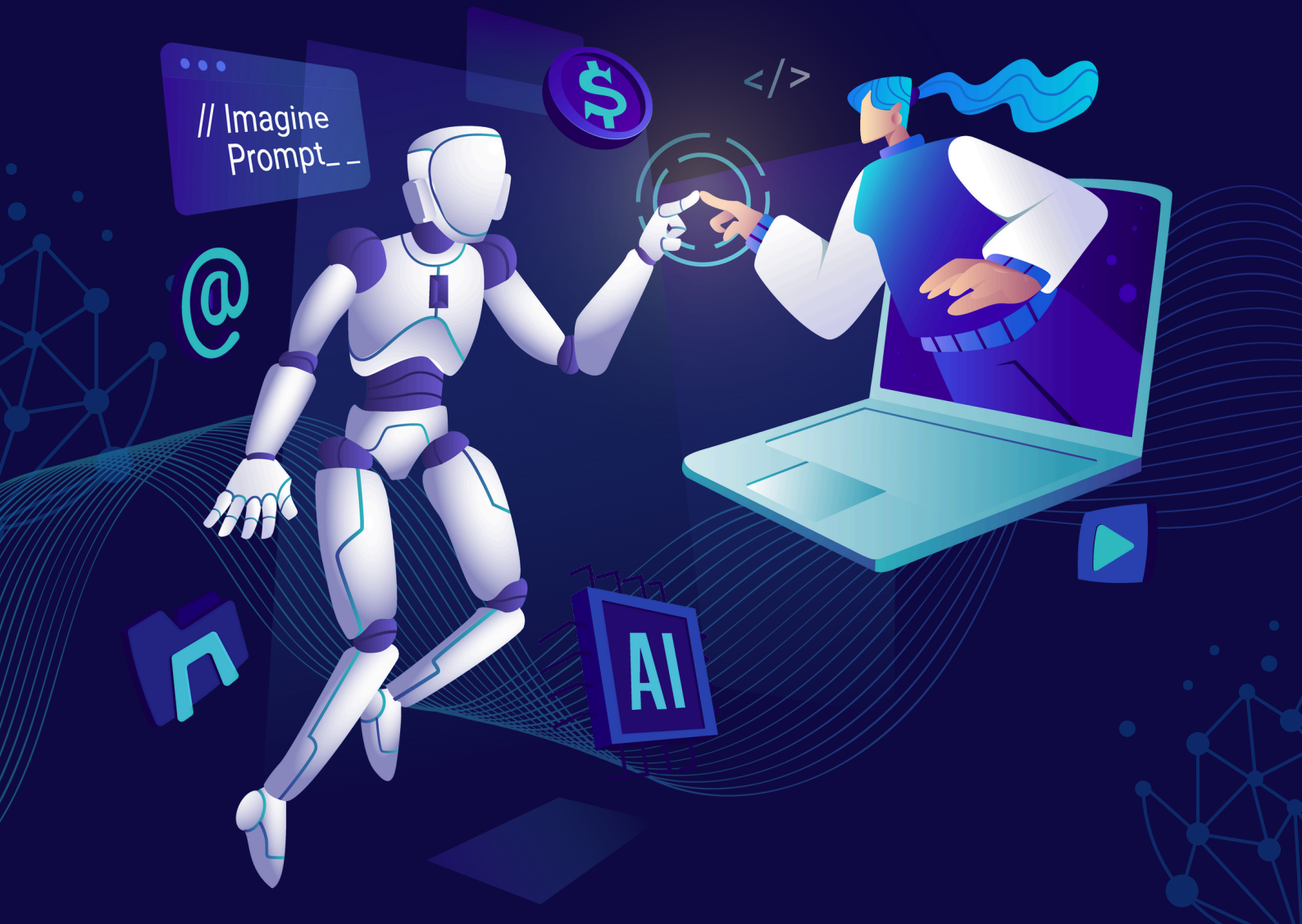
AWARE

Dicembre 2024

IA GENERATIVA:

dati sintetici tra opportunità strategiche e sfide normative



AWARE

AUTORI

Sofia Brunelli

Alessandro Brusadelli

Ginevra Denei

Martina Lasco

CURATORI

Marco Baticci

SI RINGRAZIANO

Alessia Sementilli

Alessandra Ajraldi

AWARE

CHI SIAMO

AWARE è un Think Tank nato nell'**ottobre del 2019** che si pone l'obiettivo ultimo di contribuire alla definizione del futuro che verrà stimolando la **riflessione** e il **dibattito** sulle nuove sfide del mondo contemporaneo.

Per farlo sono stati individuati due macro argomenti che meglio rappresentano le chiavi per leggere e guidare il futuro del nostro pianeta:

- l'**evoluzione delle tecnologie e del settore digitale nelle sue declinazioni**;

- lo **sviluppo sostenibile**, basato su tre pilastri (sociale, economico ed ambientale).

INDICE

I dati sintetici: verso uno sviluppo etico e responsabile dell'Intelligenza artificiale

L'Intelligenza artificiale sta rapidamente trasformando le società contemporanee in riferimento al modo in cui queste pensano e agiscono, facilitando la vita quotidiana ed il lavoro, ma sta anche facendo emergere importanti questioni etiche che attirano sempre più l'attenzione di esperti e opinione pubblica. Ormai il tema non è più tanto l'eventualità dell'impatto, ma piuttosto le conseguenze e la loro gestione.

La riflessione etica è comunemente applicata ad ambiti specifici di applicazione dell'intelligenza artificiale, come ad esempio quello sanitario o quello del lavoro e, negli ultimi anni (complici i repentini sviluppi del settore) è diventato sempre più frequente sentir parlare di **etica dell'Intelligenza artificiale**, intesa come quella branca dell'etica che studia le implicazioni economiche, sociali e culturali dello sviluppo e della diffusione di sistemi di Intelligenza artificiale, cercando di promuovere lo sviluppo di sistemi intelligenti che rispettino valori morali condivisi.

In linea con questo obiettivo, negli anni più recenti diversi autori hanno cercato di individuare dei principi etici per l'adozione di un'IA socialmente vantaggiosa e responsabile. Secondo **Floridi**[1], dal 2017 in poi, cioè dopo la pubblicazione dei Principi di Asilomar per l'IA e della Dichiarazione di Montréal per uno sviluppo responsabile dell'IA, la pratica di delineare dei principi etici per lo sviluppo dell'IA è diventata "artigianale", arrivando già nel 2020 ad oltre 160 principi proposti. Per questa ragione, il filosofo e professore ordinario di filosofia ed etica dell'informazione all'Università di Oxford e all'Università di Bologna, ha svolto un'importante analisi comparativa di diversi trattati e dei molti principi etici individuati e ha evidenziato come tra essi si possa individuare una convergenza sui 4 principi fondamentali già alla base della bioetica.

A questi quattro principi, Floridi ne ha poi aggiunto un quinto, rendendo **i principi fondamentali dell'Intelligenza Artificiale** i 5 riportati di seguito:

1. Floridi, L. (2022). Etica dell'intelligenza artificiale: Sviluppi, opportunità, sfide. Raffaello Cortina Editore.

1 BENEFICIENZA

Lo sviluppo dell'Intelligenza artificiale dovrebbe promuovere il benessere di tutte le creature senzienti e del pianeta.

2 NON MALEFICENZA

Lo sviluppo dell'Intelligenza artificiale dovrebbe avvenire prevenendo il verificarsi di danni intenzionali (legati ad un uso malevolo della tecnologia da parte dell'uomo) o imprevisti e non intenzionali (dovuti alla macchina).

3 AUTONOMIA

Lo sviluppo dell'IA dovrebbe promuovere l'autonomia di tutti gli esseri umani, non compromettendo la libertà degli esseri umani di stabilire propri standard e norme. La sfida è quella di trovare un equilibrio tra il potere decisionale che l'uomo mantiene e quello che delega ai sistemi di IA. In ogni caso, l'autonomia delle macchine dovrebbe essere sempre reversibile, così che l'uomo possa tornare sulla sua decisione di quanto e cosa delegare.

4 GIUSTIZIA

Intesa come promuovere la prosperità, preservare la solidarietà ed evitare l'iniquità. Rispondere a questo principio vuol dire promuovere lo sviluppo dell'IA a sostegno di nuove competenze mitigando allo stesso tempo l'impatto su quelle obsolete. In altri termini, implica promuovere chi sarà avvantaggiato domani senza lasciare indietro chi è svantaggiato oggi.

5 ESPLICABILITÀ

Floridi spiega che il funzionamento dell'Intelligenza Artificiale è spesso invisibile o incomprensibile ai più. Ciò è particolarmente vero per quegli approcci di IA come il deep learning in cui il risultato dell'esecuzione dell'algoritmo non è riconducibile in modo lineare all'output iniziale. Per questo motivo, il principio di esplicabilità, che include sia il senso di intelligibilità (come risposta alla domanda "Come funziona?"), sia il senso etico di responsabilità (come risposta alla domanda "Chi è responsabile del modo in cui funziona?") è il pezzo mancante del puzzle etico dell'IA[2].

In ogni caso, è bene ricordare che mitigare i rischi legati allo sviluppo di sistemi di IA è un'azione che deve procedere parallelamente all'incentivazione di tale sviluppo che, come è noto, si fonda largamente su una "materia prima" fondamentale, ovvero i dati necessari all'addestramento degli algoritmi.

