] Cfr. Tribunale UE – caso T-557-20, che ha, in parte, superato i principi espressi dal WP29 nel Parere 05/2014.

secondarie ulteriori e diverse rispetto a quelle originarie, a cui dovranno applicarsi, limitando l'analisi all'ambito sanitario e della ricerca scientifica in ambito sanitario, le prescrizioni previste dall'art. 89 del GDPR e dalla legislazione nazionale in materia.
Un ulteriore aspetto di non secondaria importanza da considerare è rappresentato dalla corretta individuazione delle basi giuridiche del trattamento, atteso che, come già evidenziato, la trasformazione dei dati reali in dati sintetici costituisce un ulteriore trattamento che necessita di un valido presupposto giuridico per rendere lecito il trattamento stesso. Sul punto si è espressa con chiarezza anche l'Autorità di controllo spagnola (*Agencia Española de Protecciòn de Datos* - AEPD), sottolineando come la creazione di dati sintetici da dati reali costituisca un trattamento regolamentato dal GDPR. Pertanto, è necessario tenere in considerazione la normativa europea sulla protezione dei dati personali - e, in particolare, il principio di accountability – e valutare il rischio di una possibile re-identificazione degli interessati partendo dal set di dati sintetici che è stato creato[82] .
Da ultimo, si rileva come, per quanto concerne il tema dell'utilizzo dei dati sintetici, ad oggi non sussista un insieme organico di regole o linee guida elaborate dai Garanti europei o dalle Autorità di controllo europee. In questo senso, limitando l'analisi al panorama italiano, è auspicabile un intervento del Garante per la protezione dei dati personali che fornisca chiarimenti e indirizzi interpretativi ed indichi regole chiare per la sintetizzazione dei dati che non si limitino alle tecnicalità informatiche, ma investano anche gli aspetti giuridici del trattamento – per esempio, individuando le basi giuridiche legittimanti la trasformazione dei dati reali in dati sintetici –, soprattutto in riferimento a quegli ambiti in cui l'utilizzo dei dati sintetici è da ritenere nevralgico (si pensi, ad esempio, ai dati sintetici ottenuti da dati relativi alla salute e utilizzati nel settore della ricerca scientifica)[83] .

3.1.2 Data Governance Act e dati sintetici: diritto al riuso (vedasi considerando 15 e art. 5 del DGA)

Il Regolamento (UE) 2022/868 sulla governance dei dati (si seguito, "**Data Governance Act**" o "**DGA**") del 30 maggio 2022 rappresenta un architrave nel "*framework* normativo" della "Strategia europea per i dati", che mira a creare uno spazio digitale europeo in cui i dati (anche di natura non personale) possano essere utilizzati indipendentemente dal loro luogo fisico di conservazione nell'Unione in modo da permetterne la condivisioni fra soggetti (pubblici e privati) ed accrescere la disponibilità di informazioni, soprattutto all'interno di settori strategici (es. sanità, ambiente, energia, agricoltura, servizi finanziari ecc.).

È importante evidenziare come la nuova regolamentazione europea non si applichi ai soli dati personali (secondo l'accezione prevista nel GDPR), ma si estenda ad ulteriori informazioni che comprendono "*...qualsiasi rappresentazione digitale di atti, fatti o informazioni e qualsiasi raccolta di tali atti, fatti o informazioni, anche sotto forma di registrazione sonora, visiva e audiovisiva*" (cfr. definizione di "dati" ai sensi dell'art. 2, n. 1) del DGA).

[82] Cfr. Comunicato AEPD dell'11.02.2023.
[83] LUCA BOLOGNINI, *"Un focus su dati sintetici e privacy, con proposte evolutive di soft regulation"*, Giuffrè Francis Lefebvre.

Il DGA si basa su tre pilastri essenziali: i. il riutilizzo[84] (o riuso) dei dati in possesso di enti pubblici[85]; ii. la condivisione volontaria dei dati sulla base del c.d. "altruismo dei dati"[86]allo scopo di creare un mercato unico dei dati e valorizzare i dati stessi; e iii. la disciplina dei servizi di intermediazione dei dati.

I principi cardine che riguardano il riutilizzo dei dati sono contenuti nel considerando 15 e nell'art. 5 del DGA, mentre le categorie di dati che rientrano nel diritto di riutilizzo sono definite nel dettaglio nell'articolo 3.

Più specificamente, il DGA si applica ai dati in possesso di enti pubblici – identificati nel Regolamento come "titolari dei dati" - che risultano protetti per motivi di: i. riservatezza commerciale (compresi i segreti commerciali, professionali o d'impresa); ii. riservatezza statistica; iii. protezione della proprietà intellettuale di terzi; e iv. protezione dei dati personali (salvo che non rientrino nell'ambito di applicazione della direttiva (UE) 2019/1024 sul riutilizzo delle informazioni del settore pubblico).

Il DGA prevede, tuttavia, alcune categorie di dati che sono invece sottratte alla possibilità di essere riutilizzate. Ci si riferisce, in particolare, alle seguenti tipologie di dati: i. dati in possesso di imprese pubbliche[87]; ii. dati in possesso di emittenti del servizio pubblico o di altri organismi e delle relative società controllate per l'adempimento di un compito di radiodiffusione di servizio pubblico; iii. dati detenuti da enti culturali e di istruzione; iv. dati in possesso di enti pubblici e protetti per motivi di pubblica sicurezza, difesa o sicurezza nazionale, e v. dati la cui fornitura non rientra nell'ambito dei compiti di servizio pubblico degli enti pubblici interessati.

Venendo al fulcro del tema oggetto del presente contributo, ovverosia il riutilizzo dei dati (in particolare dati sintetici), si osserva come il DGA abbia stabilito regole e garanzie per agevolare tale riutilizzo, permettendo l'estrazione di informazioni da tali dati, senza comprometterne la natura protetta.

La disamina giuridica del riutilizzo dei dati deve prendere le mosse dall'art. 5 del DGA - da leggere in combinato con il considerando 15 – per permettere di estrapolare le regole che devono essere rispettate per poter riutilizzare i dati.

In particolare, le disposizioni citate richiedono che le condizioni per il riutilizzo dei dati rispondano a determinati criteri – in breve, siano non discriminatorie, trasparenti e proporzionate – e, inoltre, siano supportate da un'oggettiva motivazione in relazione alla natura e alle categorie di dati nonché alle finalità del loro riutilizzo.

---

[84] Per "riutilizzo" dei dati deve intendersi, secondo l'art. 2, n. 2 del DGA: *"l'utilizzo di dati in possesso di enti pubblici da parte di persone fisiche o giuridiche a fini commerciali o non commerciali diversi dallo scopo iniziale nell'ambito dei compiti di servizio pubblico per i quali i dati sono stati prodotti, fatta eccezione per lo scambio di dati tra enti pubblici esclusivamente in adempimento dei loro compiti di servizio pubblico"*.

[85] In base all'art. 2, n. 17 del DGA sono considerati enti pubblici: *"le autorità statali, regionali o locali, gli organismi di diritto pubblico o le associazioni formate da una o più di tali autorità oppure da uno o più di tali organismi di diritto pubblico"*.

[86] L'art. 2, n. 16 del DGA fornisce la seguente definizione di "altruismo dei dati": *"la condivisione volontaria di dati sulla base del consenso accordato dagli interessati al trattamento dei dati personali che li riguardano, o sulle autorizzazioni di altri titolari dei dati volte a consentire l'uso dei loro dati non personali, senza la richiesta o la ricezione di un compenso che vada oltre la compensazione dei costi sostenuti per mettere a disposizione i propri dati, per obiettivi di interesse generale, stabiliti nel diritto nazionale, ove applicabile, quali l'assistenza sanitaria, la lotta ai cambiamenti climatici, il miglioramento della mobilità, l'agevolazione dell'elaborazione, della produzione e della divulgazione di statistiche ufficiali, il miglioramento della fornitura dei servizi pubblici, l'elaborazione delle politiche pubbliche o la ricerca scientifica nell'interesse generale"*.

[87] L'art. 2, n. 19 del DGA definisce le imprese pubbliche come: *"qualsiasi impresa su cui gli enti pubblici possono esercitare, direttamente o indirettamente, un'influenza dominante perché ne sono proprietari, vi hanno una partecipazione finanziaria, o in virtù di norme che disciplinano l'impresa in questione […]"*.

Si noti come il DGA, pur focalizzandosi sugli aspetti legati alla tutela dei dati e alle misure e garanzie da applicare al riutilizzo degli stessi, presti allo stesso tempo la massima attenzione alla concorrenza fra imprese, che non potrà essere in alcun modo limitata dalle condizioni dettate per il diritto di riutilizzo dei dati. Allo stesso modo, le medesime regole da applicare al riutilizzo dei dati non devono limitare, bensì agevolare, la promozione e lo sviluppo della ricerca scientifica.

Quanto alle misure che devono essere adottate per garantire un'adeguata protezione dei dati, viene posto a carico degli enti pubblici un obbligo di assicurare che i dati oggetto della richiesta di riutilizzo siano resi anonimi (solo nel caso di dati personali) e che, nel caso di informazioni commerciali riservate (fra cui segreti commerciali e contenuti protetti da diritti di proprietà intellettuale), queste siano adeguatamente modificate, aggregate o altrimenti gestite con opportuni controlli sulla divulgazione.

