LUISS

Corso di laurea in Giurisprudenza

Cattedra Informatica Giuridica

Data Scraping ed intelligenza artificiale generativa:

l'estensione del principio di accountability nell'adozione di azioni a contrasto per contenere gli effetti dello scraping finalizzato all'addestramento degli algoritmi

Prof. Gianluigi Ciacci

RELATORE

Prof. Filiberto E. Brozzetti

CORRELATORE

Camilla Brandimarti Matr. 144383

CANDIDATO

Anno Accademico 2024/2025

INDICE

INTRODUZIONE

Il diffondersi dell'uso dell'intelligenza artificiale generativa, e la conseguente domanda di modelli sempre più performanti, e prossimi ad una capacità di imitazione perfetta dell'interagire e delle caratteristiche del pensiero umano, ha sollevato interrogativi e perplessità da parte delle Autorità preposte alla tutela dei dati personali. Questo perché, per generare risposte sempre più pertinenti, l'intelligenza artificiale generativa ha bisogno di "nutrirsi" e di allenarsi con il prodotto dell'essenza distintiva dell'essere umano, ovvero i suoi dati. Dati che, sotto forma di pensieri, fotografie, qualifiche, parole riconducibili a stati soggettivi, geolocalizzazioni etc…, vengono giornalmente disseminati (*rectius:* resi pubblici in virtù di una base giuridica e per finalità determinate) nel *web* ed esposti, a pubblica consultazione, attraverso le "vetrine" di piattaforme, siti, pagine pubbliche di *social network,* documenti pubblicamente reperibili e via dicendo. Purtuttavia, le informazioni pubblicamente disponibili non possono essere trattate alla stregua di una *res derelicta*: non possono essere pertanto "prese" ed utilizzate a piacimento[1].

Premesse queste considerazioni, il presente lavoro focalizzerà l'indagine su uno degli aspetti più controversi dell'*iter* che porta al perfezionamento di un modello di IA generativa, ovvero la modalità automatizzata con cui vengono raccolti i dati per il suo addestramento (*web scraping*).

Come illustrato nella contestualizzazione, contenuta nel Capitolo I[2] relativa all'era dei *big data,* e poi successivamente approfondito nel Capitolo III[3], in relazione al funzionamento degli LLMs, il *web scraping*, finalizzato all'addestramento di IA generative, trae origine dalla natura stessa dell'intelligenza artificiale quale tecnologia *data- driven*, e, dunque, indissolubilmente dipendente dai dati come proprio "carburante". Questo bisogno di dati ha delle proporzioni tali da non potersi limitare solo ai dati offerti dai *datalakes,* perché, in questo modo, ai modelli in fase di addestramento mancherebbe quella varietà e quella verosimiglianza imprescindibile, affinché l'algoritmo di un modello soffra il meno possibile di allucinazioni dovute alla scarsa qualità dei dati forniti,

[1] Capitolo I, Paragrafo 3.1
[2] Capitolo I, Paragrafi 1 e 2
[3] Capitolo III, Paragrafo 1

e riesca a garantire il raggiungimento di una capacità di analisi e connessioni statistiche sufficiente a produrre *output* utili e pertinenti.

Nel Capitolo I, dunque, viene proposta una presentazione del fenomeno del *web scraping* e della sua interrelazione con la normativa sulla tutela dei dati personali. Viene poi data particolare attenzione alla questione cruciale e delicata della scelta della base giuridica per chi tratta dati raccolti tramite tecniche di *scraping* e, su tal punto, viene discussa la rilevanza di due casi concernenti l'uso di tale tecnica oggetto di provvedimento da parte del Garante*:* il caso della piattaforma Trovanumeri.com[4] e quello della società americana Cleaview AI.

Nel Capitolo II si procede quindi ad una un'analisi della cornice normativa di riferimento, con l'intenzione di sviluppare parallelamente la contestualizzazione giuridica del *web scraping* e quella dei sistemi di intelligenza artificiale. Il punto d'arrivo del Capitolo II è in realtà nel Capitolo III, nel cui paragrafo 2 si discute dei rischi concreti derivanti dal processo decisionale automatizzato, con la consapevolezza quindi, non solo che l'attività di *web scraping* non autorizzata è un'azione giuridicamente incompatibile con il GDPR, ma, quando diretta all'addestramento di algoritmi di IA generativa, è anche potenzialmente lesiva dei diritti e delle libertà fondamentali del singolo. E questo, proprio in conseguenza dei differenti livelli di rischio, insiti nel suo funzionamento e normati nell'AI Act.

Nel Capitolo III, oltre a concludersi il discorso sulla presentazione-contestualizzazione normativa e discussione dello *scraping* nell'ambito del funzionamento algoritmico, si apre la strada verso la discussione degli strumenti a disposizione del titolare del trattamento per cercare di limitare, seppur mai completamente, l'attività di *scraping*. Nella transizione verso il capitolo finale, viene argomentato[5] il valore della figura del *Data Protection Officer,* specialmente nella sua accezione di professionista avente la capacità di prestare una consulenza puntuale al titolare del trattamento, riguardo la conformità al regolamento e le prospettive di incrementare l'aderenza al principio di *accountability.*

Nel capitolo IV, come prodromo all'illustrazione di alcune misure di natura non-normativa a contrasto dello *scraping*, viene obiettato un ulteriore fattore[6] che rende la

[4] Capitolo I, paragrafo 3.1
[5] Capitolo 3, paragrafo 3
[6] Capitolo IV, paragrafo 1

tutela dei dati personali un'opera ancor più complessa, ovvero la capacità degli algoritmi di IA generativa di fare collegamenti riconducibili ad individui, nonostante l'anonimizzazione dei dati.

In questo contesto pur ricco di sfaccettature, il presente lavoro di tesi sostiene, come soluzione che riflette lo stato dell'arte e della tecnica, il valore dello sforzo attuativo del principio di *accountability*, da parte del titolare del trattamento, che voglia adeguatamente proteggere i dati resi pubblici sulle proprie piattaforme, tramite l'implementazione di misure eterogenee e multilivello.

CAPITOLO PRIMO

INTRODUZIONE AL TEMA ED AL CONTESTO

1. Il presupposto del data *scraping*: l'era dei *big data*

Prima di entrare nel vivo di questo elaborato ed affrontare la domanda di ricerca, è utile fare delle brevi riflessioni seppur in modo non esaustivo, sul presupposto logico del *data scraping*, ovvero i cosiddetti *big data*.

Negli ultimi anni i dati sono diventati uno dei maggiori motori propulsivi dell'evoluzione digitale e socio-economica nel cui vivo ci troviamo proprio ora. Secondo *l'Osservatorio Big Data & Business Analytics della School of Management* [7], in Italia il mercato dei dati vale attualmente 3,42 miliardi di euro. Per l'elaborazione di questa stima, l'Osservatorio ha esaminato i dati di spesa delle aziende italiane relativi agli investimenti fatti in infrastrutture, *software* e servizi di gestione ed analisi dei dati a partire dal 2021, anno in cui, il settore dei dati, aveva un valore di poco superiore ai 2 miliardi di euro. Quel che è emerso dal *report* quindi, è un trend di crescita attestato al 19%. Inoltre, grande polo di interesse ed investimenti è, a partire dalla rapida diffusione del 2023, l'intelligenza artificiale generativa[8].

Serve innanzitutto chiarire che non c'è né una definizione normativa né una che metta d'accorto tutti gli studiosi del settore *digital*. Si può però dire, con approssimativa certezza, a che cosa si riferisca la locuzione "*big data*", ovvero alla raccolta, all'analisi e alla conservazione di vasti e disparati *set* di dati, che in virtù dell'eterogeneità della loro

[7] Data Strategy per la valorizzazione dei Dati: mercato e maturità delle aziende italiane nel 2024 - Osservatori Digital Innovation del Politecnico di Milano-Report disponibile su sito dell'osservatorio: Reperibile su: https://www.osservatori.net/report/big-data-business-analytics/data-strategy-valorizzazione-dati-mercato-maturita-aziende-italiane-2024/. Ultima visita al sito in data 5 novembre 2024

[8] Big data, il mercato italiano vale 3,4 miliardi. Vercellis: "Ora però le aziende devono darsi una strategia" – CorCom.
Reperibile su: https://www.corrierecomunicazioni.it/digital-economy/big-data-il-mercato-italiano-vale-34-miliardi-vercellis-ora-pero-le-aziende-devono-darsi-una-strategia/ Sito consultato in data 18 Dicembre 2024.

provenienza, possono comprendere anche dati di natura personale[9][10]. Inoltre, le loro dimensioni sono tali da poter essere elaborati solo con processi automatizzati, algoritmi e l'ausilio di tecnologie avanzate come l'intelligenza artificiale. Questo perché, benché non esista una soglia dimensionale predefinita per i *big data*, un insieme di dati può essere categorizzato come tale quando gli algoritmi tradizionali di analisi, non riescono ad elaborare un risultato in tempi ragionevoli su un computer con *hardware* di fascia alta[11].
I *big data* sono generati all'interno di un ecosistema che vede l'individuo stesso, in qualità di consumatore di servizi digitali e come fonte generatrice di dati. Stante ciò, le interazioni che una persona ha con la strumentazione tecnologica e le piattaforme online, lasciano delle "impronte digitali" che vengono cedute agli operatori che utilizzano i dati. Di questo *iter* implicito che attende tutte le informazioni che vengono generate, l'individuo ne è sempre più spesso inconsapevole[12]. Gli utenti, inoltre, possono produrre ingenti quantità di dati anche nella loro vita *offline*. Questo vuol dire che le attività svolte dagli individui, seppur in assenza di interazione diretta con un dispositivo elettronico, generano dati che, se adeguatamente processati, forniscono informazioni rilevanti sui comportamenti e le preferenze dei soggetti. Un esempio pratico di ciò sono i dati prodotti dalla geolocalizzazione (quando attivata) sugli *smartphone*, che avviene a prescindere dall'interazione dell'utente con il dispositivo, oppure, quanto registrato da telecamere di sorveglianza può essere elaborato per ricavarne informazioni sui flussi di persone[13].
Nell' era dei *big data* la generazione e l'acquisizione dei dati sono spesso momenti coincidenti, e questo è dovuto alla natura dei dispositivi coinvolti: il cellulare da solo, ad

---

[9] Big data: definizione, benefici e sfide (infografica) | Tematiche | Parlamento europeo (marzo 2023) – Reperibile su: https://www.europarl.europa.eu/pdfs/news/expert/2021/2/story/20210211STO97614/20210211STO97614_it.pdf . Ultima visita al siti in data 11 gennaio 2025-
-Testi approvati - Implicazioni dei Big Data in termini di diritti fondamentali - Martedì 14 marzo 2017 Reperibile su: https://www.europarl.europa.eu/doceo/document/TA-8-2017-0076_IT.html. Ultima visita al sito in data 11 gennaio 2025

[10] Nell'accezione fornita dall'art. 4 del Regolamento (UE) 2016/679 del Parlamento europeo e del Consiglio del 27 aprile 2016, relativo alla protezione delle persone fisiche con riguardo al trattamento dei dati personali, nonché alla libera circolazione di tali dati e che abroga la direttiva 95/46/CE, di seguito anche "GDPR".

[11] Cosa sono i Big Data? | IBM – Reperibile su: https://www.ibm.com/it-it/think/topics/big-data. Ultima visita al sito in data – Ultima visita al sito in data 11 gennaio 2025

[12] "Big data" -Interim report nell'ambito dell'indagine conoscitiva di cui alla delibera n. 217/17/CONS

[13] *"Indagine conoscitiva sui big data"* è il report che ha fatto seguito ai risultati prodotti dall'indagine conoscitiva deliberata dall'Autorità per le garanzie nelle comunicazioni (AGCOM), congiuntamente al Garante per la protezione dati personali e all'Autorità garante della concorrenza e del mercato (AGCM) con la delibera n. 217/17/CONS.

esempio, come parte del suo sistema operativo, dispone di una varietà di dispositivi di *input*, come i sensori per gli usi più diversi, in un unico "luogo" (cioè il cellulare stesso), per di più connesso ad internet e che accompagna un utente lungo quasi tutta la sua quotidianità. Il concatenarsi delle implicazioni non finisce qui, perché le applicazioni pre-installate o scaricate dall'utente sono un ulteriore veicolo per l'acquisizione di dati.

Per quel che riguarda la navigazione *web*, invece, c'è lo specifico sistema di tracciamento basato sui *cookie*, che, detto in sintesi, sono *file* di testo atti a raccogliere le preferenze e le informazioni del visitatore del sito, da utilizzare in seguito per attività di profilazione. Profilazione soggetta, peraltro, ad aggiornamento dopo ogni visita al sito stesso[14]. Tuttavia, le conseguenze più sostanziose, e per certi versi difficilmente gestibili, che il momento acquisitivo porta con sé, derivano dalla circostanza della prestazione di un valido consenso a norma delle basi giuridiche stabilite all'articolo 6 del regolamento (UE) 2016/679 da parte di ciascun utente. Questo perché: se da un lato è vero che l'adeguatezza delle informazioni offerte al pubblico interessato, e la trasparenza, sono elementi fondamentali per proteggere i diritti individuali e perché no, costruirsi una pubblica immagine affidabile[15], è poi altrettanto vero che sia nell'ambito individuale che commerciale c'è ancora tanta chiarezza da conquistare e sensibilizzazione da fare sui rischi ed i diritti nel digitale. Infatti, sempre secondo i risultati della ricerca dell'Osservatorio Big Data del Politecnico di Milano[16] è vero che 7 grandi aziende su 10 hanno avviato almeno un progetto di analisi avanzata dei dati ma solo 4 su 10 hanno definito una strategia di valorizzazione dei dati, e 2 su 10 hanno inserito nel loro organico un *Data Protection Officer (DPO)*, figura di specifica esperienza nell'affrontare le nuove sfide e cogliere le occasioni offerte dall'era dei *big data*.

L'ultimo anello per chiudere questa premessa, è evidenziare in che modo i *big data* costituiscono il presupposto del *data scraping*. Questa pratica, del tutto legittima in sé, si

[14] I *cookie* sono uno strumento acquisitivo per così dire "residuale" e le informazioni da essi raccolte hanno una portata limitata rispetto a quelli generati da un utente a seguito di autenticazione sul sito. *"Cosa sono i Cookie, a cosa servono e come funzionano"*-Osservatorio Internet Media del Politecnico di Milano. Reperibile su: https://blog.osservatori.net/it_it/cookie-cosa-sono-come-funzionano -Ultima visita al sito 10 gennaio 2024.

[15] A tal proposito mi viene in mente la pronuncia della CGUE C-40/17, Fashion ID GmbH & Co. KG v Verbraucherzentrale NRW e V.

[16] Data Strategy per la valorizzazione dei Dati: mercato e maturità delle aziende italiane nel 2024 - Osservatori Digital Innovation del Politecnico di Milano- Cfr nota n. 7

riferisce alla sistematica raccolta automatizzata di dati da *internet*[17] (ma è ben possibile anche da *data base, file* e comunicazioni elettroniche). Il fine ultimo di tale attività si dirama in due direzioni: uno è ricavare informazioni utilizzabili a scopo di analisi, l'altro, come verrà più approfonditamente discusso nel capitolo terzo, è la materia prima necessaria al processo di sviluppo degli algoritmi sottesi al funzionamento dell'intelligenza artificiale.

1.1 Le caratteristiche dei *big data*

Sebbene, come affermato poc'anzi, non ci sia una definizione ufficiale di *big data*, un primo tentativo di definizione può essere fatto risalire alla società di consulenza strategica *Gartner*, la quale nel 2001 coniò un assunto che al giorno d'oggi viene utilizzato come sintesi delle principali caratteristiche dei *big data*: "*I big data sono risorse informative ad alto volume, ad alta velocità e/o ad alta varietà che richiedono tecnologie e metodi analitici specifici ed efficienti per poterne ricavare conoscenza e valore*"[18]. Dall'apparente semplicità di questa definizione si evince, come già nei primi anni Duemila si percepisse la necessità di un nuovo approccio alle sempre più grandi quantità di dati prodotte dallo sviluppo computazionale di quegli anni[19].

Le cosiddette "tre V" che caratterizzano i big data sono qualità strettamente relazionate tra loro, di seguito la breve sintesi di ognuna darà conto anche della ragione.[20]

- Il volume è un chiaro riferimento all'ammontare dei dati che fanno parte della datasfera. Come emerso dalle statistiche riportate all'inizio di questo paragrafo[21] la

[17] Daniel J. Solove e Woodrow Hartzog, *The Great Scrape: The Clash Between Scraping and Privacy* (2024). Pg.7

[18] *"Big data is high-volume, high-velocity and/or high-variety information assets that demand cost-effective, innovative forms of information processing that enable enhanced insight, decision making, and process automation."*
Definition of Big Data - IT Glossary | Gartner
Reperibile su: https://www.gartner.com/en/information-technology/glossary/big-data -Ultima visita al sito 10 gennaio 2025.

[19] "Big data" -Interim report nell'ambito dell'indagine conoscitiva di cui alla delibera n. 217/17/CONS

[20] *Ibidem*

[21] Data Strategy per la valorizzazione dei Dati: mercato e maturità delle aziende italiane nel 2024 - Osservatori Digital Innovation del Politecnico di Milano- Cfr nota n.7

crescente produzione di dati[22] ha stimolato gli investimenti volti alla loro gestione, elaborazione, ma anche archiviazione e conservazione.

- La varietà dei dati indica l'eterogeneità che ne caratterizza la grande quantità. I big data sono dunque vari in fatto di fonte di provenienza, formato, modo di acquisizione e anche nella rappresentazione ed analisi dei dati una volta raccolti.
- La velocità con cui crescenti masse di dati sempre più eterogenee aumenta, è composta da due fattori: il primo è la velocità con cui vengono raccolti i dati nelle banche; il secondo è la velocità con cui vengono elaborati per essere utilizzati per processi decisionali.

Resta ora un'ultima distinzione da fare in fatto di caratteristiche formali dei dati. I *big data* possono essere classificati secondo due conformazioni principali: dati strutturati e non.

I dati strutturati utilizzano schemi o formati predefiniti. In genere si tratta di tabelle in cui ogni dato è associato ad un campo specifico. In conseguenza di questa uniformità e coerenza nell'organizzazione, sono facilmente analizzabili ed elaborabili. Un esempio di dati strutturati possono essere i *file excel,* i *database* SQL (ovvero linguaggio di interrogazione strutturato) i quali hanno l'aspetto di tabelle con colonne righe, oppure ancora i *record* transazionali di acquisti *online*.[23]

I dati non strutturati, invece, mancano di una organicità formale, possono avere vari formati e dimensioni. Un esempio possono essere i *file* multimediali. L'identificazione di informazioni e relazioni rilevanti all'interno di una massa di dati non strutturati, richiede l'uso di algoritmi avanzati di intelligenza artificiale[24].

Nonostante questa fondamentale divisione che riguarda i dati al momento della loro generazione, a tutti i dati raccolti può essere dato, in fase di memorizzazione, un tipo di organizzazione uniforme all'interno dei *data lake.* All'interno di queste aree digitali i dati

[22] L'International Data Corporation ha previsto che nel 2025, l'aggregato di big data presente nel mondo avrà la dimensione di 163 zettabyte. Ovvero dieci volte superiore la quantità prodotta nel 2017. International Data Corporation -IDC Study_infographic_2017.
Reperibile su: https://www.seagate.com/files/www-content/our-story/trends/files/data-age-2025-infographic-2017.pdf -Sito consultato in data 10 gennaio 2025.

[23] Amazon Web Services- Cosa sono i dati strutturati? - Spiegazione dei dati strutturati - AWS-
Reperibile su: https://aws.amazon.com/it/what-is/structured-data/ -Sito visitato in data 10 gennaio 2025

[24] Oracle Cloud Italia- Tipi di dati strutturati e non strutturati | Oracle Italia
Reperibile su: https://www.oracle.com/it/big-data/structured-vs-unstructured-data/ -Sito consultato in data 10 gennaio 2025

diventano “grezzi”, ossia vengono privati di qualsiasi struttura intrinseca per essere più proficuamente elaborati[25].

2. Generalità e finalità di un fenomeno interdisciplinare

L’espressione “*web scraping”* si riferisce[26] al processo di estrazione (letteralmente “raschiamento”) di dati perlopiù non strutturati, dal *web*, con l’obiettivo di convertirli in dati strutturati. La standardizzazione dei dati avviene, in pratica, prima tramite la neutralizzazione in dati grezzi e poi la conversione nel formato più adatto allo scopo.

Vale infine la pena precisare che alcuni tipi di dati non strutturati come quelli generati dai sensori dell’*Internet of Things* o le trascrizioni delle *chat*, possono essere utilizzati anche senza averli prima archiviati in formato strutturato, questo perché tale procedura non sarebbe efficiente in termini di dispendio di risorse[27].

Come preciseremo più avanti, lo *scraping* è nato come tecnica automatizzata per organizzare in modo efficace grandi quantità di informazioni (su piccole quantità di dati può anche essere svolta manualmente) dalla natura disomogenea o semplicemente caotica, e allo stato attuale della tecnologia ha molte applicazioni utili ed è spesso utilizzato al servizio dell'interesse pubblico. Tuttavia, molti usi diffusi dello *scraping* dei dati avvantaggiano sia gli *host* di dati che gli *scraper*. I servizi di *scraping* consentono agli utenti di trovare più facilmente le informazioni che cercano. I giornalisti utilizzano la tecnologia di *scraping* per raccogliere e analizzare enormi quantità di dati statistici. Gli studiosi impiegano la tecnologia *scraping* per la loro ricerca accademica. Nel caso dei motori di metaricerca, lo *scraping* dei dati consente loro di estendere la raccolta di dati e informazioni e di confrontare, ad esempio, le informazioni sui prodotti e sui prezzi provenienti da varie fonti, esentando gli utenti dalla visita e dal controllo di più pagine *web* o motori di ricerca. Ulteriori possibili usi della tecnologia includono il monitoraggio della reputazione delle aziende o l'aggregazione di notizie e altri contenuti su siti *web*.

---

[25] Ibidem

[26] Khder, M. A. Web scraping or web crawling: State of art, techniques, approaches and application. *International Journal of Advances in Soft Computing & Its Applications*, *13*(3). (2021).

[27] Amazon Web Services- Dati strutturati e dati non strutturati: differenza tra dati collezionabili - AWS- Reperibile su: https://aws.amazon.com/it/compare/the-difference-between-structured-data-and-unstructured-data/ -Sito consultato in data 10 gennaio 2025

Nonostante queste applicazioni vantaggiose, la tecnologia di *scraping* può essere utilizzata in modo malevolo e per scopi dannosi, come lo *spamming* di account e-mail, l'arresto anomalo del sito *web* o comunque l'illecito sfruttamento dei dati raccolti. Lo *scraping* può anche essere parassitario: gli *scraper* possono trarre vantaggio dall'esclusione o dal danneggiamento degli *host* di dati; possono ridurre le entrate di un sito *web* ripubblicando i dati raccolti senza chiedere agli utenti di visualizzare annunci pubblicitari di supporto. Inoltre, possono appropriarsi di entrate, visitatori e clienti prendendo i contenuti da un altro *host* di dati. Lo *scraping* può anche incidere sulla protezione dei diritti fondamentali a causa della raccolta di informazioni di identificazione personale con implicazioni per la riservatezza dei soggetti coinvolti[28].