2. Ibidem

Proprio grazie alle applicazioni in ambito AI, il mercato dei dati è infatti in forte crescita: la Commissione europea stima che, entro il 2025, la quantità globale di dati aumenterà del 530%. Nello stesso arco temporale, il valore dell'economia dei dati all'interno dell'Unione dovrebbe salire da 301 miliardi di euro nel 2018 a 829 miliardi di euro.
È comune, se non addirittura la norma, che il progresso tecnologico avanzi più rapidamente rispetto alle istituzioni, alle amministrazioni e ai settori industriali. Così oggi, mentre la capacità di aziende, persone e PA di sfruttare efficacemente i dati è ancora limitata da ostacoli tecnologici, regolamentari e culturali e mentre le istituzioni cercano di disegnare quadri normativi che possano tenere il passo dell'innovazione, le più importanti aziende tecnologiche del mondo stanno già investendo ingenti risorse nella ricerca e sviluppo di modalità di valorizzazione di questa preziosa risorsa.

È proprio in questo contesto di "lotta" tra innovazione e regolamentazione che si inserisce, con forza dirompente, il **dato sintetico**, tecnologia che rappresenta un'opportunità senza precedenti di **sfruttare il valore dei dati per l'innovazione preservando allo stesso tempo privacy dei cittadini e fiducia pubblica nelle istituzioni e nel progresso.**

Il presente paper si pone l'obiettivo di analizzare i dati sintetici nelle loro caratteristiche e implicazioni. Il documento, partendo da un'introduzione alla loro strategicità e alle principali tecniche utilizzate per la loro generazione, approderà (passando per una panoramica delle principali aziende coinvolte nel settore) a un'analisi delle implicazioni politiche e sociali dei dati sintetici e ad alcune proposte di policy per consentire uno sviluppo florido e, allo stesso tempo, rispettoso dei diritti della comunità che ambiscono a beneficiare.

Un'introduzione ai dati sintetici e alla loro strategicità

Attualmente, la fonte predominante di dati utilizzati per l'addestramento dei sistemi di intelligenza artificiale deriva da **eventi che si verificano nel mondo fisico o da attributi individuali.** Tuttavia, la raccolta, l'elaborazione o la distribuzione di set di dati è spesso associata ad alcune complicazioni come costi di raccolta, problemi di qualità e di privacy. Infatti, sebbene i dati possano essere ampiamente disponibili, questi potrebbero essere non etichettati, altamente parziali, protetti legalmente o per lo più di bassa qualità ovvero distorti (cioè che non rispecchiano la realtà). Per superare questi problemi, le aziende e i ricercatori si stanno rivolgendo in misura sempre maggiore ai dati sintetici come soluzione di fronte alle numerose sfide legate ai dati e ai requisiti delle tecnologie di IA.
Sebbene l'ambito di studio manca ancora di un consenso generale riguardo a una definizione univoca, una proposta di definizione è stata avanzata dalla **Royal Society** e dall'**Alan Turing Institute**, che descrivono i dati sintetici come:

"dati generati utilizzando un modello matematico o un algoritmo appositamente creato, con lo scopo di risolvere un insieme di compiti"[3]**.**

Questa definizione si concentra sull'essenza tecnologica, considerando i dati sintetici come repliche delle proprietà statistiche di un'entità piuttosto che su casi applicativi.

3. Jordon J., Weller Adrian. "Synthetic Data – what, why and how?" arXiv: 2205.03257 [cs], (2022).

L'utilizzo di dati sintetici ha radici storiche, che risalgono ai lavori di Stanislaw Ulam e John von Neumann negli anni Quaranta, che impiegarono il metodo Monte Carlo per simulazioni basate su campionamenti casuali [4]. Si ha poi un'intensificazione nell'uso a partire dagli anni '70, agli albori dell'informatica, in quanto molti dei primi sistemi e algoritmi avevano bisogno di dati per funzionare. La limitata potenza di calcolo, la difficoltà di raccogliere grandi quantità di dati e i problemi di divulgazione della privacy hanno orientato gli sforzi verso la generazione di dati sintetici e, di conseguenza, hanno trasformato questo dominio in un vantaggio strategico fondamentale.

I dati sintetici sono **strumenti strategici** per diversi motivi, utili in particolare quando le organizzazioni si trovano a navigare in un panorama complesso di innovazione, regolamentazione e concorrenza. Si tratta, infatti, di uno strumento chiave per migliorare la qualità delle analisi, facilitare la condivisione delle informazioni, e non solo. Di seguito vengono riportati i maggiori benefici dei dati sintetici quali strumenti per migliorare la competitività delle aziende e il loro dinamismo nel rispondere alle sfide del contesto generale in cui operano.

4. Metropolis, N. & Ulam, S. The Monte Carlo method. J. Am. Stat. Assoc. 44, 335–341 (1949).

- **Maggiore privacy e sicurezza:** I dati sintetici consentono di creare set di dati completamente nuovi che imitano le proprietà statistiche dei dati originali senza contenere informazioni sensibili reali. Questo approccio è particolarmente importante per rispettare le severe leggi sulla protezione dei dati, come il GDPR e l'HIPAA. Utilizzando dati sintetici, le aziende possono ridurre il rischio di violazione dei dati e proteggere la privacy delle persone, consentendo una condivisione e una collaborazione più sicure. Di fatto, quando i dati sintetici non mantengono proprietà che permettono la re-identificazione degli individui attraverso tecniche di reverse engineering, le aziende sono facilitate nella condivisione dei dati tra diversi dipartimenti o con terze parti, riducendo al minimo il rischio di violazioni della privacy. Questo approccio consente alle aziende di sfruttare il valore dei dati per analisi e decisioni strategiche senza compromettere la riservatezza delle informazioni sensibili.

- **Consentire lo sviluppo dell'intelligenza artificiale e del machine learning**: l'accesso a una serie di dati diversificati e di alta qualità è fondamentale per l'addestramento di modelli robusti di IA e di apprendimento automatico. Tuttavia, l'acquisizione di tali dati può essere difficile a causa di problemi di privacy, scarsità di dati o rarità di alcuni eventi. La generazione di dati sintetici aiuta a superare questi ostacoli, fornendo insiemi di dati ampi e diversificati che migliorano le prestazioni e l'affidabilità degli algoritmi di IA, senza i vincoli e gli errori spesso presenti nei dati del mondo reale. In termini di innovazione e sperimentazione, la capacità di generare grandi volumi di dati consente di addestrare modelli di IA in modo più efficiente, facilitando la ricerca e sviluppo.

- **Riduzione dei costi ed efficienza delle risorse**: raccogliere ed etichettare grandi quantità di dati reali può essere estremamente costoso, sia in termini di risorse umane che tecnologiche, e richiede molto tempo e conformità a rigide normative sulla privacy. La generazione di dati sintetici può ridurre drasticamente questi costi creando automaticamente grandi volumi di dati etichettati. Questa efficienza non solo accelera il processo di ricerca e sviluppo, ma riduce anche la dipendenza da attività ad alta intensità di lavoro, rendendo il ciclo di sviluppo più efficiente e meno dispendioso in termini di risorse.

Inoltre, i dati sintetici possono essere generati on-demand, permettendo un accesso immediato e scalabile a informazioni utili per vari progetti e analisi. Questo approccio non solo ottimizza i processi aziendali, ma libera anche risorse che possono essere reinvestite in altre aree strategiche, contribuendo così a una maggiore efficienza e competitività sul mercato.

- **Training e testing degli algoritmi**: i dati sintetici possono essere progettati specificamente per testare casi limite o scenari rari che non sono ben rappresentati nei dati reali disponibili. Ciò è particolarmente utile in settori come la guida autonoma, i servizi finanziari e l'assistenza sanitaria, in cui sono necessari test robusti per garantire la sicurezza e l'efficacia del sistema. I dati sintetici consentono di eseguire test completi su una gamma più ampia di scenari rispetto a quanto sarebbe possibile fare con i soli dati reali. Infatti, nell'attività di training del Machine Learning, l'uso di dati sintetici permette di ampliare significativamente il set di dati, migliorando la robustezza e l'accuratezza dei modelli. Nella fase di testing, i dati sintetici forniscono scenari diversi e variabili che possono aiutare a identificare e risolvere potenziali problemi prima del lancio del prodotto o servizio.

- **Conformità normativa e formazione**: nei settori in cui la conformità normativa è fondamentale, i dati sintetici possono essere utilizzati per la formazione dei dipendenti in un ambiente privo di rischi. Ad esempio, nei servizi finanziari, i dati sintetici che assomigliano a transazioni finanziarie sensibili possono essere utilizzati per addestrare i modelli o il personale senza esporre i dati reali dei clienti, rispettando così gli standard di conformità.

- **Agilità e innovazione del mercato**: la capacità di generare e manipolare rapidamente i dati per riflettere le tendenze emergenti o le mutevoli condizioni di mercato rende i dati sintetici un potente strumento di innovazione. Le organizzazioni possono simulare diverse condizioni economiche, sociali o ambientali per prevedere i risultati, prendere decisioni strategiche e innovare i prodotti senza aspettare che i dati reali siano disponibili.

- **Mitigare e prevenire bias discriminatori nei sistemi di IA**: il pregiudizio può insinuarsi negli algoritmi in diversi modi, occorre sfatare il mito delle macchine "imparziali": le applicazioni di Intelligenza Artificiale si basano infatti sull'esecuzione di un algoritmo partendo da un dataset che può finire per riflettere indirettamente i preconcetti di chi l'ha progettato, solitamente collegati a razza, genere, sesso biologico, età e cultura. Proprio grazie ai dati sintetici, è ora possibile bilanciare categorie sottorappresentate nel dataset di partenza, garantendo robustezza e affidabilità dei risultati.