Da un'interpretazione letterale delle disposizioni contenute nell'art. 5 sembra che, per quanto concerne i dati personali, la pseudonimizzazione prevista dal GDPR quale misura di sicurezza di tipo tecnico non sia sufficiente ad ottenere l'accesso ai dati per il loro riutilizzo, atteso che il rispetto delle condizioni previste dal DGA è vincolato all'attuazione di "un'adeguata anonimizzazione". Si consideri, tuttavia, che il considerando 15 dispone invece che la trasmissione a terzi di dati (anche personali) pseudonimizzati per il loro riutilizzo può essere consentito, purché sia impossibile procedere alla re-identificazione degli interessati.

Ciò posto, nell'ottica di favorire lo scambio di dati ed il loro riutilizzo, il DGA contempla anche due opzioni, alternative fra loro, che consentono l'accesso legittimo alle informazioni detenute da un ente pubblico: i. accesso da remoto a un ambiente di trattamento sicuro, fornito o controllato da un ente pubblico; oppure ii. riutilizzo dei dati in un ambiente di trattamento sicuro ubicato presso la sede fisica di un ente pubblico. In tutti i casi, devono essere attuate rigorose misure di sicurezza.

Fra le tecniche che il DGA giudica idonee a garantire la protezione dei dati personali e ridurre il rischio di identificazione degli interessati è espressamente citato l'uso di dati sintetici (cfr. cons. 7). Sotto questo profilo, è di sicuro rilievo l'apertura della normativa europea all'uso dei dati ottenuti mediante sintetizzazione, la cui applicazione è equiparata, quanto a livello di protezione della privacy degli individui, ad altre tecniche quali l'anonimizzazione o la privacy differenziale.

In altri termini, la sintetizzazione dei dati potrebbe essere utilizzata, negli ambiti individuati dal DGA, come efficace tecnica di anonimizzazione per accedere, analizzare, condividere, riutilizzare e pubblicare i dati senza rivelare informazioni personali. La sintesi dei dati è vista, in questo senso, come uno strumento che rispetta i requisiti di protezione dei dati e allo stesso tempo stimola l'innovazione tecnologica.

Se da un lato l'applicazione di tali tecniche è già in grado di assicurare una maggiore sicurezza nell'utilizzo e riutilizzo dei dati personali nonché delle informazioni commerciali riservate a fini statistici, di ricerca e di innovazione, non può non rilevarsi come il DGA, ad ulteriore tutela degli interessati e dei dati che li riguardano, esiga, in determinati casi, l'esecuzione di ulteriori adempimenti a carico dei soggetti pubblici detentori dei dati, con l'eventuale coinvolgimento delle Autorità di controllo nazionali. Lo si evince, per esempio, dal considerando 15, laddove viene stabilito che, se la fornitura di dati anonimizzati non risponda alle esigenze del

riutilizzatore, prima di concedere il riutilizzo dei dati in loco o da remoto in un ambiente di trattamento sicuro, gli enti pubblici sono tenuti ad effettuare una valutazione d'impatto sulla protezione dei dati e, se necessario, a consultare l'Autorità di controllo competente conformemente agli artt. 35 e 36 del GDPR.

Procedendo oltre nell'analisi dei punti cardine del DGA, vengono in rilievo i compiti attribuiti a specifiche figure previste dalla nuova normativa, quali i "fornitori di servizi di intermediazione dei dati".

I servizi di intermediazione, definiti nell'art. 2, n. 11 del DGA, hanno la finalità precipua di instaurare rapporti commerciali per condividere i dati fra i titolari dei dati e gli interessati, da un lato, e gli utenti dei dati, dall'altro.

Più nel dettaglio, le tipologie di servizi di intermediazione dei dati enucleate dal DGA sono le seguenti: i. servizi di intermediazione tra i titolari dei dati e i potenziali utenti dei dati: tali servizi possono includere scambi di dati bilaterali o multilaterali o la creazione di piattaforme o banche dati che consentono lo scambio o l'utilizzo congiunto dei dati, nonché infrastrutture specifiche per l'interconnessione di titolari dei dati con gli utenti dei dati; ii. servizi di intermediazione tra interessati/persone fisiche e potenziali utenti dei dati: tali servizi, oltre ad attività di assistenza, possono comprendere la messa a disposizione di mezzi tecnici per consentire, in particolare, l'esercizio dei diritti degli interessati ai sensi del GDPR; e iii. servizi di cooperative di dati: tali servizi sono offerti da una struttura organizzativa costituita da interessati, imprese individuali o da PMI, al fine di assistere i propri membri nell'esercizio dei loro diritti in relazione a determinati dati e nel negoziare termini e condizioni per il trattamento degli stessi[88] .

È bene rilevare come il DGA richieda che i servizi di intermediazione vengano forniti attraverso una persona giuridica distinta, con ciò imponendo, di fatto, una separazione strutturale di detti servizi di intermediazione. La ratio di tale disposizione risiede evidentemente nella volontà del legislatore europeo di evitare che le società che forniscono servizi della società dell'informazione possano trarre un vantaggio dalle informazioni e dai dati raccolti mediante l'attività di intermediazione.

L'impianto normativo concepito nel DGA prevede, inoltre, che i fornitori di servizi di intermediazione di dati seguano una specifica procedura di notifica all'Autorità competente per i servizi di intermediazione individuata da ciascun Stato membro dell'UE. La notifica deve contenere le informazioni tassativamente elencate nell'art. 11, par. 6 del DGA, tra cui, a titolo esemplificativo, il nome del fornitore dei servizi di intermediazione, la forma giuridica, l'assetto proprietario, la sede legale, la descrizione del servizio di intermediazione dei dati ecc.

L'Autorità competente per i servizi di intermediazione dei dati è poi tenuta ad informare senza ritardo la Commissione europea di ogni nuova notifica ricevuta.

Le Autorità competenti, oltre ad essere responsabili del processo di notifica, devono anche occuparsi del monitoraggio della conformità dei servizi di intermediazione ai requisiti posti dal DGA, anche adottando misure sanzionatorie e ordinando la sospensione o la cessazione dell'attività, in caso di violazioni.

---

[88] ELISABETTA NUNZIANTE, *"Data Governance Act, da settembre al via la nuova disciplina"*, Risk Management 360, 05.07.2023.

Com'è facilmente intuibile, gli obblighi sopra riportati assumono un'importanza cruciale per i fornitori di servizi di intermediazione dei dati, dal momento che dall'esito positivo della procedura di notifica dipende l'effettiva possibilità per gli stessi di svolgere la propria attività ed erogare i servizi di intermediazione conformemente al DGA.

Sul punto si segnala che la Legge 21 febbraio 2024, n. 15 recante *"Delega al Governo per il recepimento delle direttive europee e l'attuazione di altri atti dell'Unione europea – Legge di delegazione europea 2022 – 2023",* all'art. 17 ha delegato il Governo italiano ad adottare, entro quattro mesi dall'entrata in vigore della predetta legge (quindi entro luglio 2024), gli atti normativi – i.e., decreti legislativi – per l'adeguamento della normativa nazionale al DGA. In adempimento alle disposizioni normative citate, il 3 luglio 2024 il Governo ha approvato lo schema di decreto legislativo che introduce norme di adeguamento della normativa nazionale alle disposizioni del DGA. In particolare, per quanto qui di interesse, il provvedimento designa l'Agenzia per l'Italia Digitale (AgID) come autorità responsabile per lo svolgimento dei compiti relativi alla procedura di notifica per i servizi di intermediazione dati nonché quale autorità competente alla registrazione delle organizzazioni per l'altruismo dei dati. AgID, inoltre, è designata quale organismo competente per assistere gli enti pubblici che concedono o rifiutano l'accesso al riutilizzo di specifiche categorie di dati.

3.1.3 Data Act: accesso ai dati

Il Regolamento europeo (UE) 2023/2854 "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" (c.d. "**Data Act**"), entrato in vigore l'11 gennaio 2024 e applicabile dal 12 settembre 2025, al pari del DGA (precedentemente esaminato), rappresenta un pilastro fondamentale della "Strategia europea sui dati", introducendo nuove norme armonizzate sull'accesso ai dati di prodotti connessi e sul loro utilizzo. L'obiettivo del Data Act - e della più ampia strategia dell'UE sui dati - è quello di facilitare un accesso affidabile e sicuro ai dati (di natura personale e non personale), promuovendone l'uso in settori economici chiave e in aree di interesse pubblico.

L'ambito di applicazione soggettivo del Data Act è molto ampio e riguarda, in primo luogo, i fabbricanti di prodotti connessi immessi sul mercato europeo, i fornitori di servizi correlati, i fornitori di servizi di trattamento dei dati (c.d. Data Processing Services – DPS) - indipendentemente dal luogo di stabilimento - i titolari dei dati, gli enti pubblici e, naturalmente, gli utenti nell'Unione Europea di tali prodotti o servizi.

Sono escluse dall'ambito di applicazione del Data Act le microimprese e le piccole imprese, salvo che si tratti di imprese facenti parte di un gruppo imprenditoriale quali imprese collegate o controllate.

Come detto, l'ampio concetto di "dati" - mutuato dal Data Governance Act – comprende anche i dati personali come definiti dall'art. 4, n. 11 del GDPR. Di conseguenza, qualora i dati dell'utente siano di natura personale, si dovrà applicare la normativa unionale in materia di protezione dei dati, unitamente alle disposizioni del Data Act, in quanto compatibili (le disposizioni del GDPR, infatti, in caso di contrasto, devono essere ritenute prevalenti).