Lo *scraping* quindi, che nella sua essenza più pura, rimane un'azione meccanica di "copia e incolla", nel contesto dei *big data* si arricchisce dell'obiettivo di ricavare nuove informazioni dai dati raschiati, e inoltre, acquisisce una complessità che deriva dai molteplici settori in cui viene usato. Le imprese possono giovare dell'estrazione mirata di informazioni dal *web* ricavandone materiale utile ad indirizzare la strategia di mercato. Altri settori in cui lo *scraping* viene utilizzato assiduamente sono la consulenza, la ricerca accademica, la ricerca di mercato, *sentiment analysis*[29][30] e l'addestramento di algoritmi di *machine learning* (l'argomento verrà estensivamente trattato in seguito). A livello macroscopico, i metadati di quasi tutti i siti *web* vengono costantemente raschiati per far sì che i motori di ricerca servano effettivamente alla loro funzione.

A tal proposito è utile sin da ora fare una distinzione concettuale tra lo *scraping* e il *crawling*. Quest'ultimo, operato da software chiamati *crawlers* o *bot* (diminutivo per il più comune robot) sono di uso regolare e quotidiano sul *web*: basti pensare che sostengono l'operato dei motori di ricerca come Google[31], i quali se ne servono per

---

[28] U. Pagallo, J. Ciani Sciolla *Anatomy of web data scraping: ethics, standards, and the troubles of the law* – European journal of Privacy Law and Technologies- 2023

[29] È una tecnica che unisce l'uso del linguaggio naturale a quello dell'intelligenza artificiale e permette di carpire le opinioni degli utenti analizzando il contenuto emotivo dei testi.
Liu, Bing-*Sentiment Analysis: Mining Opinions, Sentiments, and Emotions*.-Cambridge University Press, 2020. P. 1–17.

[30] S. D. S. Sirisuriya, *"Importance of Web Scraping as a Data Source for Machine Learning Algorithms - Review,"- IEEE* 17th International Conference on Industrial and Information Systems (ICIIS)- Peradeniya, Sri Lanka, 2023, pp. 134-139.

[31] "La ricerca di Google ha tre fasi :1. Scansione. Fase in cui i crawler scaricano testi, immagini e video da ogni pagina trovata. 2. Indicizzazione. Tutti I dati scaricati vengono analizzati e memorizzati nell'indice/database di Google. 3. Pubblicazione dei dati di ricerca.

scansionare, analizzare e salvare le informazioni riguardanti i siti con lo scopo di indicizzarli e fornire all'utente risultati ordinati e che siano il più possibili pertinenti alla sua ricerca.

Tornando allo *scraping*, esso può essere effettuato alternativamente con *software* che richiedono competenze di programmazione o meno. In particolare, e per certi versi anche piuttosto discutibilmente, sul *web* sono facilmente reperibili strumenti per effettuare *scraping* cosiddetto "visivo", cioè tramite *software* nei quali si può evidenziare il tipo di informazione che si intende raccogliere. Il linguaggio di programmazione più utilizzato per lo *scraping* è Python, e questo è dovuto alla sua praticità nel gestire processi complessi. Inoltre, grazie alla sua popolarità, sono state sviluppate molte librerie di *script*[32] in Python per facilitare l'attività degli *scraper*.

Il processo di *scraping* si può suddividere in due fasi sequenziali: la prima è l'acquisizione delle informazioni dal *web*, la seconda è l'estrazione delle informazioni *target* dai dati acquisiti. In particolare, un programma di web *scraping* inizia componendo una richiesta HTTP[33] per acquisire le informazioni da un sito *web* mirato. Questa richiesta, una volta formattata in un URL, sarà pronta per essere ricevuta ed elaborata da sito *web* di destinazione, il quale recupererà le informazioni richieste e le invierà al programma di *scraping*. Dopo aver scaricato i dati, il processo di estrazione continua ad analizzare, riformattare e organizzare i dati in modo strutturato. A questo punto, è utile precisare che i programmi di *scraping* sono formati da due moduli: uno per comporre una richiesta HTTP, e un altro per analizzare ed estrarre informazioni dal codice HTML grezzo[34].

Le tecniche di estrazione dati dal *web* si dividono in due principali approcci, benché uno dei due risulti decisamente desueto e foriero di un alto tasso di imprecisioni ed errori: da un lato c'è l'estrazione manuale, dall'altra ci sono le tecniche automatizzate. L'estrazione manuale, verosimilmente praticabile con quantità limitate di dati, rappresenta, come si

---

In-Depth Guide to How Google Search Works | Google Search Central | Documentation | Google for Developers (febbraio 2025).
Reperibile su: https://aws.amazon.com/it/compare/the-difference-between-structured-data-and-unstructured-data/ Ultima visita al sito in data 19 febbraio 2025

[32] Sono dei tipi specifici di programmi di solito abbastanza semplici oppure anche file di testo, che sono destinati ad essere portati a termine da un altro programma. Sono utilizzati per automatizzare dei compiti o eseguire operazioni specifiche.
Treccani enciclopedia online- Reperibile su: https://www.treccani.it/enciclopedia/script/. Ultima visita al sito in data 19 febbraio 2025

[33] Acronimo per *"Hypertext Transfer Protocol"*. È per l'appunto un protocollo che regola la comunicazione e le richieste tra il server ed il client.

[34] *WebScraping* -Bo Zhao College of Earth, Ocean, and Atmospheric Sciences, Oregon State University.

accennava poco sopra[35] un po' il nocciolo dell'idea di *scraping*: che banalmente è l'azione meccanica e ripetitiva del copiare e incollare delle informazioni.

I metodi di *scraping* automatizzato invece, sono programmi che hanno varie caratteristiche e vari livelli di conoscenze tecniche necessarie per poterli utilizzare. Alcuni vengono creati per riconoscere automaticamente la struttura dei dati di una pagina, per fornire un'interfaccia grafica che dispensa dalla necessità di scrivere codice, identificare e convertire i dati in formato omogeneo, o finanche fornire delle funzioni per simulare i comportamenti di navigazione che avrebbe un utente umano. Per consentire ai non programmatori di raccogliere contenuti *web* ci sono dei *crawler* che, grazie alla semplicità dell'interfaccia e all'automazione delle operazioni, consentono l'estrazione di dati dal *web* senza dover scrivere una singola riga di codice. Tali metodi sono nati per raccogliere in modo veloce ed efficace sostanziosi volumi di dati. Tra le tecniche di *scraping* automatizzato ce ne sono due principali ed entrambe lavorano sull'HTML delle pagine *web* che analizzano.

La prima è l'analisi HTML[36] che consiste nell'analisi e nell'interpretazione del contenuto e dell'organizzazione dell'HTML al fine di estrarre dati rilevanti da una pagina *web*. Questo tipo di analisi è tipicamente usata per estrarre componenti specifici come testo, immagini, collegamenti od informazioni strutturate. Durante il procedimento, il codice HTML viene sezionato e, grazie agli strumenti di *scraping* e alle librerie di programmazione, vengono individuati i componenti in base ai *tag*, agli attributi e alle relazioni gerarchiche tra gli elementi. I dati estratti sono infine soggetti ad ulteriore elaborazione, trasformazione e archiviazione in un formato strutturato come il già citato CSV, JSON o XML[37].

---

[35] Si veda Pg. 14

[36] È l'acronimo di "Hypertext Markup Language" e rappresenta uno dei linguaggi di markup basato su testo più utilizzati al mondo nel web design. Non è un linguaggio di programmazione perché tramite di esso non si possono creare né algoritmi né compiti o condizioni perché non dispone di comandi. Viene appunto utilizzato nel web design per descrivere la struttura di una pagina web con l'uso di una sintassi basata su testo.
Boolean | HTML: cos'è, come funziona e a cosa serve | Boolean Blog
Reperibile su: https://boolean.careers/blog/html-cose-come-funziona-e-a-cosa-serve .Ultima visita al sito in data 19 febbraio 2025

[37] GeeksforGeeks -Introduction to Web Scraping- Reperibile su: https://www.geeksforgeeks.org/introduction-to-web-scraping/ Ultima visita al sito in data !5 gennaio 2025

Un'altra tecnica altrettanto diffusa è l'analisi DOM, acronimo per *Document Object Model* [38]. Data la natura modificabile di questo tipo di rappresentazione di contenuti *web*, gli strumenti di *scraping* e le librerie di programmazione possono modificarne e prelevarne i dati. Per poter fare ciò, gli strumenti di *scraping* devo passare in rassegna l'albero logico tipico del DOM, basandosi su attributi e relazioni gerarchiche tra elementi, al fine di identificare i dati rilevanti.

Anche se il *data scraping* è diventato una pratica che ha acquisito popolarità, destato interesse e sollevato dibattiti solo negli ultimi anni, i suoi albori sono decisamente meno recenti. Difatti possono essere datati addirittura ai primi anni '90, poco dopo la nascita del *World Wide Web* (d'ora in poi WWW), momento in cui il dottor Mattew Gray creò il primo *web crawler*, chiamato *Wanderer*, durante il suo primo periodo di impiego presso il *Massachusetts Institute of Technology*. *Wanderer* fu ideato come strumento per misurare la grandezza dell'allora giovanissimo WWW. Nel 1993 fu anche sperimentato per indicizzare. Più avanti nello stesso anno Jonathan Fletcher, amministratore di sistema presso l'Università di Stirling in Scozia, ideò un motore di ricerca basato su crawler chiamato *Jumpstation*, il quale era in grado, tramite una ricerca lineare, di indicizzare le pagine *web* in base ai loro titoli e fornire risultati in URL. Questo progetto inoltre utilizzò per la prima volta la funzione di ricerca per parole chiave (seppur senza offrire alcuna classificazione dei risultati), che ancora oggi sta all'essenza del motore di ricerca Google. Nel 2004 poi uscì il primo vero e proprio software pensato per estrarre dati. Il programma chiamato *Beautifulsoup* è tutt'ora una grande libreria di *script* ed algoritmi pronti all'uso pensati per chi programma con Python, per aiutare a capire la struttura ed analizzare il contenuto di pagine e file in HTML e XML[39]. Questo strumento però è fruibile solamente da chi sappia programmare, ed è proprio per ovviare a questo scoglio che i *software* di *scraping* contemporaneo, oltre che consentire di evidenziare manualmente le informazioni che si desiderano estrarre, e riportarle in un foglio di calcolo o un *database*

[38] Costituisce una forma di rappresentazione, una interfaccia dei dati relativi agli oggetti che formano la struttura ed il contenuto di una pagina web. In questo modo la pagina è suscettibile di essere manipolata nella struttura, nello stile, e nel contenuto.

[39] Acronimo di *Extensible Markup Language*. È, come l'HTML, un linguaggio di markup comprensibile all'essere umano, ma la sua funzione è di descrivere la struttura dei dati di una pagina.

Excel,[40][41] possono anche comprendere funzioni di riconoscimento automatico della struttura dei dati di una pagina *web* od offrire un'interfaccia di registrazione che dispensi lo *scraper* dalla necessità di scrivere il codice di *scraping*.

3. Il *data scraping* e il diritto alla tutela dei dati: due prospettive principali ed un obiettivo comune

Concentrandoci sul contesto europeo, dal punto di vista giuridico non troviamo una legge generale che disciplini specificamente lo *scraping*. Probabilmente complice di ciò è la molteplicità di applicazioni di questo strumento e la sua neutralità tecnica.

Il Regolamento UE 2016/679 (d'ora in avanti GDPR pone regole e limiti che disciplinano il trattamento dei dati personali e conferisce, seppur entro certa misura, dignità di base giuridica all'interesse legittimo dei titolari e dei responsabili del trattamento dei dati. Nell'ambito del GDPR, chi fa *scraping* si vede teoricamente tenuto a conformarsi ad oneri precisi e relativamente stringenti sulla liceità del proprio operato. Se il modo più semplice per trattare dei dati è quello di ottenere dall'interessato un consenso che sia informato e dunque prestato per una o più specifiche finalità, l'articolo 6 del GDPR offre, alle lettere da b ad f, altre cinque basi giuridiche di liceità del trattamento, che sono: l'esecuzione di un contatto di cui l'interessato è parte; l'adempimento di un obbligo legale del titolare; salvaguardare gli interessi vitali di una persona fisica; l'esecuzione di un compito di interesse pubblico da parte del titolare del trattamento; ed infine come accennato sopra, un interesse legittimo del titolare o di terzi (ove però non prevalgano i diritti e le libertà fondamentali dell'interessato).

A questo punto, il concetto di interesse legittimo merita un breve approfondimento sotto la scorta di quanto chiarito dalle recenti linee guida sul punto, adottate dall'EDPB in

[40] Scraping Robot-*"Web Scraping History: The Origins of Web Scraping"* (2022). – https://perma.cc/K2K7-7H5H?type=standard - Data ultima visita al sito 15 gennaio 2025.
[41] Reilly, Casey. "*The implications of data scraping: it benefits big business, but what does it mean for you?"* Journal of high technology law : a student publication of Suffolk University Law School. 24.1 (2023).

Ottobre 2024[42]. Innanzitutto, l'interesse legittimo, di cui all'art. 6(1)(f)[43], gode della stessa rispettabilità delle altre cinque basi giuridiche. Questo implica che non possa essere utilizzata come "ripiego" in situazioni in cui le altre basi giuridiche non si applicherebbero e nemmeno può essere indebitamente esteso il suo ambito applicativo per eludere specifici requisiti giuridici o perché erroneamente considerata meno vincolante rispetto alle altre basi.

Affinché il trattamento possa poggiare sull'interesse legittimo, l'EDPB pone tre condizioni cumulative da soddisfare: in primo luogo, il titolare deve perseguire un interesse legittimo proprio o di un terzo; in secondo luogo, i dati personali possono essere trattati solo nella misura in cui ciò sia necessario al fine di perseguire l'interesse legittimo; e, in terzo luogo, gli interessi o le libertà e i diritti fondamentali degli interessati non devono prevalere sugli interessi legittimi del titolare del trattamento o di terzi.

Relativamente alla prima condizione da soddisfare, pare opportuno precisare che non tutti gli interessi consentono ad un titolare di invocare l'art. 6(1)(f). L'interesse da perseguire, difatti, deve poter essere qualificabile come "legittimo". Tuttavia, manca sia un elenco esaustivo di interessi che possano essere considerati legittimi cui fare riferimento, sia una definizione di tale nozione nel GDPR. Stante ciò, come affermato dalla CGUE nella senza "*SHUFA Holding*"[44], un'ampia gamma di interessi può, in linea di principio, essere considerata legittima. Alcuni esempi sono stati forniti del GDPR ai Considerando 47 e 49[45] e dalla giurisprudenza della CGUE[46]. D'altro canto, nelle sopracitate linee guida

---

[42] Guidelines 1/2024 on processing of personal data based on Article 6(1)(f) GDPR -Version 1.0 - Adopted on 8 October 2024. Pg. 4-19

[43] Art. 6(1)(f) GDPR*: "il trattamento è necessario per il perseguimento del legittimo interesse del titolare del trattamento o di terzi, a condizione che non prevalgano gli interessi o i diritti e le libertà fondamentali dell'interessato che richiedono la protezione dei dati personali, in particolare se l'interessato è un minore".*

[44] CJEU, judgment of 7 December 2023, Joined Cases C-26/22 and C-64/22, SCHUFA Holding (Libération de reliquat de dette), para. 76.

[45] Considerando 47: "[…] *Ad esempio, potrebbero sussistere tali legittimi interessi quando esista una relazione pertinente e appropriata tra l'interessato e il titolare del trattamento, ad esempio quando l'interessato è un cliente o è alle dipendenze del titolare del trattamento.* […] *Costituisce parimenti legittimo interesse del titolare del trattamento interessato trattare dati personali strettamente necessari a fini di prevenzione delle frodi. Può essere considerato legittimo interesse trattare dati personali per finalità di marketing diretto."*

Considerando 49: "[…] *Costituisce parimenti legittimo interesse del titolare del trattamento interessato trattare dati personali strettamente necessari a fini di prevenzione delle frodi. Può essere considerato legittimo interesse trattare dati personali per finalità di marketing diretto.".*

[46] Avere accesso alle informazioni online: CJEU, judgment of 13 May 2014, Case C-131/12, Google Spain and Google, para. 81. -Garantire il funzionamento continuo dei siti web accessibili al pubblico:

dell'EDPB sono contenuti, invece tre criteri cumulativi da considerare per qualificare un interesse come legittimo:

a) esso è legittimo, ossia non contrario al diritto comunitario o nazionale;
b) l'interesse è articolato in modo chiaro e preciso. Il perimetro dell'interesse legittimo perseguito deve essere chiaramente identificato al fine di garantire che esso sia adeguatamente bilanciato con gli interessi o i diritti e le libertà fondamentali dell'interessato;
c) l'interesse è reale e presente, e non speculativo. Come chiarito dalla CGUE nel caso "Asociaţia de Proprietari"[47], l'interesse legittimo deve essere presente ed effettivo alla data del trattamento dei dati.

Riflettendo però sul contesto e sulla natura dello *scraping* di dati reperiti passando in rassegna il *web*, emerge come, in realtà, quello della scelta della base giuridica sia, per chi tratta dati raccolti tramite tecniche di *scraping*, un momento cruciale e delicato.

Questo perché, in primo luogo, i dati pubblici sono beni di cui non si può disporre in modo incompatibile dalla finalità in virtù della quale sono stati trattati.

In secondo luogo, perché il GDPR prevede un sistema sanzionatorio pronto a far fronte alle infrazioni.

Per parlare di come il diritto alla tutela dei dati personali si applica all'attività di *scraping* compiuta sul *web*, conviene iniziare distinguendo i soggetti che fanno parte della scena. Il fenomeno dello *scraping* coinvolge soggetti portatori di interessi differenti e parzialmente contrastanti come possono essere:

a) gli interessati, ovvero le persone fisiche identificate o identificabili attraverso i dati;[48]

CJEU, judgment of 19 October 2016, Case C-582/14, Breyer. Para 60.- Ottenere le informazioni personali di una persona che ha danneggiato la proprietà di qualcuno al fine di citarla in giudizio per danni: CJEU, judgment of 4 May 2017, Case C-13/16, Rīgas satiksme, para. 29.

[47] CJEU, judgment of 11 December 2019, Case C-708/18, Asociaţia de Proprietari bloc M5A-Scara, para. 44.

[48] Art. 4(1) GDPR: "una persona fisica identificata o identificabile («interessato»); si considera identificabile la persona fisica che può essere identificata, direttamente o indirettamente, con particolare riferimento a un identificativo come il nome, un numero di identificazione, dati relativi all'ubicazione, un identificativo online o a uno o più elementi caratteristici della sua identità fisica, fisiologica, genetica, psichica, economica, culturale o sociale".

b) i titolari del trattamento, quali persone fisiche o giuridiche, autorità pubbliche, servizi o organismi che, singolarmente o insieme ad altri, determinano le finalità e i mezzi del trattamento di dati personali;[49]
c) i responsabili del trattamento, ovvero, persone fisiche o giuridiche, autorità pubbliche, servizi od organismi che trattano dati personali per conto del titolare del trattamento[50].

Un titolare del trattamento potrebbe essere il proprietario di un sito *web*, una piattaforma digitale, un'applicazione, che definisca le finalità, mezzi e misure di sicurezza con cui integrare il trattamento. Il responsabile invece in virtù della sua posizione subordinata ha interesse nei ricavi e nel frutto dell'utilizzo dei dati sotto il suo controllo.

Gli *scraper*, infine sono figure intermedie e nebulose in questo scenario, e che ben potrebbero coincidere con lo stesso titolare del trattamento.

Le figure di cui ho accennato, oltre a rappresentare delle posizioni rilavanti a livello concettuale e giuridico, sono all'essenza del conflitto tra interesse pubblico e privato in materia di dati, che la diffusione dello *scraping* per scopi commerciali ha innescato[51].

A livello europeo, la posizione delle Autorità in materia di tutela dati è abbastanza omogenea: lo *scraping* è illecito se coinvolge dati personali ri-utilizzati senza un'adeguata base giuridica. Se invece riguarda dati di natura non personale resta un'attività lecita, dalla quale tuttavia ci si può proteggere con una serie di azioni a contrasto, ma che comunque, dipendentemente dall'uso che viene fatto dei dati, potrebbe portare a compromissione incontrollata dei dati specialmente durante la generazione dei contenuti da parte delle IA.

Una delle prime autorità ad aver pubblicato una guida su come effettuare uno *scraping* lecito per scopi di marketing diretto anche in presenza di dati personali è stato il Garante francese (CNIL) nell'aprile del 2020. Secondo la CNIL sebbene le informazioni di contatto possano essere disponibili su siti *web* accessibili al pubblico, le persone che hanno pubblicato le informazioni non possono ragionevolmente aspettarsi che vengano

[49] Art. 4 (7) GDPR: "«titolare del trattamento»: la persona fisica o giuridica, l'autorità pubblica, il servizio o altro organismo che, singolarmente o insieme ad altri, determina le finalità e i mezzi del trattamento di dati personali. quando le finalità e i mezzi di tale trattamento sono determinati dal diritto dell'Unione o degli Stati membri, il titolare del trattamento o i criteri specifici applicabili alla sua designazione possono essere stabiliti dal diritto dell'Unione o degli Stati membri".

[50] Art. 4(8) GDPR: "la persona fisica o giuridica, l'autorità pubblica, il servizio o altro organismo che tratta dati personali per conto del titolare del trattamento".

[51] Fei L. *A comparative study on public interest considerations in data scraping dispute*. International Journal of Law in Context. Cambridge University Press. 2024

raschiate per finalità a loro sconosciute. Pertanto, tali dati personali non possono essere riutilizzati per il *marketing* senza il consenso dell'interessato, consenso che deve essere ottenuto prima di qualsiasi riutilizzo del dato stesso. La CNIL chiarisce che l'accettazione dei termini di servizio, o finanche il consenso a ricevere comunicazioni di marketing, non è sufficiente a legittimare lo *scraping*, in quanto costituirebbe un consenso non sufficientemente specifico.

Una deliberazione interessante proviene parimenti dalla CNIL ed è datata 15 Dicembre 2022[52].