- **Operazioni globali e scalabilità:** i dati sintetici eliminano molte delle barriere legali e logistiche associate al trasferimento internazionale di dati sensibili. Ad esempio, un'azienda che opera nel settore sanitario sia in Europa che negli Stati Uniti non avrebbe la facoltà di utilizzare negli Stati Uniti i dati raccolti in Europa. Tramite i dati sintetici, tali organizzazioni possono scalare le operazioni a livello globale senza violare le leggi internazionali sul trasferimento dei dati, consentendo un'espansione e una standardizzazione più agevoli oltre confine.

Perciò, sotto determinati punti di vista l'adozione di dati sintetici rappresenta una scelta strategica per le aziende che cercano di navigare le sfide del moderno panorama informativo in modo responsabile, abbracciando completamente l'approccio data-driven imprescindibile per la transizione delle imprese verso un futuro innovativo e digitale.

La strategicità di questo strumento è riflessa anche a **livello geopolitico**. Negli Stati Uniti, per esempio, il Dipartimento degli Affari dei Veterani ha impiegato dati sintetici durante la pandemia COVID-19 per proteggere le informazioni sanitarie dei veterani[5]. Di fatto, l'Ordine Esecutivo del presidente Joe Biden sull'intelligenza artificiale ha indicato la generazione di dati sintetici come un esempio di "tecnologia che migliora la privacy". Tuttavia, l'Ordine Esecutivo evidenzia chiaramente come, perfino il regolatore statunitense, sia maggiormente focalizzato sulla capacità dei dati sintetici di proteggere i consumatori piuttosto che sulla loro potenzialità di promuovere l'innovazione e la competitività delle imprese.

5. Department of Veterans Affairs, Synthetic Suicide Prevention Dataset with SDoH, Feb 2021

Questo approccio mette in luce un tema centrale e spesso dibattuto in Europa: l'eccessiva regolamentazione che favorisce la tutela dei diritti dei cittadini a scapito della promozione della competitività delle aziende. In altre parole, mentre la protezione dei dati personali e la sicurezza dei consumatori sono giustamente prioritarie, è altrettanto importante trovare un equilibrio che non soffochi l'innovazione e la crescita economica delle imprese a livello globale.

Infine, i dati sintetici rappresentano una soluzione tecnologica promettente per risolvere le incertezze normative relative all'**anonimizzazione dei dati personali**. Il GDPR definisce come "dato personale" *qualsiasi informazione riguardante una persona fisica identificata o identificabile*[6]. Per escludere il trattamento che s'intende attuare dall'ambito di applicazione del GDPR, è necessario rimuovere il carattere "personale" dei dati, rendendoli non identificabili. L'anonimizzazione dei dati implica che l'interessato non possa più essere identificato, distinguendosi dalla pseudonimizzazione che, invece, permette l'identificazione tramite informazioni aggiuntive.

Secondo l'articolo 29 del GDPR, i dati sufficientemente anonimi non sono più considerati dati personali e quindi non sono soggetti al Regolamento. Questo significa che devono essere trattati in modo tale da rendere l'identificazione dell'interessato praticamente impossibile, tenendo conto dei mezzi tecnologici disponibili. Questo criterio di ragionevolezza, delineato nell'articolo 26 del GDPR, richiede una valutazione dei costi, del tempo e delle tecnologie necessarie per una eventuale re-identificazione. Tuttavia, l'interpretazione rigida del WP29 (ora European Data Protection Board - EDPB) nel Parere 05/2014, che richiede che l'anonimizzazione sia permanente e irreversibile, ha creato incertezze su come gestire correttamente l'anonimizzazione, specialmente alla luce del rapido avanzamento delle capacità computazionali che rendono sempre più difficile garantire l'irreversibilità dell'anonimizzazione. Inoltre, quanto più si investe nell'anonimizzazione dei dati, tanto più diminuisce il loro valore intrinseco: la rimozione di elementi identificativi può compromettere la qualità dei dati per scopi statistici o scientifici, riducendone l'utilità.

I dati sintetici, che mantengono le proprietà statistiche dei dati originali senza conservare il carattere personale, rappresentano una soluzione efficace per facilitare la condivisione dei dati personali, sottraendosi alla disciplina del GDPR. Questo approccio consente di preservare l'utilità dei dati per la ricerca e l'analisi, garantendo al contempo il rispetto delle normative sulla privacy.

6. Art 4. Regolamento (UE) 2016/679

Tipologie e tecniche di generazione di dati sintetici

La classificazione dei dati sintetici può essere suddivisa in **diverse categorie**: in base alla fonte di generazione, al tipo di dati (numerici, categorici, testuali, immagini, audio), all'uso e al dominio di applicazione. Indipendentemente dal contesto applicativo, la classificazione dei dati sintetici comprende un ampio spettro che va da dati parzialmente sintetici a completamente sintetici.
Per affrontare questioni giuridiche complesse, come l'applicazione del GDPR ai dati sintetici, la trasparenza e la spiegabilità dell'algoritmo, e la proprietà intellettuale, è importante distinguere tra **dati completamente e parzialmente sintetici**.

I dati parzialmente sintetici si generano estraendo e replicando le proprietà statistiche di un set di dati del mondo reale[7]. Questo tipo di dati può essere particolarmente utile in ambito sanitario, dove possono proteggere la tutela della privacy dei pazienti consentendo allo stesso tempo ai ricercatori di condurre analisi e condividere i risultati svincolandosi dalla rigida normativa dettata dal legislatore italiano che impone l'acquisizione consenso dell'interessato per ogni uso secondario che s'intende attuare.[8]

D'altra parte, **i dati completamente sintetici** vengono creati interamente ex novo sulla base di regole, modelli o simulazioni predefiniti. Questi dati non rappresentano direttamente il mondo fisico, ma sono progettati per replicare la complessità e le variabili che potrebbero essere osservate in scenari del mondo reale. La generazione di dati completamente sintetici è particolarmente utile quando si desidera simulare sistemi complessi o quando i dati reali non sono disponibili per l'utilizzo.[9]

7. Reiter, J. Inference for partially synthetic, public use microdata sets. Surv. Methodol. 29, 181–188 (2003).
8. DECRETO LEGISLATIVO 10 agosto 2018, n. 101
9. Raghunathan, T., Reiter, J. & Rubin, D. Multiple imputation for statistical disclosure limitation. J. Stat. 19, 1–16 (2003).

Ad oggi esistono diverse tecniche per generare dati sintetici, ognuna con applicazioni specifiche.
Tra queste, i modelli di deep learning ed econometrici basati su agenti [10] rappresentano metodologie avanzate, così come le equazioni differenziali stocastiche utilizzate per simulare sistemi fisici o economici [11].

Una delle tecniche più avanzate per la generazione di dati sintetici è l'uso di modelli generativi, come le Generative Adversarial Networks (GAN) e i Variational Autoencoders (VAE).

Le GAN impiegano reti neurali per produrre dati che imitano fedelmente i dati raccolti, mentre i VAE creano rappresentazioni latenti dei dati per generare nuove informazioni, contribuendo alla varietà e alla qualità dei dati sintetici prodotti. Un'altra categoria significativa è quella dei dati sintetici basati su regole. Questi dati possono essere generati utilizzando specifiche regole o attraverso simulazioni basate su modelli matematici.[12]

Per la protezione della privacy, tecniche come la perturbazione dei dati, che aggiunge rumore ai dati reali, e il mascheramento dei dati, che sostituisce dati sensibili con valori sintetici, sono ampiamente utilizzate.[13]
Tecniche di rotazione e traslazione dei dati aumentano invece la quantità di dati di training, migliorando la robustezza dei modelli per il test di software e la formazione di modelli di machine learning. È possibile generare dati sintetici misti, che combinano dati raccolti e sintetici, mantenendo alcune caratteristiche dei dati raccolti e aggiungendo variabilità tramite dati sintetici.[14]

10. Bonabeau, E. Agent-based modeling: Methods and techniques for simulating human systems. Proc. Natl Acad. Sci. 99, 7280–7287 (2002).
11. Carmona, R. and Delarue, F. Probabilistic Theory of Mean Field Games with Applications, volume 84. Springer (2018).
2. Xu, L., Skoularidou, M., Cuesta-Infante, A. & Veeramachaneni, K. Synthesizing Tabular Data using Conditional GAN. arXiv:1907.00503 (2019).
13. Cormode, G., Procopiuc, C. M., Srivastava, D., Li, N. & Machanavajjhala, A. Differentially Private Synthetic Data Generation. IEEE International Conference on Data Engineering (2012).
14. Hilderman, R. J. & Hamilton, H. J. Generating Synthetic Data to Match Data Mining Patterns. IEEE International Conference on Data Mining (2001).

Oltre la teoria: il mercato dei dati sintetici oggi e domani

Il mercato che si fonda sulla generazione di dati sintetici ha registrato una crescita pari al 50,5% tra il 2019 e il 2023. Si stima che possa raggiungere un valore complessivo di 300 milioni di dollari alla fine del 2024. Questa crescita non è destinata ad arrestarsi: si prevede che questo "settore in via di sviluppo" raggiungerà i 13 miliardi di dollari entro il 2034, con un tasso di crescita superiore al 45% nel periodo 2024-2034.[15]

Bastano questi impressionanti numeri a rendere evidente come nella "società di domani", il valore già cruciale dei dati, assumerà una centralità ancora più assoluta nell'ambito dello sviluppo delle nuove tecnologie e, con esse, dello sviluppo di nuovi mercati e nuove economie.