Riguardo al tema dell'accessibilità dei dati disciplinato dal Data Act, è interessante notare come l'art. 3 del Regolamento introduca per la prima volta il concetto di accessibilità dei dati "by design e by default", con ciò intendendo che i prodotti connessi e i relativi servizi devono essere progettati e sviluppati, se tecnicamente possibile, in maniera tale da consentire agli utenti di accedere facilmente e direttamente alle informazioni generate da tale prodotto (es. un dispositivo) o servizio. Sotto questo profilo, una delle principali novità introdotte dal Data Act circa il diritto di accesso da parte degli utenti ai propri dati deve essere rinvenuta nell'art. 4[89] , il quale stabilisce che, qualora i dati del prodotto connesso o del relativo servizio non siano direttamente accessibili da un utente, il titolare dei dati è tenuto, senza ritardo, a rendere disponibili tali dati nonché i relativi metadati rispettando le condizioni dettate dalla norma citata (ovverosia in modo facile, sicuro, gratuito ecc.).

Fra gli aspetti rilevanti che emergono dalla disposizione sopra richiamata va citato il fatto che tutti i dati generati, per esempio, da un sistema IoT, che non siano ancora nella disponibilità dell'utente, devono adesso essere resi accessibili dal titolare nei tempi e nei modi definiti dal Regolamento. Ciò significa che, nell'ipotesi in cui i dati generati dal sistema non siano nel controllo dell'utente, quest'ultimo avrà il diritto di riceverne una copia in un formato strutturato e alle condizioni normativamente previste.

La nuova normativa europea prevede, in deroga all'ampio diritto di accesso a cui si è già accennato, che l'accesso ai dati, così come il loro utilizzo e l'ulteriore condivisione, possano essere assoggettati a una limitazione o, addirittura, a un divieto stabilito in maniera espressa dall'utente o dal titolare dei dati all'interno di un contratto, a condizione che il trattamento possa compromettere i requisiti di sicurezza del prodotto connesso e comportare gravi pregiudizi per la salute, la sicurezza o la protezione delle persone fisiche. In tale ambito, non può escludersi che l'uso di dati sintetici rappresenti una valida soluzione volta a superare tali limitazioni o divieti relativi ai trattamenti dei dati nei termini sopra evidenziati; sotto il profilo contrattuale, si potrebbe invece prevedere, all'interno di un contratto, che la limitazione o il divieto al trattamento nei termini sopra rappresentati non si applichi qualora i dati in questione siano sottoposti a un processo di sintetizzazione che consenta di evitare che dal loro uso derivi un pregiudizio ai requisiti di sicurezza del prodotto o alla salute e sicurezza delle persone fisiche.

A prescindere dalle eccezioni poc'anzi trattate, nelle previsioni normative del Data Act che disciplinano il diritto di accesso si possono scorgere alcuni vantaggi che potrebbero risolvere un problema comune a molti prodotti tecnologici: ci si riferisce, in particolare, alla necessità di effettuare riparazioni, manutenzione ed aggiornamenti solo presso i produttori stessi o i rivenditori autorizzati di tali prodotti, in quanto i dati rilevanti, che consentirebbero di comprendere il funzionamento del sistema, non sono divulgati dal produttore perché da questi ritenuti informazioni coperte da "segreto commerciale".

---

[89] L'art. 4, par. 1 del Data Act dispone che: *"Qualora l'utente non possa accedere direttamente ai dati a partire dal prodotto connesso o dal servizio correlato, i titolari dei dati mettono prontamente a disposizione dell'utente i dati, nonché i pertinenti metadati necessari per interpretare e utilizzare tali dati senza indebito ritardo, con la stessa qualità di cui dispone il titolare dei dati, in modo facile, sicuro, gratuitamente, in un formato completo, strutturato, di uso comune e leggibile da dispositivo automatico e, ove pertinente e tecnicamente possibile, in modo continuo e in tempo reale. Ciò avviene sulla base di una semplice richiesta mediante mezzi elettronici, ove tecnicamente fattibile".*

Sul punto, le norme del Data Act stabiliscono che la comunicazione di segreti commerciali possa essere effettuata a condizione che siano adottate tutte le misure specifiche necessarie a preservare la riservatezza dei dati condivisi, in particolare nei confronti di terzi. Tali misure, da concordare tra titolare dei dati e utente, possono consistere in clausole contrattuali specifiche, accordi di riservatezza, codici di condotta ecc.

L'accesso a dati che costituiscono "segreti commerciali" prevede comunque alcune eccezioni, per esempio quando il titolare dei dati, che è detentore di un segreto commerciale, sia in grado di dimostrare un'alta probabilità di subire danni patrimoniali a causa della sua divulgazione, nonostante siano adottate dall'utente misure tecniche e organizzative. In questi casi, il titolare potrà rifiutarsi di dare seguito alla richiesta di accesso ai dati ai sensi dell'art. 4, par. 7 del Data Act.

Ad ulteriore tutela del titolari dei dati, è altresì presente una disposizione di non concorrenza per prevenire l'uso improprio da parte dell'utente dei dati ottenuti ai sensi del Data Act: in base a tale previsione, all'utente è vietato utilizzare i dati a cui ha avuto accesso in seguito a una specifica richiesta secondo il paragrafo 1 dell'art. 4 per sviluppare un prodotto connesso in concorrenza con il prodotto da cui provengono i dati; né l'utente può condividere i dati con un terzo ovvero utilizzarli per ottenere informazioni sulla situazione economica, sulle risorse e sui metodi di produzione del fabbricante o del titolare dei dati. Quest'ultima disposizione è controbilanciata, in maniera simmetrica, dal par. 13 dell'art. 4, che pone un limite all'utilizzo dei dati generati dall'utente, disponendo che il titolare dei dati potrà avere accesso ed utilizzare qualsiasi dato non personale generato dall'uso di un prodotto o di un servizio correlato solo sulla base di un accordo contrattuale con l'utente e, allo stesso tempo, non potrà utilizzare tali dati per ricavare informazioni sulla situazione economica, sui beni e sui metodi di produzione dell'utente o sull'uso da parte dell'utente che potrebbero compromettere la posizione commerciale dell'utente stesso sui mercati in cui opera.

Da un lato, quindi, l'utente non può sfruttare i dati per diventare un concorrente del titolare dei dati, in quanto si tratterebbe di una pratica sleale e, in questo ambito, il Data Act ha l'intento di impedire che il diritto di accesso venga utilizzato come pretesto per ottenere illecitamente segreti commerciali e informazioni aziendali per conseguire un vantaggio commerciale. Dall'altro lato, contestualmente, viene assicurato che i dati generati dall'utente siano regolati da un contratto stipulato con il titolare dei dati, vietando a quest'ultimo di sfruttare i dati dell'utente in modo sleale, tale da comprometterne la posizione sul mercato.

È facile intuire come l'introduzione del diritto di accesso ai dati avrà un impatto significativo sia sulle grandi imprese sia sulle PMI, ma anche sulle persone fisiche.

Un settore di applicazione del Data Act di indubbia rilevanza è quello della sanità digitale (Digital Health). Al riguardo, si pensi, infatti, che qualsiasi dispositivo medico o wearable connesso all'Internet of Things (IoT) che raccolga, generi ed elabori dati della persona che lo utilizza, o relativi al suo ambiente, rientra nel perimetro di applicazione del Data Act. I dispositivi medici e sanitari sono inoltre espressamente menzionati dalla normativa e la maggior parte dei soggetti interessati alla salute digitale, come i produttori di dispositivi medici (anche indossabili), nonché i fornitori di servizi correlati, saranno assoggettati alle previsioni del Data Act quando il prodotto o il servizio è immesso sul mercato dell'Unione europea.

3.2 Futura

3.2.1 Dato sintetico nell'AI Act (artt. 10 e 59)

Dopo l'accordo provvisorio sulla proposta di Regolamento europeo contenente norme armonizzate sull'Intelligenza Artificiale (c.d. "**Artificial Intelligence Act**" o "**AI Act**") raggiunto il 9 dicembre 2023 fra la presidenza del Consiglio dell'Unione europea e i negoziatori del Parlamento europeo e la successiva intesa sul testo finale del provvedimento raggiunta il 2 febbraio 2024 dal Coreper (Comitato dei rappresentanti permanenti dei governi degli Stati membri dell'Unione europea), si è concluso l' iter legislativo relativo all'AI Act con l'approvazione del testo consolidato del provvedimento da parte del Parlamento Europeo e da parte del Consiglio dell'UE avvenuta, rispettivamente, il 6 marzo e il 21 maggio 2024. L'atto legislativo è stato pubblicato nella Gazzetta Ufficiale dell'Unione europea il 12 luglio 2024 ed entrerà in vigore nei venti giorni successivi. Il nuovo Regolamento si applicherà due anni dopo l'entrata in vigore, ovvero a partire dal 2 agosto 2026, con alcune eccezioni relative a fattispecie specifiche.

L'AI Act, che a buon diritto è destinato rappresentare una pietra miliare nella legislazione unionale, si pone l'obiettivo primario di promuovere lo sviluppo e l'adozione in tutto il mercato unico dell'Unione europea, da parte di soggetti sia pubblici che privati, di un'Intelligenza Artificiale sicura e affidabile, che dovrà sottostare a rigorose regole secondo un approccio basato sul rischio.