In tale occasione la CNIL fu investita di una richiesta di parere su un progetto di decreto relativo all'istituzione, presso il Ministero dell'Economia e delle Finanze, all'interno della Direzione Generale per la Concorrenza, il Consumo e il Controllo delle Frodi (DGCCRF) di un sistema di trattamento dei dati personali denominato *"Polygraphe*", basato proprio sull'attività di *web scraping*. La finalità di questo trattamento era quella di consentire agli agenti della DGCCRF di verificare che le opinioni pubblicate online dai consumatori su piattaforme come *"Google Maps"* o *"Tripadvisor"* fossero veritiere e perciò contribuire nello sventare le false recensioni dei consumatori, costituenti una pratica commerciale ingannevole soggetta a sanzioni amministrative e/o penali. Lo schema di decreto prevedeva, infatti, la possibilità per l'amministrazione di raccogliere e poi analizzare, mediante trattamento automatizzato, tutti i contenuti liberamente accessibili pubblicati su internet dagli utenti sulle sopracitate piattaforme online. In questa sede, la CNIL, ha ribadito che, il semplice fatto che i dati siano accessibili su internet non ne autorizza la raccolta e l'utilizzo per qualsiasi scopo. Analizzare le informazioni contenute nelle recensioni "raschiate", metterebbe la DGCCRF, nonostante le garanzie approntate (rimozione dei dati sensibili, formazione del personale coinvolto nel trattamento, distinti livelli di accesso ai dati trattati), nella posizione di aver raccolto una gran quantità di dati in modo indiscriminato, rischiando quindi di trattare dati irrilevanti ai fini del trattamento ed includerli poi nella valutazione di veridicità assegnata alle recensioni. È stato ritenuto che, lo svolgimento di tale trattamento, rischiasse di violare diritti e libertà fondamentali dell'individuo, quali la libertà di espressione e di opinione. Gli obiettivi perseguiti dalla

[52] Délibération n° 2022-125 du 15 décembre 2022 portant avis sur le projet d'arrêté relatif à la création au sein de la direction générale de la concurrence, de la consommation et de la répression des fraudes (DGCCRF) d'un traitement de données à caractère personnel dénommé « Polygraphe » (demande d'avis n° 22014966).

DGCCRF, vale a dire la lotta contro le "false recensioni" su internet e il miglioramento dell'efficacia dei controlli dei suoi agenti, sono legittimi e possono giustificare la possibilità di trattare le recensioni liberamente accessibili. Tuttavia, la raccolta di dati tramite *scraping* porta con sé rischi tali per la riservatezza, da aver bisogno di garanzie ancora più stringenti di quelle previste in sede di valutazione d'impatto. L'Autorità francese dunque, ha invitato il Ministero dell'Economia a prestare particolare attenzione nell'applicazione del principio della minimizzazione dei dati e della protezione dei dati fin dalla progettazione. Dunque, nonostante le garanzie fornite, la CNIL, ha chiuso il proprio parere richiedendo l'autorizzazione al *Conseil D'état* e suggerendo una fase sperimentale per il progetto e auspicando che il legislatore stabilisca un quadro generale entro il quale le amministrazioni potrebbero, se necessario, utilizzare tali strumenti, al fine di garantire un equilibrio tra le missioni delle amministrazioni e la tutela dei diritti delle persone, tra cui quello alla riservatezza[53].

A dimostrazione della rilevanza delle implicazioni che l'attività di *scraping* ha sulla tutela dei dati personali, è opportuno inserire nel discorso, la dichiarazione congiunta sottoscritta da dodici Autorità per la protezione dei dati internazionali, tra cui quelle di Canada, Regno Unito, Cina e Australia, del 24 agosto 2023. La dichiarazione delinea i principali rischi per la privacy associati allo *scraping* dei dati sui *social media*, tra cui figura non solo il *marketing* indesiderato e la sorveglianza individuale da parte delle autorità, ma anche la conseguente probabilità di minare la fiducia delle persone nelle aziende che si occupano di social media (le cosiddette *Social Media Companies* SMC) o di altri siti *web*, con potenziali effetti negativi sull'economia digitale[54]. Nel documento c'è altresì una sezione dedicata alle misure che sia i siti *web* che i singoli individui dovrebbero adottare per ridurre al minimo tali rischi e soddisfare le aspettative normative. Agli individui, nel loro piccolo, si suggeriscono degli sforzi tesi a consapevolizzarsi riguardo al contenuto delle *privacy policies* e alle informazioni fornite dalle aziende *social* concernenti la loro politica di condivisione delle informazioni personali e non solo, perché anche imparare a comprendere e gestire le impostazioni di sicurezza sulle piattaforme, ed essere consapevoli della mole di informazioni che vi si condividono, sono

[53] Ibidem

[54] Joint statement on data scraping and the protection of privacy -August 24, 2023- disponibile sul sito della Information Commissioner's Office (ICO): https://ico.org.uk/media/about-the-ico/documents/4026232/joint-statement-data-scraping-202308.pdf

azioni significative, in grado di integrare l'effetto delle misure tecniche approntate dalle piattaforme[55]. Stante poi la consapevolezza che nessuna misura può proteggere adeguatamente da tutti i potenziali rischi per la privacy associati allo *scraping* dei dati, le società di *social media* e altri siti dovrebbero implementare controlli tecnici e procedurali multilivello per mitigare i rischi. L'approccio multilivello, che il documento propone, si basa su una combinazione di misure e controlli che sia proporzionata alla sensibilità delle informazioni. Tra le misure considerate ce ne sono di natura organizzativa, tecnica (come le limitazioni di frequenza di visita o il monitoraggio di taluni *account* o *bot,* blocco di indirizzi IP da cui proviene attività di *scraping*) e anche giuridica.

3.1 Focus sui soggetti che trattano i dati raccolti tramite *scraping*

Di base viene quasi spontaneo pensare che quel che è liberamente accessibile sia anche liberamente fruibile, utilizzabile a nostra convenienza e secondo i nostri piani. E forse, questo è lo stesso discorso che sta all'origine della raccolta non autorizzata di dati personali dal *web*. Il *web,* che è un posto che siamo abituati, in qualità di privati, a frequentare ed a interrogare quotidianamente e che, di giorno in giorno si arricchisce di informazioni e di risposte, è un po' come se fosse l'enorme dispensa da cui chi fa *scraping* prende gli ingredienti/informazioni di cui ha bisogno. È facile però perdere di vista che la fruizione degli ingredienti provenienti da questa dispensa ha delle regole. Regole che sono più o meno stringenti e precise e variano a seconda della parte del mondo verso cui si aprono gli sportelli della dispensa.

Dunque, dal punto di vista della disciplina della tutela dei dati, chi decide di accedere alla dispensa e trattare informazioni raccolte dal *web* in virtù del fatto che siano pubblicamente disponibili, dovrebbe premurarsi di verificare se, la loro estrazione è un atto legalmente concesso: infatti pubblico non significa destinato al pubblico.

C'è un caso italiano, conosciuto con il nome della piattaforma incriminata "Trovanumeri", la cui vicenda è di seguito discussa al fine di chiarire la portata essenziale delle questioni che, già solo l'attività di *web scraping* (per un istante lasciando da parte

[55] Ibidem.

l'addestramento della IA generativa) solleva, con rispetto alla normativa dettata dal GDPR. Tali indagini risultano ancor più interessanti perché hanno impegnato il Garante lungo un arco di circa dieci anni, tempo in cui è avvenuta la transizione dalla previgente normativa europea in materia di tutela dati personali costituita dalla Direttiva 95/46/CE all'attualmente vigente Regolamento UE 2016/679. Infatti, le richieste di intervento relative alla pubblicazione non autorizzata di dati personali sulla piattaforma in questione, hanno iniziato a pervenire presso l'ufficio del Garante a partire dal 2012. Stante il fatto però che, nell'informativa privacy allora disponibile sul sito, non erano rinvenibili indicazioni riguardo l'intestatario dello stesso, le indagini volte ad identificare il titolare del trattamento hanno richiesto lungo tempo. Inoltre, le indagini sono state rese ancor più gravose dal fatto che i *server* che ospitavano il sito erano ubicati fuori dal territorio nazionale e venivano frequentemente cambiati in modo tale da rallentare ancor più l'ottenimento di informazioni utili da parte delle autorità estere interpellate. L'interrogativo che avvolgeva la figura del titolare del trattamento di "trovanumeri.com" è stato dissipato grazie a Google, al quale è stata fatta una richiesta di informazioni a norma dell'art. 157 del Codice in materia di protezione dei dati personali[56]. Questo perché, sul sito erano presenti dei *banner* pubblicitari del servizio Google *AdSense*. Il riscontro di Google, grazie al quale è stata data un'identità al titolare del sito, è giunto nel Luglio 2022. Nell'agosto 2022 il Garante ha comunicato con nota l'avvio del procedimento nei confronti del titolare. In data 3 aprile 2023 l'Ufficio del Garante ha riscontrato che il sito Trovanumeri.com era ancora *on-line* e nessuna modifica era stata apportata.

Questo caso è significativo non solo perché l'attività di *scraping* in questione è iniziata in tempi non sospetti, ossia prima della ventata di interesse e l'allarme sulle sue conseguenze su vari aspetti della società, ma anche perché riguarda quello che potrebbe rappresentare, in ordine di rilevanza, il primo vero problema dello *scraping* effettuato su dati pubblicamente accessibili sul *web*, ossia: l'assenza di un'idonea base giuridica ai sensi dell'articolo 6 GDPR.

[56] Art. 157: *"Nell'ambito dei poteri di cui all'articolo 58 del Regolamento, e per l'espletamento dei propri compiti, il Garante può richiedere al titolare, al responsabile, al rappresentante del titolare o del responsabile, all'interessato o anche a terzi di fornire informazioni e di esibire documenti anche con riferimento al contenuto di banche di dati."*

La vicenda ebbe inizio nel 2012,anno in cui un signore italiano ha deciso di iniziare a costituire un vero e proprio motore di ricerca per recapiti personali (Trovanumeri.com) facendo *scraping* per mezzo di *spider* (simili ai *crawler*) lanciati nel *web*, ogni giorno, e facendo quindi incetta di dati tra cui numeri di telefono, nominativi ed indirizzi. Le istanze pervenute hanno lamentato, tra l'altro, la totale mancanza di contatti o indicazioni su come rivolgersi al titolare del trattamento e, di fatto, l'impossibilità di ottenere la cancellazione dei dati, a causa dell'indisponibilità tecnica del *form* da compilare per la richiesta. Il funzionamento del sito si reggeva non solo sullo *scraping* ma anche sulla possibilità per chiunque di registrare sul sito gratuitamente, e senza alcuna verifica delle informazioni personali, mettendo così in pericolo titolari di utenze riservate, ed esponendo le persone, i cui recapiti figuravano reperibili sulla piattaforma, a contatti e *telemarketing* indesiderati. Come già accennato, la prima violazione in ordine logico esaminata dal Garante è la diffusione di dati personali in assenza di una idonea base giuridica (articoli 5 (1) (2) e 6 GDPR), oltre che un trattamento in violazione di legge, dato che gli elenchi telefonici possono essere realizzati a partire dalla consultazione di quanto accessibile nel *Database* Unico (DBU)[57]: fonte in grado di garantire correttezza e aggiornamento dei dati. Di conseguenza, è illegittimo formare un elenco telefonico da qualsiasi altra fonte. In questo contesto, risulta chiaro come manchi del tutto una base per il trattamento e, a maggior ragione, il requisito del consenso dell'interessato, che nel caso della creazione di elenchi telefonici è una volontà che necessita di essere espressa o revocata contattando il proprio operatore. In secondo luogo, è stato contestato il mancato rispetto dei diritti degli interessati, perché non era rinvenibile, nel sito, alcun contatto del titolare né tantomeno un identificativo. Inoltre, l'informativa sul trattamento era praticamente vacua e scritta in modo difficilmente comprensibile e le misure di garanzia manchevoli anch'esse (nessuna possibilità, ad esempio di richiedere l'accesso ai dati o la portabilità[58]). Da non dimenticare, poi, quanto disposto dall'articolo 14 GDPR che

[57] Il d.b.u. è l'archivio elettronico unico che raccoglie i numeri telefonici e i dati identificativi dei clienti di tutti gli operatori nazionali di telefonia fissa e mobile e viene utilizzato per la formazione degli elenchi telefonici. È stato istituito con le delibere dell'Agcom n. 36/02/CONS e n. 180/02/CONS ed è operativo dal 1° agosto 2005.

[58] Il diritto alla portabilità è stato introdotto dal GDPR all'articolo 20. La sua introduzione rappresenta un ampliamento concettuale di quello che era l'art. 12, rubricato: " diritto di accesso", della previgente Direttiva 95/46/CE, il quale, all'interno del diritto di accesso, prevedeva per l'interessato un generico diritto ad una comunicazione in formato intelligibile (scelto dal titolare) dei dati oggetto di trattamento:

riguarda le informazioni da fornire all'interessato, laddove i dati personali non siano stati ottenuti direttamente presso di lui. Dunque, all'art.14 comma(2)(f)[59] GDPR si prescrive che per ottemperare al principio di correttezza e trasparenza, nei confronti dell'interessato, quest'ultimo debba essere informato, dal titolare del trattamento, sulla fonte da cui provengono i dati personali, e se provengono da fonti accessibili al pubblico. Questo chiarimento sulla fonte implica che i dati sono ancora considerati personali, nonostante la loro disponibilità pubblica. L'*iter* del caso si è concluso con l'affermazione della prevalenza della mancanza di una base giuridica per il trattamento rispetto alle altre illiceità rilevate e la prescrizione di misure correttive, il tutto all'interno di un'ordinanza di ingiunzione per l'applicazione di una sanzione amministrativa pecuniaria[60].

L'uso che viene fatto delle informazioni ricavate tramite questa attività è l'aspetto più rilevante della questione perché permette di differenziare il tipo di *scraping* intrapreso e la posizione giuridica che ne consegue. Perciò, trattare dei dati raccolti sfruttando le vulnerabilità di un sistema informatico non adeguatamente protetto, è un comportamento qualificabile come *scraping* "malevolo". Per chiarire meglio il concetto di *scraping*

"la comunicazione informa intelligibile dei dati che sono oggetto dei trattamenti, nonché di tutte le informazioni disponibili sull'origine dei dati".
Il diritto alla portabilità introdotto dall'art. 20 GDPR prevede il diritto dell'interessato a vedersi comunicare i dati che lo riguardano e che sono oggetto di trattamento, non solo in un formato di uso comune, ma che sia anche strutturato e leggibile da un dispositivo automatico. Inoltre, nell'esercitare i propri diritti relativamente alla portabilità dei dati l'interessato ha il diritto di ottenere la trasmissione diretta dei dati da un titolare del trattamento all'altro, se tecnicamente fattibile. Tale diritto, poi, si applica qualora l'interessato abbia fornito i dati personali acconsentendo al trattamento o se il trattamento è necessario per l'esecuzione di un contratto. Non si applica, invece, qualora il trattamento si basi su un altro motivo legittimo diverso dal consenso o contratto: *" L'interessato ha il diritto di ricevere in un formato strutturato, di uso comune e leggibile da dispositivo automatico i dati personali che lo riguardano forniti a un titolare del trattamento e ha il diritto di trasmettere tali dati a un altro titolare del trattamento senza impedimenti da parte del titolare del trattamento cui li ha forniti qualora: a) il trattamento si basi sul consenso ai sensi dell'articolo 6, paragrafo 1, lettera a), o dell'articolo 9, paragrafo 2, lettera a), o su un contratto ai sensi dell'articolo 6, paragrafo 1, lettera b); e b) il trattamento sia effettuato con mezzi automatizzati.".*
Cfr. Altalex- *Il diritto alla portabilità dei dati*- M. Iaselli- 2017. Reperibile su: https://www.altalex.com/documents/news/2017/11/02/il-diritto-alla-portabilita-dei-dati-personali - Ultima visita al sito in data 15 gennaio 2025.

[59] Articolo 14(2) (f) GDPR: "*2. Oltre alle informazioni di cui al paragrafo 1, il titolare del trattamento fornisce all'interessato le seguenti informazioni necessarie per garantire un trattamento corretto e trasparente nei confronti dell'interessato: […]*
(f) la fonte da cui hanno origine i dati personali e, se del caso, l'eventualità che i dati provengano da fonti accessibili al pubblico;"

[60] "*pur dovendo ritenere determinante tra essi la violazione di cui al punto 2.1. Questa infatti, riguardando il presupposto stesso del trattamento, è da considerarsi assorbente e già di per sé sufficiente ad inficiare l'intero trattamento, tenuto conto del fatto che l'eventuale rettifica delle illiceità descritte ai punti successivi, che pure aggrava la condotta, non sarebbe sufficiente a sanare il fatto che il trattamento stesso è posto in essere in assenza di una idonea base giuridica e soprattutto in violazione di legge".*
Provvedimento del 17 maggio 2023 [9903067] - Garante Privacy

malevolo, e distinguerlo da quello effettuato dai *bot,* che senza mirare allo sfruttamento di alcuna vulnerabilità di sistema, si limitano a "raccogliere" informazioni in modo massivo ed indiscriminato, al fine di reperire materiale per l'addestramento di modelli di IA, si può richiamare il caso[61] che riguardò *Meta Platforms Ireland* Ltd, società che nel 2022 venne sanzionata dall'Autorità irlandese per la protezione dei dati (il nome irlandese è *"An Coimisiùn um Chosaint Sonraí"*,ma è conosciuta anche con la sigla DPC, dalla abbreviazione inglese di Data Protection Commission) con una multa di 265 milioni di euro ed una serie di misure correttive.

Tutto ebbe inizio nell'aprile 2021, quando hanno iniziato a diffondersi nei *media* notizie che evidenziavano era stato oggetto di *scraping*, e poi dischiuso sul *web*, un sostanzioso *set* di dati contenenti anche informazioni personali relativi a circa 533 milioni di utenti di Facebook in tutto il mondo. L'indagine dell'Autorità irlandese DPC ha avuto come obiettivo quello di verificare l'adempimento degli obblighi di Meta, in quanto titolare del trattamento ai sensi del GDPR, in particolare sulla corretta applicazione del principio della protezione dati fin dalla progettazione dell'articolo 25 del GDPR, e sull'efficacia e l'integrazione delle misure tecniche e organizzative di alcune funzionalità del social Facebook (come *Facebook Search, Facebook Contact Importer, Messenger Contact Importer* e *Instagram Contact Importer*). Più specificamente, l'indagine condotta dall'Autorità irlandese per la protezione dei dati, ha preso in considerazione anche il fatto che le funzionalità di Facebook, grazie alle quali gli *scraper* sono riusciti a carpire i dati degli utenti, erano suscettibili di essere sfruttate da *account* e *bot* falsi, nonché il fatto che Meta aveva chiaramente identificato casi di *scraping* di massa, con relativa attività di *bot* e *account* falsi nelle funzionalità incriminate[62]. L'articolo 25(1) del GDPR, relativo alla protezione dei dati fin dalla progettazione, considera infatti l'efficacia delle misure e delle garanzie adottate dal titolare del trattamento per garantire i diritti dell'interessato come un elemento chiave. Ciò implica che tali misure e salvaguardie dovrebbero essere concepite in modo da essere solide e in modo che il titolare del trattamento possa integrarle con ulteriori misure al fine di affrontare correttamente qualsiasi aumento del rischio. Il peso

[61] Data Protection Commission in the matter of Meta Platforms Ireland Ltd- *Decision of the Data Protection Commission made pursuant to Section 111 of the Data Protection Act 2018 and Article 60 of the General Data Protection Regulation.* – Reperibile su: https://www.dataprotection.ie/sites/default/files/uploads/2022-12/Final%20Decision_IN-21-4-2_Redacted.pdf

[62] Ibidem

della responsabilità del titolare del trattamento implica che egli dovrebbe sempre essere in grado di dimostrare che il principio della protezione dei dati fin dalla progettazione è stato mantenuto grazie alle misure attuate, e ciò a maggior ragione in caso di fuga di dati. La protezione dei dati per impostazione predefinita, delineata all'articolo 25(2), richiede che il titolare del trattamento attui misure tecniche e organizzative adeguate per garantire che vengano trattati solo i dati personali necessari per ogni specifica finalità. È importante sottolineare che le misure, così adottate dal titolare del trattamento, devono garantire che, per impostazione predefinita, a meno di un intervento proveniente dall'interessato, i dati personali non siano resi accessibili a un numero indefinito di persone fisiche[63]. La protezione dei dati fin dalla progettazione e per impostazione predefinita sono entrambe espressioni del principio di responsabilizzazione (*accountability)*, al centro del sistema di responsabilità istituito dal GDPR. Nel caso che coinvolge Meta, l'Autorità irlandese DPC ha identificato il trattamento come prono a gravi rischi per gli interessati come frode, furto d'identità e truffa. Inoltre, ha sottolineato che Meta, in qualità di titolare del trattamento, non ha adottato misure adeguate per valutare e prevenire il rischio rappresentato dagli *scraper*, e che i provvedimenti presi per contrastare la loro attività, quali il *rate limiting* ed il rilevamento dei *bot*[64], erano comunque insufficienti. Il DPC ha poi sottolineato come Meta avrebbe potuto implementare, all'epoca dei fatti, una serie di misure aggiuntive di tipo tecnico (tra cui l'evitare di fornire riscontri precisi in risposta alle ricerche per numero

[63] Vedi *EDPB- Linee guida 4/2019* sull'articolo 25(2): *Protezione dei dati fin dalla progettazione e per impostazione predefinita:* "Il nucleo della disposizione è garantire una protezione dei dati adeguata ed efficace sia fin dalla progettazione che per impostazione predefinita, il che significa che i titolari del trattamento dovrebbero essere in grado di dimostrare di disporre delle misure e delle garanzie adeguate nel trattamento per garantire che i principi di protezione dei dati e i diritti e le libertà degli interessati siano effettivi".

[64]Data Protection Commission Reference: IN-21-4-2 In the matter of Meta Platforms Ireland Ltd. (Formerly Facebook Ireland Ltd.) - Punto 140 della decisione: "*Having particular regard to the risk of varying likelihood and severity for rights and freedoms of natural persons posed by MPIL's processing of personal data in the Relevant Features, I do not consider that the rate limiting and bot detection measures implemented by MPIL were sufficient for the purposes of Article 25(1) GDPR. As outlined above, the scope of MPIL's processing in the Relevant Features concerned a very large number of data subjects. This resulted in a particular risk of bad actors using the Relevant Features to acquire personal data due to the higher likelihood that random numbers and email addresses would match real Facebook users. There are a range of additional measures that MPIL could have implemented at the time to prevent such misuse of the Relevant Features during the Temporal Scope. While MPIL's non implementation of any specific measure or group of measures does not in and of itself constitute an infringement of Article 25(1) GDPR, these measures provide context to the consideration of whether the measures implemented by MPIL were appropriate at the relevant time under consideration, and as to whether, taking into account the state of the art, the cost of implementation, the particular risks, MPIL, both at the time of the determination of the means for processing and at the time of the processing itself, implemented appropriate technical and organisational measures". […]*

di telefono o l'utilizzo di CAPTCHA[65]) per prevenire l'uso improprio delle funzioni in questione. Il DPC ha poi proseguito evidenziando che, seppur la mancata attuazione di tali misure non possa ritenersi di per sé una violazione dell'articolo 25(1) GDPR, la loro implementazione avrebbe restituito un contesto più favorevole alla posizione di Meta nel considerare la sua condotta generale come espressiva di una tutela adeguata allo stato dell'arte e della tecnica.