Come evidenziato nel corso di questa ricerca, nonostante i numeri e il valore economico - sia attuale che potenziale - che caratterizzano questo mercato emergente, la regolamentazione sull'utilizzo e lo sviluppo dei dati sintetici rimane ancora ampiamente carente. Da un lato, vi sono le istituzioni, con l'UE che potrebbe assumere un ruolo di leadership dal punto di vista normativo, come ha fatto spesso negli ultimi anni soprattutto nel settore del digitale. Dall'altro lato, vi sono le aziende, in particolare i grandi player del mercato digitale globale, i cui team di ricerca, come analizzeremo nel paragrafo seguente, sono già ampiamente impegnati nella corsa all'oro: riuscire a garantirsi una fonte inesauribile di dati, senza incappare - almeno per il momento - nelle stringenti normative occidentali in materia di protezione dei dati personali.

Le potenzialità di questa tecnologia, oltre ad essere dimostrate dalle stime economiche che ne prevedono, come già illustrato, una crescita esponenziale nel medio e lungo periodo, sono provate anche da un dato ancor più concreto e tangibile: gli investimenti che grandi e medi player del settore stanno portando avanti a ritmi elevatissimi: aziende tecnologiche del calibro di Amazon, Microsoft, OpenAI e IBM hanno infatti già avviato notevoli progetti di sviluppo, proponendosi ancora una volta di

15. Synthetic Data Generation Market: Forecast Analysis [2030] (no date) Synthetic Data Generation Market | Forecast Analysis [2030]. Available at: https://www.fortunebusinessinsights.com/synthetic-data-generation-market-108433 (Accessed: 06 August 2024).

guidare l'innovazione tecnologica, rispondendo così, ben prima della regolamentazione, alle nuove esigenze del mercato e stabilendo standard, prassi e trend economici che modelleranno l'intero settore nei decenni a venire.

Si riporta dunque di seguito una breve panoramica delle **aziende leader del settore** nonché dei loro progetti rispetto allo sviluppo dell'intelligenza artificiale generativa e con questa, dei dati sintetici.

Google

"Do the right thing, don't be evil", un motto ventennale che è una vera e propria dichiarazione di intenti: indica in modo chiaro l'impegno inequivocabile di Google per la protezione della privacy degli utenti, mantenendo un codice di condotta leale e - per dirla con un'esagerazione archetipica - "dalla parte dei buoni".
Non stupisce, dunque, che la protezione della privacy sia il propulsore dell'impegno di Google e della sua divisione AI anche nel campo della ricerca e dell'utilizzo dei dati sintetici.

Ormai simbolico dell'impegno di Google in materia di privacy è il concetto di "*differential privacy*"; si tratta di un modello matematico basato sull'aggiunta di rumore casuale ai dati e sulla garanzia formale che l'output di un dataset non cambia significamente se un singolo dato viene aggiunto o rimosso.
Il modello di fatto "anonimizza" dei dati sensibili, consentendone un utilizzo protetto; è facile, dunque, capire che nella nostra "società dei dati" si tratta di un sistema imprescindibile che ha consentito e continua a consentire la raccolta e lo studio di dati sensibili, la costruzione di dataset poi destinati alla pubblicazione e, in ultimo ma non per importanza, maggiori opportunità di training di sistemi di intelligenza artificiale generativa.

Come ormai noto, l'obiettivo della generazione di dati sintetici è la produzione di dati statisticamente simili alle fonti di dati del mondo reale. Tuttavia, quando i dati sono troppo simili e replicano in modo univoco i dettagli identificativi dei dati di origine, la promessa di preservare la privacy viene compromessa. È qui che la Differential Privacy entra in gioco.

L'evoluzione naturale dell'impegno di Google in materia di privacy è la ricerca aziendale in materia di dati sintetici che rispettino l'assunto base di differential privacy[16]: ovvero dati che hanno le caratteristiche chiave dei dati originali, ma sono del tutto artificiali e, mediante il modello della privacy differenziale, e dunque l'aggiunta di rumore casuale, offrono una garanzia forte e matematicamente rigorosa di protezione dei "dati fonte".

Recentemente, dunque, la ricerca di Google tenta di rispondere al duplice problema della quantità e della riservatezza del dato ricorrendo a nuovi metodi per generare dati sintetici. Il colosso americano ha infatti sviluppato modelli generativi avanzati e metriche di valutazione per garantire che i dati sintetici prodotti siano affidabili e utilizzabili in contesti reali.

Oltre alle più ovvie applicazioni nell'ambito del machine learning e del miglioramento delle potenzialità dei sistemi di IA, tra le più recenti applicazioni, Google ha recentemente lanciato un update del software di base per i dispositivi mobile, volto a rilevare contenuti inappropriati mediante l'uso di un classificatore di sicurezza addestrato proprio su set di dati sintetici.

Infine, al fine di indagare possibili sviluppi e utilizzi futuri, vale la pena menzionare il "Project Nightingale" [17] di Google Cloud e Ascension, avviato nel 2019. Sebbene non sia esclusivamente focalizzato sui dati sintetici, questo progetto pone l'obiettivo di utilizzare grandi quantità di dati sanitari per sviluppare algoritmi. Come noto, i dati sanitari sono oggetto di normative di privacy solitamente più stringenti, il progetto, dunque, potrebbe in futuro beneficiare dell'uso di dati sintetici con l'obiettivo di garantire la privacy dei pazienti e senza rinunciare a campioni che siano matematicamente significativi.

16. Synthetic Data Generation Market: Forecast Analysis [2030] (no date) Synthetic Data Generation Market | Forecast Analysis [2030]. Available at: https://www.fortunebusinessinsights.com/synthetic-data-generation-market-108433 (Accessed: 06 August 2024).
17. Per ulteriori approfondimenti: Brush, K. (2020) What is project Nightingale? everything you need to know, Health IT and EHR. Available at: https://www.techtarget.com/searchhealthit/definition/Project-Nightingale (Accessed: 26 September 2024).

Microsoft

Azienda che da quarant'anni guida e direziona lo sviluppo tecnologico globale, anche nella competizione per lo sviluppo di sistemi di intelligenza artificiale generativa Microsoft non sta certamente rimanendo in secondo piano: Co-Pilot, l'assistente di IA Generativa pensata per supportare e assistere chi lavora in maniera integrata, è considerata una delle IA che più rivoluzionerà il nostro modo di intendere e strutturare il lavoro.

L'azienda fondata da Bill Gates e Paul Allen è infatti stata una delle più lungimiranti da questo punto di vista: a differenza di altri grandi player come Google, partito con rincorsa nella ricerca e nello sviluppo della propria IA e Meta, con la (per ora) perdente scommessa di Mark Zuckerberg sul Metaverso, già nel 2019 Microsoft è stata una dei maggiori finanziatori dell'ancora semi sconosciuta OpenAI, società che ha sviluppato il Chatbot che tutti abbiamo imparato a conoscere negli ultimi anni: **ChatGPT**. Su questo versante, si ricorda da ultimo l'accordo pluriennale stretto tra le due società nel gennaio 2023 che, si mormora, graviti attorno a un valore complessivo di circa 10 miliardi di dollari. Accordo che ha tra l'altro catturato l'attenzione - oltre che dell'industria - delle istituzioni europee, con la CE che ha avviato un processo di verifica volto a stabilire se la partnership sia compatibile con la normativa europea in materia di concentrazioni.[18]

Non sorprende dunque che una parte dei programmi di ricerca che Microsoft ha all'attivo, riguardino anche sullo sviluppo e l'utilizzo di dati sintetici. Attraverso il programma Microsoft Research IA, Microsoft investe risorse anche in tecnologie volte a creare ambienti di simulazione basati su dati sintetici per testare e sviluppare algoritmi di IA. L'azienda ha inoltre sviluppato la piattaforma "**Microsoft Synthetic Data Showcase**"[19], per la generazione di dati sintetici utilizzati per formare modelli di IA più generalizzabili e meno soggetti a bias.

In linea con il suo noto approccio eticamente responsabile Microsoft fa infatti un grande uso di dati sintetici per migliorare i suoi stessi sistemi di GenAI: dati sintetici sono stati ad esempio utilizzati con la finalità di migliorare gli algoritmi di riconoscimento facciale e vocale. Un altro esempio virtuoso di utilizzo dei

18. Microsoft partner di OpenAI (2023) WindowsBlogItalia. Available at: https://www.windowsblogitalia.com/2024/07/microsoft-10-miliardi-openai/ (Accessed: 26 September 2024).
19. SDS - Synthetic Data Showcase. Available at: https://microsoft.github.io/synthetic-data-showcase/#/prepare (Accessed: 26 September 2024).

dati sintetici è Phi-3, lo Small Language Model di Microsoft, addestrato su un dataset che include dati sintetici generati da modelli di IA e dati filtrati da siti web pubblicamente disponibili, con un'attenzione particolare alla qualità e alla densità di ragionamento dei dati.

In linea con il suo noto approccio eticamente responsabile Microsoft fa infatti un grande uso di dati sintetici per migliorare i suoi stessi sistemi di GenAI: dati sintetici sono stati ad esempio utilizzati con la finalità di migliorare gli algoritmi di riconoscimento facciale e vocale[20]. Un altro esempio virtuoso di utilizzo dei dati sintetici è Phi-3[21], lo Small Language Model di Microsoft, addestrato su un dataset che include dati sintetici generati da modelli di IA e dati filtrati da siti web pubblicamente disponibili, con un'attenzione particolare alla qualità e alla densità di ragionamento dei dati.

