Dati sintetici: quando l'intelligenza artificiale apprende da se stessa…

di Valentina Cavani

Abstract: *Synthetic data: when artificial intelligence learns from itself …* - Data is the lifeblood of modern artificial intelligence. Although we are living in the age of Big Data, data for training intelligent systems is lacking. For this reason, the idea was to create an artificial version of it. Synthetic data is not a new idea, but it has never been as relevant as it is in the current technological scenario, where it is often presented as the game-changer for the inherent trade-off between data utility and data protection. The purpose of the following reflections is to verify whether this balance actually holds.

Keywords: Synthetic data; Data protection; GDPR; Hyperreality; Artificial intelligence

1. La "scorciatoia"

In principio si pensava che le macchine, per diventare intelligenti, dovessero pensare in modo umano.

I primi sistemi di intelligenza artificiale venivano definiti "Sistemi Esperti" ed erano progettati per ragionare su una base di "conoscenze".

Dopo anni di entusiasmo e tante promesse, ci si scontrò, però, con un problema fondamentale: l'ambiguità del mondo reale. Le regole da fornire ai Sistemi Esperti non erano mai abbastanza, divenivano sempre più complesse e comportavano sempre più eccezioni.

Fu a quel punto che i ricercatori decisero di smettere di tentare di inquadrare la complessità del mondo reale in regole e modelli e di fare uso del migliore alleato che abbiamo per comprenderla: i dati.

Sostituire la comprensione teorica con relazioni statistiche ottenute da grandi quantità di dati fu la prima, vera, "scorciatoia"[1] intrapresa lungo la strada verso la produzione di comportamento intelligente.

I dati sono la linfa vitale della moderna intelligenza artificiale. Dalla manipolazione del linguaggio agli onnipresenti sistemi di raccomandazione, oggi tutte le principali applicazioni dell'AI sono rese possibili da pattern imparati dai dati.

Un chiaro problema che si crea sostituendo le teorie con dei dati, naturalmente, è quello di trovare i dati necessari, un compito che potrebbe essere tanto costoso quanto creare la teoria stessa.

[1] Si prende, qui, in prestito il titolo del libro di N. Cristianini, *La Scorciatoia*, Bologna, 2023.

DPCE online
ISSN: 2037-6677

Ottenere dati di qualità è la parte più importante e allo stesso tempo più impegnativa della creazione di un'intelligenza artificiale "intelligente".

Le problematiche legate alla raccolta e alla preparazione dei dati sono molteplici. Eppure, possono essere risolte dalla stessa intelligenza artificiale, che, oggi, è in grado di creare "da sé" i dati necessari al proprio addestramento, attraverso un processo che prende il nome di "sintesi".

Nel corso del presente scritto si cercherà di analizzare l'impatto dei c.d. dati sintetici sulla privacy degli individui e, soprattutto, sul loro rapporto con la realtà che li circonda.

2. I dati scarseggiano

Prima di entrare nel vivo dell'argomento che occuperà le prossime riflessioni (i dati sintetici), vale la pena considerare il motivo per cui c'è interesse a produrre ancora più dati, dal momento che – sentiamo sempre dire… – viviamo in un'abbondanza o sovrabbondanza di dati senza precedenti[2]. In breve, dal punto di vista dell'apprendimento automatico e della ricerca informatica, non c'è un eccesso, ma piuttosto una terribile mancanza di dati. La raccolta di dati di qualità dal mondo reale è complicata, costosa e richiede tempo. Gli informatici lamentano una «carenza di dati»[3] e una «scarsità di dati»[4] e affermano che «many problems of modern AI come down to insufficient data»[5].

I pessimisti prevedono che esauriremo i dati testuali "freschi" nel 2050 e quelli relativi alle immagini nel 2060[6].

Ci sono tre cause della carenza di dati citate frequentemente sia nella letteratura tecnica che in quella commerciale.

La prima deriva dal fatto che i dati per l'apprendimento supervisionato, l'approccio più popolare all'apprendimento automatico, devono essere etichettati manualmente. Ciò richiede molto lavoro e rende la creazione di set di dati etichettati difficile e il relativo acquisto costoso[7].

La seconda causa citata della carenza di dati è che per alcune applicazioni semplicemente non esistono dati sufficienti. Siamo portati a pensare ai dati raccolti dalle grandi piattaforme digitali, che ogni giorno

---

[2] Cfr. M. Andrejevic, *Infoglut: How Too Much Information Is Changing the Way We Think and Know*, New York, 2013.

[3] Cfr. Q. Yang, *GDPR, Data shortage and AI*, 2019, https://www.swissre.com/dam/jcr:2edc3963-3226-4ce4-aee6-8496c0247b63/Smart_resilience_Qiang%20Yang_ENGLISH.pdf.

[4] Cfr. A. Bansal, R. Sharma, M. Kathuria, *A systematic review on data scarcity problem in deep learning: solution and applications*, in *ACM Computing Surveys*, 54(10s), 2022; R. Babbar, B. Schölkopf, *Data scarcity, robustness and extreme multi-label classification*, in *Machine Learning*, 108(8), 2019, 1329–1351.

[5] S. Nikolenko, *Synthetic Data for Deep Learning*, Cham, 2021, 12.

[6] P. Villalobos, J. Sevilla, L. Heim, T. Besiroglu, M. Hobbhahn, A. Ho, *Will we run out of data? An analysis of the limits of scaling datasets in machine learning*, 2022, https://arxiv.org/abs/2211.04325.

[7] Cfr. A. Shtylenko, *The Advantages of Synthetic Data*, in *LinkedIn*, 23 novembre 2022, secondo il quale la procedura di *labelling* «is often costly, generally time-consuming, and error-prone». «Manual data annotation can also introduce bias to the data as annotators might make errors in judgment while performing data labelling».

catturano milioni di interazioni, ma per molte potenziali applicazioni dell'apprendimento automatico ci sono, in realtà, pochissimi dati.

La terza causa riguarda l'accessibilità dei dati. In questo caso i dati esistono, ma sono resi inaccessibili da fattori sociali, tecnici, o normativi[8].

3. Da dove provengono i dati. Dati, persone e privacy

Se la qualità dei dati è importante, la provenienza è fondamentale.

Ma da dove provengono i dati?

Le prime ricerche critiche sui dati hanno evidenziato come gli stessi non "esistano" semplicemente, completamente formati, in natura: devono essere creati. Per Manovich, «i creatori di dati devono raccogliere dati e organizzarli o crearli da zero»[9]. Per Bowker, «"raw data" è sia un ossimoro che una cattiva idea». Piuttosto, i dati devono essere «cucinati» o preparati[10].

Anche le successive ricerche rifiutano la nozione di dati grezzi[11]; tuttavia, abbandonano il riferimento suggestivo di Manovich alla creazione di dati da zero.