Il Garante irlandese ha quindi riscontrato la violazione dell'articolo 25, paragrafi 1 e 2 del GDPR da parte di Meta e ha emesso un ordine di adeguamento del trattamento dei dati al GDPR, un rimprovero a Meta e due sanzioni amministrative pecuniarie. Le preoccupazioni sollevate durante il processo di maturazione di questa decisione, sono state rilevanti in quanto attuali e concrete, tanto da essere successivamente riproposte anche nella dichiarazione congiunta sui rischi e azioni a contrasto per lo *scraping*, redatto delle 12 autorità garanti internazionali di cui abbiamo detto nel paragrafo 3[66].

3.1.1 Lo *scraping* di dati biometrici ed il caso Clearview AI

I dati biometrici vengono definiti dal GDPR, all' art 4 (14), come: "*dati personali ottenuti da un trattamento tecnico specifico, relativi alle caratteristiche fisiche, fisiologiche o comportamentali di una persona fisica che ne consentono o confermano l'identificazione univoca, quali l'immagine facciale o i dati dattiloscopici".* I dati biometrici sono classificati come categoria particolare di dati personali in virtù della loro particolare capacità di consentire l'esatta identificazione di una persona. Conformemente all'articolo 9(1) GDPR è fatto divieto di trattare le categorie particolari di dati, nei quali sono ricompresi i dati biometrici. Questo divieto è però soggetto alle deroghe elencate nel punto (2) e la prima di queste (lett. a) è il consenso esplicito dell'interessato, salvi i casi in cui, in forza di legge, tale divieto di trattamento non sia obiettabile da parte dell'interessato. Un divieto sovrapponibile per natura, ma diverso per scopo, è quello contenuto all'articolo 10 della direttiva UE 2016/680 che, in qualità di *lex specialis* nei

[65] Ibidem. Punti 141 e 143

[66] Joint statement on data scraping and the protection of privacy -August 24, 2023- disponibile sul sito della Information Commissioner's Office (ICO): Reperibile su: https://ico.org.uk/media/about-the-ico/documents/4026232/joint-statement-data-scraping-202308.pdf

confronti del GDPR, copre le "*persone fisiche con riguardo al trattamento dei dati personali da parte delle autorità competenti a fini di prevenzione, indagine, accertamento e perseguimento di reati o esecuzione di sanzioni penali, incluse la salvaguardia e la prevenzione di minacce alla sicurezza pubblica*"[67].

In ambito nazionale, l'art. 2 *septies* del Codice Privacy, attua l'art. 9, par. 4 del GDPR, prevedendo che il trattamento dei dati biometrici, genetici e relativi alla salute sia subordinato all'osservanza di misure di garanzia, stabilite dal Garante, tenendo in particolare considerazione, oltre alle linee guida, raccomandazioni e migliori prassi pubblicate dal Comitato Europeo per la Protezione dei Dati, anche l'evoluzione tecnologica e scientifica del settore a cui tali misure sono rivolte, nonché l'interesse alla libera circolazione dei dati nel territorio europeo.

I dati biometrici, in virtù della loro potenzialità di identificare in modo univoco gli individui, sono stati il motore per lo sviluppo di tecnologie a base biometrica, tra le quali, hanno avuto particolare diffusione le tecnologie di riconoscimento facciale (*Facial Recognition Technologies* abbreviato con l'acronimo FRT), le quali sono , tra l'altro, sempre più spesso potenziate dall'uso dell'IA[68]. Attraverso le FRT è possibile individuare volti specifici all'interno di immagini, comparare un volto con altri presenti in un *database* per rilevarne l'identità, non in ultimo, queste tecnologie si prestano anche ad operazioni di categorizzazione dei volti in base a criteri quali età, etnia o stato emotivo desumibile dai tratti caratteristici delle espressioni facciali[69]. I vantaggi e la praticità che derivano dall'utilizzo delle FRT nel quotidiano, non può sfuggire: infatti, la troviamo spesso come metodo opzionale di accesso a dispositivi elettronici o aree riservate all'interno di piattaforme, come autorizzazione ai pagamenti con carta e all'invio di bonifici bancari entrambi tramite *smartphone*, nel campo della sicurezza le FRT vengono utilizzate per il controllo delle frontiere. Tuttavia, alla base dell'utilizzo delle FRT, c'è il presupposto della raccolta - talvolta anche in tempo reale - di una quantità considerevole ed indiscriminata di immagini che vengono poi conservate, al fine di poter effettuare comparazioni ed ottenere riscontri, questo accade spesso senza ottenere il previo consenso

[67] Articolo 1 Direttiva (Ue) 2016/680 Del Parlamento Europeo E Del Consiglio del 27 aprile 2016

[68] Di Matteo F. *"La riservatezza dei dati biometrici nello Spazio europeo dei diritti fondamentali: sui limiti all'utilizzo delle tecnologie di riconoscimento facciale*." Freedom, security & justice: european legal studies: 1, 2023- Pg. 80-87

[69] Agenzia dell'Unione europea per i diritti fondamentali (FRA), *Facial recognition technology: fundamental rights considerations in the context of law enforcement*, del 27 novembre 2019, p. 7.

da parte degli individui sottoposti a riconoscimento facciale o, addirittura, a loro insaputa, come nel caso dello *scraping* di immagini dal *web* .

Lo *scraping* di foto di persone non equivale automaticamente a un trattamento di categorie speciali di dati, ma ciò accade solo quando le foto vengono elaborate attraverso un mezzo tecnico specifico, che consente l'univoca identificazione o autenticazione di una persona fisica[70]. Seguendo quindi i principi giuridici in materia di protezione dati, il trattamento delle fotografie deve essere lecito, equo, trasparente, essere motivato e sostenuto da una finalità specifica, esplicita e legittima. Inoltre, dovrebbe essere conforme agli ulteriori requisiti di minimizzazione dei dati, accuratezza, limitazione della conservazione, sicurezza e responsabilità. Sul punto della relazione tra diritti umani- e sviluppo della IA, con particolare riferimento alla messa in discussione dei pilastri internazionali del diritto alla protezione dei dati personali, già si espresse David Kaye nel 2018, che fu relatore speciale delle Nazioni Unite sulla libertà di opinione ed espressione dal 2014 al 2020, in un *report*[71] nel quale formulò delle raccomandazioni tese da un lato a prendere in considerazione l'impatto che il solo sviluppo delle FRT può avere sulla effettiva tutela dei dati in rete (onde la necessità di valutazioni di impatto preliminari l'immissione sul mercato della tecnologia o *audit* periodici operati da periti terzi ed imparziali), dall'altro a spronare l'elaborazione di rimedi esperibili dai soggetti i cui diritti sono interessati da tale tecnologia.

A tal proposito, il GDPR offre al titolare, all'art. 35 lo strumento della valutazione d'impatto sulla protezione dei dati (DPIA), la quale diventa espressamente obbligatoria, tra gli altri casi elencati, anche in caso di trattamento su larga scala di categorie particolari di dati di cui all'articolo 9 punto (3)(b). Seppur la DPIA non sia stata concepita come strumento specificamente atto a contrastare lo *scraping*, e nemmeno a consentirlo, resta una via di autoanalisi con cui il titolare viene obbligato a riflettere e a rendersi responsabile del trattamento dati sotteso alle tecnologie di cui si è avvalso.

---

[70] Considerando 51 GDPR: "*Il trattamento di fotografie non dovrebbe costituire sistematicamente un trattamento di categorie particolari di dati personali, poiché esse rientrano nella definizione di dati biometrici soltanto quando saranno trattate attraverso un dispositivo tecnico specifico che consente l'identificazione univoca o l'autenticazione di una persona fisica."*

[71] David Kaye, -*Promotion and protection of the right to freedom of opinion and expression*, 29 agosto 2018,- A/73/348

Il caso di Clearview AI[72], ha fatto luce sui rischi per la riservatezza, la protezione dei dati e sulle relative questioni concernenti i diritti e le libertà fondamentali degli interessati, derivanti dallo *scraping* massiccio ed indiscriminato da parte di aziende private, di immagini da cui ricavare dati biometrici.

Clearview AI Inc. è una startup statunitense che ha utilizzato il *web scraping* su *social network*, *blog*, video e siti, al fine di creare un *database* con miliardi di immagini per sviluppare un *software* di intelligenza artificiale *in house* per il riconoscimento facciale, ad uso poi delle forze dell'ordine americane, per supportarle nell'identificazione dei criminali. Il software si sostanzia in un'applicazione per la ricerca di immagini, che fornisce risultati di ricerca con collegamenti ai siti *web* d'origine di terze parti. Lo strumento di riconoscimento facciale di Clearview si basa su quattro passaggi chiave, in sequenza. Il primo è rappresentato dallo *scraping* di immagini di volti e relativi metadati da fonti online accessibili al pubblico, e dalla successiva archiviazione di tali informazioni nel suo *database*. In secondo luogo, l'azienda crea, per ogni immagine nel suo *database*, degli identificatori biometrici sotto forma di rappresentazioni vettoriali atte a ricalcare le diverse linee uniche di un volto. Il sistema, poi, consente di comparare via *server*, le immagini caricate dagli utenti con gli identificatori biometrici indicizzati nel *database*, ed eventualmente trovare come risposta un abbinamento. Infine, il sistema fornisce un elenco di risultati, contenente tutte le immagini e i metadati corrispondenti. Facendo *clic* su uno di questi risultati, gli utenti vengono indirizzati alla pagina di origine dell'immagine. Di fatto Clearview non raccoglie solamente le immagini con lo scopo di renderle accessibili ai clienti, ma offre un vero e proprio servizio di ricerca biometrica ed un archivio di risorse sviluppatosi attraverso il tempo. C'è da aggiungere anche che tale servizio è destinato ad una specifica categoria di clienti, ossia le forze di polizia. L'azienda, oltre a commercializzare la sua applicazione attraverso la concessione di licenze alle forze dell'ordine statunitensi per un breve periodo di tempo, alla fine del 2019, ha affermato di aver aperto degli *account* di prova a degli enti europei che avevano manifestato interesse per il suo prodotto. Proprio questa evenienza ha portato il Garante italiano a pronunciarsi sulla vicenda, all'esito di una complessa istruttoria a cui hanno contribuito, riferendo informazioni, anche altre autorità di controllo europee. Un aspetto

[72] Ordinanza ingiunzione nei confronti di Clearview AI-Registro dei provvedimenti n. 50 del 10 febbraio 2022- [doc. web n. 9751362]

rilevante è che le foto raccolte erano destinate a nutrire l'archivio del sistema, e rimanere nel *database* anche in caso di rimozione o privatizzazione nel sito d'origine; e nel caso di nuovi riscontri a distanza di tempo, le informazioni acquisite venivano integrate con quelle estratte più recentemente. Un tale meccanismo risulta quindi idoneo a riflettere i cambiamenti fisici subiti da un soggetto nell'arco del tempo.

La decisione del Garante italiano, che peraltro si inserisce in un contesto di pronunce concordanti, come le decisioni sul caso prese dalla Autorità Garante tedesca del *land* di Amburgo[73] e dalla francese CNIL[74], conferma innanzitutto la sussistenza della competenza del Garante, qualificando quello di Clearview come un trattamento transfrontaliero di dati personali, perché appunto capace di incidere su soggetti in più di uno Stato Membro[75]. Ma, ancora, il fulcro della decisione sta proprio nel fatto che il Garante ha confermato la posizione di Clearview come titolare di un trattamento di dati personali comuni e biometrici eseguito in violazione dell'articolo 5 GDPR.[76][77] infatti gli interessati le cui foto sono state prelevate dal *web*, non avevano nessuna relazione con l'azienda americana, ed è perciò verosimile che non sia aspettassero che le loro immagini venissero utilizzate per sviluppare un *software* di riconoscimento facciale di una piattaforma privata, a loro sconosciuta, e per di più stabilità al di fuori dell'Unione Europea[78]. Su questo aspetto c'è da precisare che Clearview non aveva alcuno stabilimento in Europa. Tuttavia l'applicabilità del GDPR è stata assodata seguendo il criterio di *targeting* previsto all'art. 3(2). L'applicazione del criterio di *targeting* richiede il verificarsi di due condizioni: la prima sub art. 3(2)(a) richiede l'intenzione del titolare del trattamento di rivolgersi al mercato europeo. L'applicazione di tale criterio viene

---

[73] Decisione dell'Autorità di controllo tedesca del Land di Amburgo (decisione 545/2020; 32.02-102) 545_2020_Anhörung_CVAI_DE_Redacted.pdf

[74] Decisione n° MED 2021-134 of 1st November 2021 issuing an order to comply to the company Clearview AI- Decision n° MED 2021-134 of 1st November 2021 issuing an order to comply to the company CLEARVIEW AI

[75] Ai sensi dell'art. 4(1) n. 23 GDPR

[76] Articolo 5 (1) lett. a-b-e GDPR

[77] Ordinanza ingiunzione nei confronti di Clearview AI-Registro dei provvedimenti n. 50 del 10 febbraio 2022- [doc. web n. 9751362]

[78] Art. 3(2) GDPR: "*Il presente regolamento si applica al trattamento dei dati personali di interessati che si trovano nell'Unione, effettuato da un titolare del trattamento o da un responsabile del trattamento che non è stabilito nell'Unione, quando le attività di trattamento riguardano:*
a) l'offerta di beni o la prestazione di servizi ai suddetti interessati nell'Unione, indipendentemente dall'obbligatorietà di un pagamento dell'interessato; oppure
b) il monitoraggio del loro comportamento nella misura in cui tale comportamento ha luogo all'interno dell'Unione."

confermata da una serie di elementi raccolti nel provvedimento del Garante italiano[79]:la decisione dell'Autorità Garante svedese IMY di sanzionare[80] le proprie forze dell'ordine per aver utilizzato il software di Clearview; una nota difensiva della stessa Clearview in cui dichiara di aver voluto chiudere, nel corso del 2020, gli *account* europei e di voler cessare l'offerta del suo prodotto in quest'area; i termini in cui era originariamente formulata la *privacy policy* della società, i quali menzionavano la base giuridica del trattamento, l'adeguamento alle norme sulla protezione dati, l'eventualità del trasferimento dei dati al di fuori del SEE, la possibilità di presentare reclamo ad un'Autorità di protezione dati competente, l'attribuzione nei *ToS* agli utilizzatori, della qualità di "utenti".

Il secondo dei criteri di *targeting* rinvenibile sub art. 3(2)(b), collega l'applicazione del GDPR alle attività di trattamento correlate al monitoraggio del comportamento di interessati nell'Unione Europea che avvenga all'interno dell'Unione stessa. A tal riguardo, l'intero processo di raccolta ed elaborazione posto in essere da Clearview mira a costituire un *dataset* al quale comparare le immagini caricate dall'utente ed estrarre poi, dal proprio archivio, le immagini associabili ad esse da un punto di vista biometrico, nonché i metadati associati. Tale meccanismo di ricerca si sostanzia, quindi, in un processo di comparazione. C'è poi da aggiungere che le informazioni raccolte da Clearview vengono archiviate nel suo *database* ed arricchite nel tempo con il frutto di ulteriore attività di *scraping*, fatto che si è rivelato (grazie all'esame di alcuni dei reclami proposti al Garante italiano) idoneo a riflettere i cambiamenti fisici degli interessati nel tempo. La considerazione di queste circostanze ha portato il Garante italiano a ritenere quest'attività *"idonea ad integrare, come richiesto nel Considerando 24, un'attività assimilabile al controllo del comportamento dell'interessato in quanto posta in essere tramite il tracciamento in internet e la successiva profilazione."*

È stato inoltre escluso che Clearview potesse vantare una valida base giuridica su cui fondare il trattamento: infatti, pacificamente esclusa l'avvenuta acquisizione del

[79] Ordinanza ingiunzione nei confronti di Clearview AI-Registro dei provvedimenti n. 50 del 10 febbraio 2022- [doc. web n. 9751362]

[80] IMY-DI-2020-2719:A126.614/2020 del 10 febbraio 2021. Provvedimento disponibile sul sito dell'Autorità (in svedese): https://www.imy.se/globalassets/dokument/beslut/beslut-tillsyn-polismyndigheten-cvai.pdf
Comunicato stampa (in inglese) riassuntivo del provvedimento reperibile sul sito dell'Autorità: https://www.imy.se/en/news/police-unlawfully-used-facial-recognition-app/

consenso, non è stato possibile ricorrere nemmeno al legittimo interesse della società che, spinta da fini di lucro, ha agito operando un trattamento decisamente intrusivo della sfera privata individuale di un elevato numero di persone. Atteso inoltre che, il trattamento illecito operato da Clearview ha coinvolto anche dei dati biometrici, i quali godono all'interno del Regolamento di tutele più stringenti, il fondamento giuridico da invocare, nel caso di una società nella stessa posizione di ClearviewAI, avrebbe dovuto ,essere in maniera cumulativa, anche quanto previsto dall'articolo 9 [81][82].

3.2 Il ruolo dei titolari del trattamento

Tornando ora sulla posizione dei titolari, verrà però cambiata la prospettiva. Fin qui si è ragionato infatti sulla figura del titolare sia come attore ed autore di *web scraping*, sia come agente provocatore della resilienza ai principi del trattamento dei dati. Per questo è stato necessario analizzare e contestualizzare le criticità che più sovente la pratica dello *scraping* ha sollevato rispetto alla normativa sulla tutela dei dati personali.

Nel caso in cui i titolari o responsabili si trovino nella posizione di rendere accessibili al pubblico, attraverso le piattaforme o i siti *web* sotto loro gestione, dati personali, entra ora in gioco il dovere di responsabilità su quei dati, che, tra le altre accezioni, comprende anche il dover fare il possibile per proteggere tali dati da azioni di raccolta indiscriminata non autorizzata. Il principio di responsabilizzazione, chiamata in inglese *accountability*, prevede che titolari e responsabili si impegnino in comportamenti proattivi atti a dimostrare di aver adottato misure tali da assicurare l'applicazione del Regolamento[83].

[81] EDPB- Linee guida 8/2020 sul targeting degli utenti di social media pag 35 punto 114: *"Oltre alle condizioni dell'articolo 9 GDPR, il trattamento di categorie particolari di dati deve fondarsi su una base giuridica stabilita nell'articolo 6 GDPR ed essere effettuato in conformità con i principi fondamentali di cui all'articolo 5 GDPR".*

[82] Ordinanza ingiunzione nei confronti di Clearview AI-Registro dei provvedimenti n. 50 del 10 febbraio 2022- [doc. web n. 9751362]

[83]In particolare quanto contenuto nel capo IV del GDPR su titolari, responsabili e meccanismi di sicurezza e certificazione; ed in aggiunta le disposizioni agli artt.. 23-25.
Per una panoramica sui tratti principali dell'accountability, si veda la sintesi fatta da Garante sul sito dell'Autorità, reperibile su: https://www.garanteprivacy.it/regolamentoue/approccio-basato-sul-rischio-e-misure-di-accountability-responsabilizzazione-di-titolari-e-responsabili

3.2.1 L'attuazione del principio di *Accountability*

Il sistema di disposizioni, attraverso cui si articolano i doveri che caratterizzano il soddisfacimento più possibile completo del principio di *accountability*, è una novità introdotta con la promulgazione del GDPR. Difatti, il suo predecessore, ovvero la Direttiva 95/46/CE, ne era carente[84]. Per l'appunto, nel suo parere sul Futuro della Privacy del 2009, il Gruppo di Lavoro dell'Articolo 29 ha ritenuto che la legislazione in vigore all'epoca non fosse in grado di fornire un'efficace protezione dei dati, offrendo piuttosto, come soluzioni, la responsabilizzazione ed un approccio basato sul rischio: "*un approccio uniforme che garantisca poteri specifici e stabilisca sanzioni pecuniarie per titolari e responsabili"*[85]. Nel suo parere sull'*accountability*, il Gruppo di Lavoro ha presentato il suo punto di vista sul concetto di responsabilità come un modo per migliorare l'efficienza della normativa sulla protezione dei dati, in particolare nei confronti dei titolari del trattamento nelle attività ad alta intensità di dati. Titolari e responsabili di tali[86] attività, il cui trattamento dei dati rappresenta un rischio elevato per le persone, dovrebbero attuare misure di protezione dei dati più estese per adempiere al loro obbligo di responsabilità[87].

D'altro canto, l'approccio basato sul rischio implica anche la possibilità di scalare gli obblighi di protezione, in modo che le attività a basso rischio possano essere soggette a requisiti meno rigorosi. Ciò significa non solo che le misure di gestione dei rischi possono essere meno rigide, ma, anche, che l'identificazione e la valutazione dei rischi possono essere meno approfondite.

All'articolo 5(2)[88] viene disposto come primo compito del titolare quale soggetto che determina le finalità e i mezzi del trattamento, sia quello di rispettare i principi generali

---

[84] Karjalainen, Tuulia. "*All Talk, No Action? The Effect of the GDPR Accountability Principle on the EU Data Protection Paradigm.*" European Data Protection Law Review (EDPL), vol. 8, no. 1, 2022, pp. 19-30.

[85] Article 29 Working Party 'The Future of Privacy, Joint contribution to the Consultation of the European Commission on the legal framework for the fundamental right to protection of personal data adopted on 01 December 2009' (WP 168)- punti 31 e 90.

[86] Il Gruppo di Lavoro art. 29 era un organismo consultivo e indipendente, istituito, per l'appunto, dall'art. 29 della Direttiva 95/46. Il Gruppo, che si è occupato delle questioni relative alla tutela della privacy e dei dati personali è stato attivo fino al 25 maggio 2018, data in cui è entrato in vigore il GDPR. A partire da questa data è stato sostituito dal Comitato Europeo per la Protezione dei dati (EDPB).

[87] Article 29 Data Protection Working Party-WP 173- Opinion 3/2010 on the principle of accountability. Punto 14.

[88] Art.5(2) GDPR: *"Il titolare del trattamento è competente per il rispetto del paragrafo 1 e in grado di comprovarlo."*

del trattamento dei dati, definiti dallo stesso articolo nel paragrafo precedente, e di dimostrare tale conformità. L'obiettivo della responsabilizzazione è anche quello di incentivare i titolari del trattamento a cercare, tramite il proprio operato, di raggiungere un livello sempre più alto di conformità alla normativa. Il GDPR definisce la responsabilizzazione come un obbligo esteso ad ogni fase del trattamento. Nella sua formulazione lungo il capo IV del Regolamento, seppure non vengano specificate le misure concrete, sono descritti i parametri a cui i titolari del trattamento dovrebbero attenersi per dimostrare la propria *compliance*.

