I dati sintetici: panacea della privacy?

di Filiberto E. Brozzetti
Of Counsel, Baker McKenzie Italia
Assistant Professor of AI, Law and Ethics, LUISS Guido Carli

1. Cosa sono, perché tutti ne parlano e come si generano

I dati sintetici costituiscono un'alternativa artificiale ai dati impiegata in diversi settori, inclusi lo sviluppo di algoritmi di *machine learning* e la ricerca in ambiti quali la medicina e la farmaceutica, le diverse scienze sociali, nonché per l'elaborazione di strategie economiche, politiche e militari[1]. A differenza dei dati reali, che derivano da fenomeni (fisici, biologici, sociali, etc.) empiricamente osservati, i dati sintetici sono generati computazionalmente allo scopo di emulare le caratteristiche di tali fenomeni. Il loro utilizzo, in effetti, è particolarmente apprezzabile, allorché ricavare dati reali risulta difficile, costoso o eticamente problematico.

Nel campo del *machine training*, ad esempio, i dati sintetici aiutano a superare limiti pratici, come la scarsità di *labeled data* (i.e. dati opportunamente caratterizzati attraverso *tag* significativi), ma costituiscono una valida opzione per superare anche limitazioni giuridiche, quali le restrizioni all'accesso ed al trattamento di dati reali in ragione della disciplina in materia di protezione dei dati personali. Così, i dati sintetici consentono di simulare scenari diversi per la validazione delle decisioni algoritmiche, riducendo i rischi d'impatto negativo, permettendo di esplorare mondi e sperimentare casi fittizi, scevri peraltro da *bias* di origine umana, e valutare ipotesi innovative per la scienza e la conoscenza umane. Ovviamente, l'efficacia ed affidabilità degli scenari d'uso dei dati sintetici dipende dall'accuratezza delle ipotesi usate nella loro generazione.

La pratica di usare modelli matematici per simulare sistemi fisici o economici con dati basati su distribuzioni di probabilità è consolidata da decenni. La rilevanza attuale dei dati sintetici e l'interesse che suscitano risiedono nella loro pretesa conformità con le norme in materia di protezione dei dati, come il GDPR europeo. I dati sintetici vengono infatti principalmente pensati ed impiegati al fine di ridurre i rischi di esposizione all'impatto per i diritti e le libertà degli individui, inserendosi in un contesto di *accountability* e *data protection* che in tal caso risulterebbe sia *by design* che *by default*. Sono a tutti gli effetti considerati una *Privacy Enhancing Technology* (PET) e permettono di effettuare trattamenti anche molto avanzati di dati in modo sicuro, riducendo o eliminando del tutto la necessità di utilizzare informazioni personali reali.

Le istituzioni competenti nazionali e sovra-nazionali promuovono l'uso dei dati sintetici, specialmente nel campo di sviluppo ed implementazione dell'AI, per la protezione di categorie speciali di dati – ed in particolar modo quelli biometrici e genetici – all'interno delle *regulatory sandbox*[2]. Anche negli Stati Uniti l'Executive Order 14110 sull'uso e sviluppo sicuri ed affidabili dell'Intelligenza Artificiale menziona espressamente e riconosce i dati sintetici come strumento e soluzione per tutelare la privacy degl'individui,

---

[1] K. EL EMAM, L. MOSQUERA, R. HOPTROFF, *Practical Synthetic Data Generation: Balancing Privacy and the Broad Availability of Data*, O'Reilly Media, Sebastopol (CA) 2020.

[2] Cfr. artt. 53, par. 1 e 54, par. 1 della Proposta di Regolamento del Parlamento europeo e del Consiglio che stabilisce regole armonizzate sull'Intelligenza Artificiale (*AI Act*) e ddi moddifica di alcuni atti legislativi dell'Unione (COM/2021/206 final).

specialmente nei settori ad alto rischio[3]. È quindi comprensibile come questo crescente *endorsement* pubblico all'impiego dei dati sintetici stia fungendo da stimolo pet ulteriori e crescenti investimenti e ricerche in tale ambito.