I dati sono ormai riconosciuti come il prodotto di sensori che raccolgono le tracce delle azioni umane; una copia del mondo, non qualcosa che può essere creato da zero. Per Sadowski, i dati sono una «astrazione registrata del mondo», gran parte della quale riguarda «le persone»[12]. Per Gregory, i dati «sono fatti di persone» o, più precisamente, «degli stessi ritmi, circolazioni, palpitazioni e mutazioni dei nostri corpi»[13]. Per Lemov, i big data «sono persone» perché «non sono solo generati sugli individui, ma sono anche costituiti da individui»[14]. I dati possiedono una «umanità intrinseca», poiché si basano sull'«estrazione quasi letterale della soggettività»[15] o sull'estrazione dell'«esperienza umana come materia prima gratuita»[16].

Queste premesse sono state trasferite negli studi critici sull'intelligenza artificiale, dove si sostiene che «i dati utilizzati oggi nel ML, e in particolare nel DL, sono fin troppo umani»[17] e che «l'intelligenza

[8] K. El Emam, *Could synthetic data be the future of data sharing?*, in *CPO Magazine*, 2021.
[9] L. Manovich, *The Language of New Media*, Cambridge, MA, 2021, 224.
[10] G.C. Bowker, *Memory Practices in the Sciences*, Cambridge, MA, 2005, 184.
[11] L. Gitelman (ed.), *"Raw Data" Is an Oxymoron*, Cambridge, MA, 2013.
[12] J. Sadowski, *When data is capital: datafication, accumulation, and extraction*, in *Big Data & Society*, 6(1), 2019, rispettivamente 2 e 6.
[13] K. Gregory, *Big Data, Like Soylent Green, Is Made of People*, Digital Labor Working Group, 2019.
[14] R. Lemov, *Why big data is actually small, personal and very human*, in *Aeon*, 2016.
[15] *Ivi.*
[16] S. Zuboff, *The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power*, London, 2019, 5.
[17] J. Roberge, M. Castelle, *Toward an end-to-end sociology of 21st-century machine learning*, in Eadem (eds.), *The Cultural Life of Machine Learning*, Cham, 2021, 10.

DPCE online
ISSN: 2037-6677

artificiale dovrebbe essere di proprietà del popolo, perché è prevalentemente "nutrita" dal popolo»[18].

Se i dati *ineriscono* alle persone tanto da *appartenervi*, si crea una fin troppo logica frizione tra il loro libero utilizzo e la tutela di quello che è ormai considerato un diritto fondamentale degli individui, il diritto alla privacy, nella sua specifica accezione di diritto alla protezione dei dati personali.

Nella loro essenza, le normative in materia di privacy «limitano il potere che le informazioni umane conferiscono»[19] e, per tale motivo, rappresentano uno dei più rilevanti ostacoli alla libera utilizzabilità dei dati e uno dei principali motivi di preoccupazione per gli sviluppatori delle tecnologie di AI che devono reperire materiale per l'addestramento dei propri sistemi.

4. Dati personali e dati anonimi

Ai sensi del Regolamento (UE) 2016/679 (General Data Protection Regulation, "GDPR"), i dati che riguardano persone fisiche, identificate o identificabili, sono "dati personali" e, in quanto tali, il loro utilizzo è soggetto alle regole e ai limiti previsti dal Regolamento stesso.

Per non ricadere nell'ambito materiale della normativa UE, i dati devono essere "non personali" o, secondo la definizione del GDPR, "anonimi".

Una delle principali tecniche per rendere "anonimi" i dati è proprio l'"anonimizzazione", che consiste nell'eliminare dal dato tutti gli identificatori che potrebbero individuare in modo univoco la persona cui il dato si riferisce.

Il problema dell'anonimizzazione è che interferisce con l'utilità dei dati. In parole povere, si crea un muro attorno ai dati ma, così facendo, li si rende ingombranti e poco pratici per l'intelligenza artificiale complessa o lo sviluppo di software.

Inoltre, il progresso tecnologico ha reso molto difficile garantire l'assoluta irreversibilità dei dati ottenuti attraverso l'anonimizzazione. Dalla disfatta delle *query* di ricerca di AOL alla vicenda del dataset di Netflix[20], è divenuto sempre più semplice, anche per i meno esperti, "unire" le informazioni ausiliarie con una serie di informazioni "perturbate"[21] e svelare proprio i dati che l'anonimizzazione intendeva proteggere.

---

[18] A. Dippel, *Metaphors we live by*, in A. Sudmann (ed.), *The Democratization of Artificial Intelligence: Net Politics in the Era of Learning Algorithms*, Bielefeld: transcript, 2021, 40.

[19] Cfr. N. Richards, *Why Privacy Matters*, Oxford, 2022.

[20] Per una ricostruzione delle vicende si veda: M. Barbaro, T. Zeller Jr., *A Face Is Exposed for AOL Searcher No. 4417749*, in *The New York Times*, 9 agosto 2006; A. Pierre Pallotta, *La de-anonimizzazione dei dati personali. Il caso del dataset Netflix*, in *ICT Security Magazine*, 24 febbraio 2021.

[21] Il termine è usato e spiegato da S.M. Bellovin, P.K. Dutta, N. Reitinger, *Privacy and Synthetic Datasets*, in *22 Stan. Tech. L. Rev*, 1, 2019, 4, nt. 9.

L'anonimizzazione, insomma, sebbene sia ancora *una* soluzione per tutelare la privacy degli individui, non è più *la* soluzione migliore[22].

5. La sintesi dei dati

Preso atto del «fallimento»[23] dell'anonimizzazione come tecnica di de-identificazione, si è posto il problema di reperire diversamente i dati per l'addestramento dell'intelligenza artificiale.

Come spesso accade, quando una materia prima scarseggia, gli esseri umani rispondono creandone una versione artificiale.

I dati sintetici sono una vecchia tecnica di de-identificazione che ha recentemente subìto un cambiamento epocale in termini di funzionalità e ambito di applicazione.

Nella primavera del 1993, un professore di statistica di Harvard, Donald Rubin, scrisse un articolo che avrebbe cambiato il modo in cui l'intelligenza artificiale viene pensata e fatta funzionare[24]. Tuttavia, il suo obiettivo dichiarato era più modesto: analizzare i dati del censimento statunitense del 1990, preservando l'anonimato dei suoi intervistati.

Non era possibile anonimizzare semplicemente i dati, perché gli individui potevano ancora essere identificati tramite il loro indirizzo di casa, il numero di telefono o il numero di previdenza sociale; tutti elementi cruciali per le analisi che i colleghi di Rubin volevano eseguire. Per risolvere il problema, Rubin ha generato una serie di risposte al censimento anonime le cui statistiche sulla popolazione rispecchiavano quelle del set di dati originale. In questo modo, i colleghi di Rubin hanno potuto trarre valide inferenze statistiche sulla struttura degli Stati Uniti senza compromettere l'identità dei loro cittadini.