Seguendo sempre il filone dello sviluppo etico, vale la pena ricordare che nel 2022, l'azienda ha pubblicato il suo **Microsoft Responsible AI Standard**[22]: un quadro che guida lo sviluppo di sistemi di IA nel rispetto dei principi fondamentali di equità, affidabilità, sicurezza, privacy, inclusività, trasparenza e responsabilità e impegna l'azienda alla pubblicazione annuale di un Responsible AI Trasparency Report[23]. Sulla scorta di questo impegno dichiarato, Microsoft collabora attivamente con istituzioni, organizzazioni internazionali e altre Big Tech per promuovere pratiche di IA responsabili, come testimonia l'impegno proattivo all'interno del Frontier Model Forum, istituito insieme a Google, OpenAI e Anthropic e volto a condividere know-how e sviluppare linee guida comuni sull'uso etico dell'IA.

Recentemente, l'azienda ha chiarito anche le sue indicazioni etiche in materia di dati sintetici: come Google, anche Microsoft sta concentrando i suoi sforzi nella generazione di dati sintetici che siano "differenzialmente privati". Ad esempio, la piattaforma SmartNoise[24] di Microsoft utilizza tecniche di privacy differenziale per creare dataset sintetici che mantengono le caratteristiche statistiche dei dati originali senza compromettere la privacy degli individui coinvolti.

20. Afonja, G. et al. (2024) Improving synthetic data without compromising privacy protection, Microsoft Research. Available at: https://www.microsoft.com/en-us/research/blog/the-crossroads-of-innovation-and-privacy-private-synthetic-data-for-generative-ai/ (Accessed: 30 August 2024).
21. Per ulteriori informazioni: Cisternino, A. (2024) Phi3, ecco la guida alla più piccola ia di Microsoft, Agenda Digitale. Available at: https://www.agendadigitale.eu/cultura-digitale/phi3-prestazioni-e-comportamento-del-nuovo-modello-llm-di-microsoft/ (Accessed: 15 September 2024).
22. Thomson, J. (2024) Providing further transparency on our responsible AI efforts, Microsoft On the Issues. Available at: https://blogs.microsoft.com/on-the-issues/2024/05/01/responsible-ai-transparency-report-2024/ (Accessed: 26 September 2024).
23. (2022) Microsoft Research. rep. Microsoft. Available at: https://blogs.microsoft.com/wp-content/uploads/prod/sites/5/2022/06/Microsoft-Responsible-AI-Standard-v2-General-Requirements-3.pdf
24. Smart Noise - Digital Tools For Smarter Music Marketing. Available at: https://smart-noise.com/ (Accessed: 17 September 2024)

OpenAI

OpenAI, società leader nel settore dello sviluppo della GenAI, ha mostrato un interesse particolare nell'uso dei dati sintetici per migliorare la performance e l'equità dei suoi modelli linguistici avanzati, come GPT-3. Come Microsoft, anche OpenAI si è fatta promotrice dell'idea che i dati sintetici possono rivelarsi strumenti fondamentali per ridurre i bias e migliorare la qualità, l'efficienza e l'equilibrio dei LLM.

In quest'ottica, i dati sintetici diverrebbero, oltre che fonte di valore tecnologico e dunque economico, anche strumenti che potrebbero agevolare gli sviluppatori di intelligenza artificiale ad allenare e dunque "costruire" sistemi di IA più solidi dal punto di vista etico e meno soggetti a bias e discriminazioni.

Lo stesso Sam Altman - Ceo e co-founder di OpenAI, sostenitore della necessità di regolamentazione in materia di IA ha affermato che un modello di GenIA, in futuro, dovranno essere in grado di generare dati sintetici sufficienti per il proprio addestramento [25] ; dunque, in un'ottica di "circolarità" ed efficienza, appare chiaro come i dati utilizzati per l'addestramento rappresentino una delle risorse più preziose del boom dell'IA.

I dati sintetici, quindi, risultano la chiave di volta per il futuro di OpenAI e dell'IA in generale, anche grazie a una riduzione di dipendenza dei sistemi dai dati protetti da copyright e privacy.

Amazon

Nella corsa ai dati sintetici, anche Amazon, attraverso Amazon Web Services (AWS) e il suo reparto di ricerca, sta investendo significativamente per semplificare l'addestramento dei sistemi di IA e facilitare l'accesso e l'utilizzo dei suoi servizi. Tra questi, uno dei più famosi, l'assistente virtuale Alexa, è già da tempo andato incontro ad un processo di miglioramento del riconoscimento vocale multilingue proprio grazie all'utilizzo di dati sintetici, i quali sono stati utilizzati per modelli di discorso, sintassi e semantica.

25. Rif: Sam Altman: How to Build the Future. Available at: https://www.youtube.com/watch?v=sYMqVwsewSg

Un esempio significativo è poi Amazon SageMaker Ground Truth[26], un tool introdotto da AWS nel 2022, progettato per semplificare il processo di etichettatura dei dati necessari per addestrare modelli di machine learning. Questo servizio è particolarmente utilizzato in applicazioni di visione artificiale, dove la raccolta e l'etichettatura manuale dei dati possono essere costose e lunghe. Ground Truth supporta anche la creazione di dati sintetici, arricchendo i dataset e migliorando la varietà e la qualità dei dati disponibili per l'addestramento, riducendo bias e bilanciando dataset che altrimenti potrebbero non essere rappresentativi.

Recentemente, nel novembre 2023, AWS ha formalizzato una collaborazione con Gretel, integrando la piattaforma Gretel con Amazon SageMaker Pipelines[27] attraverso il programma Synthetic Data Accelerator. Questo programma fornisce alle organizzazioni l'accesso a dati sintetici di alta qualità che replicano le caratteristiche statistiche dei dati reali, ma senza legami diretti con dati sensibili. L'iniziativa sta già avendo una certa risonanza in settori particolarmente sensibili come i servizi finanziari e la sanità.

Inoltre, Amazon utilizza dati sintetici in vari ambiti della gestione e ottimizzazione della logistica e del magazzino. Modelli di simulazione basati su dati sintetici, come l'Adaptive Transportation Optimization Service (ATROPS)[28], sono impiegati per prevedere scenari di traffico, ottimizzare le rotte di consegna e migliorare l'efficienza del magazzino, con strumenti come l'Amazon Bin Packing Algorithm.

Infine, nel Marketplace di AWS per il Machine Learning, sono disponibili diverse soluzioni di terze parti che offrono strumenti per la creazione e l'utilizzo di dati sintetici. Le aziende possono trovare e implementare software di generazione di dati sintetici che meglio si adattano alle loro specifiche esigenze industriali e di conformità.

26. (2022a) New – amazon sagemaker ground truth now supports Synthetic Data Generation | AWS News Blog. Available at: https://aws.amazon.com/blogs/aws/new-amazon-sagemaker-ground-truth-now-supports-synthetic-data-generation/ (Accessed: 06 September 2024).
27. Workflows for Machine Learning - Amazon sagemaker pipelines. Available at: https://aws.amazon.com/sagemaker/pipelines/ (Accessed: 06 September 2024).
28. O'Neill, S. (2024) Sizing down to scale up: How Amazon reworked its fulfillment network to meet customer demand, Amazon Science. Available at: https://www.amazon.science/news-and-features/how-amazon-reworked-its-fulfillment-network-to-meet-customer-demand (Accessed: 26 September 2024).

Altri protagonisti del mercato

In aggiunta ai principali player menzionati, un numero sempre maggiore di società, più o meno grandi, sta assumendo un ruolo altrettanto decisivo nello sviluppo e nell'adozione dei dati sintetici. In questo contesto, aziende come Apple, Meta, IBM, Intel, NVIDIA, Oracle, SAP, Salesforce e Tencent stanno investendo risorse significative in Ricerca e Sviluppo.

Apple

L'azienda di Cupertino sta esplorando l'utilizzo dei dati sintetici per potenziare le sue tecnologie di intelligenza artificiale e apprendimento automatico senza compromettere la riservatezza dei dati personali. Apple impiega tecniche avanzate di machine learning, come il trasferimento di conoscenza e la privacy differenziale, per sviluppare modelli che garantiscano la privacy degli utenti. La ricerca di Apple si focalizza sull'impiego di dati sintetici per perfezionare le sue tecnologie di riconoscimento facciale e vocale, come Siri e Face ID.

Meta

Anche Meta è attivamente impegnata nella ricerca e nell'uso dei dati sintetici. La piattaforma si è focalizzata in particolare sull'utilizzo di questi dati per migliorare la sicurezza e l'efficacia dei suoi modelli di IA che l'azienda utilizza per una varietà di applicazioni: si spazia dai miglioramenti in realtà aumentata (AR) e virtuale (VR) al miglioramento dell'interazione utente nei suoi servizi social all'uso dell'IA per la moderazione dei contenuti. Da menzionare tra le altre cose il programma "Data for Good" [29] che prevede l'uso di dati sintetici per modellare e prevedere scenari di crisi, come la diffusione di malattie o disastri naturali e per aiutare le organizzazioni umanitarie a pianificare interventi più efficaci.

IBM

La "Big Blue" utilizza e genera dati sintetici attraverso diverse piattaforme e strumenti per migliorare l'efficienza e la sicurezza nella gestione dei dati e nell'addestramento dei modelli di IA. IBM Watsonx [30] , piattaforma di dati e IA che supporta le aziende nella creazione di applicazioni di IA personalizzate, offre strumenti per generare dati tabulari sintetici a partire da dati di produzione o da schemi di dati personalizzati mediante algoritmi di modellazione visiva.

29. Data for good at Meta Home (no date) Data For Good at Meta Home. Available at: https://dataforgood.facebook.com/ (Accessed: 12 September 2024).
30. *IBM watsonx - an AI and data platform built for business* (2024a) *IBM watsonx - An AI and data platform built for business*. Available at: https://www.ibm.com/watsonx (Accessed: 26 September 2024).