Il tema che si intende approfondire in questa sede attiene all'utilizzo dei dati sintetici e alla loro disciplina contenuta nell'AI Act, in particolare negli artt. 10 – *"Dati e governance dei dati"* – e 59 – *"Ulteriore trattamento dei dati personali per lo sviluppo nello spazio di sperimentazione normativa per l'IA di determinati sistemi di IA nell'interesse pubblico"*.

In linea di premessa deve rilevarsi come la diffusione dei sistemi di Intelligenza Artificiale abbia fatto emergere in maniera preponderante l'esigenza del legislatore europeo di regolamentare l'utilizzo di tali strumenti, ed in particolare di prevedere norme specifiche tese a disciplinare puntualmente gli aspetti connessi alla protezione dei dati personali, ponendo al centro la protezione dei diritti e delle libertà fondamentali degli individui.

Come rilevato anche dal Comitato consultivo della Convenzione per la protezione delle persone rispetto al trattamento automatizzato di dati a carattere personale (Convenzione 108), l'innovazione responsabile nel campo dell'Intelligenza Artificiale richiede un approccio incentrato sulla mitigazione e sull'eliminazione dei rischi potenziali del trattamento dei dati personali[90] .

Già nelle proprie Linee guida, il Comitato consultivo sottolineava l'importanza per sviluppatori, produttori e fornitori di servizi di IA di adottare un approccio volto a tutelare i diritti umani fin dalla progettazione di tali servizi (c.d. "*human rights design*"), evitando qualsiasi potenziale pregiudizio (*bias*), anche involontario od occulto, nonché rischi di discriminazione o altri effetti negativi sui diritti e le libertà fondamentali delle persone fisiche. Da qui la necessità, in capo agli sviluppatori, di utilizzare solo set di dati accurati e di qualità, riducendo i dati eccedenti, inutili o marginali durante lo sviluppo e le fasi di addestramento del sistema.

[90] Cfr. anche *"Guidelines on Artificial Intelligence and data protection"* del 25 gennaio 2019.

Nel proprio documento di indirizzo il Comitato menziona i dati sintetici, indicando il loro utilizzo come una soluzione adeguata a "minimizzare" la quantità di dati personali reali trattati dalle applicazioni "intelligenti".

Si aggiunga che attraverso i dati sintetici – a seconda della metodologia utilizzata nel processo di sintetizzazione - è inoltre possibile impedire il recupero di informazioni che consentirebbero di risalire agli interessati (garantendo quindi la non reversibilità della de-identificazione) e superare certe prescrizioni restrittive poste dalla normativa in materia di protezione dei dati personali.

Sul punto, si può notare come l'AI Act si riferisca ai dati sintetici e a quelli anonimi in termini di equivalenza. Lo si evince dall'art. 59, laddove vengono disciplinate le condizioni che, nelle c.d. "AI regulatory sandbox"[91], devono essere soddisfatte per poter riutilizzare ai soli fini di sviluppo, addestramento e collaudo dei sistemi di AI i dati personali precedentemente raccolti per altre finalità. In sostanza, da un'interpretazione letterale del par. 1, lett. b) della norma richiamata, l'utilizzo dei dati sintetici (al pari di quelli anonimi e non personali) sembra doversi considerare prevalente – e preferibile, salvo eccezioni – nel soddisfare i requisiti previsti dall'AI Act per i sistemi di IA classificati "ad alto rischio"[92]. In verità non sfugge all'interprete come il legislatore europeo sembri equiparare i dati sintetici ai dati di natura non personale: si veda, a questo proposito, l'inserimento nello stesso art. 59, par. 1, lett. b) della locuzione "*altri dati non personali*", preceduto dal riferimento ai dati sintetici (nonché a quelli anonimi), da cui si può evincere l'assimilazione di tali ultimi tipi di dati alla categoria dei dati non personali.

Per poter immettere sul mercato detti sistemi intelligenti, l'AI Act prevede una serie di obblighi stringenti a carico di sviluppatori e produttori dovuti alle loro particolari caratteristiche. Tali obblighi comprendono, innanzitutto, l'istituzione e il mantenimento di un adeguato sistema di gestione del rischio e di data governance, volto a garantire solidità, sicurezza e accuratezza dei sistemi di Intelligenza Artificiale. Deve inoltre essere garantita la possibilità di correggere o disattivare il sistema in caso di errori o rischi e devono essere previsti meccanismi di supervisione nonché la possibilità di assicurare l'intervento umano per prevenire o mitigare danni o impatti negativi sui diritti e le libertà delle persone fisiche.

Prima dell'immissione del sistema sul mercato è altresì necessario predisporre la documentazione tecnica completa per dimostrare la conformità del sistema di Intelligenza Artificiale ai requisiti dettati dalla normativa europea e, contestualmente, istituire un registro che garantisca la tracciabilità del funzionamento del sistema per l'intera permanenza dello stesso sul mercato.

L'articolo 10 dell'AI Act stabilisce criteri di qualità rigorosi che i dati utilizzati come base per lo sviluppo di sistemi di Intelligenza Artificiale ad alto rischio nelle fasi di addestramento, convalida e test devono possedere.

---

[91] Con il termine *"AI Regulatory sandbox"* si fa riferimento ad ambienti "controllati", stabiliti da un'autorità competente in materia, in cui vengono sviluppati, addestrati e testati i sistemi di AI secondo un progetto concordato fra lo sviluppatore di sistemi di AI e l'autorità stessa.

[92] Con l'espressione sistemi di Intelligenza Artificiale ad *"alto rischio"* ci si riferisce ai sistemi AI utilizzati come componenti di sicurezza di un prodotto, quelli che sono essi stessi un prodotto, coperto dalla legislazione UE in materia di salute e sicurezza e, infine, quelli che rientrano nei settori elencati nell'Allegato III dell'AI Act (es. sistemi di identificazione biometrica a distanza).

Gli obblighi che attengono alla gestione e alla verifica dei dati utilizzati per lo sviluppo dei sistemi di Intelligenza Artificiale comprendono l'analisi di possibili pregiudizi che possano incidere sulla salute e sulla sicurezza delle persone, oppure avere un impatto negativo sui diritti fondamentali o, ancora, portare a discriminazioni. Per ridurre detti rischi, è richiesta l'applicazione di misure appropriate volte ad individuare, prevenire e attenuare i possibili pregiudizi. Gli insiemi di dati utilizzati devono inoltre essere pertinenti, sufficientemente rappresentativi, e, nei limiti del possibile, privi di errori e il più possibile completi in considerazione delle finalità perseguite e devono possedere le proprietà statistiche appropriate.

L'articolo 10, paragrafo 5, autorizza i fornitori di sistemi di Intelligenza Artificiale ad alto rischio, nella misura strettamente necessaria a garantire la rilevazione e la correzione dei pregiudizi negativi o distorsioni (c.d. *bias*), a trattare eccezionalmente dati personali appartenenti alle categorie particolari di cui agli artt. 9(1) of Regolamento (UE) 2016/679 (GDPR), 10 della Direttiva (UE) 2016/680 e 10(1) del Regolamento (UE) 2018/1725. Tuttavia, tale trattamento deve essere soggetto ad adeguate garanzie per i diritti e le libertà fondamentali delle persone fisiche, comprese limitazioni tecniche al riutilizzo e all'uso di misure di sicurezza e di tutela della privacy all'avanguardia.

La norma in commento, in sintesi, individua le condizioni tassative che costituiscono il presupposto per il trattamento dei dati particolari per le finalità suindicate.

La prima di tali condizioni - che devono ricorrere cumulativamente - stabilisce che il trattamento dei dati particolari ai sensi dell'art. 10 è consentito solo se il suo obiettivo, ossia l'individuazione e la correzione dei pregiudizi, non può essere efficacemente conseguito mediante il trattamento di dati sintetici o anonimizzati. Al contrario, se un fornitore di sistemi di Intelligenza Artificiale è in grado di rilevare e correggere i pregiudizi utilizzando dati sintetici o anonimizzati, è tenuto a farlo e non può fare affidamento su altre "garanzie adeguate".

La seconda condizione necessaria si riferisce, invece, all'applicazione di limitazioni tecniche relative al riutilizzo delle categorie particolari di dati personali, nonché a misure più avanzate di sicurezza e di tutela della vita privata, compresa la pseudonimizzazione.

È altresì necessario che i dati particolari siano sottoposti a misure tese a garantire un'adeguata sicurezza del trattamento nonché a garanzie appropriate – fra cui controlli tracciabili degli accessi ai dati – al fine di evitare abusi e assicurare che solo le persone autorizzate, vincolate a specifici obblighi di riservatezza, abbiano accesso a tali dati personali (il riferimento è, evidentemente, ai soggetti autorizzati al trattamento di cui agli artt. 29, 32, par. 4 del GDPR e 2-*quaterdecies* del Codice Privacy).

Quali ulteriori condizioni da soddisfare per ricadere nella deroga al divieto di trattamento dei dati particolari, la norma in commento prevede un esplicito divieto di trasmissione, trasferimento o consultazione da parte di terzi di tale categoria di dati personali nonché l'obbligo di cancellare i dati personali dopo che sia stata corretta la distorsione (o *bias*) oppure una volta decorso il periodo di conservazione individuato (a seconda di quale condizione si verifichi per prima).

Infine, si rileva l'ulteriore obbligo di annotare nel registro dei trattamenti i motivi che hanno reso necessario il trattamento delle categorie particolari di dati personali, specificando le ragioni per le quali il medesimo risultato non poteva essere raggiunto mediante il trattamento di altri tipi di dati.