La soluzione di Rubin era originale e rivoluzionaria. Aveva prodotto dati "sintetici", contribuendo così a introdurre il termine nel lessico tecnico e accademico[25].

Come molte altre "invenzioni", la tecnica di Rubin sfrutta oggi i progressi dell'intelligenza artificiale e dell'apprendimento automatico per aumentare le proprie capacità di elaborazione e analisi. Gli strumenti di AI generativa hanno consentito ai dati sintetici di compiere significativi miglioramenti nell'affrontare la sfida relativa alla de-identificazione.

In generale, il termine 'sintesi' descrive un atto di combinazione; riunire elementi per generare un nuovo insieme; un «processo che, mediante

---

[22] Cfr. A. Narayanan, E.W. Felton, *No Silver Bullet: De-Identification Still Doesn't Work*, 2014, https://perma.cc/X6PZ-X9EP.

[23] Così, C.A. Trovato, C. Rauccio, *L'anonimizzazione è morta? Un'analisi dei dati sintetici come proposta per superare la dicotomia "dato personale-non personale"*, in *Ciberspazio e diritto*, 2, 2022. Cfr. anche E.A. Brasher, *Addressing the failure of anonymization: guidance from the European Union's General Data Protection Regulation*, in *Columbia Business Law Review*, 1, 2018, 209-253; P. Ohm, *Broken Promises of Privacy: Responding to the Surprising Failure of Anonymization*, in *57 Ucla L. Rev.*, 2010.

[24] D.B. Rubin, *Discussion Statistical Disclosure Limitation*, in *Journal of Official Statistics*, 9(2), 1993, 461-468.

[25] Nello scritto, Rubin descrive i dati sintetici come dati «constructed using multiple imputation so that they can be validly analyzed using standard statistical software».

DPCE online
ISSN: 2037-6677

l'intervento umano, emula determinate proprietà di un materiale presente in natura»[26].

Gli esseri umani si avvalgono della "sintesi" per creare artificialmente qualcosa che già esiste in natura. Anche se la gomma naturale, ricavata dal lattice dell'albero Hevea brasiliensis, rimane un componente degli pneumatici per automobili, è stata sostituita dalla gomma sintetica, ottenuta mediante sintesi chimica di prodotti petroliferi. La gomma sintetica emula il lattice naturale e rende l'Hevea brasiliensis sempre meno indispensabile per gli pneumatici.

Anche i dati sintetici fanno parte di questo processo di riproduzione, di "simulazione" del mondo reale. Si tratta, infatti, di dati che non sono traccia, copia o registrazione di fatti storici, ma il prodotto di un processo computazionale[27]. Essi sono descritti come «dati creati artificialmente anziché generati da eventi reali»[28].

I dati sintetici pretendono, quindi, di attenuare la connessione tra dati e persone.

Sebbene, come detto, tutti i dati siano in una certa misura sintetici poiché non si trovano completamente formati in natura, il fenomeno contemporaneo dei dati sintetici si differenzia proprio per la sua (più o meno marcata) disconnessione dal mondo reale.

Nonostante i dati sintetici siano lungi dal sostituire quelli convenzionali e lungi dall'eliminare completamente la componente umana dalla produzione dei dati, sono anche lontani dall'essere meramente teorici.

I dati sintetici per l'apprendimento automatico sono stati considerati una delle «10 migliori tecnologie rivoluzionarie» del 2022 dalla MIT Technology Review[29]. Nel luglio 2021, Gartner, una società di ricerca e consulenza, ha scritto un pezzo provocatoriamente intitolato «Forget About Your Real Data. Synthetic Data Is the Future of AI»[30].

È, quindi, fondamentale esaminare le implicazioni che gli algoritmi di apprendimento automatico addestrati su dati sintetici potranno avere sulla vita degli individui.

5.1 Come avviene la sintesi dei dati

Da un punto di vista tecnologico, il processo di generazione dei dati sintetici (appunto, la "sintesi") può essere eseguito utilizzando diverse tecniche che

[26] T.E. Raghunathan, *Synthetic data*, in *Annual Review of Statistics and Its Applications*, 8, 2021, 131.

[27] Data generated *by* algorithms and *for* algorithms: B.N. Jacobsen, *Machine learning and the politics of synthetic data*, in *Big Data & Society*, 10(1), 2023.

[28] C. Dilemgani, *Synthetic data generation: techniques, best practices & tools*, 14 febbraio 2024, https://research.aimultiple.com/synthetic-data-generation/.

[29] W.D. Heaven, *Our weird behavior during the pandemic is messing with AI models*, in *MIT Technology Review*, 22 febbraio 2022.

[30] https://www.gartner.com/en/documents/4002912#:~:text=Synthetic%20data%20is%20often%20seen,AI%20models%20without%20synthetic%20data.

sfruttano l'intelligenza artificiale e il "machine learning", compresi gli algoritmi di "deep learning"[31].

Più nel dettaglio, il processo di sintesi si compone di due, fondamentali, elementi.

Il primo. Un data set di informazioni personali. Si tratta dei dati originali e reali, di cui i dati sintetici riproducono le proprietà statistiche. Il data set di partenza può includere identificatori indiretti, quali, ad esempio, sesso, razza, figli, stato di fumatore, orientamento sessuale, o identificatori diretti, quali, ad esempio, i dati anagrafici, l'immagine del viso, il profilo genetico.

Il secondo. Un sistema di intelligenza artificiale generativa, composto da un algoritmo utilizzato per generare dati sintetici artificiali partendo dal data set di dati originali di cui sopra.

A questi elementi si deve necessariamente aggiungere anche un sistema di controllo finale sui dati generati, in grado di garantire che i dati sintetici risultanti non siano dati personali effettivi (ovvero appartenenti a persone realmente esistenti).

6. Peculiarità e vantaggi dei dati sintetici

Alla luce di quanto detto, risultano evidenti le peculiarità del processo di sintesi, rispetto al processo di anonimizzazione dei dati: con quest'ultima, infatti, si rende più difficile o si impedisce l'identificazione dell'interessato, mentre con i dati sintetici si estrae valore da un determinato set di informazioni personali, creando una nuova generazione di dati che non risultano riconducibili ad alcun soggetto interessato realmente esistente[32].

I vantaggi dei dati sintetici sono innumerevoli.

Innanzitutto, rispetto ai dati storici, i dati sintetici sono più economici e veloci da raccogliere, dal momento che non necessitano di essere etichettati a mano[33].