Inoltre, il contenuto del principio può comunque essere considerato come direttamente proporzionale al crescere dei rischi per la protezione dei dati coinvolti nel trattamento. Maggiore è il rischio che un trattamento comporta, maggiore è la responsabilità che ci si dovrebbe attendere dal titolare del trattamento. Ad ogni modo, ed indipendentemente dal rischio, tutti i titolari sono tenuti a rispondere del proprio operato. Ciò che varia sono le misure necessarie per dimostrare che l'obbligo è stato soddisfatto. Il livello di responsabilizzazione richiesto varia da caso a caso, lasciando un notevole margine di riflessione sul caso concreto. In linea di principio, il GDPR richiede ai titolari del trattamento di attuare misure adeguate, mentre non sono necessari sforzi sproporzionati. Al considerando 74 del GDPR viene specificato che il titolare del trattamento "*dovrebbe essere tenuto ad attuare misure adeguate ed efficaci"* mentre il considerando 76 invita a valutare l'adeguatezza di una misura avendo riguardo della "*natura, dell'ambito di applicazione, del contesto e delle finalità del trattamento*".

Nel contesto dei *big data* che ricordiamo essere dati eterogenei, non strutturati e raccolti con *software* in grado di processarne le dimensioni, una serie di fattori strutturali può limitare la capacità di un titolare del trattamento di tenere adeguatamente in conto dei rischi per i diritti e le libertà degli interessati. Questo potrebbe accadere sebbene il titolare agisca presumibilmente in buona fede e avendo tecnicamente ottemperato con misure tali da garantire teoricamente sicurezza e protezione massima al trattamento. Ad esempio, la valutazione dei rischi connessi a operazioni tecniche complesse può essere estremamente difficile, il che significa che anche il controllore più meticoloso potrebbe non riuscire ad identificare e mitigare adeguatamente i rischi.

Anche volendo tralasciare considerazioni di carattere psicologico, dai casi discussi nei paragrafi precedenti si può notare come l'elaborazione dei *big data* sia caratterizzata da

tentativi ed errori. E molto spesso i dati vengono elaborati forse anche con troppa leggerezza e poca contemperazione delle seppur giuste prospettive economiche dei titolari, con l'obbligo di responsabilità verso gli interessati.

Tornando ai criteri offerti dal GDPR ai titolari, essi costituiscono una guida per decidere autonomamente modalità, garanzie e limiti del trattamento. Il primo in ordine di disposizione viene riassunto dall'espressione inglese "*data protection by design and by default"* con cui è rubricato l'articolo 25 del GDPR. Questa espressione racchiude la necessità di approntare garanzie adeguate per il trattamento dei dati fin dall'inizio, e con inizio si intende il momento in cui si determinano i mezzi del trattamento. È richiesto poi che quest'impegno preventivo venga portato avanti da parte del titolare lungo tutto il trattamento. Il secondo criterio, dettagliato ai considerando 75-77 del GDPR, fa riferimento al rischio di un possibile impatto negativo, sui diritti e libertà dell'interessato, inerente al trattamento. Agli articoli 35-36 del GDPR vengono a tal proposito offerti due strumenti (la valutazione di impatto sulla protezione dei dati e la consultazione preventiva) per rendere il più possibile comprensibile la valutazione dei rischi da parte del titolare e, di conseguenza permettergli di decidere se iniziare il trattamento con le dovute cautele o consultare l'autorità di controllo per una consultazione sulla gestione dei rischi residuali

Fatte queste premesse sul principio di *accountability*, diventa agevole osservare come esso sia non solo uno dei principi guida della normativa in materia di protezione dei dati personali, ma soprattutto una solida base a partire dalla quale trovare modi sempre più adatti a rispondere alle sfide poste dall'evoluzione dell'era dei *big data.*

Grazie quindi alla sua naturale "apertura" verso l'adozione da parte dei titolari, di misure pratiche commisurate all'entità della minaccia alla salvaguardia del diritto alla riservatezza ed al rispetto dei principi del trattamento, l'*accountability* permette di agire in continuità con il GDPR.

È all'interno di questo solco che si colloca, dapprima, la dichiarazione congiunta delle dodici Autorità garanti internazionali a contrasto dello *scraping* a fini generici[89], e

[89] Joint statement on data scraping and the protection of privacy -August 24, 2023- disponibile sul sito della Information Commissioner's Office (ICO): Reperibile su: https://ico.org.uk/media/about-the-ico/documents/4026232/joint-statement-data-scraping-202308.pdf

successivamente, la nota informativa del Garante italiano[90] che mira ad offrire delle misure concrete per arginare lo *scraping* finalizzato allo sviluppo dell'intelligenza artificiale generativa.

[90] Nota informativa emanata con il provvedimento del 20 maggio 2024 [doc. web n. 10020316] e disponibile sul sito del Garante.

CAPITOLO SECONDO

UNA PANORAMICA SULLA REGOLAZIONE IN AMBITO EUROPEO E NAZIONALE

Nel precedente capitolo è stata fornita un'ampia contestualizzazione dello *scraping* e degli attriti che l'uso di questa modalità di raccolta automatizzata di dati provoca rispetto all'osservazione della normativa a tutela dei dati personali. Dopo aver quindi introdotto i temi principali di discussione dell'elaborato, il presente secondo capitolo intende illustrare il quadro normativo di riferimento da cui trae origine la discussione di tali temi. A livello europeo, il quadro normativo, è composto dal Regolamento UE 2016/679 sulla protezione dei dati personali e dal recentissimo Regolamento UE 2024/1689, anche conosciuto come "AI Act". A livello nazionale è invece opportuno considerare il disegno di legge n. 1146 in materia di intelligenza artificiale presentato in Senato in data 20 maggio 2024, ed attualmente in corso di esame in commissione. Nel corso del capitolo verranno anche evidenziate le sfide che la diffusione intelligenza artificiale ha sollevato nell'ambito della protezione dei dati ed il rapporto tra le due normative europee.

1. La definizione di IA

La definizione giuridica di IA , che oggi troviamo all'articolo 3 numero (1) dell'AI Act[91], ha costituito, per le Istituzioni dell'Unione un vero e proprio impegno di sintesi, di ricerca della neutralità verbale ed è stato teso il più possibile a costruire una locuzione che comprendesse il maggior numero di tipi di sistemi di IA[92]. Una prima definizione era stata

[91] Articolo 3(1) AI Act: *"Ai fini del presente regolamento si applicano le definizioni seguenti:*
1)«sistema di IA»: un sistema automatizzato progettato per funzionare con livelli di autonomia variabili e che può presentare adattabilità dopo la diffusione e che, per obiettivi espliciti o impliciti, deduce dall'input che riceve come generare output quali previsioni, contenuti, raccomandazioni o decisioni che possono influenzare ambienti fisici o virtuali; […]"

[92] G. Contissa, F. Galli, F. Gordano, G. Sartor, Il Regolamento europeo sull'intelligenza artificiale, in "i-lex. Scienze Giuridiche, Scienze Cognitive e Intelligenza Artificiale", Rivista semestrale online. Fascicolo 2. Dicembre 2021. ISSN 1825-1927. Pg 8

data dalla Commissione nella sua comunicazione *"L'intelligenza artificiale per l'Europa"* del 2018[93]. Questa definizione, seppur ingenua nella sua chiarezza, contiene in nuce dei riferimenti a caratteristiche distintive dell'IA che riscontriamo oggi nell'art.3 AI Act. quali: l'autonomia nel compimento di azioni e la presenza di obiettivi specifici al funzionamento.

Nel 2020 è stato pubblicato dalla Commissione Europea il Libro Bianco[94] sull'intelligenza artificiale, documento con cui la Commissione intendeva promuovere gli investimenti per una più ampia adozione dell'IA a favore dello sviluppo del settore digitale in Europa, pur mantenendo l'impegno per affrontare i rischi associati al suo impiego. Per quel che riguarda l'ambito normativo, il Libro Bianco ha definito come prioritari degli interventi in tema di governance di dati, regimi di responsabilità e diritti fondamentali. Il documento della Commissione, seppur non contenendo una definizione di IA, è riuscito a porre l'accento in modo efficace sui suoi elementi principali, ovvero: algoritmi e dati che attraverso le strutture di calcolo generano valore aggiunto. Nello stesso documento, trattando l'ambito di applicazione del futuro quadro normativo UE in tema IA, la Commissione ha espresso l'obiettivo a cui una futura definizione avrebbe dovuto mirare: *"qualunque nuovo strumento giuridico dovrà comprendere una definizione di IA abbastanza flessibile da accogliere il progresso tecnico, ma anche sufficientemente precisa da garantire la necessaria certezza del diritto*[95]".

La definizione giuridica di intelligenza artificiale, nella proposta di legge europea sull'IA (di seguito anche "AI Act"), rispecchia l'obiettivo di bilanciare l'urgenza di trovare una definizione in grado di resistere alla rapida evoluzione tecnologica, con l'esigenza di tenere un approccio analitico. Questa stessa definizione ha subito vari affinamenti durante l'iter legislativo. Inizialmente, nella proposta di regolamento, la definizione era articolata in due parti. La prima, contenuta nel vecchio articolo 3, forniva un'ampia definizione

[93] Commissione Europea-Comunicazione Della Commissione Al Parlamento Europeo, Al Consiglio, Al Comitato Economico E Sociale Europeo E Al Comitato Delle Regioni. L'intelligenza artificiale per l'Europa-COM(2018) 237 final. Bruxelles 25 aprile 2018. Pg. 1: "*Intelligenza artificiale (IA) indica sistemi che mostrano un comportamento intelligente analizzando il proprio ambiente e compiendo azioni, con un certo grado di autonomia, per raggiungere specifici obiettivi. I sistemi basati sull'IA possono consistere solo in software che agiscono nel mondo virtuale (ad esempio assistenti vocali, software per l'analisi delle immagini, motori di ricerca, sistemi di riconoscimento vocale e facciale), oppure incorporare l'IA in dispositivi hardware (per esempio in robot avanzati, auto a guida autonoma, droni o applicazioni dell'Internet delle cose)."*

[94]Commissione Europea Bruxelles, - Libro Bianco sull'Intelligenza Artificiale - Un approccio europeo all'eccellenza e alla fiducia1-9.2.2020 COM(2020) 65 final

[95] Ibidem. Cit.pg. 19

dell'IA quale un *software* capace di generare una varietà di *output* sulla base di obiettivi definiti dall'uomo, ed in grado di influenzare l'ambiente con il quale interagisce. La seconda parte era costituita dall'Allegato I (che nella versione attuale del testo normativo ha cambiato contenuto), la quale prevedeva un dettagliato elenco di tecniche, a loro volta raggruppate in tre modelli generali: il *machine learning*, inclusivo dell'apprendimento supervisionato e non quello per rinforzo; gli approcci basati su logica e conoscenza, tra cui i sistemi simbolici e infine gli approcci statistici, comprensivi di metodi di ottimizzazione e di ricerca[96]. Seguendo quest'impostazione il legislatore aveva previsto, per la Commissione, il potere delegato di modificare l'elenco all'Allegato I, per permettere alla definizione di essere sempre adeguata agli sviluppi tecnologici e di mercato. Tuttavia, a quest'approccio non è stato dato un seguito nella versione definitiva. Infatti, è stato criticato, in quanto foriero di incertezza e vuoti normativi; questo sia perché un elenco di tecniche avrebbe portato ad un approccio troppo letterale al suo contenuto, sia per il timore della difficoltà di catalogare gli sviluppi tecnologici futuri nelle tre categorie approntate.

In conseguenza del dibattito avvenuto in sede di discussione del testo, la versione corrente risulta più essenziale ed incentrata sulle caratteristiche fondamentali dei sistemi di IA, richiamando quindi la loro autonomia nel funzionamento, l'adattabilità loro conferita dagli algoritmi, e la loro capacità di deduzione e generazione a partire da un determinato *input*. È stato congedato ogni riferimento ai *software* perché ritenuto un temine potenzialmente portatore di confusione con i programmi di calcolo tradizionali; così come è stato ritenuto superfluo specificare che i compiti siano predefiniti dall'uomo[97].

Dunque, secondo l'attuale definizione contenuta nel sopracitato articolo 3(1) AI Act, un sistema di IA è: *"Un sistema automatizzato progettato per funzionare con livelli di autonomia variabili e che può presentare adattabilità dopo la diffusione e che, per obiettivi espliciti o impliciti, deduce dall'input che riceve come generare output quali previsioni, contenuti, raccomandazioni o decisioni che possono influenzare ambienti fisici o virtuali"*.

[96] G. Contissa, F. Galli, F. Gordano, G. Sartor, Il Regolamento europeo sull'intelligenza artificiale, in "i-lex. Scienze Giuridiche, Scienze Cognitive e Intelligenza Artificiale", Rivista semestrale online. Fascicolo 2. Dicembre 2021. ISSN 1825-1927

[97] Lo Sapio G, Intelligenza artificiale: rischi, modelli regolatori, metafore, in federalismi.it – ISSN 1826-3534, n. 27/2022. P. 2

2. AI Act: i sistemi di IA nell'approccio basato sul rischio

Nell'aprile 2021 la Commissione Europea ha presentato la proposta di legge europea sull'IA, avviando così un processo legislativo per creare un quadro normativo che regolasse le tecnologie di IA operanti nell'Unione. Dopo molti mesi di lavori che hanno coinvolto i tre legislatori (Parlamento Europeo, Commissione e Consiglio), è stato raggiunto un consenso sul testo del Regolamento. Dopo la sua pubblicazione formale nella Gazzetta Ufficiale dell'UE il 12 luglio 2024, la sua applicazione, generalmente prevista a partire dall'agosto 2026, è stata anticipata e dilazionata in varie fasi: a partire dal sesto mese *post*-entrata in vigore fino ad arrivare ai 36 mesi per taluni capi che contengono le previsioni di più urgente applicazione (quali i capi I e II riguardanti disposizioni generali e pratiche vietate). Sebbene l'ambito di applicazione del Regolamento sull'IA comprenda tutte le applicazioni dell'intelligenza artificiale tranne, come specificamente previsto all'articolo 2, i sistemi destinati a scopo di ricerca scientifica, scopi militari o sicurezza nazionale, solo alcuni usi dell'IA sono soggetti ad una regolamentazione vincolante. Per l'appunto, il Regolamento sull'IA riconduce i sistemi di IA in quattro categorie di rischio (pratiche vietate, alto rischio, limitato e basso), che a loro volta definiscono il rigore delle norme di conformità.

Le quattro categorie di rischio dipendono nella pratica, dai campi di applicazione dei sistemi di intelligenza artificiale. Per fare un esempio: le IA sviluppate per la creazione o l'ampliamento di banche dati di riconoscimento facciale o per il punteggio sociale sono considerate inaccettabilmente rischiose e pertanto vietate[98]. Analogamente, le IA utilizzate in settori quali l'istruzione, l'occupazione, la migrazione, la giustizia e l'applicazione della legge sono considerate ad alto rischio e, pertanto, sono sottoposte a procedure di valutazione della conformità e richiedono garanzie aggiuntive (di cui si

[98] Vedi art.5 lett. (c)(e) AI Act: "*1. Sono vietate le pratiche di IA seguenti: […]*
(c) l'immissione sul mercato, la messa in servizio o l'uso di sistemi di IA per la valutazione o la classificazione delle persone fisiche o di gruppi di persone per un determinato periodo di tempo sulla base del loro comportamento sociale o di caratteristiche personali o della personalità note, inferite o previste, in cui il punteggio sociale così ottenuto comporti il verificarsi di uno o di entrambi gli scenari seguenti:
i) un trattamento pregiudizievole o sfavorevole di determinate persone fisiche o di gruppi di persone in contesti sociali che non sono collegati ai contesti in cui i dati sono stati originariamente generati o raccolti;
ii) un trattamento pregiudizievole o sfavorevole di determinate persone fisiche o di gruppi di persone che sia ingiustificato o sproporzionato rispetto al loro comportamento sociale o alla sua gravità;
[…]
(e) l'immissione sul mercato, la messa in servizio per tale finalità specifica o l'uso di sistemi di IA che creano o ampliano le banche dati di riconoscimento facciale mediante scraping non mirato di immagini facciali da internet o da filmati di telecamere a circuito chiuso;"

occupa la sezione 2 del Capo III del regolamento sull'IA contenente i requisiti per i sistemi di IA ad alto rischio).

2.1 Il rischio inaccettabile e le pratiche di IA vietate

La categoria che implica rischi inaccettabili riguarda tutti gli usi vietati delle tecnologie di intelligenza artificiale, compresi i sistemi che utilizzano tecniche subliminali per manipolare il comportamento umano, quelli utilizzati per la classificazione sociale, per lo *scraping* di immagini facciali da Internet o da telecamere a circuito chiuso e per l'identificazione biometrica remota in tempo reale a fini di attività di contrasto a reati.
L'articolo 5 dell'AI Act, si apre con un elenco dettagliato delle pratiche vietate.
Sub a) e b) vengono distinte due forme di manipolazione, di cui la seconda - sub b) - specificamente indirizzata ad un *target* di soggetti ritenuti a vario titolo vulnerabili (per età, disabilità o situazione sociale). Per parlare di manipolazione effettiva è necessario che sussista, in entrambe le ipotesi, l'intenzione manipolativa, che, nel caso sub a), si concretizza in tecniche subliminali e, per loro natura, nascoste; nel caso sub b) si avvale della specifica vulnerabilità della vittima. Tali tecniche manipolative devono essere concretamente in grado di produrre un danno fisico o psicologico, definito "significativo". Vale la pena evidenziare che non è richiesta la prova dell'esistenza di un nesso causale fra la manipolazione ed il danno, questo perché anche la mera potenzialità produttiva di un danno, rileva ai fini della sussistenza della manipolazione. Risulta poi implicito, nella norma, che il soggetto, a cui fa capo la responsabilità di ottemperare al divieto, è colui che utilizza, commercializza e mette in funzione questi sistemi di IA. Secondo l'analisi di Raposo[99], però, questa impostazione soffre di alcune incertezze: innanzitutto, non è chiaro se il danno potenziale, richiesto da entrambi i paragrafi, debba essere causato da un singolo evento, in grado di produrre, di per sé, un effetto manipolativo, o se possa essere causato da una pluralità di eventi, nessuno dei quali, se

[99] RAPOSO V. L., Ex Machina: preliminary critical assessment of the European Draft Act on artificial intelligence, *in International Journal of Law and Information Technology*, n.30/2022.
Che a sua volta trae la sua analisi da:
Veale, Michael and Zuiderveen Borgesius, Frederik, Demystifying the Draft EU Artificial Intelligence Act (July 31, 2021). Computer Law Review International (2021) 22(4) 97-112, Available at SSRN: https://ssrn.com/abstract=3896852

preso singolarmente, sufficiente a causare un tale effetto, ma in grado di farlo quando operanti insieme e ripetutamente nel tempo.

Al comma 1 lett. c) dell'art. 5 AI Act viene fatto divieto generalizzato dell'impiego di sistemi di *social scoring*, ovvero sistemi di IA destinati a generare punteggi valutativi del comportamento o di caratteristiche di una persona fisica. Il divieto in questione è legato, però, alla concretizzazione, anche alternativa, di due scenari. Il primo prevede che la classificazione del soggetto comporti, per lui, un trattamento pregiudizievole, che vada ad inficiare contesti sociali diversi da quello da cui provengono i dati. Il secondo scenario, che può quindi anche verificarsi simultaneamente al primo, prevede che il soggetto subisca, in conseguenza del suo comportamento, un trattamento pregiudizievole che risulti sproporzionato o finanche ingiusto. Nello stesso comma, sub d), troviamo una ulteriore specificazione del divieto di *social scoring* che, in questo caso, riguarda sistemi di IA che, attraverso la profilazione e la valutazione di tratti di personalità, riescono a stimare il rischio criminogeno di un individuo. La formulazione attuale del divieto di *social scoring* risulta, rispetto alla sua originaria formulazione nella proposta fatta dalla Commissione, certamente ampliato nella sua portata[100], tuttavia conserva l'implicita esclusione del divieto dello *scoring* episodico; infatti, la norma richiede che il sistema funzioni "*per un certo periodo di tempo*". Resta dunque aperta la questione del perché lo *scoring* "una tantum" sia stato ritenuto così poco preoccupante, per la tutela dei diritti fondamentali, da essere stato escluso dallo spettro applicativo della norma.

Infine, resta da trattare brevemente quali impieghi dei sistemi di identificazione biometrica siano vietati. L'articolo 5 sub (1)(e) dell'AI Act vieta la creazione e l'ampliamento di banche dati per il riconoscimento facciale, quando siano realizzate a mediante *scraping* indiscriminato di immagini da internet o filmati di CCTV. Alla lett. (g) del medesimo articolo, vengono vietati sia a livello di utilizzo, che di messa in commercio, i sistemi di categorizzazione biometrica, che abbiano la finalità di dedurre ed inferire dettagli umanamente sensibili sui soggetti analizzati. Sono ritenute informazioni sensibili, e perciò destinate a rimanere nell'assoluta disposizione della persona fisica a cui appartengono, quelle che riguardano: "*razza, opinioni politiche, appartenenza*

[100] La versione contenuta nella proposta restringeva il campo del divieto ai soli sistemi di IA utilizzati da o per conto di autorità pubbliche: "*L'immissione sul mercato, la messa in servizio o l'uso di sistemi di IA da parte delle autorità pubbliche o per loro conto ai fini della valutazione o della classificazione dell'affidabilità delle persone fisiche per un determinato periodo di tempo […]"*

sindacale, convinzioni religiose o filosofiche, vita sessuale o orientamento sessuale". Resta fuori dal divieto la categorizzazione, nel contesto di attività di contrasto, se basata di dati biometrici legalmente ottenuti. Come ultimo aspetto, merita attenzione quanto previsto per i sistemi di identificazione biometrica remota in tempo reale, quando sia operata in spazi pubblici a fini di contrasto. Di base è fatto divieto di ogni uso. A questo punto, tuttavia, l'intersecarsi di eccezioni e specificazioni si fa complesso ed il divieto viene ulteriormente dettagliato. Sono quindi utilizzabili sistemi di IA in ambito penale per: 1) ricerca mirata di vittime di reati particolarmente gravi come tratta di esseri umani e rapimento; 2) prevenzione di una minaccia specifica ed imminente alla vita o all'incolumità delle persone fisiche o di un atto terroristico; 3) localizzazione o identificazione di una persona sospettata di aver commesso uno dei reati contenuti nella lista dell'Allegato II dell'AI Act[101]. Al comma 2 dello stesso articolo 5, viene specificato che l'uso di detti sistemi di identificazione biometrica nelle ipotesi poc'anzi illustrate, è ammesso al solo scopo di confermare l'identità di una persona specifica (quindi sulla base di una precedente ipotesi, formulata sulla base di altri strumenti). In aggiunta, l'autorità di contrasto, prima di procedere con il sistema biometrico dovrebbe soppesare due elementi: a) cosa succederebbe in caso di mancato uso del sistema; b) quali conseguenze ci sarebbero per le libertà ed i diritti delle persone coinvolte. Queste considerazioni fanno parte della tappa obbligatoria che l'autorità di contrasto deve superare per essere autorizzata all'uso del sistema biometrico. Tali considerazioni fanno parte della valutazione d'impatto sui diritti fondamentali, disciplinata all'articolo 27 AI Act (di cui si parlera più avanti) e sono il prodromo necessario per poter registrare il sistema di IA nella banca dati UE a norma degli artt. 49 e 71 AI Act. Ai commi 3 e 4 viene prescritta un'altra autorizzazione preliminare, la quale deve provenire da un'autorità giudiziaria dello stato membro in cui avverrà l'uso; in secondo luogo, l'uso del sistema biometrico in questione deve essere notificato alle autorità di vigilanza del mercato e della protezione dei dati pertinenti. A conclusione di questo sguardo d'insieme sulle tecnologie di

[101]Allegato II AI Act: *"Reati di cui all'articolo 5, comma 1, primo comma, lettera h), punto iii): — terrorismo, — tratta di esseri umani, — sfruttamento sessuale di minori e pornografia minorile, — traffico illecito di stupefacenti o sostanze psicotrope, — traffico illecito di armi, munizioni ed esplosivi, — omicidio volontario, lesioni gravi, — traffico illecito di organi e tessuti umani, — traffico illecito di materie nucleari e radioattive, — sequestro, detenzione illegale e presa di ostaggi, — reati che rientrano nella competenza giurisdizionale della Corte penale internazionale, — illecita cattura di aeromobile o nave, — violenza sessuale, — reato ambientale, — rapina organizzata o a mano armata, — sabotaggio, — partecipazione a un'organizzazione criminale coinvolta in uno o più dei reati elencati sopra*."

intelligenza artificiale che il Regolamento sull'IA ha ritenuto foriere di rischi inaccettabili, vale la pena fare un piccolo salto indietro nel testo della norma considerata e notare che, la fine del primo comma dell'articolo 5, sembra quasi voler dare una chiara conferma del dubbio che potrebbe sorgere proseguendo nella lettura dell'articolo, ovvero, che cosa ne sia del trattamento dei dati biometrici a fini diversi dall'attività di contrasto a reati di cui per l'appunto al comma (1) lett. (h). Al riguardo, chiarezza viene fatta dal rimando diretto[102] a quanto stabilito dall'articolo 9 del regolamento (UE) 2016/679 nel suo articolato divieto di trattamento delle categorie particolari di dati personali.