Altro esempio dell'impegno di IBM nel settore dei dati sintetici è Task2Sim[31], sviluppato in collaborazione con la Boston University: si tratta di un modello AI che genera dati sintetici specifici per compiti di pretraining di modelli di classificazione delle immagini, permettendo di controllare parametri specifici e facilitando la creazione di dataset illimitati con etichette gratuite, al fine di aumentare l'efficienza e la precisione dei modelli addestrati. Infine, SenseGen[32] è un'architettura di deep learning sviluppata da IBM per generare dati sintetici da sensori.

Intel

Sta investendo significativamente nella ricerca sui dati sintetici per migliorare l'efficienza e la sicurezza dei suoi processori e delle sue piattaforme di intelligenza artificiale. Intel Labs, in particolare, sta progressivamente concentrando i suoi sforzi nell'uso dei dati sintetici per simulare scenari complessi e addestrare modelli di intelligenza artificiale in modo sicuro e scalabile.

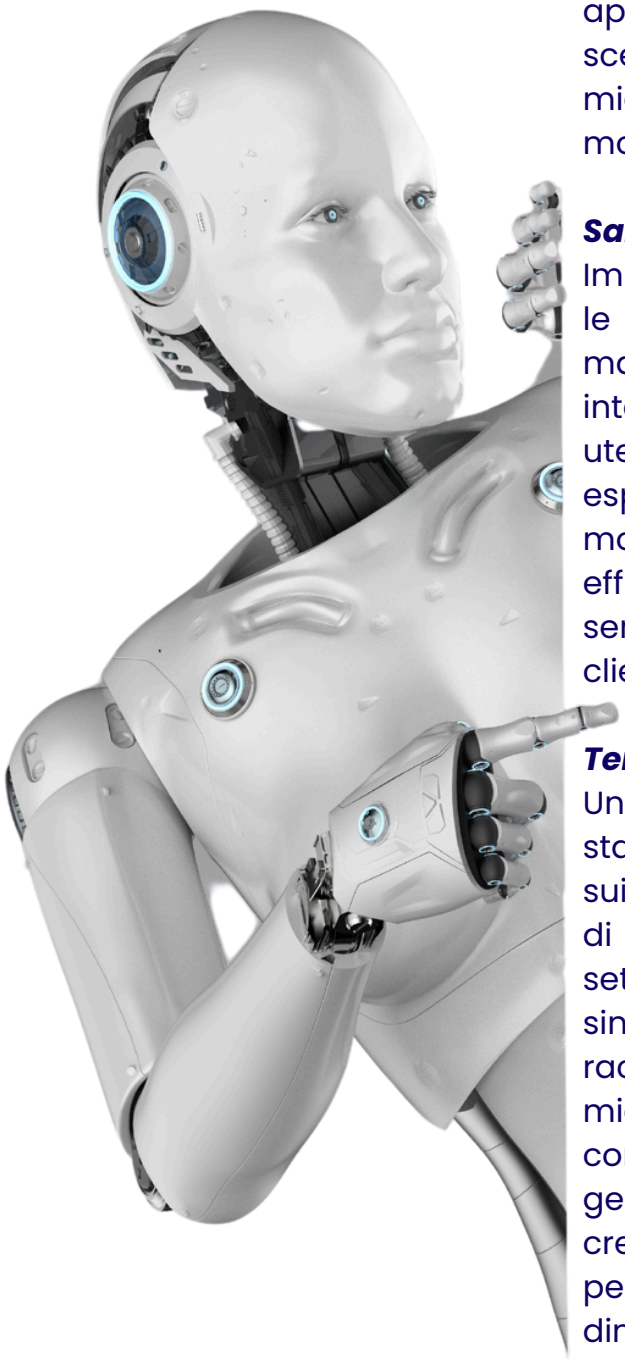
NVIDIA

Azienda leader nel campo della grafica computazionale e dell'intelligenza artificiale, utilizza dati sintetici per migliorare le sue tecnologie di simulazione e visualizzazione. Si tratta sicuramente di un player destinato a giocare un ruolo sempre più significativo nell'ambito delle tecnologie innovative. Le piattaforme di NVIDIA, come Omniverse, sfruttano i dati sintetici per creare simulazioni realistiche utili alla formazione di algoritmi di intelligenza artificiale, specialmente in settori come l'automotive e la robotica. NVIDIA collabora con numerose aziende e istituti di ricerca per sviluppare tecnologie avanzate di generazione di dati sintetici.

Oracle

Attore chiave nel settore dei database e dei servizi cloud, impiega i dati sintetici per migliorare la sicurezza e la scalabilità delle sue soluzioni di gestione dei dati. Oracle Cloud Infrastructure è un ottimo esempio di tale utilizzo poiché offre strumenti per la generazione di dati sintetici, aiutando direttamente le aziende, mediante un'interfaccia agile, a creare modelli di machine learning robusti e scalabili senza compromettere la qualità dei dati.

31. Martineau, K. (2023) A new way to generate synthetic data for Pretraining Computer Vision models, IBM Research. Available at: https://research.ibm.com/blog/task2sim-synthetic-images (Accessed: 26 September 2024).
32. Alzantot, M., Chakraborty, S. and Srivastava, M. (2017) SenseGen: A deep learning architecture for synthetic sensor data generation for PERCOM workshops 2017, IBM Research. Available at: https://research.ibm.com/publications/sensegen-a-deep-learning-architecture-for-synthetic-sensor-data-generation (Accessed: 26 September 2024).



SAP

Utilizza dati sintetici per migliorare le sue soluzioni di Enterprise Resource Planning (ERP) e altre applicazioni aziendali, mediante la creazione di scenari aziendali complessi e la possibilità di migliorare costantemente la formazione dei modelli di intelligenza artificiale.

Salesforce

Impiega dati sintetici soprattutto per perfezionare le sue soluzioni di customer relationship management (CRM) e per addestrare modelli di intelligenza artificiale rispettosi della privacy degli utenti. Anche Salesforce AI Research sta esplorando l'uso dei dati sintetici per creare modelli di machine learning capaci di operare efficacemente in contesti aziendali complessi senza compromettere la riservatezza dei dati dei clienti.

Tencent

Una delle più grandi aziende tecnologiche in Cina, sta investendo considerevolmente nella ricerca sui dati sintetici per migliorare le sue piattaforme di social media, giochi e servizi finanziari. Nel settore dei social media, Tencent utilizza i dati sintetici per potenziare gli algoritmi di raccomandazione e personalizzazione, migliorando l'esperienza utente senza compromettere la privacy. In ambito gaming, la generazione di dati sintetici è fondamentale per creare ambienti di simulazione realistici che permettono di addestrare IA capaci di rispondere dinamicamente alle azioni dei giocatori, elevando così il livello di interattività e immersione. Una delle innovazioni più significative sviluppate dal Tencent AI Lab è il framework AlphaLLM, che integra tecniche di simulazione e feedback interno per migliorare le capacità di ragionamento dei modelli di linguaggio naturale senza la necessità di ulteriori annotazioni di dati.

L'ecosistema italiano

Nell'attuale e dinamico contesto dell'intelligenza artificiale, anche le aziende italiane stanno esplorando soluzioni innovative per sfruttare al meglio i propri asset informativi.
Riguardo la generazione di set sintetici, in Italia troviamo principalmente due startup di rilievo: **AINDO e Clearbox AI**.

AINDO

Aindo, nata nel 2018 dalla Scuola Internazionale Superiore di Studi Avanzati di Trieste, rappresenta il risultato di approfondite ricerche condotte da Daniele Panfilo e il suo team nel campo dei modelli di machine learning generativi. Attualmente composta da quasi 30 talenti, la startup ha sviluppato e brevettato una tecnologia all'avanguardia per la generazione di dati sintetici. Di recente, ha concluso con successo un round di finanziamento da 6 milioni di euro, che supporterà ulteriori sviluppi mirati a facilitare l'utilizzo dell'intelligenza artificiale in settori chiave come la sanità, la finanza e la pubblica amministrazione.

Aindo si distingue per la sua innovativa piattaforma incentrata sui dati sintetici. I modelli sviluppati dall'azienda sono in grado di creare dati artificiali che replicano con precisione pattern e comportamenti della realtà fisica, consentendo alle imprese con un vasto patrimonio di dati, finora inutilizzabili per motivi di privacy, di estrarne valore. L'impiego etico dell'intelligenza artificiale da parte di Aindo si concretizza nella **generazione di dati sintetici che mantengono l'utilità statistica dei dati originali, ma che, generati da algoritmi, sono privi di informazioni sensibili**, eliminando così i rischi di re-identificazione e salvaguardando la privacy dei soggetti coinvolti.

Le potenzialità di questa tecnologia si manifestano principalmente nel **settore sanitario**, dove è possibile potenziare la ricerca sulle patologie rare attraverso la creazione di versioni artificiali dei dati clinici, permettendo così lo svolgimento di studi senza compromettere la privacy dei pazienti. In ambito finanziario, invece, le applicazioni si concentrano sulla prevenzione delle frodi, senza compromettere le informazioni sensibili relative ai correntisti e alle transazioni bancarie.

Clearbox AI

ClearBox AI, fondata nel 2019, nasce dall'Incubatore del Politecnico di Torino e vede come co-fondatori Shalini Kurapati, Luca Gilli, Matteo Giovannetti e Federico Tomassetti. L'azienda propone un prodotto innovativo basato su tecnologia proprietaria che genera dati sintetici per supportare le imprese che desiderano avviare e potenziare i propri progetti di Intelligenza Artificiale. I dati sintetici offerti da ClearBox AI risultano utili per varie figure professionali aziendali, tra cui reparti di data science e AI, innovazione, software engineering e privacy.