I dati sintetici sono, poi, particolarmente utili quando i set di dati sono difficili da reperire. Prendiamo ad esempio il settore automobilistico. Attraverso set di dati sintetici, i produttori possono imitare il comportamento del conducente nelle simulazioni virtuali per addestrare i loro modelli in una vasta e più ricca serie di situazioni e rendere, così, le auto più sicure[34]. In un altro esempio importante, Amazon ha utilizzato dati sintetici per addestrare Alexa, il suo assistente digitale, ad applicare il

[31] Per una descrizione completa dei diversi metodi di generazione dei dati sintetici cfr. K. El Eman, L. Mosquera, R. Hoptroff, *Practical Synthesic Data Generation*, Sebastopol, 2020.

[32] S. Shaked, *Synthetic Test Data Vs. Data Masking: What Are the Main Differences?*, 22 marzo 2022, https://www.datomize.com/synthetic-test-data-vs-data-masking-what-are-the-main-differences/.

[33] Un imprenditore ha stimato che «[a] single image that could cost [six dollars] from a labeling service can be artificially generated for six cents»: cfr. G. Andrews, *What Is Synthetic Data?*, 8 giugno 2021, https://blogs.nvidia.com/blog/2021/06/08/what-is-synthetic-data.

[34] Cfr. J. Tordable, *Synthetic Data Creates Real Results*, in *Forbes*, 26 agosto 2020.

DPCE online
ISSN: 2037-6677

riconoscimento vocale in hindi, spagnolo statunitense e portoghese brasiliano, lingue per le quali non vi erano sufficienti dati raccolti[35].

Non da ultimo, le aziende sanitarie stanno sempre più utilizzando i dati sintetici per testare casi medici per i quali non esistono dati sufficienti[36].

I dati sintetici hanno anche il potenziale per correggere alcune delle evidenti incoerenze e pregiudizi nei nostri attuali set di dati. Secondo Gartner, ad esempio, circa l'85% degli algoritmi attualmente in uso sono soggetti a errori, in gran parte a causa di pregiudizi, spesso dovuti alla sottorappresentazione nel campione di dati di donne, persone di colore o altri gruppi minoritari. Con i dati sintetici, gli ingegneri possono aumentare artificialmente il numero di minoranze sottorappresentate all'interno di un set di dati, semplicemente generando nuove caratteristiche sintetiche rappresentative del gruppo minoritario in questione.

I dati sintetici, infine, "democratizzano"[37] il campo dell'AI: riducendo le barriere di accesso ai dati, gli stessi possono aumentare la concorrenza tra i fornitori e facilitare l'innovazione basata sui dati[38].

7. Svantaggi e limiti dei dati sintetici

Così presentati i dati sintetici sono un concetto elegantemente semplice, una di quelle idee che sembrano quasi troppo belle per essere vere.

E, in effetti, non è tutto oro quello che luccica.

Anche con la loro capacità di minimizzare le distorsioni storiche conosciute, sarebbe un errore pensare che i dati sintetici siano privi di distorsioni. I dati senza pregiudizi sono generalmente un'illusione: le persone prendono decisioni su quali dati includere, escludere e come analizzarli, e tali scelte si basano su ciò che è ritenuto importante o rilevante, che di solito è parziale. Questo continua a essere il caso anche quando si tratta di prendere decisioni sui set di dati sintetici. Gli ingegneri generano questi dati sulla base di un campione più piccolo di dati reali etichettati con tutti gli aspetti ritenuti rilevanti per l'addestramento dell'intelligenza artificiale e una serie di regole che cercano di contrastare eventuali pregiudizi evidenti e noti nel set di dati originale. Ma il punto centrale del pregiudizio è che tutti ne soffriamo e spesso non riusciamo a vederlo da soli. E nella realtà ci sono più complessità e sfumature di quanto saremo mai in grado di riflettere e tenere conto sistematicamente nei set di dati sintetici. Finché saranno gli esseri umani a decidere quali di questi set di dati dovrebbero essere costruiti, quali problemi dovrebbero risolvere e quali dati

[35] J. Slifka, *Tools for Generating Synthetic Data Helped Bootstrap Alexa's New-Language Releases*, in *Amazon Science Blog*, 11 ottobre 2019.

[36] Cfr. M. Giuffrè, D.L. Shung, *Harnessing the power of synthetic data in healthcare: innovation, application, and privacy*, in *npj Digit. Med*, 2023.

[37] L'espressione è utilizzata e spiegata da P. Lee, *Synthetic Data and the Future of AI*, in *110 Cornell Law Review*, 2024.

[38] Cfr. D.L. Rubinfeld, M.S. Gal, *Access Barriers to Big Data*, in *59 Ariz. L. Rev.*, 2017. Secondo Toews, «The net effect of the rise of synthetic data will be to empower a whole new generation of AI upstarts and unleash a wave of AI innovation by lowering the data barriers to building AI-first products»: R. Toews, *Synthetic Data is About to Transform Artificial Intelligence*, in *Forbes*, 12 giugno 2022.

del mondo reale dovrebbero essere la loro base, non saremo mai in grado di rimuovere completamente i pregiudizi. E come tali, i dati sintetici possono riprodurre modelli e distorsioni dai dati da cui vengono ricavati e persino amplificarli[39].

Un'ulteriore preoccupazione riguarda il mercato dei dati sintetici. Come detto, è prevedibile che gli stessi "democratizzeranno" l'accesso all'innovazione, ma non può escludersi il rischio opposto, ovvero che i dati sintetici conducano a un rafforzamento del potere degli operatori storici del settore, i quali possiedono un impareggiabile volume di dati reali da sottoporre al processo di sintesi.

Per di più, aumentando potenzialmente l'accuratezza delle informazioni in possesso dei fornitori sui consumatori/utenti, si aprono maggiori opportunità di utilizzo vantaggioso, ma anche di sfruttamento, manipolazione e abuso.

Ultimo, ma non per importanza, il rischio di re-identificazione. Così come evidenziato per le tecniche di anonimizzazione, anche nel caso della sintesi artificiale di dati il rischio di re-identificazione degli interessati permane, anche se in una forma diversa.

Valutazioni empiriche dimostrano che alcuni strumenti di sintesi producono dati che sono preoccupantemente vicini ai dati di origine[40]. Se il modello generativo apprende le proprietà statistiche dei dati di origine in modo troppo accurato o troppo esatto, ovvero se si "adatta eccessivamente" ai dati, i dati sintetici replicheranno semplicemente i dati di origine, facilitando la re-identificazione. Inoltre, anche nel caso in cui il modello generativo non soffra di "overfitting", la replicazione dei record potrebbe comunque avvenire per caso, anche se con una probabilità inferiore[41].

8. Dati sintetici e normative privacy

Come visto, a seconda del contesto, i dati sintetici possono contribuire ad alleviare o a rafforzare le sfide legate alla *governance* dei dati.