2.2 Alto Rischio

Il capo III dell'AI Act contiene delle disposizioni mirate ai sistemi di IA cosiddetti ad alto rischio. Il rischio da cui si intendono salvaguardare le persone fisiche è quello per la loro salute, sicurezza e diritti fondamentali. A differenza dei sistemi che comportano un rischio inaccettabile ai sensi dell'AI Act, i sistemi ad alto rischio sono ammessi nel mercato europeo solo se in possesso di specifici requisiti, accompagnati da una valutazione di conformità preliminare. La classificazione di un sistema di IA come ad alto rischio dipende non solo dalla sua funzione, ma anche dalla sua finalità e modalità di utilizzo. La sezione 1 del capo III AI Act divide i sistemi di IA ad alto rischio in due categorie principali: da un lato i sistemi che sono componenti di sicurezza di prodotti soggetti a valutazione di conformità *ex ante*; dall'altro sistemi *stand alone* (indipendenti), anche loro soggetti a valutazione *ex ante*. Nell'allegato III sono contenuti in forma di elenco e raggruppati per settori di applicazione, vari esempi di sistemi ad alto rischio tra cui: quelli utilizzati nella biometria, nell'istruzione, nelle infrastrutture critiche, nei servizi pubblici essenziali e nelle attività di contrasto. L'elenco di cui all'allegato III non è fatto per essere esaustivo; perciò, per massimizzare le possibilità che il Regolamento si adatti ai mutevoli usi e applicazioni dell'IA, all'articolo 7 AI Act è previsto che la Commissione possa ampliare tale lista seppur dovendo tenere conto dei settori dell'allegato III e operando una valutazione di rischi per i nuovi sistemi da inserire.

[102] Articolo 5 comma(1) (h) AI Act: *"La lettera h) del primo comma lascia impregiudicato l'articolo 9 del regolamento (UE) 2016/679 per quanto riguarda il trattamento dei dati biometrici a fini diversi dall'attività di contrasto" […]*

Nella sezione 2 dell'AI Act sono contenute disposizioni che arricchiscono di ulteriori requisiti giuridici le condizioni d'uso dei sistemi di IA ad alto rischio, tra cui l'istituzione di un sistema di gestione dei rischi, la redazione e conservazione della documentazione tecnica, gli obblighi di trasparenza e la fornitura di informazioni agli utenti da parte dei distributori e, ma non in ultimo, la *governance* dei dati utilizzati per l'addestramento di modelli di IA. La sezione 2 costituisce un piccolo gruppo coerente di norme che in fatto di requisiti minimi e richiede agli operatori sforzi che costituiscono già lo stato dell'arte per le media degli operatori diligenti. Dato il rapido sviluppo delle abilità tecnologiche ed ingegneristiche, viene lasciato spazio ad una certa flessibilità nel raggiungere la conformità a tali requisiti, questo tramite norme o specifiche tecniche.

La sezione 3 dell'AI Act definisce una serie di obblighi orizzontali per i fornitori di sistemi di IA ad alto rischio. Sono previsti obblighi commisurati all'entità della loro partecipazione lungo la catena del valore dell'IA, anche per soggetti terzi, tra cui: importatori, distributori e rappresentanti autorizzati[103].

2.3 Rischio limitato e basso

Tali sistemi non sono tenuti a sottoporsi alla procedura in due fasi. I pochi requisiti a cui sono soggetti i sistemi di IA a rischio basso si limitano principalmente agli obblighi di trasparenza. Tuttavia, i fornitori possono volontariamente creare i propri codici di condotta (per i sistemi a rischio basso) che incorporano alcuni dei requisiti stabiliti nella normativa. Poiché non è stato istituito alcun sistema per controllare il rispetto di tali codici, la loro applicazione potrebbe essere di fatto limitata.

Tra i sistemi a rischio limitato e basso rientrano applicazioni di intelligenza artificiale come *chatbot* e *deepfake*. Si ritiene che questi sistemi comportino un rischio limitato e che gli utilizzatori debbano solo assicurarsi che gli utenti finali siano consapevoli del fatto che stiano interagendo con una tecnologia di IA o che siano comunque esposti ad essa (articolo 50 AI Act).

[103] Commissione Europea-Proposta Di Regolamento Del Parlamento Europeo E Del Consiglio Che Stabilisce Regole Armonizzate Sull'intelligenza Artificiale (Legge Sull'intelligenza Artificiale) E Modifica Alcuni Atti Legislativi Dell'unione.- Bruxelles, 21.4.2021 COM(2021) 206 final 2021/0106 (COD) Reperibile su: https://eur-lex.europa.eu/resource.html?uri=cellar:e0649735-a372-11eb-9585-01aa75ed71a1.0006.02/DOC_1&format=PDF

2.4 I diversi aspetti alla base della regolazione della IA

L'emanazione del AI Act è stato frutto del complesso trilogo tra Commissione, Consiglio e Parlamento ed è una normativa complessa dal punto di vista delle ragioni e degli obiettivi da cui è sostenuta. A tale riguardo C.Schepisi[104] ha prodotto un interessante approfondimento riguardante il bilanciamento dei diritti sotteso alla struttura multilivello del rischio nel Regolamento.

Come emerge dalla relazione, premessa alla allora proposta di regolamento[105] e dal Libro Bianco sull'IA[106], l'interesse a sostenere lo sviluppo della competitività tecnologica dell'Unione, e quindi di regolarne gli aspetti tecnici dei sistemi di IA, fa da sfondo all'obiettivo precipuo del Regolamento. L'auspicio infatti è quello di contribuire, con questo quadro giuridico a creare "*un ecosistema di fiducia per un'IA affidabile*". Il fattore umano o, meglio, mutuando l'aggettivo dall'articolo 1(1) AI Act: "antropocentrico" risulta essere l'obiettivo trainante che ha guidato il lavoro di elaborazione del Regolamento sull'IA. La regolazione del mercato basato sui sistemi di IA passa quindi necessariamente attraverso la lente del rispetto dei diritti fondamentali. Rispetto che viene incarnato dal diverso livello di tutela dei diritti sotteso ai diversi livelli di rischio. E ancora: un rispetto che consiste nell'assegnare a soggetti diversi (ad esempio gli utilizzatori o i produttori), divieti od obblighi tecnico-procedimentali graduati a seconda del livello di tutela considerato. In questa maniera la tutela dei diritti fondamentali risulta incorporata fin dalla progettazione nel sistema di IA.

A livello europeo, la regolazione dell'IA è stata uno sforzo di generalizzazione normativa per trovare una soluzione *passe-partout* alla minaccia della frammentazione del mercato interno, che vedeva ciascuno Stato Membro nella posizione di porre unilateralmente limiti e divieti all'uso di sistemi di IA, con la potenzialità di ostacolare la libera circolazione dei servizi e dei prodotti. La scelta calibrare le differenti applicazioni di IA seguendo l'entità

---

[104] C. Schepisi, Le "dimensioni della regolazione dell'intelligenza artificiale nella proposta di regolamento della Commissione, in Quaderni AISDUE, ISSN 2723-9969, Sezione "Atti convegni AISDUE", n. 16/2022, p. 334

[105]Commissione Europea-Proposta Di Regolamento Del Parlamento Europeo E Del Consiglio Che Stabilisce Regole Armonizzate Sull'intelligenza Artificiale (Legge Sull'intelligenza Artificiale) E Modifica Alcuni Atti Legislativi Dell'unione.- Bruxelles, 21.4.2021 COM(2021) 206 final 2021/0106 (COD) Reperibile su: https://eur-lex.europa.eu/resource.html?uri=cellar:e0649735-a372-11eb-9585-01aa75ed71a1.0006.02/DOC_1&format=PDF

[106]Commissione Europea Bruxelles, - Libro Bianco sull'Intelligenza Artificiale - Un approccio europeo all'eccellenza e alla fiducia1-9.2.2020 COM(2020) 65 final.Consultabile a:eur-lex.europa.eu/legal-content/IT/TXT/PDF/?uri=CELEX:52020DC0065

del rischio che comporta per l'integrità dei diritti fondamentali, va sempre letta in quest'ottica di unitarietà di comportamento da parte degli Stati Membri. L'approccio basato sul rischio garantisce un'azione trasversale che copre qualunque (e di ogni dimensione) settore, uso, operatore. Anche perché, optare per una disciplina verticale, ossia diretta a disciplinare i diversi settori in cui i sistemi di IA possono destare preoccupazione, o addirittura una disciplina mirata ai vari sistemi (ad esempio al solo *machine learning*), sarebbe risultata decisamente risicata viste le innumerevoli applicazioni ed impatti che un sistema di IA può avere.

L'approccio basato sul rischio multilivello implica l'intenzione di difendere i diritti fondamentali da potenziali lesioni talmente gravi da risultare irreparabili o non compensabili a posteriori. Si pensi ai danni cagionati all'integrità psichica ma anche fisica di un soggetto. Ci troviamo di fronte all'applicazione del principio di prevenzione del danno, il quale è riconosciuto a livello nazionale e internazionale; in ambito di IA è stato richiamato anche dal *Comitato ad hoc per l'intelligenza artificiale*[107][108]. È all'interno di questo ragionamento che si possono comprendere, ad esempio, le radici della scelta di creare una classe di rischio ritenuta inaccettabile e perciò arrivare a vietare la commercializzazione e, a monte, anche lo sviluppo dei sistemi di IA che ne sono ricompresi. Tra i diritti fondamentali a cui si è accennato poco sopra e che in rapporto a questa categoria divengono irriducibili e prevalenti rispetto a esigenze di carattere economico e ad altri diritti, ci sono: la tutela della vita, della salute e della dignità della persona. In questa prospettiva di bilanciamento dei diritti in campo, si comprende meglio anche il senso delle deroghe, ossia i casi in cui il rischio di lesione di un diritto non prevale sui benefici derivanti dall'applicazione di un sistema di IA; finanche ad arrivare al

[107] La CAHAI ha adempiuto al suo mandato (2019-2021) ed è stata sostituita dal Comitato per l'Intelligenza Artificiale (CAI). Per approfondire l'attività dell'organo consultare la pagina dedicata sul sito del Consiglio d'Europa: https://www.coe.int/en/web/artificial-intelligence/cai

[108] Consiglio d'Europa (Ad Hoc Committee On Artificial Intelligence)- CAHAI (2020)23- *Studio di fattibilità su un quadro giuridico per la progettazione, lo sviluppo e l'applicazione dell'IA basato sulle norme del Consiglio d'Europa,* adottato dalla CAHAI a Strasburgo il 17 dicembre 2020. P. 28: *"La prevenzione del danno è un principio fondamentale che dovrebbe essere sostenuto, sia nella dimensione individuale che in quella collettiva, soprattutto quando tale danno riguarda l'impatto negativo sui diritti umani, sulla democrazia e sullo Stato di diritto. L'integrità fisica e mentale degli esseri umani deve essere adeguatamente tutelata, con ulteriori garanzie per le persone e i gruppi più vulnerabili. Particolare attenzione deve essere prestata anche alle situazioni in cui l'uso di sistemi di IA può causare o aggravare impatti negativi dovuti ad asimmetrie di potere o di informazione, ad esempio tra datori di lavoro e lavoratori, imprese e consumatori o governi e cittadini."*
Reperibile su: https://www.coe.int/en/web/artificial-intelligence/cahai

sacrificio di un diritto per avere la possibilità di proteggerne e rafforzarne uno prevalente: si pensi ai sistemi utilizzati per le diagnosi mediche o a quelli in grado di prevenire gli attacchi terroristici. In sunto, la scelta di strutturare l'AI Act sul "peso" dei diritti fondamentali (a tal punto da surclassare non solo le esigenze di mercato ma anche certuni diritti rispetto ad altri), è importante perché riesce a creare un bagaglio valoriale comune a livello europeo e superiore rispetto a quelli di ciascuno Stato membro. Troviamo concretizzazione del meccanismo appena descritto nel modo in cui sono regolati i sistemi di riconoscimento biometrico in tempo reale da parte delle autorità di contrasto: sono generalmente vietati perché lesivi del diritto alla vita privata ed alla riservatezza; tuttavia, l'utilizzo di questi sistemi viene concesso se utilizzati in ambito penale sia per l'individuazione delle vittime di reato sia per i loro autori, ecco allora che vediamo cedere il diritto alla vita privata e alla riservatezza di fronte al diritto tutela dell'integrità fisica. Di fronte, nel caso in esempio, al doveroso obiettivo di un'autorità di contrasto, diventa giustificabile la momentanea compressione di un diritto fondamentale. Seguendo questa scia appare contraddittoria la scelta del legislatore di "scontare" ai sistemi di *deep fake* una disciplina più limitante, nonostante essi possano rappresentare una minaccia alla dignità umana e al diritto all'informazione. Anche il Parlamento europeo si è espresso in tal senso nella sua risoluzione del 20 gennaio 2021[109]. A norma dell'articolo 50 comma 4 AI Act per i sistemi di *deep fake* vige un obbligo informativo da parte dei distributori di rendere noto che si tratta di un contenuto manipolato artificialmente; è previsa però una duplice deroga a tale obbligo: la prima va nel senso di garantire alla collettività la protezione da parte delle autorità di contrasto, e quindi l'uso di *deep fake* può non essere dichiarato in caso di autorizzazione *ex lege* al fine di accertare, prevenire o perseguire

[109] Parlamento Europeo- "Intelligenza artificiale: questioni relative all'interpretazione e applicazione del diritto internazionale". Risoluzione del Parlamento europeo del 20 gennaio 2021 sull'intelligenza artificiale: questioni relative all'interpretazione e applicazione del diritto internazionale nella misura in cui l'UE è interessata relativamente agli impieghi civili e militari e all'autorità dello Stato al di fuori dell'ambito della giustizia penale (2020/2013(INI). G.U. (2021/C 456/04) Paragrafo 76. *"[…] profonda preoccupazione per le tecnologie di deepfake, che consentono di produrre foto, audio e video falsificati sempre più realistici che potrebbero essere utilizzati per compiere ricatti, creare notizie false o minare la fiducia dei cittadini e influenzare il dibattito pubblico; ritiene che tali pratiche siano in grado di destabilizzare paesi, diffondere la disinformazione e influenzare le consultazioni elettorali; chiede pertanto l'introduzione di un obbligo in base al quale tutti i materiali deepfake o altri video artificiali realizzati in modo realistico debbano essere etichettati come «non originali» dal loro creatore, con severi limiti al loro utilizzo a fini elettorali, e che tale obbligo sia applicato rigorosamente; chiede che siano svolte adeguate attività di ricerca in questo campo per garantire che le tecnologie di contrasto dei suddetti fenomeni siano al passo con gli utilizzi dolosi dell'IA.*"

reati; la seconda deroga segue la strada della tutela alla libertà di espressione. Infatti, nell'ambito di programmi artistici o comunque manifestamente fittizi, l'obbligo informativo si limita a rendere nota l'esistenza di tali contenuti. Nell'esempio fatto, la libertà di espressione spicca sugli altri diritti considerati dalla norma. Questa impostazione appare tuttavia in contrasto con l'importanza che l'Unione riserva alla salvaguardia della dignità della persona.

3. L'IA generativa nella disciplina del AI Act

L'intelligenza artificiale generativa è un ramo significativo dell'IA; la sua peculiarità è quella di poter creare contenuti nuovi, ossia non precedentemente esistenti nei dati utilizzati per lo sviluppo, e generati riproducendo nel modo più dettagliato possibile i dati di *input*. Seguendo l'ordine della normativa, la quarta distinzione operata all'interno dei sistemi di IA ha ad oggetto la cosiddetta IA per uso generale (abbreviato in GPAI), ovvero un tipo di sistemi o modelli di IA, che sono stati inclusi nel testo del Regolamento solo nelle ultime fasi dell'*iter* legislativo. La definizione data dal Regolamento sull'IA all'art. 3 (63) di questi modelli fa riferimento ad un insieme di tecnologie di IA caratterizzate da una "*significativa generalità*" e "*in grado di svolgere con competenza un'ampia gamma di compiti distinti*" Questa definizione viene in certa misura dettagliata e circoscritta dal dettato dei Considerando 98 e 99, i quali forniscono un esempio di cosa determini la "*generalità significativa*" di un modello. Nello specifico, il Considerando 98 AI Act afferma che: sebbene la generalità di un modello possa essere determinata da vari parametri, c'è una caratteristica in particolare che dimostra la "*generalità significativa*" del modello e la sua attitudine a svolgere "*con competenza un'ampia gamma di compiti distinti*". Tale caratteristica concerne la quantità di parametri e di dati utilizzati per addestrarlo, infatti: "*i modelli con almeno un miliardo di parametri e addestrati con grandi quantità di dati utilizzando l'aut- supervisione su larga scala dovrebbero ritenersi caratterizzati da una generalità significativa e in grado di svolgere con competenza un'ampia gamma di compiti distinti*" [110].

[110] Considerando 98 AI Act

Il Considerando 99 AI Act aggiunge che i modelli generativi sono un esempio di modello di IA per uso generale perché "*consentono una generazione flessibile di contenuti, ad esempio sotto forma di testo, audio, immagini o video, che possono prontamente rispondere a un'ampia gamma di compiti distinti*"[111].

In questa definizione sono ricompresi i modelli linguistici di grandi dimensioni (LLM) e altri modelli di base. Questo gruppo di tecnologie di IA generativa è stato sottoposto a una regolamentazione più rigorosa a causa della loro capacità di influenzare il raggiungimento degli obiettivi dell'AI Act. In questo caso, si possono distinguere due sottocategorie, in quanto è stato previsto un regime normativo separato e più rigoroso per la GPAI che può comportare un rischio sistemico. Le tecnologie GPAI possono essere classificate come modelli con rischio sistemico se presentano *"capacità di impatto elevato valutate sulla base di strumenti e metodologie tecniche adeguati*", come specificato nel Regolamento sull'IA all'art. 51(1) lett. (a)(b) o sulla base di una decisione della Commissione[112]. Vige infine, una presunzione che il modello GPAI abbia capacità considerate ad alto impatto quando la quantità di potenza di calcolo utilizzata per l'addestramento e misurata in operazioni in virgola mobile, sia maggiore di 10^{25} (articolo 51 (2))[113]. Questa misura di riferimento è tuttavia una mera presunzione, perché può essere disattesa sia in senso negativo che positivo dalla Commissione, la quale ha la possibilità di adeguare la soglia classificatoria agli sviluppi dell'evoluzione tecnologica (ad esempio miglioramenti degli algoritmi ma anche aumento dell'efficienza degli *hardware*). Nonostante sia diffusa, anche a così pochi mesi dalla piena entrata in vigore dell' AI Act, la convinzione che nel futuro a breve raggio la potenza dei modelli di GPAI non dipenderà solo dalla potenza di calcolo, ma la predisposizione di questo primo criterio iniziale, potrebbe essere utile per capire meglio come definire la rischiosità di un

[111] Considerando 99 AI Act

[112] Articolo 51 (1) lett. (a) (b) AI Act: *"1. Un modello di IA per finalità generali è classificato come modello di IA per finalità generali con rischio sistemico se soddisfa una delle condizioni seguenti:*

a) presenta capacità di impatto elevato valutate sulla base di strumenti tecnici e metodologie adeguati, compresi indicatori e parametri di riferimento;

b) sulla base di una decisione della Commissione, ex officio o a seguito di una segnalazione qualificata del gruppo di esperti scientifici, presenta capacità o un impatto equivalenti a quelli di cui alla lettera a), tenendo conto dei criteri di cui all'allegato XIII." […]

[113] Articolo 51 (2) AI Act: […] "*sulla base di una decisione della Commissione, ex officio o a seguito di una segnalazione qualificata del gruppo di esperti scientifici, presenta capacità o un impatto equivalenti a quelli di cui alla lettera a), tenendo conto dei criteri di cui all'allegato XIII."*

sistema[114]. A questo punto, i fornitori di sistemi che soddisfano i requisiti di impatto e potenza devono notificarlo alla Commissione al più tardi entro due settimane dal raggiungimento della soglia, o anche preliminarmente, se il fornitore ne ha previsto il raggiungimento[115]. La procedura di cui all'articolo 52 AI Act prevede, oltre al potere unilaterale della Commissione di categorizzare un modello per finalità generali come *"a rischio sistemico"*, anche la possibilità da parte del fornitore di dimostrare, con argomentazioni fondate, che il proprio modello non concretizza rischio sistemico sebbene ne siano presenti i requisiti. L'articolo 52 si conclude stabilendo che la Commissione deve garantire la pubblicazione di un elenco aggiornato di GPAI con rischio sistemico.

I restanti usi dell'IA, che non sono né vietati, né classificati come ad alto rischio o GPAI con rischio sistemico, rientrano in un'ultima categoria residuale e generica. Questo gruppo costituisce, infatti, la maggior parte delle applicazioni di IA e non è soggetto ad alcun obbligo significativo, a parte l'aderenza a codici di condotta volontari (art. 95 AI Act).