Sembra pertanto che, in tale contesto, il legislatore europeo identifichi l'uso dei dati sintetici – unitamente ai dati anonimizzati - come metodo preferenziale per rilevare e risolvere i *bias* presenti nei dati utilizzati per lo sviluppo dei sistemi di Intelligenza Artificiale ad alto rischio. Va inoltre soggiunto che i dati sintetici, se impiegati per sostituire i dati personali, sono in grado di fornire un importante ausilio allo sviluppo dei processi di *data governance* e cybersicurezza in ottica di mitigazione dei rischi derivanti dall'utilizzo dei sistemi di Intelligenza Artificiale, compresi quelli ad alto rischio. È altresì opportuno rimarcare come il riconoscimento del dato sintetico come dato non personale incentivi l'impiego di questa tecnologia generativa in un'ampia gamma di applicazioni, a beneficio della ricerca scientifica e della programmazione delle politiche pubbliche - in particolare in materia sanitaria -, considerato che la sintetizzazione consente di condividere, archiviare e riutilizzare i dati per finalità di interesse pubblico diverse da quelle perseguite inizialmente – ovverosia al momento dell'acquisizione dei dati -, garantendo nello stesso tempo la tutela del diritto alla protezione dei dati personali.

I fornitori e gli utenti di sistemi di Intelligenza Artificiale dovrebbero implementare misure tecniche e organizzative all'avanguardia per proteggere tali diritti: tali misure dovrebbero includere non solo l'anonimizzazione e la crittografia, ma anche l'uso di tecnologie che permettano di applicare algoritmi ai dati e di ricavarne informazioni preziose senza la trasmissione tra le parti o la copia non necessaria dei dati grezzi o strutturati.

Da ultimo, si rileva come la valenza giuridica attribuita ai dati sintetici mediante l'inclusione nell'AI Act di previsioni normative che esplicitamente li riguardano, è da considerare una conferma della crescente consapevolezza politica del potenziale di questa tecnologia. Ciò è sottolineato da un recente *report* del Centro comune per la ricerca della Commissione europea (*Joint Research Center* - JRC), in cui si legge che i dati sintetici non solo possono essere condivisi liberamente, ma possono anche aiutare a riequilibrare le classi sottorappresentate negli studi di ricerca attraverso un sovracampionamento, rendendoli un input perfetto per i modelli di apprendimento automatico e di Intelligenza Artificiale[93] .

3.2.2 European Health Data Space: proposta di regolamento

Il 24 aprile 2024, dopo l'accordo raggiunto con il Consiglio dell'Unione europea, il Parlamento europeo ha approvato il testo del Regolamento che istituisce l'European Health Data Space (di seguito, solo "EHDS"), destinato ad apportare cambiamenti radicali nella gestione dei dati sanitari all'interno degli Stati membri dell'Unione.

La normativa europea contenuta nell'EHDS rappresenta infatti una notevole opportunità per rendere lecito un miglior uso di grandi quantità di dati sanitari ad oggi di difficile accesso, avendo quale scopo prioritario la creazione di uno spazio comune europeo, con standard interoperabili e un accesso facilitato ai dati sanitari sia per l'uso primario (fornitura di servizi sanitari) sia per gli usi secondari.

Nella relazione alla proposta di Regolamento è stato evidenziato come, attualmente, le persone fisiche incontrino difficoltà nell'esercizio dei propri diritti sui dati sanitari elettronici che li riguardano (es. nell'esercizio del diritto di accesso e nella loro trasmissione nazionale e transfrontaliera), dovute principalmente alla disomogeneità di attuazione ed interpretazione, da parte degli Stati membri, delle norme relative al trattamento dei dati sanitari contenute nel

[93] Cfr. JRC Technical Report, *"Multipurpose synthetic population for policy applications"*, 13.06.2022.

GDPR, con la conseguente determinazione di incertezze giuridiche in grado di creare ostacoli all'uso secondario dei dati sanitari elettronici. Inoltre, anche a causa dell'eterogeneità delle legislazioni nazionali e dell'interoperabilità limitata, i fabbricanti di prodotti di sanità digitale ed i prestatori di servizi di sanità digitale che operano in uno Stato membro incontrano difficoltà e costi aggiuntivi nell'accedere ai dati conservati in un altro Stato membro.

Consentire un più ampio riutilizzo dei dati sanitari significherebbe, dunque, apportare benefici e miglioramenti, fra gli altri, nei settori della ricerca scientifica, dell'innovazione e della sicurezza dei pazienti.

Tanto premesso, si osserva come il Regolamento contempli requisiti piuttosto stringenti per poter accedere ai dati sanitari elettronici detenuti da un "titolare dei dati"[94] e procedere al loro uso secondario.

Tra le diverse finalità secondarie perseguibili dal richiedente attraverso il trattamento dei dati di natura sanitaria - in particolare, quelli contenuti nelle cartelle cliniche elettroniche – l'EHDS annovera espressamente anche le attività di addestramento, prova e valutazione degli algoritmi, anche relativi a dispositivi medici, sistemi di Intelligenza Artificiale e applicazioni di sanità digitale, che contribuiscano alla sanità pubblica o alla sicurezza sociale, oppure garantiscano elevati livelli di qualità e sicurezza dell'assistenza sanitaria, dei medicinali o dei dispositivi medici.

Secondo i precetti normativi sopra richiamati, il richiedente (utilizzatore dei dati) che intenda accedere a dati sanitari elettronici (di natura personale e non personale) in possesso di un titolare dei dati per l'uso secondario deve essere in grado di dimostrare che l'accesso è necessario a perseguire una delle finalità esplicitamente richiamate dall'EHDS (v. art. 34).

L'EHDS prevede, inoltre, un divieto di trattamento dei dati sanitari elettronici per usi secondari per una serie di finalità, quali: i. prendere decisioni pregiudizievoli per una persona fisica sulla base dei suoi dati sanitari elettronici; ii. escludere una persona fisica o un gruppo di persone fisiche dai benefici derivanti da un contratto di assicurazione o modificare i loro contributi e premi assicurativi; iii. sviluppare prodotti o servizi che possono danneggiare gli individui e la società in generale, tra cui tabacco, bevande alcoliche; iv. svolgere attività pubblicitarie o di marketing rivolte a professionisti sanitari, organizzazioni sanitarie o persone fisiche; v. fornire l'accesso ai dati sanitari elettronici o renderli disponibili in altro modo, a terzi non menzionati nell'autorizzazione all'uso dei dati.

L'EHDS prevede un elenco piuttosto ampio di categorie minime di dati sanitari elettronici che i titolari dei dati devono mettere a disposizione per il riutilizzo, tra cui le cartelle cliniche elettroniche, i dati genetici, genomici e proteomici umani, i dati identificativi relativi agli operatori sanitari coinvolti nella cura dei pazienti, i dati sanitari elettronici provenienti da sperimentazioni cliniche, i dati sanitari elettronici provenienti da biobanche ecc.

---

[94] L'art. 2, lett. y) della proposta di EHDS definisce il "titolare dei dati" come *"una persona fisica o giuridica che è un soggetto o un organismo del settore sanitario o dell'assistenza, o che svolge attività di ricerca in relazione a tali settori, nonché le istituzioni, gli organi e gli organismi dell'Unione che hanno il diritto o l'obbligo, conformemente al presente regolamento, al diritto dell'Unione applicabile o alla legislazione nazionale di attuazione del diritto dell'Unione, o, nel caso di dati non personali, mediante il controllo della progettazione tecnica di un prodotto e dei servizi correlati, la capacità, di rendere disponibili determinati dati, anche in termini di registrazione, fornitura, limitazione dell'accesso o scambio"*.

Nella richiesta di accesso ai dati sanitari elettronici per uso secondario, da presentare all'organismo responsabile per l'accesso ai dati sanitari istituito presso ciascuno Stato membro, l'utente è obbligato a fornire una serie di elementi, tra cui una spiegazione dettagliata dell'uso previsto di tali dati e l'indicazione delle finalità perseguite, fra quelle previste dal Regolamento; inoltre, tra le informazioni che devono essere contenute nella richiesta di accesso, è necessario indicare se questa si riferisca a dati anonimizzati ovvero pseudonimizzati. La distinzione non è di scarsa rilevanza, atteso che, se i dati a cui l'utente intende accedere sono in forma anonimizzata, l'EHDS prevede una procedura "semplificata" per ottenere l'autorizzazione da parte dell'organismo responsabile dell'accesso ai dati sanitari. Al contrario, qualora l'utente abbia la necessità di accedere a dati sanitari solo pseudonimizzati – quindi dati di natura personale a cui è applicabile la normativa in materia di protezione dei dati personali -, il richiedente è tenuto a comunicare alcune informazioni integrative, tra cui la motivazione su cui si basa la domanda di accesso, esplicitando le ragioni per le quali i dati anonimi non sono sufficienti al perseguimento delle finalità manifestate. Qualora siano resi disponibili dati sanitari elettronici personali in forma pseudonimizzata, la chiave di cifratura può essere detenuta solo dall'organismo responsabile dell'accesso ai dati sanitari. Inoltre, nella propria richiesta di accesso ai dati sanitari, l'utente ha l'obbligo di identificare la base giuridica del trattamento ai sensi dell'art. 6 del GDPR (segnatamente, l'esercizio di un compito di interesse pubblico o legittimo interesse), nonché, eventualmente, una valutazione etica basata sulla legge nazionale.