Alla luce delle caratteristiche appena descritte, occorre domandarsi se i dati sintetici siano effettivamente esclusi dal rispetto delle normative in materia di privacy.

Secondo alcuni, i dati sintetici sono dati anonimi e, come tali, non coperti dalle leggi sulla privacy. In un rapporto commissionato da *Mostly AI*,

[39] Si veda, ad esempio, N. Jain, A. Olmo, S. Sengupta, L. Manikonda, S. Kambhampati, *Imperfect ImaGANation: Implications of GANs exacerbating bias on facial data augmentation and snapchat face lens*, in *Artificial Intelligence*, 304, 2022.

[40] M. Hittmeir, A. Ekelhart, R. Mayer, *Utility and Privacy Assessments of Synthetic Data for Regression Tasks*, in *2019 IEEE International Conference on Big Data (Big Data)*, 2019, 5763-72; Eadem, *On the Utility of Synthetic Data: An Empirical Evaluation on Machine Learning Tasks*, in *ARES '19: Proceedings of the 14th International Conference on Availability, Reliability and Security*, 29, 2019, 1-6.

[41] Come osserva un dirigente di Gartner, «[i]f you are creating data for a rural area and it's one person per [one hundred] miles, even though I can create a synthetic person, it doesn't hide anything»: cfr. *The Executive's Guide to Accelerating Artificial Intelligence and Data Innovation with Synthetic Data*, Briefing Paper, in *Harvard Business Review*, 6, 2021.

DPCE online
ISSN: 2037-6677

un fornitore di dati sintetici, e pubblicato da *Harvard Business Review Analytics Services*, si legge: «With […] emerging privacy regulations around the world making the sharing of personal information so complicated, if not impossible, synthetic data is vital to support collaboration. As it is fully anonymous, it is exempt from these rules»[42].

Secondo altri, tali affermazioni categoriche dovrebbero essere respinte[43].

Come per le altre tecniche di anonimizzazione, per stabilire l'assoggettabilità o meno dei dati sintetici alle leggi sulla privacy, occorrerebbe valutare il rischio di re-identificazione dell'interessato che gli stessi comportano. La possibilità di dedurre informazioni personali da insiemi di dati sintetici comporterebbe, infatti, la qualificazione degli stessi come dati personali, soggetti ai vincoli in tema di riservatezza.

A ben vedere, tale affermazione si rivela problematica per diversi motivi. In primo luogo, alcuni esperti sostengono che esista *sempre* un rischio residuo di re-identificazione, in tutti i tipi di dati sintetici[44], e, pertanto, gli stessi dovrebbero essere soggetti di *default* alle leggi sulla privacy. In secondo luogo, trattandosi appunto di un "rischio", lo stesso potrebbe anche non concretizzarsi e le informazioni contenute nel set di dati potrebbero rimanere anonime, creando così una notevole incertezza giuridica e un inutile aggravio normativo per coloro che elaborano set di dati sintetici.

Al fine di superare le questioni discusse sopra, è stato proposto di integrare le normative sulla protezione dei dati con una regola *de minimis* per la re-identificazione, un livello accettabile di rischio al di sotto del quale le normative stesse non si applicherebbero[45].

Anche tale strategia, tuttavia, non è priva di complicazioni. La sua attuazione potrebbe comportare difficoltà pratiche per i responsabili del trattamento dei dati sintetici, i quali avrebbero bisogno di metodi chiari per stabilire i livelli di rischio ed evitare di incorrere in sanzioni.

Ancora più drastica è la proposta di superare la distinzione binaria tra dati personali e dati anonimi[46], come nel caso dell'EU Data Act (Regolamento 2023/2854). In tale ottica, i dati sono sempre più intrecciati

---

[42] *The Executive's Guide to Accelerating Artificial Intelligence and Data Innovation with Synthetic Data*, cit. Analogamente, la ricerca sostenuta da Nvidia suggerisce che «synthetic data would not be considered identifiable personal data, privacy regulations would not apply, and obligations of additional consent to use the data for secondary purposes would not be required»: K. El Emam, *Accelerating AI with Synthetic Data: Generating Data for AI Projects*, https://www.nvidia.com/content/dam/en-zz/Solutions/deep-learning/resources/accelerating-ai-with-synthetic-data-ebook/accelerating-ai-with-synthetic-data-nvidia_web.pdf.

[43] M.S. Gal, O. Lynskey, *Synthetic Data: Legal Implications of the Data-Generation Revolution*, in *Iowa Law Review*, 109, 2023.

[44] T. Stadler, B. Oprisanu, C. Troncoso, *Synthetic data – anonymisation Groundhog Day*, in *Proceedings of the 31st USENIX security symposium*, 2019.

[45] S.M. Bellovin, P.K. Dutta, N. Reitinger, *Privacy and Synthetic Datasets*, cit., 50.

[46] Cfr. A. Beduschi, *Synthetic data protection: Towards a paradigm change in data regulation?*, in *Big Data & Society*, 2024; N. Purtova, *The Law of Everything: Broad Concept of Personal Data and Future of EU Data Protection Law*, in *Law, Innovation & Tech.*, 10, 2018.

in set grandi e complessi e ciò renderebbe la loro classificazione difficile o addirittura ridondante.

Anche l'idea di superare l'attuale dualismo dei dati è stata oggetto di forti critiche, in quanto potrebbe creare confusione e diminuire l'efficacia delle leggi sulla privacy, come il GDPR[47].

Alla luce di queste riflessioni, non sorprende che le autorità di protezione dei dati abbiano posizioni moderate quando si tratta di dati sintetici.

In un post sul blog del novembre 2023[48], l'Agencia Española de Protección de Datos ha chiarito che, pur considerando i dati sintetici uno «strumento potente», è comunque necessario rispettare le disposizioni normative del GDPR, dal momento che «la creazione di dati sintetici a partire da dati personali reali costituisce essa stessa un'attività di trattamento» ai sensi del Regolamento.

In un documento di ricerca del giugno 2022[49], l'Ufficio del Commissario per l'informazione del Regno Unito ha sottolineato che per ridurre il rischio di re-identificazione da un set di dati sintetici, le organizzazioni che intendono farne uso dovrebbero continuare ad allinearsi ai principi del GDPR in tema di minimizzazione dei dati e limitazione delle finalità, includendo nei loro set «solo le proprietà necessarie per soddisfare il loro caso d'uso specifico e nient'altro».

Infine, anche il Garante europeo raccomanda ai fornitori di dati sintetici di eseguire, prima della loro generazione, un «privacy assurance assessment» al fine di garantire che i risultati non siano dati personali effettivi[50].

Anche l'utilizzo di dati sintetici, pertanto, non autorizza gli operatori all'esonero dalla normale *due diligence* quanto si tratta di un diritto tanto fondamentale come quello alla privacy.