Il Data Engine di ClearBox AI si basa su modelli generativi avanzati, adattandosi trasversalmente a diversi settori. Il loro prodotto, Enterprise Solution, già operativo in collaborazioni finanziarie, sanitarie, nel retail, nell'energia e nella mobilità, si integra facilmente nei processi aziendali esistenti, garantendo un'installazione agevole sia in loco che su cloud e fornisce un dettagliato report sulle metriche dei dati generati.

La missione di ClearBox AI è **comprendere le sfide che le aziende affrontano nello sviluppo e nell'implementazione di progetti di intelligenza artificiale, offrendo il proprio supporto attraverso tecnologie all'avanguardia.**

Elementi di diritto

L'emergere dei dati sintetici ha aperto nuove frontiere nel trattamento delle informazioni, ma comporta anche sfide significative che necessitano di attenzione e regolamentazione. La generazione di dati sintetici, sebbene promettente per molteplici applicazioni, richiede un approccio rigoroso per garantirne l'affidabilità e la conformità alle normative vigenti. In questo contesto, è cruciale affrontare vari aspetti, tra cui la validazione della qualità dei dati, la trasparenza degli algoritmi di generazione, le implicazioni legali relative al GDPR, e la questione della proprietà intellettuale. Solo attraverso l'implementazione di pratiche e normative adeguate sarà possibile sfruttare appieno il potenziale dei dati sintetici, promuovendo un ecosistema che favorisca l'innovazione e tuteli i diritti degli individui.

In primo luogo, è fondamentale implementare pratiche rigorose di **validazione e verifica durante la generazione dei dati sintetici**, al fine di garantire la **qualità** e l'**accuratezza** dei dati prodotti. Senza tali misure, l'affidabilità dei dati sintetici potrebbe essere compromessa, con potenziali ripercussioni negative su tutte le applicazioni che ne fanno uso.

In secondo luogo, è essenziale sviluppare **meccanismi di valutazione della spiegabilità dei procedimenti algoritmici utilizzati per la generazione dei dati sintetici**. La trasparenza nell'operato degli algoritmi non solo aumenta la fiducia degli utenti, ma è anche fondamentale per il rispetto di norme etiche e legali. A tal proposito, l'intervento delle autorità di controllo è necessario per stabilire linee guida e standard di spiegabilità che possano essere adottati universalmente.

Terzo, è imperativo chiarire le **circostanze applicative del GDPR ai dati sintetici**. Poiché i dati sintetici possono derivare da dati personali, è di vitale importanza stabilire regolamenti chiari per garantire che i diritti degli individui siano adeguatamente protetti, evitando al contempo oneri normativi eccessivi per le imprese.

Infine, la definizione della **proprietà intellettuale dei dati sintetici** rappresenta un altro aspetto cruciale. La mancanza di linee guida chiare su chi detiene i diritti sui dati sintetici potrebbe ostacolare l'innovazione e la collaborazione tra le diverse entità coinvolte. Pertanto, una normativa ben definita in questo ambito è necessaria per promuovere un ecosistema di dati sintetici equo e innovativo.

a) Qualità e accuratezza

Un aspetto cruciale è il delicato equilibrio che deve essere mantenuto tra l'accuratezza statistica e la protezione della privacy. Un'eccessiva anonimizzazione dei dati può infatti ridurre il loro valore analitico, compromettendo la capacità dei modelli di IA di fare previsioni accurate e utili. Questo rischio è amplificato dalla possibilità di introdurre bias nei modelli di IA se i dati sintetici non sono ben progettati. I bias possono derivare da una mancanza di variabilità e correlazione nei dati sintetici, portando a risultati distorti, ingannevoli o persino discriminatori.
Per mitigare questi rischi, è fondamentale stabilire degli standard di qualità rigorosi nella generazione dei dati sintetici. Questi standard devono includere metriche per valutare l'accuratezza statistica dei dati sintetici rispetto ai dati reali, nonché metodologie per garantire che i dati generati siano sufficientemente diversi da proteggere la privacy degli individui, senza però perdere le caratteristiche essenziali necessarie per l'analisi.
Inoltre, è importante implementare pratiche di validazione e verifica continue durante il processo di generazione dei dati sintetici. Questo include l'uso di tecniche di benchmark per confrontare le prestazioni dei modelli di IA addestrati con dati sintetici rispetto a quelli addestrati con dati reali. Solo attraverso un approccio metodico e ben strutturato le aziende possono garantire che i dati sintetici siano non solo sicuri, ma anche di alta qualità e accurati, massimizzando così il loro valore per l'analisi e la decisione strategica.

b) Responsabilità

La questione della responsabilità nella generazione di dati sintetici è di cruciale importanza, particolarmente in settori sensibili come la ricerca clinica e l'ambito sanitario, dove l'accuratezza dei dati ha implicazioni significative. Per affrontare adeguatamente questo tema, è necessaria una trasparenza e spiegabilità dell'algoritmo che genera i dati sintetici. Gli algoritmi devono essere spiegabili in modo tale da permettere alla persona di comprendere il percorso che, partendo dai dati reali (input), ha portato alla generazione dei dati sintetici (output). Questo è essenziale per ovviare il problema della "scatola nera", dove il funzionamento interno dell'algoritmo non è comprensibile agli utenti finali.

Sebbene la spiegabilità dell'intero processo che genera i dati sintetici possa essere complessa e, in alcuni casi, irraggiungibile, è fondamentale almeno la spiegazione del procedimento o della formula che permette il funzionamento del modello di generazione. Questo livello di trasparenza non solo aumenterebbe la fiducia degli utenti nei dati sintetici, ma consentirebbe anche una valutazione indipendente della qualità e dell'affidabilità degli algoritmi utilizzati.

Per garantire questa trasparenza, si potrebbe pensare all'implementazione di meccanismi di valutazione della spiegabilità degli algoritmi, possibilmente anche in modalità open source. La disponibilità di tali meccanismi permetterebbe una revisione continua e indipendente da parte di autorità di controllo e della comunità scientifica, contribuendo a migliorare la fiducia nel processo di generazione dei dati sintetici. Tali autorità di controllo potrebbero stabilire linee guida e standard per la spiegabilità, assicurando che gli algoritmi non solo siano tecnicamente validi, ma anche eticamente responsabili.

c) Privacy

La generazione di dati sintetici, intesa come il processo di sintetizzazione, rientra nella definizione di "trattamento" ai sensi dell'art. 4 del GDPR. Pertanto, è un errore comune pensare che i dati sintetici siano automaticamente esenti dalla regolamentazione del GDPR per loro natura. Il GDPR non si applica ai dati completamente sintetici o generati da dataset contenenti esclusivamente dati non personali. Tuttavia, la situazione cambia se questi dati sono stati generati a partire da dati personali e non possono essere considerati "totalmente" anonimi.

Per determinare l'applicabilità del GDPR ai dati sintetici, è fondamentale innanzitutto esaminare la natura dati trattati. Se sono stati generati interamente sulla base di regole predefinite e non derivano da dati reali, il GDPR non si applica, poiché tali non sono collegabili a persone fisiche e non contengono informazioni che possano condurre all'identificazione di individui. Se, invece, i dati sintetici sono stati generati a partire da dati esistenti, è necessario un ulteriore livello di analisi. A questo punto, bisogna determinare se i dati di partenza utilizzati per la generazione siano dati personali. Se i dati di partenza non sono dati personali, il GDPR non si applica. Tuttavia, se i dati originali sono dati personali, è indispensabile considerare la spiegabilità dell'algoritmo utilizzato per generare i dati sintetici.

Se l'algoritmo non è spiegabile, ovvero se non è possibile comprendere come i dati originali abbiano influenzato la generazione dei dati sintetici, allora non è possibile ottenere la ri-identificazione dell'interessato. In questo caso, il GDPR non si applica, ma sorge il problema della "scatola nera": tali dati potrebbero essere inutilizzabili in contesti che richiedono la trasparenza del modello.
Se l'algoritmo è spiegabile, ovvero se è comprensibile il procedimento di sintetizzazione, il passaggio finale è determinare se il processo di spiegabilità identifichi il carattere personale del dato. Se il processo non lo identifica, i dati sintetici possono considerarsi anonimi e il GDPR non si applica. Tuttavia, se il processo di spiegabilità rivela che i dati sintetici contengono informazioni personali, allora i dati sono considerati pseudoanonimizzati e il GDPR si applica.

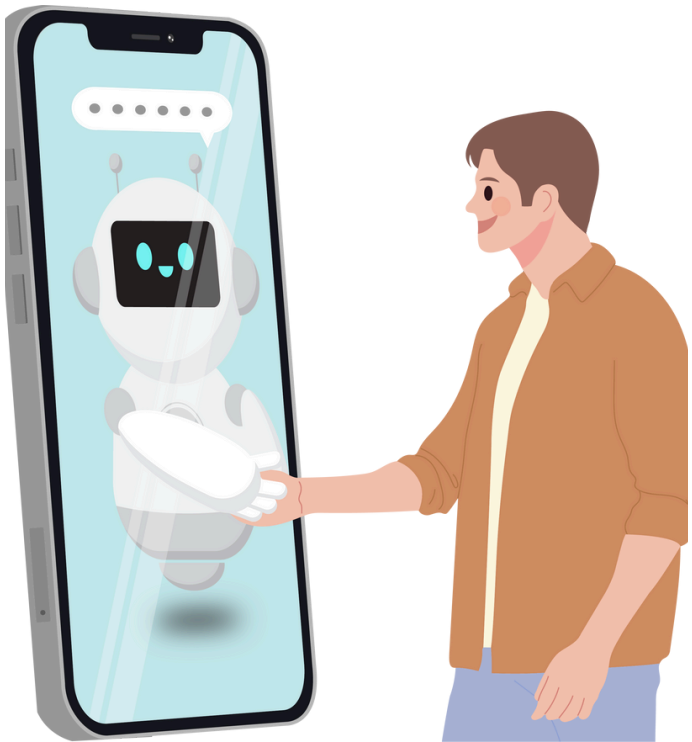
d) Proprietà intellettuale

La natura dei dati sintetici solleva questioni su chi detiene i diritti sui dati stessi. A differenza dei dati personali, che sono soggetti a rigide regolamentazioni legali riguardanti la loro raccolta e utilizzo, i dati sintetici non rappresentano direttamente individui reali. La legge non è ancora chiara su come trattare i dati sintetici in termini di proprietà intellettuale, rendendo la questione di chi detiene i diritti su questi dati un'area grigia.