Esistono varie tecniche e metodologie per generare dati sintetici. Uno di questi è l'uso delle regressioni, laddove un *training set* costituito da un insieme di osservazioni viene adattato a una funzione di regressione di più variabili reali e possibilmente non lineare, per generare dati che approssimano quelli osservati. Un altro metodo è quello della massima verosimiglianza (*Maximun Likelihood Extimation* – MLE), che determina i valori dei parametri di una distribuzione di probabilità di più variabili reali basandosi su un criterio di somiglianza coi dati osservati, al fine di massimizzare la probabilità che la distribuzione abbia generato quei dati. Entrambi questi metodi sono tradizionali e producono un sintetizzatore in grado di generare un numero illimitato di nuovi dati sintetici[4].

Più recentemente, sono emersi approcci basati su algoritmi di intelligenza artificiale, in particolare l'uso di reti generative avversarie (*Generative Adversarial Networks* – GAN), in cui una rete neurale generativa (G) ed una rete neurale discriminativa (D) competono fra loro: G crea dati sintetici che D cerca di distinguere dai dati reali. Attraverso questo processo iterativo, G impara a produrre dati sempre più realistici, mentre D incrementa la sua abilità nel riconoscere i dati sintetici. Alla fine, G diventa capace di simulare dati che appaiono naturali e realistici (tanto che D non è più in grado di riconoscerli e distinguerli da quelli reali), utili in vari contesti[5]. Questi metodi costituiscono un significativo passo avanti nelle tecniche di generazione di dati sintetici, offrendo maggiori flessibilità e realismo.

2. Luci ed ombre

I dati sintetici sono diventati una risorsa versatile con applicazioni nei più svariati campi. Nel marketing, i dati sintetici permettono l'elaborazione di profilazioni artificiali per finalità per le quali il trattamento di dati personali originali è limitato da restrizioni normative o *policy* interne. In ambito sanitario, vengono usati per sviluppare e allenare strumenti diagnostici basati sull'AI, tipicamente nell'*imaging* radiologico[6]. Anche la ricerca farmaceutica e genetica beneficia di dati sintetici per l'addestramento di modelli di *machine learning* per l'individuazione di *pattern* significativi, senza compromettere dati sensibili di individui reali[7].

Nel settore automobilistico, sono utili, ad esempio, per l'addestramento di veicoli a guida autonoma in condizioni varie e complesse[8], o per sintetizzare le risultanti di sensori sempre più presenti e sofisticati. Un altro caso d'impiego comune è l'addestramento degli assistenti vocali, allo scopo di migliorare la comprensione del linguaggio naturale. Nel

---

[3] Cfr. in particolar modo la sec. 3, lett. (z), dell'E.O. 14110 on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence, October 30th 2023.

[4] G. D'ACQUISTO, *Synthetic Data and Data Protection Laws*, in corso di pubblicazione.

[5] v. E. LEZMI, J. ROCHE, T. RONCALLI, J. XU, *Improving the Robustness of Trading Strategy Backtesting with Boltzmann Machines and Generative Adversarial Networks*, 2020.

[6] v. fra gli altri R.J. CHEN, M.Y. LU, T.Y. CHEN, D.F.K. WILLIAMSON, F. MAHMOOD, *Synthetic data in machine learning for medicine and healthcare*, in «Nature Biomedical Engineering» 5, 2021.

[7] v. fra gli altri B. OPRISANU, G. GANEV, E. DE CRISTOFARO, *On Utility and Privacy in Synthetic Genomic Data*, presentato a «The Network and Distributed System Security Symposium», 2022.