9. Dati sintetici tra Europa e Stati Uniti

Pur con i limiti appena descritti, non si può non concludere che i dati sintetici siano, oggi, considerati una delle migliori strategie in termini di tutela delle informazioni personali.

Da una parte all'altra dell'Oceano, i dati sintetici sono accreditati dai principali sforzi normativi in tema di regolamentazione, generale, dell'intelligenza artificiale.

Nello specifico, l'AI Act – il Regolamento europeo sull'intelligenza artificiale, attualmente allo stadio finale della sua approvazione – valida i dati sintetici come la prima soluzione da adottare per rilevare e correggere le

[47] B. Da Rosa Lazarotto, G. Malgieri, *The Data Act: a (slippery) third way beyond personal/non-personal data dualism?*, in *European Law Blog*, 4 maggio 2023.

[48] https://www.aepd.es/en/prensa-y-comunicacion/blog/synthetic-data-and-data-protection.

[49] https://ico.org.uk/media/for-organisations/documents/4025484/sythetic-data-roundtable-202306.pdf.

[50] https://www.edps.europa.eu/press-publications/publications/techsonar/synthetic-data_en.

DPCE online
ISSN: 2037-6677

distorsioni presenti nei set di dati di addestramento dell'AI (art. 10, par. 5). Solo nel caso in cui simili dati non siano disponibili, il Regolamento autorizza il fornitore a trattare "categorie particolari di dati personali", i c.d. dati sensibili.

L'utilità dei dati sintetici per la tutela della privacy non è sfuggita nemmeno all'attenzione del governo degli Stati Uniti. Nell'Executive Order (EO) del presidente Biden (emanato il 30 ottobre 2023) in tema di «Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence», gli strumenti di generazione dei dati sintetici (synthetic-data-generation tools) sono ricompresi tra le c.d. tecnologie di miglioramento della privacy ("Privacy-Enhancing Technologies" (PETs)) ed espressamente qualificati come una "tecnologia di tutela della privacy"[51].

10. Dati *realmente* sintetici

A questo punto occorre introdurre una distinzione linguistica, che ha implicazioni non solo formali ma del tutto sostanziali.

I dati sintetici di cui si è parlato sino a questo momento sono "sintetici" solo nel senso che sono stati "sintetizzati" da dati storici. Hanno una risoluzione inferiore, ma non sono generati da una fonte artificiale.

Da questo punto in poi utilizzeremo l'espressione "dati realmente sintetici" per sottolineare la provenienza interamente ed esclusivamente artificiale dei dati di cui andremo a parlare. In questa prospettiva, i dati sono sintetici non perché estrapolati da dati convenzionali esistenti, ma perché raccolti in un mondo sintetico. È, come dice L. Floridi, «una distinzione ontologica, che può avere importanti implicazioni in termini epistemologici»[52].

Se i dati sintetizzati attenuano la connessione tra uomo e dato, i dati realmente sintetici fanno sparire qualsiasi componente umana dal processo di generazione dell'informazione.

Strati compositi di mediazione tecnologica – generazione procedurale, ambienti simulati, apprendimento automatico – possono emulare l'essere umano come fonte di dati.

A quanto pare, quindi, i dati non rappresentano necessariamente le persone.

Per comprendere il concetto è utile riportare l'esempio di L. Floridi relativo al gioco degli scacchi[53].

---

[51] In esecuzione dell'Ordine, il 15 dicembre 2023, il Federal Chief Data Officers Council ha pubblicato una richiesta di informazioni per stabilire le migliori pratiche in tema di generazione dei dati sintetici. Secondo il Consiglio, «la capacità di generare e utilizzare dati sintetici rappresenterebbe un punto di svolta nell'uso da parte del dipartimento di tecnologie complesse e in rapida evoluzione per soddisfare la sua missione critica proteggendo al tempo stesso la privacy»: https://public-inspection.federalregister.gov/2024-00036.pdf.

[52] L. Floridi, *Etica dell'intelligenza artificiale. Sviluppi, opportunità, sfide*, Milano, Raffaello Cortina Editore, 2022, 70. Sulle sfide epistemologiche legale agli algoritmi cfr. R. Kitchin, *Big Data, new epistemologies and paradigm shifts*, in *Big Data & Society*, 1(1), 2014; I. Lowrie, *Algorithmic rationality: Epistemology and efficiency in the data sciences*, in *Big Data & Society*, 4(1), 2017.

[53] *Ibidem.*

In passato, giocare a scacchi contro un computer significava giocare contro i migliori giocatori umani che avessero mai preso parte al gioco. Perciò, una delle caratteristiche di Deep Blue, il programma di scacchi della IBM che aveva sconfitto il campione del mondo Garry Kasparov, consisteva in un uso efficace di un database delle partite di un grande maestro. Ma AlphaZero, l'ultima versione del sistema di AI sviluppato da DeepMind, ha imparato a giocare meglio di chiunque altro, e in effetti di qualsiasi altro software, facendo affidamento soltanto sulle regole del gioco, senza alcun input di dati da alcuna fonte esterna. Non aveva alcuna memoria storica: AlphaZero ha imparato giocando contro se stesso, generando così i propri dati sintetici relativi agli scacchi.

I dati realmente sintetici hanno alcune straordinarie proprietà.

Innanzitutto, condividono quelle già descritte per i dati sintetizzati: sono veloci, economici, durevoli, riutilizzabili, rapidamente trasportabili, facilmente duplicabili, simultaneamente condivisibili senza fine, ecc. Sono anche puliti e affidabili (in termini di accuratezza), non violano privacy o riservatezza nella fase di sviluppo, se vengono persi non è un disastro perché possono essere ricreati e sono perfettamente formattati per essere utilizzati dal sistema che li genera.

Oltre a queste caratteristiche, i dati realmente sintetici hanno una straordinaria qualità: con gli stessi, l'intelligenza artificiale non è mai costretta ad abbandonare il suo spazio digitale, dove può esercitare il controllo completo su qualsiasi input e output. In termini più epistemologici, con i dati sintetici «l'AI gode della posizione privilegiata della conoscenza del costruttore, che conosce la natura intrinseca e il funzionamento di qualcosa perché lo ha costruito»[54].

In questa prospettiva, i dati realmente sintetici potrebbero, addirittura, ampliare il mondo reale. Mentre i sistemi di AI allenati con dati sintetizzati si limitano a "imparare dagli esempi"[55], l'utilizzo di dati totalmente sintetici consente di costruire gli esempi secondo necessità. Sotto questo aspetto, i dati sintetici migliorano i dati convenzionali in quanto hanno una portata più ampia: sono «senza cornice»[56] e, pertanto, possono potenzialmente rappresentare qualsiasi cosa, almeno tutto ciò che può essere simulato.

11. Le problematiche dei dati realmente sintetici

Anche i dati realmente sintetici non sono privi di problematiche.