3.1 Regolazione dei modelli GPAI di base e con rischio sistemico

La sezione 2 del Capo V AI Act si occupa di stabilire degli obblighi rivolti a tutti i fornitori di modelli di GPAI, a prescindere dal rischio che portano con sé. Tali obblighi minimi riguardano:

a) la trasparenza, e quindi la redazione e la messa a disposizione della documentazione tecnica del modello, ivi compresa la fase di addestramento e valutazione dei risultati;

b) la collaborazione, che ha una dimensione sia orizzontale che verticale; in orizzontale, i fornitori devono condividere tra di loro la documentazione necessaria per integrare i

[114] Artificial Intelligence Act, con il voto del Coreper il tempo dei cambiamenti è finito - Federprivacy- a cura di Innocenzo Genna. Febbraio 2024. Articolo Reperibile su: https://www.federprivacy.org/informazione/primo-piano/artificial-intelligence-act-con-il-voto-del-coreper-il-tempo-dei-cambiamenti-e-finito. Ultima visita al sito in data 19 febbraio 2025

[115] Articolo 52(1) AI Act: "1. *Se un modello di IA per finalità generali soddisfa la condizione di cui all'articolo 51, paragrafo 1, lettera a), il fornitore pertinente informa la Commissione senza ritardo e in ogni caso entro due settimane dal soddisfacimento di tale requisito o dal momento in cui viene a conoscenza che tale requisito sarà soddisfatto. Tale notifica comprende le informazioni necessarie a dimostrare che il requisito in questione è stato soddisfatto. Se la Commissione viene a conoscenza di un modello di IA per finalità generali che presenta rischi sistemici di cui non è stata informata, può decidere di designarlo come modello con rischio sistemico"* […]

GPAI in altri sistemi di IA; in verticale, i fornitori collaborano con la Commissione e le autorità nazionali in caso di necessità;

c) la divulgazione, e la predisposizione al pubblico di sintesi dettagliate ma comprensibili, riguardanti il materiale usato per l'addestramento;

d) l'attuazione di una politica in armonia con il diritto d'autore[116].

Si noti che, come stabilito all'art. 53(2) AI Act[117] ed esplicato al Considerando 104[118], nel caso dei modelli di GPAI con licenza libera e *open source,* la regolamentazione che riguarda i requisiti di trasparenza deve ritenersi alleggerita e perciò, i requisiti non-imposti, a meno che, tali modelli di GPAI presentino un rischio sistemico. Infatti, la licenza libera non può essere considerata un modo per eludere la conformità al Regolamento, né tantomeno può essere in alcun modo "abbonato" l'obbligo di rendere pubblici i contenuti usati per l'addestramento.

La sezione 3 del capo V dell' AI Act è composta da un solo articolo, il quale è dedicato agli obblighi dei fornitori di modelli di GPAI che comportano un rischio sistemico. La disciplina di questi modelli è costruita sulla base e, quindi, come integrazione di quanto detto fin qui per i modelli per finalità generali. In particolare, i fornitori:

---

[116] Articolo 53 (1) AI Act: "*1.I fornitori di modelli di IA per finalità generali:*
a) redigono e mantengono aggiornata la documentazione tecnica del modello, compresi il processo di addestramento e prova e i risultati della sua valutazione, che contiene almeno le informazioni di cui all'allegato XI affinché possa essere trasmessa, su richiesta, all'ufficio per l'IA e alle autorità nazionali competenti;
b)elaborano, mantengono aggiornate e mettono a disposizione informazioni e documentazione per i fornitori di sistemi di IA che intendono integrare il modello di IA per finalità generali nei loro sistemi di IA. Fatta salva la necessità di rispettare e proteggere i diritti di proprietà intellettuale e le informazioni commerciali riservate o i segreti commerciali conformemente al diritto dell'Unione e nazionale, le informazioni e la documentazione […]
[…] c) attuano una politica volta ad adempiere al diritto dell'Unione in materia di diritto d'autore e diritti ad esso collegati e, in particolare, a individuare e rispettare, anche attraverso tecnologie all'avanguardia, una riserva di diritti espressa a norma dell'articolo 4, paragrafo 3, della direttiva (UE) 2019/790;
d)redigono e mettono a disposizione del pubblico una sintesi sufficientemente dettagliata dei contenuti utilizzati per l'addestramento del modello di IA per finalità generali, secondo un modello fornito dall'ufficio per l'IA".[…]

[117] Articolo 53 (2) AI Act: *" 2. Gli obblighi di cui al paragrafo 1, lettere a) e b), non si applicano ai fornitori di modelli di IA rilasciati con licenza libera e open source che consentono l'accesso, l'uso, la modifica e la distribuzione del modello e i cui parametri, compresi i pesi, le informazioni sull'architettura del modello e le informazioni sull'uso del modello, sono resi pubblici. Tale eccezione non si applica ai modelli di IA per finalità generali con rischi sistemici."* […]

[118] Considerando 104 AI Act: […] *"dovrebbero essere soggetti ad eccezioni per quanto riguarda i requisiti relativi alla trasparenza imposti ai modelli di IA per finalità generali, a meno che non si possa ritenere che presentino un rischio sistemico, nel qual caso la circostanza che il modello sia trasparente e corredato di una licenza open source non dovrebbe ritenersi un motivo sufficiente per escludere la conformità agli obblighi di cui al presente regolamento."* […]

a) hanno l'obbligo di valutare i modelli con l'ausilio di strumenti standardizzati e comparandone la conformità agli appositi protocolli reperibili al momento (allo stato dell'arte). Se dal caso svolgono e documentano test contraddittori[119] per individuare ed attenuare i rischi;
b) devono occuparsi di valutare e nel caso attenuare i rischi per l'Unione derivanti dallo sviluppo, commercializzazione ed uso di modelli a rischio;
c) mantengono un dialogo collaborativo con l'ufficio per l'IA e con le autorità nazionali, tenendo traccia e documentando il verificarsi di incidenti pregiudizievoli, con l'obiettivo di elaborare unitamente delle misure rimediali e correttive;
d) devono garantire la sicurezza del modello sia a livello informatico sia a livello della sua infrastruttura fisica.

3.2 I codici di condotta per i modelli di IA per finalità generali

Un codice di condotta è uno strumento facoltativo, atto a fissare degli *standard*. Nel caso specifico sono documenti tecnici elaborati dai fornitori tramite cui essi si impegnano a garantire livelli di sicurezza, trasparenza e gestione del rischio perlomeno pari o superiori rispetto a quelli previsti dalla normativa per la categoria di riferimento. L'AI Act, a conclusione del capo V, nella sezione 4, prevede la possibilità sia per i modelli GPAI generici che per quelli a rischio sistemico, di utilizzare dei codici di condotta per dimostrare il soddisfacimento dei propri obblighi normativi. Al momento presente i codici di condotta sono un espediente per andare a compensare, durante il tempo previsto per la completa entrata in vigore della normativa, la mancanza di legislazioni nazionali integrative del da parte degli Stati Membri.

[119] Il test contraddittorio, chiamato anche *"adversarial testing"* è una pratica per garantire la sicurezza dei sistemi. Lo svolgimento prevede che i tester simulino attacchi ad un sistema per identificare vulnerabilità e punti deboli.
I tester, ovvero chi opera il test, spesso indicati anche come "avversari", utilizzano tecniche e tattiche simili a quelle utilizzate dagli aggressori reali per sfruttare potenziali falle nella sicurezza. L'obiettivo principale è capire in che modo un sistema potrebbe essere violato e migliorarne le difese sanando i problemi identificati. Nel contesto dei test dell'intelligenza artificiale, i test avversari consistono nel cercare deliberatamente di ingannare il modello con input specifici per vedere come reagisce.
Adversarial Testing for Generative AI | Machine Learning | Google for Developers. Reperibile su: https://developers.google.com/machine-learning/guides/adv-testing. Ultimo accesso al sito in data 10 febbraio 2025.

L'elaborazione avviene nell'alveo della supervisione dell'Ufficio AI (organo di controllo istituito all'interno della Commissione che ha il compito sostenere l'applicazione dell' AI Act ed indagare su possibili violazioni della normativa) in collaborazione con il Comitato per l'IA (organo che raggruppa i rappresentanti degli stati membri); qualora sia ritenuto necessario dall'Ufficio per l'IA possono essere coinvolti nel processo di redazione del codice anche le autorità nazionali competenti. Esperti indipendenti ed esponenti del mondo accademico posso essere chiamati a svolgere un ruolo supportivo. L'importanza del ruolo rivestito dall'Ufficio AI resta anche parzialmente sottinteso nella normativa: infatti, seppur non viene prescritta la necessità di un accordo formale sul testo del codice tra quest'ultimo e fornitore, in caso di disaccordo o contestazioni si otterrebbe di fatto un documento non approvato dall'organo di supervisione. Quindi, in una certa maniera, il valore del documento verrebbe sminuito. C'è da precisare che il rispetto di un codice di condotta comporta solo una semplice presunzione di rispetto degli obblighi legislativi sull'AI. Secondo una logica che va nel senso opposto, invece, violare un codice di condotta equivale a violazione degli obblighi di legge. E di conseguenza aziona l'irrogazione di sanzioni in parte previste dalla normativa ed in parte delegate allo Stato Membro[120][121].

4. L'IA generativa trova uno spazio nel GDPR?

Sebbene l'IA generativa non sia esplicitamente menzionata nel GDPR, e nonostante alcune sue disposizioni siano state messe alla prova dalle nuove modalità di trattamento dei dati rese possibili dall'utilizzo dell'IA, questa trova comunque terreno fertile per fruire dell'applicazione dei principi fondamentali del trattamento[122]. Se si pensa all'opinione dell' EDPB che definisce il quadro normativo del GDPR come "*tecnologicamente neutrale*" ed in virtù di ciò "*in grado di fronteggiare ogni cambiamento tecnologico o*

[120] Articolo 56. AI Act

[121] Artificial Intelligence Act, con il voto del Coreper il tempo dei cambiamenti è finito - Federprivacy- a cura di Innocenzo Genna. Febbraio 2024- Articolo Reperibile su: https://www.federprivacy.org/informazione/primo-piano/artificial-intelligence-act-con-il-voto-del-coreper-il-tempo-dei-cambiamenti-e-finito Ultimo accesso al sito in data 19 febbraio 2025

[122] Sartor G., The Impact Of The General Data Protection Regulation (Gdpr) On Artificial Intelligence,- European Parliamentary research service (EPRS), panel for the future of science and technology, June 2020. Reperibile sul sito del Parlamento europeo: https://www.europarl.europa.eu/RegData/etudes/STUD/2020/641530/EPRS_STU(2020)641530_EN.pdf

evoluzione", quest'impostazione non sorprende. Le parole dell'EDPB traggono origine dal considerando 15 del GDPR in cui si sancisce che: *"Al fine di evitare l'insorgere di gravi rischi di elusione, la protezione delle persone fisiche dovrebbe essere neutrale sotto il profilo tecnologico e non dovrebbe dipendere dalle tecniche impiegate"*. Seguendo questa logica, qualsiasi trattamento di dati personali per mezzo di un algoritmo rientra nell'ambito di applicazione del GDPR. Ciò comporta la competenza sulla creazione e l'uso della maggior parte degli algoritmi. Grazie all'approccio basato sul rischio, al principio di minimizzazione dei dati e alla protezione dei dati fin dalla progettazione e per impostazione predefinita, il GDPR offre, quantomeno a livello teorico, un quadro giuridico entro cui attrarre molti dei potenziali rischi associati al trattamento dei dati personali tramite algoritmi, tra cui: l'ingiustizia, la discriminazione, il trattamento di massicce quantità di dati che crea una spirale in cui c'è un bisogno di dati su larga scala sempre crescente e di pari passo aumenta il rischio di operazioni di raccolta illecite; la sempre crescente complessità degli algoritmi che rende difficile il riuscire a rispondere ai requisiti di trasparenza con ragionevoli livelli di chiarezza e comprensibilità. Naturalmente "l'aiuto" che il GDPR può dare resta nella prospettiva della tutela dati. Anche la CEDPO[123], nel documento prodotto dal suo gruppo di lavoro dedicato all'AI, sostiene l'applicabilità del GDPR al contesto dell'IA generativa, con particolare riferimento, nel silenzio del IA Act, alla scelta della base giuridica per il trattamento e alla *Privacy by Design,* ritenuta uno strumento cruciale per garantire una solida *accountability* e conformità al Regolamento. Tuttavia, nel documento si esorta a prestare attenzione ad un aspetto importante, ovvero quello della somministrazione di dati a scopo di apprendimento. Normalmente, questi dati vengono assimilati dall'IA sotto forma di modelli e caratteristiche e vengono poi utilizzati per generare un *output*. Ciò implica che tramite il processo di apprendimento i dati entrano far parte del modello interno dell'IA e ne influenzeranno i risultati. Questo aspetto diventa ancor più preoccupante se lo si rapporta all' IA generativa, poiché nei contenuti che crea potrebbe "ri-estrarre" un numero indefinito di volte, dati personali utilizzati per l'apprendimento, in risposta alle ricerche degli utenti. Questa situazione, chiaramente, espone gli interessati a potenziali attività illegali, finanche a frodi. Da questo scenario si trae un assaggio della complessità (con

[123] Acronimo inglese per: *Confederation of European Data Protection Organisations*. Ovvero la Confederazione delle Organizzazioni Europee per la Protezione dei dati.

dei seri connotati di impossibilità) della missione del mantenimento del controllo sui dati personali una volta che siano entrati a far parte dei modelli elaborativi e memorizzativi di un'IA generativa[124].

5. Il principio di *accountability* è diviso tra due mondi? Piccola analisi comparata tra AI Act e GDPR

Dopo aver definito a livello giuridico cos'è l'IA, aver fatto una panoramica sul metodo classificatorio dei sistemi ed aver discusso la posizione ed il trattamento dell'IA generativa all'interno dell'AI Act, è opportuno ora riavvicinare lo sguardo al tema della responsabilità che costituisce l'antecedente logico di un comportamento che possa dirsi virtuoso. Affrontare il tema della responsabilità in questo elaborato non vuol dire solamente ripercorrere, per gli aspetti pertinenti, la giovane "carriera" del Regolamento 2016/679 dalla sua entrata in vigore fino al momento attuale, ma significa anche cercare di capire quanto e cosa il nuovo Regolamento sull'IA abbia mutuato dalla normativa sulla tutela dati. I due provvedimenti in realtà sono connessi, e la loro connessione è rappresentata dalla "materia prima" dell'era dei big data: i dati. Questo si spiega ragionando sul fatto che alla base del funzionamento dell'intelligenza artificiale ci siano i dati, personali e non. Questo fa dell'IA una tecnologia annoverabile tra quelle *data driven.*

Inoltre, il rapporto tra il Regolamento sull'IA e quello in materia di tutela dati personali rientra all'interno dell' esigenza di preservare l'assetto europeo sulla protezione dei dati in modo molto più ampio. Tale quadro legislativo si compone del Regolamento sulla protezione dei dati personali per le istituzioni, gli organi e gli organismi dell'Unione[125] e della Direttiva sulla protezione dei dati nelle attività di contrasto[126]. Come suggerito nel

[124] CEDPO AI Working Group-Generative AI: The Data Protection Implications, 16 Ottobre 2023. Reperibile su: https://cedpo.eu/wp-content/uploads/generative-ai-the-data-protection-implications-16-10-2023.pdf

[125] Regolamento (UE) 2018/1725 del Parlamento europeo e del Consiglio, del 23 ottobre 2018, sulla tutela delle persone fisiche in relazione al trattamento dei dati personali da parte delle istituzioni, degli organi e degli organismi dell'Unione e sulla libera circolazione di tali dati, e che abroga il regolamento (CE) n. 45/2001 e la decisione n. 1247/2002/CE

[126] Direttiva (UE) 2016/680 del Parlamento europeo e del Consiglio, del 27 aprile 2016, relativa alla protezione delle persone fisiche con riguardo al trattamento dei dati personali da parte delle autorità competenti a fini di prevenzione, indagine, accertamento e perseguimento di reati o esecuzione di sanzioni

parere congiunto di EDPB ed EDPS[127], quanto detto va considerato alla stregua di una cornice che mira a consentire l'inserimento di nuove normative (come da ultimo l'AI Act) senza però interferire con la sfera applicativa delle altre già presenti. Quanto detto non si limita ad essere solo un auspicio per quieto vivere, ma si pone a garanzia del rispetto del diritto fondamentale alla protezione dei dati personali, così come stabilito agli artt. 16 TFUE e 8 della Carta dei Diritti UE.

Tornando ora a ciò che più interessa ai fini della logica di questo elaborato, è certamente necessaria una disamina della disciplina della responsabilità, per individuare i punti di contatto, di continuità e di eventuale contrasto tra le due normative prese in esame. Come già accennato nel capitolo primo[128], il concetto della responsabilità, nel testo del GDPR assume la veste dell'*accountability.*

5.1 La dimensione patrimoniale della responsabilità nel GDPR

Riprendendo le fila del discorso lasciato in sospeso nel capitolo precedente, conviene adesso proseguire l'analisi del principio di *accountability,* di cui finora si è parlato solo in termini teorici di approccio al trattamento. La responsabilità comprende anche ulteriori dimensioni, tra le quali c'è la sicurezza tecnica del trattamento, le misure atte a contrastare i rischi di accesso e trasmissione non autorizzata dei dati trattati (ma anche evitare la loro perdita o accidentale distruzione), ed infine la dimensione patrimoniale legata alle violazioni normative e al diritto al risarcimento. Le differenti dimensioni in cui si articola il principio di *accountability,* non solo nel GDPR ma, a modo proprio, anche nell'AI Act, riflettono una realtà in cui garantire la sicurezza nell'utilizzo e nella produzione di dati, richiede uno sforzo sinergico diviso tra IT, legislazione e organizzazione pratica.

Per quel che concerne la dimensione patrimoniale della responsabilità nel GDPR si può fare riferimento al capo VIII, rubricato *"mezzi di ricorso, responsabilità e sanzioni"*.

penali, nonché alla libera circolazione di tali dati e che abroga la decisione quadro 2008/977/GAI del Consiglio

[127] EDPB-EDPS, Parere congiunto 5/2021 sulla proposta di regolamento del Parlamento europeo e del Consiglio che stabilisce regole armonizzate sull'intelligenza artificiale (legge sull'intelligenza artificiale) - 18 giugno 2021Punti 55 e 56

[128] Si veda il Capitolo I, paragrafo 3.2.1.

In particolare, all'articolo 82(1)[129] viene prevista la possibilità di chiedere un risarcimento per danni materiali od immateriali, in conseguenza della violazione di una delle disposizioni del Regolamento.

Da questo punto poi la norma opera, al secondo comma, una necessaria distinzione tra le implicazioni della posizione del titolare e quelle del responsabile del trattamento. Le due figure, infatti, portano con sé livelli di imputazione della responsabilità differenti. Perciò, alla figura del titolare viene agganciata una responsabilità qualificabile come oggettiva[130], infatti egli risponde del danno causato dal suo trattamento qualora eseguito in violazione del Regolamento. Per quanto riguarda, ove sia presente, il responsabile del trattamento, la sua responsabilità viene richiamata invece in due circostanze: la prima nel caso in cui manchi di adempiere agli obblighi a lui legislativamente assegnati; la seconda si realizza quando il responsabile agisca in modo difforme od in violazione delle istruzioni dettate dal titolare.

La dimensione patrimoniale del principio di *accountability*, costituisce di fatto solo uno dei suoi volti. Con ciò si intende la distinzione tra responsabilizzazione antecedente il danno e quella susseguente, più propriamente definita come responsabilità civile[131].

5.1.1 Ricognizione del quadro normativo del titolare: obblighi, adempimenti e cautele che dettagliano un agire responsabile

Come accennato poco sopra ed ancor prima nel capitolo primo[132], quello dell'*accountability* è un principio fondamentale e pervasivo nella disciplina della protezione dei dati, e lo è altrettanto, anche se con "vesti" diverse, nell'AI act.

La dimostrazione della conformità al GDPR e quindi ancora la dimostrazione di responsabilità e affidabilità, si riversa per la maggior parte sul titolare del trattamento (in parte necessariamente minore sul responsabile del trattamento ove presente), il quale è

[129] Articolo 82 (1) GDPR: *"1. Chiunque subisca un danno materiale o immateriale causato da una violazione del presente regolamento ha il diritto di ottenere il risarcimento del danno dal titolare del trattamento o dal responsabile del trattamento."* […]

[130] Cocuccio M., Dimensione "patrimoniale" del dato personale e tutele risarcitorie, in Diritto di Famiglia e delle Persone(II), fasc. 1, 1 marzo 2022-pg. 251.

[131] Barbierato D. , Trattamento dei dati personali e "nuova" e responsabilità civile, in Responsabilità civile e previdenza, n.6/2019. pg. 2153

[132] Si veda il Capitolo I, paragrafo 3.2.1.

circondato da una serie di obblighi, adempimenti e cautele che possono essere divise a seconda della loro natura di: 1) decisione volontaria, 2) di obbligo *ex ante* una violazione normativa o se richiesto dall'autorità di controllo; 3) di obbligo successivo ad una violazione normativa, 4) di obblighi ed adempimenti che devono essere attivati in conseguenza di specifiche circostanze.

Alla prima categoria menzionata, e che troviamo nel capo IV del GDPR, appartiene la decisione volontaria del titolare di aderire a codici di condotta e sistemi di certificazione. Tali adempimenti, previsti agli artt. 40-42 sono stati dettati nell'ottica di offrire al titolare del trattamento degli strumenti ulteriori per dimostrare la conformità al GDPR delle proprie attività.

Nella seconda categoria figurano degli obblighi che attengono ad aspetti fondamentali ed imprescindibili di qualsivoglia trattamento. Tra questi c'è il necessario rispetto dei principi fondamentali applicabili al trattamento; l'obbligo di acquisire il consenso da parte dell'interessato (tolte le casistiche di esonero); l'obbligo (cui all'art. 12) di gestire il trattamento in modo trasparente, il che comporta di fornire all'interessato tutte le informazioni riguardanti il suo trattamento con parole chiare e linguaggio accessibile; la predisposizione di un'adeguata informativa all'interessato; il rispetto dell'ampio catalogo dei diritti (agli artt. 15 e ss.) che spettano all'interessato; la predisposizione di misure di sicurezza adeguate (tra cui quelle proposte all'art. 32) a dimostrare conformità al Regolamento; la messa in atto di misure tecniche ed organizzative che tengono conto dei specifici rischi del trattamento atte a proteggere i diritti degli interessati fin dalla progettazione (art. 25(1)); la predisposizione, per impostazione predefinita al trattamento, di misure tecniche che mirano a circoscrivere la quantità dei dati raccolti a quella strettamente necessaria alla finalità concordata (art. 25(2)); il titolare infine non deve dimenticare di istruire adeguatamente sui loro compiti tutti coloro che sono autorizzati ai dati. Qui con particolare riferimento alla subordinazione del responsabile del trattamento come sancito all'art.29[133].