La generazione di dati sintetici avviene tramite l'uso di software e algoritmi che possono essere protetti da diritti d'autore, brevetti o altri diritti di proprietà intellettuale. È possibile che il proprietario della licenza abbia diritti sulla proprietà intellettuale riguardante il processo di creazione dei dati sintetici. L'uso di algoritmi proprietari comporta il rispetto delle condizioni di licenza del software, che possono imporre restrizioni su come i dati sintetici possono essere utilizzati o condivisi.

Allo stesso tempo, se la generazione dei dati sintetici coinvolge un significativo contributo intellettuale umano, questo può influenzare la determinazione dei diritti di proprietà intellettuale sui dati stessi. È chiaro che, se gli algoritmi impiegati per generare dati sintetici fossero open source, l'uso e la distribuzione dei dati generati sarebbe facilitato.

In ogni caso, le organizzazioni che utilizzano dati sintetici devono considerare attentamente sia le licenze associate ai software utilizzati per la generazione dei dati che alla proprietà dati originali da cui derivano i dati sintetici, specialmente se questi ultimi si basano su dati reali.



Proposte di Policy

Sulla base delle analisi condotte nel capitolo precedente ("Le implicazioni giuridiche dei dati sintetici"), questo capitolo si concentra sulla formulazione di proposte di policy, organizzate in due principali aree di analisi: l'affidabilità e la robustezza dei dati sintetici generati e il loro rapporto con il GDPR.

Le riflessioni precedenti hanno evidenziato questioni cruciali quali la distinzione giuridica tra dati anonimi e dati pseudonimizzati, l'incertezza normativa in tema di privacy e le persistenti ambiguità del legislatore europeo riguardo alle tecniche di anonimizzazione.

Partendo da queste considerazioni, questo capitolo propone due specifiche direttrici di intervento politico e normativo.

Proposta 1:
Certificazione delle modalità e tecniche per la sintesi dei dati

L'articolo 42 del GDPR incoraggia l'adozione di meccanismi di certificazione per garantire la protezione dei dati personali durante le operazioni di trattamento. In linea con questo principio, **si propone l'introduzione di un sistema di certificazione specifico per l'utilizzo di modelli di intelligenza artificiale generativa nella creazione di dati sintetici**. Tale certificazione avrebbe il duplice scopo di salvaguardare i diritti degli interessati e di assicurare la qualità e la conformità dei processi di sintesi.

Il sistema di certificazione potrebbe servire come strumento per verificare la correttezza delle operazioni di sintetizzazione, rivolgendosi sia ai titolari del trattamento (responsabili della decisione di sintetizzare i dati) **sia ai responsabili del trattamento** (sviluppatori e fornitori delle tecnologie di sintesi). Questo approccio consentirebbe di ridurre i rischi legati a potenziali violazioni della privacy e alla perdita di qualità dei dati generati. L'adozione di criteri approvati per la certificazione faciliterebbe, inoltre, la documentazione e la verifica della conformità alle disposizioni del GDPR, agevolando audit e monitoraggi.

Gli standard di certificazione potrebbero includere linee guida pratiche per definire, ad esempio quando un dato può essere considerato sufficientemente anonimo o quando un dato sintetico soddisfa requisiti di robustezza e affidabilità. Tale sistema, potenzialmente strutturato sotto forma di manuali o protocolli, offrirebbe agli sviluppatori un riferimento chiaro e sistematico per garantire la conformità normativa. In questo modo, la valutazione della conformità sarebbe semplificata, riducendo il rischio per i titolari del trattamento di generare dati sintetici inaffidabili o non conformi alle disposizioni in materia di protezione dei dati.

Proposta 2:
Informativa sulla generazione del dato sintetico

Un ulteriore strumento di policy consiste nell'**obbligo**, per i produttori di modelli generativi, **di fornire un'informativa dettagliata** agli utilizzatori. Questa informativa dovrebbe includere:

1 > La **natura dei dati impiegati** nonché delle metodologie utilizzate per testare e validare il sistema.
2 > Una **valutazione dei rischi associati** all'utilizzo del modello.

L'obiettivo principale è garantire trasparenza e consapevolezza da parte degli utilizzatori, promuovendo un uso responsabile dei dati sintetici e delle tecnologie sottostanti. Fornire informazioni dettagliate sul funzionamento del sistema, comprese le sue potenziali vulnerabilità e limiti, consente agli utenti di valutare in modo critico i rischi connessi e di utilizzarlo con maggiore prudenza e responsabilità.
In linea con il principio di trasparenza, sancito dal GDPR, si propone la formulazione di una informativa da fornire agli utilizzatori di dati sintetici e dei relativi modelli di intelligenza artificiale. Essi dovrebbero ricevere un'informativa chiara, completa e comprensibile sul funzionamento dei sistemi generativi, al fine di comprenderne le implicazioni operative, etiche e legali.

In contesti critici, come quello medico, questa trasparenza assume un'importanza cruciale. Ad esempio, se un sistema diagnostico basato su dati sintetici producesse un risultato errato o discriminatorio, l'utilizzatore avrebbe l'obbligo di verificarne l'affidabilità e di contestualizzare i risultati nel quadro clinico. L'affidarsi ciecamente a un algoritmo senza considerare la qualità dei dati sintetici utilizzati potrebbe configurare un caso di negligenza professionale. La riflessione sul rapporto tra uomo e macchina, specialmente in ambiti ad alta responsabilità come quello medico, evidenzia l'urgenza di fornire agli utilizzatori strumenti adeguati per comprendere la natura dei dati e dei modelli utilizzati. Nel caso dei dati sintetici, è fondamentale chiarire i principi adottati per la loro generazione e le metodologie per garantirne l'accuratezza per mettere una valutazione critica dei rischi legati all'utilizzo dei modelli generativi, una **maggiore sicurezza e responsabilità nelle decisioni basate su tali dati**, a beneficio di tutti gli interessati, come i pazienti nel caso clinico, e un **supporto alla tutela legale degli attori coinvolti**, fornendo una base per l'analisi della responsabilità in caso di errori.
In un'epoca in cui le decisioni, anche quelle cruciali, sono sempre più orientate dai dati, è imprescindibile dotare gli utilizzatori di conoscenze approfondite sui sistemi che generano tali dati. **Un'informativa dettagliata sui modelli generativi non solo tutela i diritti degli utilizzatori e degli interessati, ma contribuisce a elevare gli standard etici e legali nella progettazione e nell'impiego di queste tecnologie.**

Conclusioni

L'analisi dei dati sintetici condotta in questo studio evidenzia sia le opportunità strategiche sia le sfide normative che questa tecnologia presenta. I dati sintetici emergono come una soluzione efficace per superare i problemi legati alla disponibilità, alla privacy e alla qualità dei dati raccolti, offrendo un'alternativa preziosa per il training di modelli di IA in diversi settori come la sanità, la finanza e la ricerca scientifica. Essi permettono di mantenere alta la qualità dei risultati e, allo stesso tempo, di rispettare le normative sulla protezione dei dati, in particolare quelle delineate dal GDPR.
Tuttavia, non si possono trascurare le **criticità normative** inerenti all'utilizzo dei dati sintetici. È auspicabile un quadro normativo più chiaro e strutturato, che tenga conto delle peculiarità di questa nuova tipologia di dati.

In particolare, occorre risolvere le **incertezze legate alla loro classificazione all'interno del GDPR** e stabilire principi più definiti riguardo alla proprietà intellettuale, alla trasparenza degli algoritmi di generazione e alla qualità dei dati sintetici stessi.

Parallelamente, **è imprescindibile l'adozione di una governance etica e trasparente**, necessaria per mitigare i rischi legati ai bias e garantire che lo sviluppo dei sistemi di IA basati su dati sintetici sia conforme ai principi di equità, giustizia e spiegabilità. La promozione di strumenti come il bollino digitale per identificare i dati sintetici e un'informativa di trasparenza possono rappresentare passi significativi per costruire un ecosistema affidabile e sicuro per l'uso dei dati sintetici. Inoltre, tali proposte, per il loro basso grado di invasività, rappresenterebbero un esempio di compromesso ideale tra le esigenze regolative menzionate e l'esigenza (spesso manifestata dalle società) di un contesto normativo che favorisca, invece di opprimere, la sperimentazione di tecnologie innovative.

Le proposte di policy esaminate in questo documento offrono un punto di partenza per regolatori e stakeholder che desiderano adottare un approccio bilanciato, capace di favorire l'innovazione tecnologica senza compromettere la protezione dei diritti degli individui e la fiducia pubblica. È necessario un impegno continuo per **monitorare e adeguare le normative** alla rapida evoluzione tecnologica, al fine di garantire che i **dati sintetici** possano essere **utilizzati in modo responsabile e sostenibile nel lungo termine**.

AWARE

www.awarethinktank.it

info@awarethinktank.it