La prima ha natura squisitamente tecnica.

Alcuni ricercatori[57] ritengono, infatti, che la proliferazione di questi contenuti potrebbe creare un cortocircuito definito "MAD" (Model Autophagy Disorder), un processo "autofagico" per cui i modelli che si nutrono delle informazioni che hanno creato finiscono per collassare.

---

[54] *Ivi*, 72.

[55] F. Offert, *Latent deep space: Generative Adversarial Networks (GANs) in the sciences*, in *Media+Environment*, 3(2), 2021, 16.

[56] M. Andrejevic, *Automated Media*, New York, Routledge, 2020, 106.

[57] S. Alemohammad, J. Casco-Rodriguez, L. Luz, A. Imtiaz Humayun, H. Babaei, D. LeJeune, A. Siahkoohi, R.G. Baraniuk, *Self-Consuming Generative Models Go MAD*, 4 giugno 2023, https://arxiv.org/pdf/2307.01850.pdf.

DPCE online
ISSN: 2037-6677

Ciò è dovuto a una sorta di consanguineità tra i dati in input e quelli di output[58], che porta a risultati sempre più distorti, insipidi e complessivamente scadenti. Addestrando un sistema di AI, ripetutamente, con contenuti sintetici, le informazioni marginali e meno rappresentate inizieranno a scomparire. Il modello inizierà quindi a basarsi su dati sempre più convergenti e meno vari e questo lo porterà a sgretolarsi su se stesso.

In altre parole, senza "dati reali freschi" per nutrire l'AI, possiamo aspettarci che i suoi risultati ne risentano drasticamente[59].

Al di là degli aspetti tecnici, vi sono preoccupazioni molto più profonde che accompagnano la proliferazione dei contenuti interamente sintetici.

Un set di dati completamente generato da un algoritmo va contro tutto ciò che sappiamo sull'empirismo, in particolare sul fatto che i fatti comprovati costituiscono il nucleo della scienza, del processo decisionale e della logica stessa.

Sostituire i dati reali con quelli sintetici diventerà sempre più una tentazione. Ma poiché una percentuale maggiore di informazioni è generata dal computer, c'è il rischio che più decisioni siano prese dalla visione di un algoritmo. Ciò significa che siamo sull'orlo di un mondo in cui molte delle tecnologie che ci circondano potrebbero non essere costruite in risposta alla realtà, ma a ciò che una macchina immagina che quella realtà sia.

E questo è pericoloso, perché anche il miglior set di dati sintetici non sarà mai una rappresentazione esatta della realtà in cui viviamo, estremamente varia e in costante cambiamento.

Se il set di dati di addestramento non è fondato su (o, forse, non è costituito da) una comprensione rigorosa dei più recenti fenomeni umani sottostanti, come le differenze tra ciò che le persone dicono e fanno, o l'influenza inaspettata di variabili tangenziali nelle nostre vite e nelle azioni che compiamo, il sistema di AI che lo utilizza rischia di alterare la realtà in modi che potrebbero essere non solo assurdi ma anche dannosi.

Si è detto, è vero, che uno dei presupposti alla base della generazione e dell'utilizzo dei dati sintetici è che la c.d. "variabilità" possa essere generata sinteticamente. In altri termini, se la diversità manca, gli algoritmi possono crearla.

---

[58] «Ci piace l'analogia con il morbo della mucca pazza: nutrire mucche con altre mucche giovani, in un ciclo che si ripete e che porta ad agenti patogeni che distruggono il cervello», ha dichiarato il professor Richard G. Baraniuk, tra gli autori dello studio citato. Jonathan Sadowski, data researcher della Monash University di Melbourne, ha utilizzato invece la curiosa espressione «AI asburgica», riferendosi all'antica famiglia austriaca che praticava il matrimonio tra parenti piuttosto stretti. Lo stesso Sadowski parla di «un sistema che viene addestrato così pesantemente sui risultati di altre AI generative da diventare un mutante consanguineo [...] con caratteristiche esagerate e grottesche»: https://twitter.com/jathansadowski/status/1625245803211272194.

[59] Sul punto: M. van Rijmenam, *The danger of AI model collapse: when LLMS are trained on synthetic data*, in *The Digital Speaker*, 28 giugno 2023, https://www.thedigitalspeaker.com/danger-of-ai-model-collapse-llms-trained-synthetic-data/; R. Groh, *AI's Achilles Heel: Confronting the Threat of Model Autophagy Disorder (MAD)*, 6 dicembre 2023, https://medium.com/@empa.consulting/ais-achilles-heel-confronting-the-threat-of-model-autophagy-disorder-mad-9d1f9262d79e; M. Harrison Dupré, *AI Loses Its Mind After Being Trained on AI-Generated Data*, 7 dicembre 2023, https://futurism.com/ai-trained-ai-generated-data.

Si tratta, tuttavia, di un presupposto denso di implicazioni. Per comprenderne alcune, consideriamo, ad esempio, la strategia di marketing della società Datagen, che fornisce dati sintetici per l'addestramento di diversi modelli di AI, generando cataloghi di quelle che vengono definite «identità umane uniche». Sul sito dell'azienda si legge: «I nostri set di dati possono essere personalizzati per includere qualsiasi variazione di cui hai bisogno, dall'età, al sesso, alla razza e alla massa corporea fino alle imperfezioni, alla struttura della pelle e ad altri casi limite»[60].

L'intelligenza artificiale, quindi, può creare qualsiasi diversità "di cui hai bisogno".

Ciò solleva alcuni interrogativi: quali tipi di diversità possono essere generate? E come saranno utilizzate?

Queste domande acquisiscono ulteriore profondità se si considera la promessa di aziende come Datagen di generare tutta la diversità necessaria in termini di identità umana. Dovremmo, pertanto, domandarci quali nuove modalità e tecniche di razzializzazione e profilazione siano rese possibili dai dati sintetici per algoritmi di apprendimento automatico[61].

12. Oltre l'iperrealtà

Un'altra peculiarità dei dati sintetici porta a una riflessione ancora più "essenziale".

Al centro del fascino dei dati sintetici c'è l'evocazione dell'artificialità o della finzione, il fatto che le informazioni "non appartengano" a nessuna persona specifica, che siano completamente fittizie.

Questa nozione di dati sintetici come forma di "non appartenenza" è quindi un aspetto centrale della logica emergente del rischio associato a tale tipologia di dati: rischio che, in questa prospettiva, sarebbe talmente ridotto da proiettare i dati sintetici "oltre" l'ambito stesso del rischio algoritmico[62].

Tuttavia, affinché i dati sintetici possano essere utilizzati come valida sostituzione o aggiunta ai dati reali, devono essere resi "somiglianti" ai dati reali.

Cosa si intende per somiglianza in questo contesto?