[8] v. fra gli altri Y. TIAN, K. PEI, S. JANA, B. RAY, *DeepTest: automated testing of deep-neural-network-driven autonomous cars*, presentato all'«*International Conference on Software Engineering*», 2018.

settore assicurativo e finanziario, risultano formidabili nello sviluppo e nel test di modelli di valutazione del rischio e strategie di investimento, simulando condizioni di mercato e scenari economici.

Possiamo ancora menzionare il loro impiego per il *training* di sistemi di rilevamento delle frodi, algoritmi di riconoscimento facciale in diverse condizioni[9], e sistemi di cybersicurezza di reti o, nel campo dello sviluppo informatico, aiutando a creare scenari di test realistici per identificare e risolvere problemi prima del rilascio di software.

Chiaramente, i dati sintetici non possono sostituire universalmente i dati reali, in ragione del rischio persistente di generare falsi positivi e negativi, influenzando decisioni di allocazione delle risorse e creando problemi anche a livello individuale[10]. Ad esempio, nella previsione della propensione all'investimento dei propri correntisti da parte di una banca, l'uso di dati sintetici potrebbe portare a decisioni errate, come lo sviluppo di prodotti di investimento inappropriati in ragione di errori nella simulazione, causando insoddisfazione dei clienti e perdita di fiducia, oltre a inefficienze per la banca sia in termini di risorse che di reputazione. Esempi analoghi possono ipotizzarsi anche nel caso della formulazione di medicinali da parte di una casa farmaceutica, o dell'elaborazione di strategie di marketing basate su profili ottenuti trattando dati sintetici. L'uso impreciso di questi ultimi, pertanto, conduce a decisioni strategiche sbagliate con impatti negativi tutte le parti coinvolte.

Inoltre, l'analisi comparata di dati reali e sintetici mostra come questi non si sovrappongano perfettamente, indicando che le anomalie che si riscontrano nel mondo reale potrebbero non venire rappresentate dai dati sintetici e viceversa. Ciò potrebbe avere implicazioni importanti, in particolare nell'applicazione dei principi di protezione dei dati personali, comportando decisioni discriminatorie o fondate su dati inesatti.

La natura personale o anonima dei dati non è determinata dal modo in cui vengono raccolti o generati. I dati sintetici, benché generati artificialmente, non sono per ciò stesso automaticamente anonimi[11]. Valutare se si è ancora in presenza di dati considerabili come personali richiede un'analisi del rischio basata su identificabilità e impatto diretto o indiretto sugli individui. Un dato è infatti considerato personale non solo qualora identifichi direttamente, ma anche laddove permetta di identificare indirettamente uno specifico individuo all'interno di un gruppo.

Pertanto, nel caso in cui i dati sintetici riproducano fin troppo fedelmente i dati reali, replicando caratteristiche in modo riconoscibile e consentendo d'identificare una persona fisica, o laddove le decisioni basate sui dati sintetici hanno un impatto su individui reali, i dati sintetici devono essere considerati alla stregua di dati personali, potendo insorgere problemi di identificazione o comunque relativi alla protezione dei diritti e delle libertà degli individui, con particolare riferimento alla riservatezza ed alla sicurezza dei dati personali di costoro. Sicché, anche l'impiego di dati sintetici può esigere l'applicazione ed il rispetto delle regole in materia di protezione dei dati personali, specialmente qualora non soddisfino i criteri di anonimizzazione (irreversibile), identificando un individuo specifico, collegando i *record* a persone reali, o facendo inferenze facilmente prevedibili.

---

[9] v. fra gli altri F. BOUTROS, V. STRUC, J. FIÉRREZ, N. SER DAMER, *Synthetic data for face recognition: Current state and future prospects*, in «Image and Vision Computing», 135, 2023.

[10] G. D'ACQUISTO, *Synthetic Data and Data Protection Laws*, cit..

[11] *Ibidem*.