A ciò si aggiunga che l'EHDS attribuisce compiti di interesse pubblico agli organismi responsabili dell'accesso ai dati sanitari ai sensi dell'art. 6, par. 1, lett. e) del GDPR e, allo stesso tempo, costituisce la base giuridica idonea a soddisfare le condizioni per il trattamento ai sensi dell'art. 9, par. 2, lett. g), h), i) e j) del GDPR. Pertanto, se la domanda presentata dall'utente per accedere ai dati sanitari elettronici si fonda sull'art. 6, par. 1, lett. e) del GDPR, questi dovrà indicare la legislazione nazionale o dell'Unione Europea che rende legittimo il trattamento dei dati sanitari elettronici, mentre, se la richiesta sia basata sull'art. 6, par. 1 lett. f) del GDPR, è lo stesso EHDS a fornire le necessarie garanzie, dal momento che le autorizzazioni all'uso secondario dei dati rilasciate dagli organismi responsabili dell'accesso ai dati sanitari possiedono la natura giuridica di un provvedimento amministrativo che definisce puntualmente le condizioni per accedere ai dati (cfr. Cons. 37).

Come sopra accennato, le domande di accesso sono gestite da un unico organismo di accesso ai dati sanitari. Gli utenti dei dati che desiderano accedere ai dati sanitari elettronici di più di uno Stato membro sono tenuti a presentare un'unica domanda a uno degli organismi di accesso ai dati sanitari di loro scelta. Tuttavia, se un richiedente intenda accedere ai dati sanitari elettronici detenuti da un unico titolare dei dati, può presentare una domanda di accesso ai dati o una richiesta di dati direttamente a tale titolare dei dati.

L'organismo di accesso ai dati sanitari deve rilasciare o rifiutare l'autorizzazione entro due mesi dal ricevimento della domanda di accesso ai dati (prorogabile di ulteriori due mesi). Se l'organismo non prende una decisione entro il termine stabilito, l'autorizzazione viene rilasciata, applicandosi quindi una forma di silenzio - assenso.

L'autorizzazione ai dati stabilisce le condizioni generali applicabili all'utente dei dati, tra cui tipi e formato dei dati sanitari elettronici a cui si è avuto accesso, finalità della messa a disposizione, durata dell'autorizzazione ecc.

Una volta accolta la richiesta, l'organismo di accesso ai dati sanitari richiederà i dati al titolare, che dovrà fornirli all'utente entro due mesi.

Infine, gli utenti hanno l'obbligo di rendere pubblici i risultati o gli output relativi all'uso secondario dei dati sanitari elettronici, comprese le informazioni rilevanti per la fornitura di assistenza sanitaria, entro diciotto mesi dal completamento delle attività di trattamento di tali dati.

3.2.3 Dato sintetico e sanità digitale: prospettive di utilizzo nel panorama normativo italiano

Nell'ambito della sanità digitale italiana, emerge un'inefficienza strutturale attribuibile principalmente a disparità regionali e a una subottimale gestione dei dati personali. Tali deficienze rappresentano un ostacolo allo sviluppo efficace del sistema, nonostante la pressante necessità di avanzamenti digitali. In tale contesto, l'adozione di dati sintetici si profila come una soluzione potenzialmente rivoluzionaria, capace di armonizzare la programmazione sanitaria con la realtà operativa, senza compromettere, in linea di principio, la privacy dei cittadini

I dati sintetici, elaborati mediante sofisticate metodologie di intelligenza artificiale, mantengono le proprietà statistiche dei dati originari, escludendo, però, gli elementi potenzialmente identificativi. Tale processo, che assicura la protezione del dato personale, offre molteplici vantaggi: permette, da una parte, di effettuare analisi approfondite e di formulare politiche sanitarie su una base dati realistica e aggiornata; dall'altra, riduce i rischi connessi alla gestione diretta dei dati sensibili dei pazienti. Tuttavia, l'impiego di dati sintetici richiede un esame attento dal punto di vista etico e normativo per assicurare che l'utilizzo di tali tecnologie sia in linea con i principi di giustizia ed equità[95].

Il panorama normativo relativo all'uso dei dati sintetici nel settore sanitario italiano si inserisce in un contesto più ampio di regolamentazione europea e nazionale, mirato a bilanciare l'innovazione tecnologica con la tutela della privacy e la sicurezza dei dati.

La generazione e l'utilizzo di dati sintetici, pone nuove sfide normative, specialmente in termini di privacy e protezione dei dati. Queste tecnologie, pur promettendo notevoli benefici in ambito sanitario, richiedono un'attenta valutazione dei rischi e dei benefici associati, nonché l'adozione di un quadro legale adeguato che ne regoli l'impiego in maniera efficace e sicura.

Il Professore Pasquale Stanzione, Presidente del Garante per la protezione dei dati personali, ha evidenziato la tensione tra il bisogno di riservatezza e le necessità di ricerca e innovazione nel settore sanitario, enfatizzando la ricerca di un equilibrio nel contesto digitale attuale. La riflessione sull'utilizzo dei dati sanitari si presenta come una moneta a due facce: da una parte, l'impiego di questi dati è fondamentale per il progresso e l'innovazione nel campo della salute;

[95] SCORZA, G. (2024, 5 febbraio). Il diritto a non dover scegliere tra più privacy e più salute. Il Sole 24 Ore.

dall'altra, emerge l'impellente necessità di proteggere i dati sanitari, considerati *"ipersensibili"*[96].

La riflessione del Professore Stanzione enfatizza l'importanza di un'intensa sinergia tra avanzamento tecnologico, regolamentazioni a sostegno della protezione dati e politiche di gestione sanitaria. Il bilanciamento tra necessità contrapposte diventa, quindi, *conditio sine qua non* per assicurare che lo sviluppo del settore sanitario proceda nel rispetto dei diritti e della dignità dei pazienti.

Nel contesto sanitario italiano, l'impiego di dati sintetici si rivela uno strumento prezioso per incrementare la qualità e l'efficienza delle prestazioni sanitarie, assicurando un accesso equo alle innovazioni. Le tecniche di sintetizzazione contribuiscono a superare diversi scenari che hanno un impatto significativo sui modelli di apprendimento automatico, come quello relativo alla carenza di informazioni, specialmente per malattie rare o poco studiate. Al contempo, l'utilizzo di dati sintetici permette di progettare e sviluppare ricerche più ampie e inclusive. A tal riguardo, la generazione di ampi set di dati facilita lo sviluppo di modelli predittivi e diagnostici precisi, abbreviando i tempi di trasferimento delle scoperte scientifiche in ambito clinico e favorendo un approccio più personalizzato e preventivo nella medicina[97].

Sul tema, nel 2020 è stato condotto uno studio per elaborare una metodologia capace di generare immagini TC (Tomografia Computerizzata) legate al COVID-19 attraverso l'uso di reti generative avversarie condizionali (cGAN)[98]. La ricerca di Jiang si poneva l'obiettivo di produrre immagini TC di elevata qualità e realismo, essenziali nell'addestramento delle tecniche di deep learning in ambito di imaging medico. La motivazione principale che ha indotto i ricercatori a integrare l'uso di dati sintetici con quelli reali deriva dall'elevata diffusione e contagiosità del virus, fattori che hanno reso difficile la raccolta di un numero adeguato di dati reali per finalità di *training*. Pertanto, l'impiego di dati sintetici si è rivelato cruciale, compensando la scarsità di informazioni. A ciò si aggiunge che i dati sintetici hanno permesso di mitigare i pericoli per gli operatori sanitari derivanti dalla contrazione del virus. L'indagine ha dimostrato l'efficacia della strategia basata su cGAN rispetto ad altre metodologie di punta nella sintesi d'immagini, dando vita a immagini di eccezionale qualità. Questo studio evidenzia il valore e le possibilità offerte dai dati sintetici in ambito sanitario. Il risultato ottenuto in questo studio, che evidenzia il notevole avanzamento nella generazione di dati sintetici per il rafforzamento dei modelli di apprendimento automatico, rispecchia successi simili raggiunti da altre ricerche condotte durante il periodo della pandemia, sottolineando un interesse e un progresso collettivo nell'utilizzo di dati sintetici[99].

L'adozione di modelli predittivi basati su machine learning, alimentati da dati sintetici, introduce una nuova era di capacità adattive e di calcolo precedentemente inesplorabili. Queste tecnologie non solo potenziano la capacità di previsione e di intervento del sistema sanitario ma aprono anche la strada a una personalizzazione del trattamento medico, elevando il livello

---

[96] Camera dei Deputati - Commissione XII Affari Sociali. (2024, February 13). Audizione informale del Presidente del Garante per la protezione dei dati personali, Prof. Pasquale Stanzione - Esame delle risoluzioni Loizzo n. 7-00183 e Girelli n. 7-00183, in materia di raccolta e utilizzo dei dati sanitari..

[97] MISCHITELLI, L. (2023, Maggio 16). *"Dati sintetici in Sanità: ecco i problemi che possono risolvere"*. Agenda Digitale. https://www.agendadigitale.eu/sanita/dati-sintetici-in-sanita-ecco-i-problemi-che-possono-risolvere/.

[98] JIANG, Y., CHEN, H., LOEW, M., & KO, H. (2020). *"COVID-19 CT image synthesis with a conditional generative adversarial network"* [arXiv:2007.14638].