Ovviamente, l'idea ha un significato fluttuante nella misura in cui differisce a seconda del dominio, del compito o del problema per il quale i dati sintetici vengono generati e utilizzati. Ad esempio, in contesti come i software di riconoscimento facciale, è fondamentale che le immagini sintetiche siano altamente fotorealistiche e catturino le varie sfumature, contorni e complessità del volto umano. In altri contesti, come il rilevamento delle frodi con carte di credito, è importante che i dati sintetici abbiano le stesse proprietà e caratteristiche statistiche dei dati reali, poiché gli informatici desiderano sviluppare modelli che imparino a rilevare

[60] https://datagen.tech/.

[61] Cfr. R. Amaro, *Threshold Value. E-Flux Architecture*, 2020, disponibile online: https://www.e-flux.com/architecture/education/322664/threshold-value/; T. Phan, S. Wark, *What personalisation can do for you! or: How to do racial discrimination without 'race'?*, in *Culture Machine*, 20, 2021, 1–29.

[62] B.N. Jacobsen, *Machine learning and the politics of synthetic data*, in *Big Data & Society*, 10(1), 2023.

DPCE online
ISSN: 2037-6677

irregolarità e anomalie nei flussi di input di dati reali. Insomma, affinché i dati sintetici come volti, impronte digitali, iridi o texture della pelle siano utili, devono differire sufficientemente dal set di dati del mondo reale per essere sintetici, ma non differire così tanto da impedire che gli algoritmi vengano effettivamente utilizzati nel mondo reale. In altre parole, un certo divario deve esserci, ma distanze e prossimità devono comunque essere gestite con attenzione.

Questa interazione tra realtà e simulazione richiama per forza alla mente le riflessioni del sociologo e filosofo francese Jean Baudrillard.

Nella sua opera fondamentale, *Simulacra and Simulation*, egli ha introdotto il concetto di "simulacri", una nozione che ha avuto profonde implicazioni sulla nostra comprensione della realtà e della rappresentazione. Secondo Baudrillard, nel mondo postmoderno, la realtà è stata sostituita da simboli e segni e che hanno perso ogni connessione con il mondo reale che rappresentano.

Un "simulacro" è una copia senza originale, una rappresentazione senza referente nella realtà. Non si tratta di una semplice imitazione, ma piuttosto di una perversione della realtà, una distorsione che crea una propria "iperrealtà", che finisce per sostituirsi alla realtà che imita.

Man mano che si approfondisce il mondo della "sintesi" artificiale, emerge un affascinante collegamento con i simulacri di Baudrillard. Anzi, direi di più: un'estensione del concetto. I dati sintetici, a loro volta prodotti da dati sintetici, possono essere concepiti come la copia del simulacro, una eco dell'eco. Non sono una copia del "vero", ma una copia di una copia senza originale. Questo fenomeno introduce un altro livello di separazione dalla realtà, trascinandoci ulteriormente nell'iperreale.

Si tratta di una situazione pericolosa, specie se iniziamo a contemplare gli usi più nefasti dei dati sintetici, come i *deepfake* o la disinformazione su vasta scala.

Non è, quindi, affatto vero che se i dati non ci appartengono non sono rischiosi.

13. Alcune riflessioni conclusive

È in corso una rivoluzione nella generazione dei dati. Fino a poco tempo fa, la maggior parte dei dati utilizzati per il processo decisionale algoritmico erano raccolti da eventi che avevano luogo nel mondo fisico. Tuttavia, si prevede che, entro il 2024, il 60% dei dati utilizzati per addestrare i sistemi di AI in tutto il mondo sarà sintetico[63].

Questa rivoluzione nella generazione dei dati ci impone di rivalutare e potenzialmente ristrutturare il nostro regime di *governance* dei dati.

Come visto, i dati sintetici accentuano le sfide esistenti in merito all'efficacia delle leggi sulla privacy, mettendo in discussione le loro motivazioni e i presupposti su cui si basano.

Nonostante il fatto che nessun individuo reale sia incluso in un rilascio di dati, almeno per quanto riguarda i dati completamente sintetici, gli individui sintetici e quelli reali rimangono comunque collegati dalle

[63] A. White, *By 2024, 60% of the data used for the development of AI and analytics projects will be synthetically generated*, in *Gartner Blog*, 2021.

informazioni che trasmettono. Se un aggressore è in grado di recuperare informazioni corrette su individui reali, non importa che tali informazioni siano basate su dati simulati. Anche se tale divulgazione non rientra nell'ambito della legislazione sulla privacy, può comunque essere considerata non etica nella misura in cui riguarda individui reali.

Per questo motivo, dovrebbero esserci linee guida più chiare per tutti i tipi di dati sintetici, compresi quelli che si qualificano come dati non personali.

Queste linee guida dovrebbero riflettere i principi di trasparenza, responsabilità ed equità. La trasparenza richiederebbe, ad esempio, che i dati sintetici siano chiaramente etichettati come tali e che le informazioni sulla loro generazione siano fornite agli utenti. La responsabilità comporterebbe la definizione di procedure chiare per chiamare a rispondere i responsabili della generazione e del trattamento dei dati sintetici. L'equità dovrebbe includere garanzie che i dati sintetici non vengano generati e utilizzati in modi che provochino effetti negativi sugli individui e sulla società, come perpetuando pregiudizi esistenti o creandone di nuovi.

Inoltre, abbiamo visto come – specie in relazione ai dati completamente sintetici – rimanga aperta una «vulnerabilità epistemologica»[64] che non può non essere affrontata.

I dati sintetici si prefiggono di sostituire il reale con il realistico, senza tuttavia che sia stato elaborato (e nell'incertezza sulla possibilità stessa di farlo) un metodo condiviso per arrivare a convalidare tale nozione.

Sembra potersi concludere che i dati sintetici siano sì una valida alternativa ai dati storici, ma non una panacea.

Per evitare di creare danni alla società, si auspica che i progressi della sintesi artificiale siano accompagnati da una sana politica, in grado di problematizzare maggiormente il rapporto tra appartenenza e rischio connesso ai dati.

È un percorso pieno di tensione e contingenza, perché il processo stesso di generazione dei dati sintetici si basa su una sorta di "spazio di gioco", una continua negoziazione socio-tecnica tra distanza e prossimità, tra ciò che è "troppo reale" e ciò che "non è abbastanza reale".

Per non perdere contatto con la realtà bisognerà, forse, abbracciare una visione del mondo in cui la qualità, il contesto e l'origine dei dati contino più della quantità.

Valentina Cavani
Dipartimento di Giurisprudenza
Università di Modena e Reggio Emilia
valentina.cavani@unimore.it

[64] Così, C. Accoto, *Il mondo in sintesi*, Milano, 2022, 58.

DPCE online
ISSN: 2037-6677