3. Dati sintetici e privacy compliance

Il trattamento dei dati personali pretende l'adempimento di obblighi precisi, inclusa l'individuazione di una base giuridica appropriata per l'ottimizzazione dei parametri di sintetizzazione. Ciò potrebbe anche esigere che il trattamento di dati sintetici sia effettuato comunque per il perseguimento di una finalità che risulti compatibile con la base giuridica originaria, o, da un punto di vista tecnico di sicurezza, una procedura di anonimizzazione effettiva per mitigare i rischi di identificazione individuale. Diventano inoltre cruciali, nel caso di generazione di dati sintetici a partire da dati reali, la garanzia di trasparenza ed il rispetto dei diritti degli interessati, ed in particolar modo il diritto di opposizione e il diritto a non essere oggetto di decisioni basate soltanto su trattamenti automatizzati, soprattutto quando i dati sintetici replicano con un alto grado di verosimiglianza i dati personali o influenzano processi decisionali, automatizzati o meno, che possono avere un impatto rilevante sulla vita degli individui.

Per garantire un'opposizione (e quindi un *opt-out*) efficace, si possono esplorare due opzioni: la c.d. *differential privacy* (o anche *randomization*), che aggiunge "rumore" ai dati personali reali da sintetizzare, rendendoli anonimi già al momento della raccolta, minimizzando l'impatto e garantendo così la riservatezza degli individui interessati da questo trattamento[12]; o il *machine unlearning*, che permette un *opt-out* selettivo *ex-post*, riflettendo la capacità umana di dimenticare e conferendo all'AI la capacità di rimuovere o ridurre informazioni specifiche dal proprio addestramento[13]. Mentre la prima garantisce che l'aggiunta o la rimozione di dati non influenzi significativamente il risultato e richiede quindi cautela per non compromettere la qualità dei dati sintetici, il secondo, d'altra parte, offre una maggiore accuratezza, utilizzando dati reali, ma presenta particolari problematiche nel determinare l'influenza specifica di ciascuna informazione sul processo di ottimizzazione sintetica.

Per concludere, da tutto quanto fin qui esposto, in maniera estremamente sintetica, si evince come il dibattito sull'uso dei dati sintetici supera di misura la disamina squisitamente tecnica, ponendo una questione giuridica sostanziale circa la possibilità di impiegare un surrogato artificiale nelle analisi computazionali, basate sui dati personali. Benché la riproduzione fedele dei dati reali non implichi necessariamente una violazione della *privacy*, la stocasticità intrinseca in ogni processo di generazione artificiale di dati può indurre ad identificazioni individuali non intenzionali, o anche discriminazioni algoritmiche. Tuttavia, allo stesso modo, soppesare tali fenomeni unicamente dal punto di vista giuridico potrebbe non costituire l'approccio normativo più virtuoso, col rischio di creare un ambiente regolamentare eccessivamente prudente. È fondamentale quindi bilanciare l'ontologica stocasticità dell'AI con l'*accountability* deontologica nel trattamento dei dati personali[14]. Le norme in materia di protezione dati e le *Privacy Enhancing Technology* (PET) offrono congiuntamente strumenti più che validi per affrontare i rischi per la riservatezza degl'individui, ma una semplificazione eccessiva nel trattare i dati sintetici come dati non

---

[12] v. G. D'ACQUISTO, *Big Data e Privacy by Design. Anonimizzazione Pseudonimizzazione Sicurezza*, Giappichelli, Torino 2017, p. 67 ss.

[13] v. fra gli altri T. SHAIK, X. TAO, H. XIE, L. LI, X. ZHU, Q. LI, *Exploring the Landscape of Machine Unlearning: A Comprehensive Survey and Taxonomy*, 2023.

[14] Cfr. G. D'ACQUISTO, *Synthetic Data and Data Protection Laws*, cit..

personali potrebbe portare a incertezze giuridiche esiziali, ostacolando un profilo d'innovazione salubre nel settore della *data analysis*, un aspetto sempre più cruciale delle strategie digitali governative globali.