[99] DAS, H. P., TRAN, R., SINGH, J., YUE, X., TISON, G., SANGIOVANNI-VINCENTELLI, A., & SPANOS, C. J. (2021). *"Conditional synthetic data generation for robust machine learning applications with limited pandemic data"*. arXiv preprint arXiv:2109.06486.

di assistenza e ottimizzando l'allocazione delle risorse sanitarie. Tuttavia, l'utilizzo di dati sintetici non è privo di sfide. La regolamentazione adeguata, la trasparenza nell'uso dei dati e l'accuratezza delle simulazioni sono aspetti critici che richiedono un'attenta valutazione. La costruzione di un quadro normativo solido e la formazione di una cultura digitale informata sono indispensabili per integrare efficacemente i dati sintetici nella sanità digitale, assicurando che il progresso tecnologico si traduca in benefici concreti per tutti i cittadini.

In conclusione, mentre la sanità digitale italiana naviga tra le complessità di una transizione tecnologica e le inerenti disparità territoriali, i dati sintetici rappresentano una leva strategica per catalizzare il cambiamento, promettendo di trasformare le sfide attuali in opportunità per un futuro sanitario più equo e efficiente.

4. Politica del diritto con proposte interpretative ed evolutive

4.1 Sintetizzazione del dato in ambiti diversi da quello di ricerca scientifica

La generazione di dati sintetici rappresenta un processo innovativo mediante il quale si danno vita a informazioni artificiali che conservano le caratteristiche statistiche fondamentali dei dati originali. Tale processo riveste un ruolo cruciale in svariati ambiti di applicazione, inclusi ma non limitati a, settore finanziario, ambito sanitario, e analisi statistica.

Nel contesto aziendale, ad esempio, la generazione di dati sintetici si manifesta come uno strumento strategico per l'analisi di tendenze commerciali, l'ottimizzazione delle strategie di marketing e l'amplificazione della redditività aziendale.

I vantaggi apportati dalla sintesi dei dati si traducono in un aumento dell'efficienza, una maggiore flessibilità e una robusta tutela della privacy.

Nel settore finanziario, i dati sintetici contribuiscono significativamente alla valutazione di rischi e alla scoperta di nuove opportunità di investimento, fornendo agli analisti strumenti avanzati per la gestione del portafoglio. L'utilizzo, infatti, di tale tecnologia rappresenta una risposta efficace a necessità impellenti, sia interne che esterne alle organizzazioni, che permette di affrontare annose restrizioni normative e sfide tecniche con un approccio innovativo. Questo processo consente di superare le limitazioni imposte dalla regolamentazione sulla condivisione dei dati tra diverse aree aziendali. Si rivela inoltre cruciale nella compensazione della mancanza di dati storici per analizzare eventi rari o emergenti, facilitando la creazione di scenari controfattuali per testare strategie e inferenze.

A ciò si aggiunge che, numerosi studi sottolineano l'importanza dell'utilizzo dei dati sintetici nel settore finanziario. Tra questi, la ricerca di Potluru si concentra su applicazioni prototipiche di dati sintetici, esplorando come queste possano rivoluzionare la gestione, l'analisi e la condivisione dei dati finanziari[100].

Con riferimento all'utilizzo dei dati sintetici in ambiti diversi dalla ricerca, occorre prendere in considerazione l'impiego di dati sintetici per finalità statistiche. A questo proposito, l'Office for National Statistics (ONS) del Regno Unito ha esplorato i vari scenari d'impiego dei dati sintetici, sviluppando progetti pilota, tra cui la preparazione per il censimento del 2021 e la generazione di una versione sintetica del sondaggio sull'infezione da Covid, che sfruttano tecnologie come le reti generative avversarie (GAN) e la privacy differenziale. Questi progetti mirano a testare le pipeline analitiche e ingegneristiche, utilizzando dataset come quello del Censimento del Regno Unito e sondaggi sull'infezione da COVID per valutare l'efficacia e la sicurezza nell'uso di dati sintetici[101][102].

Dall'ottimizzazione delle strategie aziendali al rafforzamento dei modelli predittivi nel settore finanziario, fino alla generazione di statistiche per obiettivi sociali, i dati sintetici si affermano come strumento versatile in molteplici scenari differenti.

---

[100] POTLURU, V. K., ET AL. (2023). *"Synthetic Data Applications in Finance"*. J.P. Morgan AI Research.
[101] Office for National Statistics: Trialling the use of synthetic data at the United Kingdom's national statistics institute - UN GWG on Big Data - Privacy Preserving Techniques Wiki - UN Statistics Wiki. (n.d.).
[102] ONS methodology working paper series number 16 - Synthetic data pilot - Office for National Statistics. (n.d.)

4.2 Parere 05/2014 sulle tecniche di anonimizzazione (WP 216): perché dovrebbe cambiare l'approccio da parte delle autorità europee

Il documento del 10 aprile 2014 dal Gruppo di Lavoro ex articolo 29, denominato WP 216, rappresenta un punto di riferimento per la tutela dei dati personali nell'Unione Europea. Il parere in esame offre indicazioni pratiche su come rendere anonimi i dati personali per assicurarsi che le attività di trattamento siano in linea con le normative europee sulla privacy e protezione dei dati, tra cui la Direttiva 95/46/EC e il Regolamento Generale sulla Protezione dei Dati (GDPR).

L'anonimizzazione viene descritta come un processo attraverso il quale i dati personali vengono trasformati in modo tale da impedire l'identificazione della persona a cui si riferiscono. Questo implica una progettazione attenta per eliminare qualsiasi possibilità che i dati, anche indirettamente, possano essere ricondotti all'individuo originale.

A ciò si aggiunge che, il testo in esame chiarisce la differenza tra anonimizzazione e pseudonimizzazione. Anche se la pseudonimizzazione può aumentare la sicurezza dei dati, riducendo le possibilità di collegamento diretto con l'identità della persona, essa non elimina completamente il rischio di identificazione. A differenza dell'anonimizzazione, che mira a rimuovere definitivamente ogni collegamento tra i dati e l'individuo, la pseudonimizzazione lascia aperta la possibilità di ricollegare i dati all'individuo, se disponibili informazioni aggiuntive.

L'applicazione di queste tecniche deve essere considerata attentamente, tenendo conto del contesto specifico, della natura dei dati trattati, delle potenziali minacce alla privacy e delle tecnologie esistenti. Il WP 216 sottolinea l'importanza di adottare un approccio dinamico e aggiornato nella gestione della sicurezza dei dati, in risposta ai cambiamenti tecnologici e alle nuove modalità di attacco. Proteggere la privacy dei dati richiede dunque un impegno costante per adeguare e rivedere le strategie di sicurezza[103].

Tanto premesso, la Sentenza "Deloitte" del Tribunale della Corte di Giustizia dell'Unione Europea (CGUE), come analizzata al precedente paragrafo 2.6, offre un'interpretazione innovativa e potenzialmente trasformativa in merito alla distinzione tra dati anonimi e pseudonimizzati.

Nel caso in esame, la Corte ha affrontato questioni legate alla nozione di dati personali, fornendo chiarimenti su quando un dato può essere considerato anonimo o pseudonimizzato ai sensi del Regolamento Generale sulla Protezione dei Dati (GDPR). Secondo il GDPR, i dati anonimi sono esclusi dal suo ambito di applicazione, mentre i dati pseudonimizzati restano dati personali e quindi soggetti alla regolamentazione.

La sentenza evidenzia l'importanza di valutare la reversibilità del processo di anonimizzazione o pseudonimizzazione, nonché la disponibilità e l'uso di mezzi aggiuntivi che potrebbero consentire l'identificazione dell'interessato. Questo aspetto chiave indica che la valutazione deve considerare non solo la tecnica utilizzata per modificare i dati, ma anche il contesto operativo, inclusa la capacità di accedere a informazioni aggiuntive che potrebbero rendere possibile l'identificazione.

---

[103] Gruppo di Lavoro ex articolo 29 (2014, 10 aprile). Parere 05/2014 sul trattamento dei dati personali nell'ambito delle attività di polizia e di cooperazione giudiziaria in materia penale.

Alla luce delle considerazioni di cui sopra, la sintetizzazione dei dati quale metodologia avanzata per la generazione dei dati sintetici offre ulteriori spunti di riflessione sui temi di pseudonimizzazione e anonimizzazione. Nel merito, i dati sintetici generati attraverso meccanismi avanzati di sintetizzazione, come quelli esaminati nel paragrafo 2.7.5 della presente analisi, riducono la possibilità di re-identificazione ad un livello sufficientemente remoto[104]. Pertanto, secondo quanto indicato dal WP29 nel parere 05/2014 WP216, nella misura in cui i dati sono aggregati a "*un livello in cui i singoli eventi non sono più identificabili si può definire anonimo l'insieme di dati risultante*". Quindi anche i dati sintetici generati in base a tali metodologie avanzate, potrebbero essere qualificati come dati anonimi[105].

Con l'avvento e il perfezionamento delle tecniche di sintetizzazione dei dati, sostenute da avanzate capacità di machine learning e algoritmi di tipo generativo, sono state sviluppate numerose metodologie volte alla tutela dei dati. I dati sintetici rappresentano una soluzione promettente per affrontare le sfide poste dalla necessità di conciliare innovazione tecnologica e tutela della privacy. Questa tecnologia consente la generazione di dati sintetici che, pur mantenendo le stesse caratteristiche statistiche dei dati originali, non comportano rischi per l'identificabilità degli individui dai quali tali dati sono derivati.

I benefici derivanti dall'uso dei dati sintetici sono molteplici e si estendono attraverso vari settori. In particolare, offrono vantaggi significativi in termini di privacy, consentendo allo stesso tempo di sfruttare ampie moli di dati per l'addestramento di algoritmi di machine learning. Questo aspetto è cruciale, ad esempio, per il settore medico, dove la disponibilità di grandi dataset può accelerare la ricerca e lo sviluppo di nuove cure e terapie senza compromettere la riservatezza dei dati dei pazienti.

In conclusione, l'evoluzione normativa e giurisprudenziale nell'ambito della protezione dei dati personali sottolinea l'importanza di adottare un approccio più flessibile e innovativo verso l'anonimizzazione e la pseudonimizzazione. A ciò si aggiunge che, la sintetizzazione dei dati emerge come una soluzione promettente, capace di bilanciare efficacemente l'esigenza di tutela della privacy con le opportunità offerte dall'utilizzo avanzato dei dati. Attraverso tecniche evolute di sintetizzazione, è possibile superare i vincoli imposti da interpretazioni rigide, spianando la strada a nuove modalità di elaborazione dei dati che rispettano la privacy senza ostacolare l'innovazione. Questo orientamento non solo allinea la normativa sulla protezione dei dati alle sfide poste dall'evoluzione tecnologica, ma promuove anche una cultura del dato che considera la privacy non come un ostacolo, ma come un valore aggiunto nell'era digitale.

---

[104] Information Commissioner's Office (ICO). (n.d.). *"Chapter 2: How do we ensure anonymisation is effective?"* https://ico.org.uk/media/about-the-ico/documents/4018606/chapter-2-anonymisation-draft.pdf.

[105] Gruppo di Lavoro ex articolo 29 (2014, 10 aprile). Parere 05/2014 sul trattamento dei dati personali nell'ambito delle attività di polizia e di cooperazione giudiziaria in materia penale.

5. Conclusioni

5.1 La certificazione come strumento per garantire la conformità della sintetizzazione?

Le certificazioni Europee, emesse ai sensi degli artt. 42 e 43 del GDPR, le quali rappresentano una novità storica nell'ambito della protezione dei dati, potrebbero offrire uno strumento efficace per garantire la conformità anche delle tecniche di sintetizzazione dei dati, promuovendo l'uso sicuro e regolamentato dei dati sintetici.

Innanzitutto, occorre ricordare che uno schema di certificazione paneuropeo approvato, include dei criteri di verifica della conformità dei trattamenti posti in essere, riconosciuti in svariati Paesi dell'UE e dello Spazio Economico Europeo, incluse le rispettive autorità di controllo nazionali. Questo riconoscimento esteso favorisce una maggiore uniformità e fiducia nel panorama della protezione dei dati, che potrebbe includere anche i processi di sintetizzazione dei dati certificati.
Di conseguenza, sia i titolari del trattamento (che decidono di sottoporre a sintetizzazione i loro dati), sia i responsabili del trattamento (che sviluppano e forniscono soluzioni di sintetizzazione) possono verificare sistematicamente la conformità dei loro trattamenti, riducendo i rischi associati a eventuali violazioni. Inoltre, i soggetti interessati possono valutare la conformità dei loro fornitori di servizi, garantendo così una maggiore trasparenza. Pertanto, l'uso dei criteri approvati di uno schema di certificazione, consente di documentare la conformità al GDPR in modo strutturato, facilitando il processo di audit e monitoraggio della conformità, inclusa la conformità delle tecniche di sintetizzazione.
La certificazione semplifica la valutazione della conformità e riduce i rischi per i titolari del trattamento. Inoltre, crea un vantaggio competitivo per i responsabili del trattamento (ad esempio i fornitori delle soluzioni di sintetizzazione), i cui servizi sono certificati, poiché l'articolo 28 del GDPR riconosce la certificazione come strumento adeguato per validare le misure tecniche e organizzative adottate dai responsabili del trattamento. Questo è particolarmente rilevante per le tecniche di sintetizzazione, che richiedono standard elevati di sicurezza e protezione dei dati.

La certificazione garantisce che la conformità non sia solo una formalità, ma una realtà effettiva ed è riconosciuta come un mezzo per dimostrare la protezione dei dati fin dalla progettazione e per impostazione predefinita, come previsto dall'articolo 25 del GDPR. Inoltre, in caso di incidenti di sicurezza, come una violazione dei dati personali, la certificazione (ufficialmente riconosciuta) costituisce un fattore mitigante nell'eventualità di sanzioni amministrative.

Pertanto, l'utilizzo della certificazione rappresenterebbe un passo significativo verso una maggiore sicurezza e trasparenza nella gestione dei dati personali. Questo strumento non solo facilita la conformità al GDPR, ma promuove anche l'innovazione e la competitività, offrendo vantaggi tangibili a tutte le parti coinvolte. In particolare, essa offre una soluzione robusta per garantire la conformità anche delle tecniche di sintetizzazione, assicurando che tali tecniche siano utilizzate in modo sicuro e conforme alle normative vigenti.

5.2 Conclusioni: Verso una Sintetizzazione Sicura e Conforme delle Informazioni Personali

Le analisi condotte nel presente paper dimostrano che la sintetizzazione dei dati rappresenta, se realizzata attraverso determinate metodologie e addirittura certificata, una tecnica di anonimizzazione avanzata e conforme ai requisiti normativi attuali e futuri in materia di protezione dei dati personali. La valutazione della singola tecnica di sintetizzazione è necessaria per verificare come questa possa essere utilizzata efficacemente per generare dati artificiali che mantengono le proprietà statistiche dei dati reali, senza comprometterne la privacy degli individui.
Difatti, determinate tecniche di sintetizzazione possono soddisfare i requisiti richiesti per l'anonimizzazione, descritti dal Considerando 26 del GDPR, i quali escludono i dati anonimi dall'ambito di applicazione del Regolamento, ad esempio quelle basate sulla verifica della distanza tra *datapoint* reali e sintetici, riescono a garantire che i dati sintetici generati non consentano la re-identificazione dei dati originari, rispettando quindi i requisiti richiesti dall'anonimizzazione.

Come analizzato nei precedenti paragrafi, anche le linee guida fornite *dall'Information Commissioner's Office* ("ICO") e dalle autorità di protezione dei dati europee, come l'AEPD e l'EDPS, incoraggiano l'uso di tecniche di "anonimizzazione robuste" per la protezione dei dati personali.
L'utilizzo dei dati sintetici non solo preserva la privacy, ma permette anche un'innovazione continua nel campo dell'intelligenza artificiale, adattandosi alle esigenze del futuro contesto normativo delineato dal legislatore europeo.
Pure il regolamento sull'European Health Data Space (EHDS) mira a facilitare l'accesso e la condivisione dei dati sanitari, promuovendo l'uso di tecniche di anonimizzazione per garantire la sicurezza dei dati. La sintetizzazione dei dati, grazie alla sua capacità di generare dati artificiali che mantengono le proprietà statistiche dei dati originali, senza contenere informazioni personali connesse agli interessati, si presenterebbe come una soluzione ideale per soddisfare anche i requisiti del EHDS. In linea generale, la possibilità di utilizzare dati sintetici per scopi di ricerca scientifica, senza compromettere la privacy dei pazienti rappresenta un avanzamento significativo nella protezione dei dati sanitari.

La sintetizzazione dei dati emerge quindi come una *Privacy-Enhancing Technology* ("PET") valida e conforme sia alla normativa attuale che a quella futura in materia di protezione dei dati personali. Le tecniche avanzate di sintetizzazione, descritte nel presente *paper*, dimostrano che i dati sintetici possono essere generati in modo sicuro, preservando la privacy degli individui e garantendo al contempo l'utilità dei dati per l'analisi statistica e il training dei modelli di intelligenza artificiale.
Le capacità dei dati sintetici di mitigare i rischi di re-identificazione sono cruciali in un'epoca in cui le minacce alla privacy sono in costante evoluzione. La possibilità di generare dati che non possono essere ricondotti a individui specifici, offre una soluzione robusta per la protezione delle informazioni personali, riducendo al minimo i rischi associati alla gestione e alla condivisione dei dati. Questo aspetto è particolarmente significativo nel contesto della ricerca medica, dove la protezione dei dati "sensibili" è essenziale per garantire la fiducia dei pazienti e la sicurezza delle informazioni. Inoltre, la sintetizzazione dei dati supporta l'innovazione tecnologica, permettendo alle aziende e ai ricercatori di sviluppare nuovi modelli

e soluzioni basate su dati di alta qualità senza compromettere la privacy degli individui. Questa tecnica facilita la creazione di *dataset* ricchi e diversificati, necessari per l'addestramento di algoritmi di *machine learning* e altre applicazioni avanzate, superando le limitazioni dei dati reali che possono essere di bassa qualità o soggetti a restrizioni di utilizzo.

In conclusione, la sintetizzazione dei dati rappresenta una soluzione preziosa e strategica nel panorama della protezione dei dati. La sua capacità di rispondere ai requisiti normativi, mitigare i rischi di re-identificazione e supportare l'innovazione tecnologica ne fanno una tecnologia di grande valore per il futuro della gestione dei dati personali. Con l'evolversi delle normative e delle esigenze di protezione dei dati, la sintetizzazione si afferma come uno strumento utile per conciliare la necessità di proteggere la privacy con l'obiettivo di promuovere l'innovazione e l'efficienza nell'utilizzo dei dati.

https://www.dataintermediariesalliance.org/

Data Intermediaries Alliance

ISTITUTO ITALIANO PER LA PRIVACY E LA VALORIZZAZIONE DEI DATI