Alla terza categoria appartiene quindi l'obbligo per il titolare di cooperare con l'autorità di controllo quando gliene venga fatta richiesta; l'obbligo di notificare a quest'ultima ogni

[133] Articolo 29 GDPR: *"Il responsabile del trattamento, o chiunque agisca sotto la sua autorità o sotto quella del titolare del trattamento, che abbia accesso a dati personali non può trattare tali dati se non è istruito in tal senso dal titolare del trattamento, salvo che lo richieda il diritto dell'Unione o degli Stati membri."*

violazione dei dati intervenuta; la necessità di rendere edotto della violazione anche l'interessato; ed infine il sopracitato obbligo di risarcimento danni. Egli è peraltro esonerabile da tale obbligo solo nel caso in cui l'evento dannoso non sia a lui imputabile. In quarto luogo, tra gli adempimenti dovuti in conseguenza di specifiche circostanze troviamo ad esempio gli obblighi informativi verso gli interessati nel caso di contitolarità del trattamento o la redazione della valutazione d'impatto sulla protezione dati da fare quando il titolare si accinga ad intraprendere un trattamento rischioso a livello di diritti e libertà fondamentali[134].

5.2 La dimensione teorica del principio di *accountability*

Il Gruppo di lavoro "Articolo 29" istituito dalla direttiva 95/46/CE, nell'opinione 3/2010 sul principio di *accountability*, riconosce che quello di responsabilizzazione è un principio che offre due livelli: il primo vincolante ed il secondo volontario.

La dimensione vincolante è considerata tale perché consiste in tutti quegli obblighi ed adempimenti rivolti ai titolari del trattamento. Il contenuto del requisito comprenderebbe a sua volta due elementi: l'attuazione di misure e procedure, da un lato, e la dimostrabilità dell'adempimento dall'altro.

Il secondo livello, come accennato, comprende sistemi volontari di responsabilità che vanno al di là dei requisiti giuridici minimi per la protezione dei dati, fornendo quindi garanzie più elevate di quelle richieste dalle norme applicabili sia in termini di modalità di attuazione che di garanzia dell'efficacia delle misure[135].

5.3 Il principio di *accountability* tra GDPR e AI Act

Come già si è avuto modo di dire, i dati sono il comune denominatore che unisce i destini del GDPR e dell'AI Act. Nell'uno, i dati (personali) sono oggetto di protezione ed i relativi interessati sono "resi forti" da una serie di diritti. Nell'altro invece, viene destinata ai dati una tutela più tecnica ed indiretta che si comporta come riflesso del livello di

[134]Categorizzazione tratta da: Di Paolo Marini, GDPR: la tabella degli adempimenti del titolare. -Altalex.- Aprile 2018. Reperibile su: https://www.altalex.com/documents/news/2017/03/10/privacy-e-regolamento-ue-la-tabella-degli-obblighi-e-degli-adempimenti-del-titolare -Ultima visita al sito in data 10 febbraio 2025

[135] Considerazione tratta da: Gruppo Di Lavoro Articolo 29 Per La Protezione Dei Dati -00062/10/EN WP 173 -Opinion 3/2010 on the principle of accountability- Adopted on 13 July 2010.

rischio insito nel sistema di IA. L'atteggiamento che le due normative assumono nella gestione ed assegnazione della responsabilità è, seppur con diverse intensità, sovrapponibile e identificabile con due modelli: quello basato sui diritti poc'anzi citato, e quello basato sul rischio che assegna la responsabilità direttamente ai soggetti coinvolti [136].

Più nello specifico, nel GDPR, l'*accountability* può essere dimostrata ed implementata in larga parte, attraverso misure scelte discrezionalmente dal titolare in relazione al rischio stimato del trattamento. Nel Regolamento sull'intelligenza artificiale invece, l'approccio basato sul rischio costituisce una struttura alquanto rigida ed impermeabile a spazi di discrezionalità per i fornitori. A tal proposito, è probabile che la flessibilità offerta dalla struttura del GDPR sia la ragione per cui sembra adatto ad adeguarsi a fattori sociali ed evoluzione tecnologica[137]. Sebbene le due normative possano essere in certa misura confrontate riguardo l'estrinsecazione del principio di *accountability*, sono forse le differenze quelle che rendono più giustizia al loro scopo. Come già evidenziato, il GDPR è riuscito a coniugare la gestione dei trattamenti dei dati personali con la tutela dei singoli individui interessati. Questo è stato fatto grazie alla schiera di diritti esercitabili dagli interessati e dai principi fondamentali del trattamento da cui si irraggia il dovere per i titolari di rendersi *ex ante accountable* di un trattamento contornato da tutte le cautele necessarie; e se possibile, non mancano le esortazioni a far di più. Secondo l'analisi proposta da Capuzzo [138] è fuorviante cercare di estendere tale ragionamento alla disciplina dell'intelligenza artificiale. Se è vero che i sistemi di intelligenza artificiale si nutrono di dati sin dai primi momenti del loro sviluppo (pensiamo ai *set* di dati di addestramento a cui poi ne fanno seguito altri di convalida e verifica) è parimenti vero che solo una porzione di essi è di natura personale. Naturalmente, la maggiore o minore quantità di dati personali dipende anche dalla provenienza di tali dati. Che gli sviluppatori attingano da *dataset* frutto della propria attività di *web scraping* o attingano da *datalake* di terze parti (che a loro volta hanno utilizzato lo *scraping* per formarli), se i dati provengono a

[136] O.Pollicino, G. De Gregorio, Intelligenza artificiale, data protection e responsabilità. In Pajno, Alessandro, et al. *Intelligenza artificiale e diritto : una rivoluzione?* Il Mulino, 2022.
[137] *Ibidem*
[138] Capuzzo G., *A(i)Minority Report. Uno studio su intelligenza artificiale e comparazione giuridica tra UE, USA e Cina.* Rivista Critica del Diritto Privato, Anno XL- 4 Dicembre 2022 Trimestrale. pp. 493-494

monte da un'attività di *scraping* malevolo[139], la mole di dati personali indebitamente sottratti sarà maggiore. Perciò, tornando a ragionare sulla disciplina dell'AI Act, se in linea generale solo una piccola parte dei dati usati sono riconducibili ad un soggetto identificato o identificabile, la tutela giuridica approntata dall'AI Act risulta divergere dalla disciplina del GDPR proprio sulla natura dei dati in oggetto. In questo contesto, appare meno incoerente con gli intenti di tutela dei diritti fondamentali dichiarati nella proposta[140] il loro effettivo esiguo riscontro nell'articolato, dove infatti si rinvengono rimandi minimi alla tutela dei diritti fondamentali. Questo perché il precipuo scopo dell'AI Act è regolamentare in modo specifico l'ambito dell'intelligenza artificiale che prima d'ora era un settore in grande crescita ma privo di disciplina[141].

6. Il disegno di legge del Senato 1146

Il disegno di legge del Senato n. 1146[142], al momento in corso di esame in commissione, si colloca perfettamente nello spazio lasciato dall'AI Act alle norme attuative degli Stati Membri, in virtù di ciò al suo interno troviamo un sistema di principi, disposizioni di *governance*, e misure elaborate specificamente per settori quali: il sanitario, la pubblica amministrazione ed il mondo del lavoro.

L'intervento è stato giudicato necessario in assenza di una normativa nazionale organica e di strumenti di tutela a favore di cittadini ed imprese.

Lo schema del d.d.l. si divide in sei capi dedicati a sei aree tematiche di intervento diverse e ventisei articoli. Le sei aree tematiche sono affrontate nel corso dell'articolato nel

---

[139] Si intende lo scraping che mira a sottrarre dati tramite lo sfruttamento di vulnerabilità di sistemi non adeguatamente securizzati.

[140] Motivi ed obiettivi della proposta: "*La presente proposta mira a […] sviluppare un ecosistema di fiducia proponendo un quadro giuridico per un'IA affidabile. La proposta si basa sui valori e sui diritti fondamentali dell'UE e si prefigge di dare alle persone e agli altri utenti la fiducia per adottare le soluzioni basate sull'IA, incoraggiando al contempo le imprese a svilupparle. L'IA dovrebbe rappresentare uno strumento per le persone e un fattore positivo per la società, con il fine ultimo di migliorare il benessere degli esseri umani. Le regole per l'IA disponibili sul mercato dell'Unione o che comunque interessano le persone nell'Unione dovrebbero pertanto essere incentrate sulle persone, affinché queste ultime possano confidare nel fatto che la tecnologia sia usata in modo sicuro e conforme alla legge, anche in termini di rispetto dei diritti fondamentali."* […]
Proposta Di Regolamento Del Parlamento Europeo E Del Consiglio Che Stabilisce Regole Armonizzate Sull'intelligenza Artificiale (Legge Sull'intelligenza Artificiale) E Modifica Alcuni Atti Legislativi Dell'unione-COM/2021/206 Final- 2021/0106(COD)

[141] Capuzzo G., *A(i)Minority Report. Uno studio su intelligenza artificiale e comparazione giuridica tra UE, USA e Cina*. Rivista Critica del Diritto Privato, Anno XL- 4 Dicembre 2022 Trimestrale pp. 493-494

[142] Reperibile sul sito del Senato:
https://www.senato.it/leg/19/BGT/Schede/FascicoloSchedeDDL/ebook/58262.pdf

seguente ordine: la normativa di principio, le disposizioni di settore, *governance*- autorità nazionali- azioni di promozione, la tutela del diritto d'autore, le sanzioni penali ed infine le disposizioni finanziarie.

Le finalità della normativa vengono dichiarate al capo I e divise nei sei articoli che lo compongono seguendo la materia di applicazione (sviluppo economico, sicurezza e difesa nazionale, riservatezza dati personali). I sei articoli iniziali hanno lo scopo di riportare il *focus* sulla dimensione umana dello sviluppo dei modelli di IA nei settori di attività nazionali considerati.

L'ambito d'applicazione dei principi sottesi alla normativa seguono una duplice linea d'azione: la prima è quella antropocentrica imperniata sull'utilizzo trasparente e responsabile dell'IA; La seconda è quella legata alla vigilanza sui potenziali rischi economico-sociali derivanti dall'utilizzo dell'IA.

L'articolo 3 condensa, rubricandoli come *"principi generali"*, un elenco di capisaldi strutturali per tutte le disposizioni di dettaglio seguenti.

L'ispirazione per la formulazione dei principi in questione è stata presa dalle elaborazioni del *"Gruppo Indipendente di Esperti ad Alto Livello sull'Intelligenza Artificiale"* istituito dalla Commissione Europea, il quale ha sintetizzato sette requisiti chiave da rispettare per poter ritenere un'IA come affidabile (*trustworthy*) : 1) Intervento e sorveglianza umana; 2) Robustezza tecnica e sicurezza 3) Riservatezza e governance dei dati 4) Trasparenza 5) Diversità, non discriminazione ed equità 6) Benessere sociale ed ambientale 7) *Accountability*[143].

Al primo comma dell'articolo 3 è racchiusa tutta la dimensione umanistica che per questione di scopo e direzione, manca nell'AI Act: *"La ricerca, la sperimentazione, lo sviluppo, l'adozione, l'applicazione e l'utilizzo di sistemi e di modelli di intelligenza artificiale avvengono nel rispetto dei diritti fondamentali e delle libertà previste dalla Costituzione, del diritto dell'Unione europea e dei princìpi di trasparenza, proporzionalità, sicurezza, protezione dei dati personali, riservatezza, accuratezza, non discriminazione, parità dei sessi e sostenibilità"*.

All'articolo 4 viene ribadita la conformità al diritto dell'Unione in materia di riservatezza dei dati personali che si intende imprimere al trattamento operato nel contesto dei sistemi

[143]Commissione Europea, Communication From The Commission To The European Parliament, The Council, The European Economic And Social Committee And The Committee Of The Regions- Building Trust in Human-Centric Artificial Intelligence- Brussels, 8.4.2019 COM(2019) 168 final. Pg.3

di IA. Perciò si dispone che l'utilizzo dell'IA nei mezzi di comunicazione sia contornato da obiettività, completezza, imparzialità e lealtà dell'informazione. Al comma 3 si stabilisce che lo sviluppo dei sistemi si evolva in modo conoscibile e spiegabile.

All'articolo 22 è contenuta una delega al governo che funge da raccordo[144] tra l'AI Act e tutta quella materia normativa nazionale che, pur non facendo espresso riferimento agli algoritmi e alla IA, con essa tuttavia interagisce (un esempio congruo al tema dell'elaborato è l'inevitabile interazione con le disposizioni in materia di privacy e tutela dei dati personali). Tale delega assume ancor più rilevanza perché permette all'AI Act di raggiungere ed adeguarsi in tutti quei settori non direttamente disciplinati dal Regolamento stesso, il quale si occupa di tracciare percorsi diversi in fatto di gestione e *governance* della IA in relazione al livello di rischio derivante dal suo utilizzo.

Di particolare interesse risulta essere il contenuto del parere richiesto dal Senato al Garante sul disegno di legge in questione[145], essendo la protezione dei dati, il diritto fondamentale maggiormente coinvolto dall'utilizzo dei sistemi di IA, che, come si vedrà meglio più avanti, nella maggior parte dei casi si avvalgono di dati personali. Nella memoria presentata, il Prof Stanzione, attuale presidente del Garante, ha sottolineato la stretta interrelazione tra protezione dati ed intelligenza artificiale e di conseguenza l'importanza che anche il presente d.d.l. si inserisca in questa prospettiva. Tale interrelazione passa non solo attraverso l'affinità nell'analizzare l'incidenza del trattamento su diritti e libertà, tra valutazione di impatto ex art. 35(1) GDPR e lo strumento previsto all'art. 27 dell'AI Act (valutazione d'impatto sui diritti fondamentali); ma passa anche attraverso la specifica riserva di competenza per i sistemi di IA ad alto rischio stabilita dall'AI Act all'art. 74(8), ovvero: "*nella misura in cui tali sistemi sono utilizzati a fini di attività di contrasto, gestione delle frontiere, giustizia e democrazia*" .

Seguendo queste osservazioni, il Garante ha ritenuto opportuno suggerire la modifica

---

[144]Dall'intervento del Prof. Alberto Gambino durante gli "Gli Stati Generali del Diritto di Internet e della Intelligenza artificiale" -IV edizione. In *Diritto di Internet. Digital Copyright e Data Protection*. A cura di Giuseppe Cassano e Francesco Di Ciommo. 28 Novembre 2024
Video reperibile su: https://www.youtube.com/watch?v=aKylH1Zm2Hw&t=1770s "Gli Stati Generali del Diritto di Internet e della Intelligenza Artificiale", IV edizione Giorno 1. Minuto 30:00-42:00.

[145] Senato della Repubblica, Commissioni 8^ e 10^ riunite-AS 1146, Disposizioni e delega al Governo in materia di intelligenza artificiale. Audizione del Presidente del Garante per la protezione dei dati personali Prof. Pasquale Stanzione. Reperibile sul sito del Senato a:
https://www.senato.it/leg/19/BGT/Schede/FascicoloSchedeDDL/ebook/58262.pdf

dell'articolo 22 comma 2(a)[146], nel senso di designare il Garante quale Autorità nazionale di vigilanza del mercato ai sensi dell'art. 70(1) dell'AI Act[147]. Questo risponderebbe ad esigenze di semplificazione a livello amministrativo, stante che l'Autorità dovrebbe esercitare le proprie attribuzioni tutte le numerose volte in cui il processo algoritmico coinvolga dati personali, fermi restando i casi espressamente previsti dall'AI Act.

[146] Art.22 comma 2(a) d.d.l. 1146: *"2. Nell'esercizio della delega di cui al comma 1 il Governo si attiene, oltre che ai princìpi e criteri direttivi generali di cui al l'articolo 32 della legge 24 dicembre 2012, n. 234, ai seguenti principi e criteri direttivi specifici: a) designazione, in coerenza con quanto previsto dall'articolo 18 della pre sente legge, come autorità nazionali compe tenti ai fini dell'attuazione del regolamento di cui al comma 1, di un'autorità di vigi lanza del mercato, di un'autorità di notifica, nonché del punto di contatto con le istitu zioni dell'Unione europea;[...]*

[147] Art. 70(1) AI Act*: "1. Ciascuno Stato membro istituisce o designa come autorità nazionali competenti ai fini del presente regolamento almeno un'autorità di notifica e almeno un'autorità di vigilanza del mercato. Tali autorità nazionali competenti esercitano i loro poteri in modo indipendente, imparziale e senza pregiudizi, in modo da salvaguardare i principi di obiettività delle loro attività e dei loro compiti e garantire l'applicazione e l'attuazione del presente regolamento. I membri di tali autorità si astengono da qualsiasi atto incompatibile con le loro funzioni. A condizione che siano rispettati detti principi, tali compiti e attività possono essere svolti da una o più autorità designate, conformemente alle esigenze organizzative dello Stato membro."*

CAPITOLO TERZO

QUANDO E COME VALUTARE (ED EVITARE) I RISCHI POSTI DAL TRATTAMENTO DEI DATI OPERATO DALLA IA?

Dopo aver provveduto a fornire un contesto introduttivo del *web scraping*, come fenomeno da cui si dipana la domanda di ricerca di questo elaborato, il capitolo secondo ha proseguito con l'intento di tracciare un quadro giuridico, seppur necessariamente riassuntivo, che fosse di riferimento ed entro cui collocare l'asservimento del *web scraping* allo sviluppo della IA generativa. A questo punto, si rende opportuno, innanzi tutto, una delucidazione su che tipo di tecnologia stia alla base del funzionamento della IA di cui fin qui si è parlato in termini normativi e nozionistici. Verrà quindi proposta una sintesi tecnica del funzionamento dei modelli linguistici di grandi dimensioni, e degli attuali limiti tecnici che spesso rendono l'impiego della IA potenzialmente lesivo, non solo dei diritti e delle libertà di ogni persona ma anche della salvaguardia del diritto alla riservatezza così come impostato dal Regolamento (UE) 2016/679. In seguito, il secondo paragrafo verrà dedicato ad individuare ed analizzare criticità e soluzioni nell'ambito del processo decisionale automatizzato. Inoltre, a conclusione del paragrafo, verrà svolta una disamina delle difficoltà applicative, e del loro superamento, relativamente ai principi generali del GDPR, con riguardo alle tecnologie *data driven* come i sistemi algoritmici. Sarà prestata particolare attenzione al principio di *accountability*.
Infine, nel terzo paragrafo, sarà discussa la figura del *Data Protection Officer* in relazione al principio di *accountability*.

1. I modelli linguistici di grandi dimensioni e loro funzionamento

Per comprendere al meglio la dinamica su cui si impernia la discussione dell'elaborato, è bene aver chiara l'entità della tecnologia che soggiace al funzionamento della IA, ovvero i modelli linguistici di grandi dimensioni (la cui abbreviazione dall'inglese è LLM)[148].

I modelli linguistici di grandi dimensioni sono un sottoinsieme dei modelli di base (*foundation models*), i quali a loro volta, sono una classe di modelli di *machine learning* studiati e sviluppati per essere versatili ed impiegabili per l'esecuzione di *task* diversificate, senza dover essere riprogrammati specificamente per ogni compito. Come è agevole intuire, la vastità degli insiemi di dati richiesti per ottenere un algoritmo performante, richiede continue attività di raccolta[149].

Come fu nel caso degli albori delle primissime civiltà umane, in cui lo sviluppo delle prime forme di linguaggio svolse un ruolo essenziale nel favorire la comunicazione tra gli esseri umani e lo sviluppo di società sempre più avanzate, nel caso degli LLM la loro elaborazione è stata resa possibile da una forma di linguaggio apposita: gli algoritmi, in grado di abilitare l'interazione tra esseri umani e macchine[150]. La rilevanza del linguaggio risiede proprio nella sua suscettibilità di veicolare idee, concetti e istruzioni in maniera chiara ed efficiente, fungendo da collegamento tra le persone e le tecnologie. Negli ultimi anni, i modelli linguistici di grandi dimensioni hanno attirato una crescente attenzione nel campo del *Machine Learning* (ML) grazie alle loro straordinarie potenzialità che li rende adatti a sopperire alla crescente domanda di macchine per eseguire compiti linguistici complessi, tra i quali ci sono: traduzione, sintesi e recupero di informazioni, ma anche le interazioni conversazionali (pensiamo ai *bot* che nelle varie piattaforme fungono da primo strumento di assistenza per gli utenti). L'addestramento basato su vastissime raccolte di dati di derivazione umana, ha fatto sì che questi modelli si siano dimostrati in grado di comprendere e generare linguaggio naturale con notevole precisione e fluidità.

Un modello linguistico di grandi dimensioni è un algoritmo basato sul *Deep Learning* (DL) progettato per svolgere una vasta gamma di attività legate all'elaborazione del

---

148 "*Large language models"*

149 IBM- Murphy Mike-What are foundation models? - IBM Research (2022). Reperibile su: (https://research.ibm.com/blog/what-are-foundation-models) Ultima visita al sito in data 15 febbraio 2025

150 HUMZA N., *A Comprehensive Overview of Large Language Models,* in *arXiv preprint arXiv:2307.06435,* 2024.

linguaggio naturale (*Natural Language Processing*, abbreviato in NLP), come interpretare, tradurre, fare previsioni o creare testi e altri tipi di contenuti. Questi modelli si avvalgono dell'architettura di rete neurale detta *Transformer*, che tramite un *Encoder*, per processare l'*input*, e un *Decoder*, per generare l'*output*, è progettata per trasformare o modificare una sequenza di *input* in una sequenza di *output*.[151] L'elevata precisione nelle risposte, prodotte da questi modelli, deriva principalmente dall'addestramento dell'algoritmo su *dataset* di dimensioni considerevoli[152].

Gli LLM, sono noti anche come reti neurali (*Neural Networks, NN*), e si ispirano al funzionamento del cervello umano. Queste reti operano attraverso una struttura di nodi stratificati, paragonabili ai neuroni, che elaborano e trasmettono informazioni tramite connessioni con pesi variabili. Questo meccanismo consente ai modelli di apprendere e migliorare le proprie prestazioni, affrontando con successo compiti complessi nell'ambito dell'elaborazione del linguaggio naturale.

I più avanzati LLM vengono addestrati su enormi quantità di dati provenienti da diverse fonti e sono in grado di generare testi che spesso risultano, indistinguibili (o pressappoco) da quelli prodotti da un essere umano[153]. Inoltre, possono processare *input* acustici o visivi, come ad esempio immagini, i quali possono essere trasformati in testo con un elevato grado di accuratezza. Come testimoniato dall'altissimo livello di verosimiglianza toccato dai *deep fakes*, anche le risposte vocali generate dall'IA sono assimilabili alla voce umana in modo impressionante.

Tornando a considerazioni più generali: rispetto ai modelli linguistici pre-addestrati, gli LLM riescono ad offrire prestazioni caratterizzate da migliore generalizzazione e adattamento. Inoltre, a giustificare la loro rapida diffusione, sta il fatto che gli LLM sembrano avere delle specifiche abilità emergenti, come il ragionamento, la pianificazione, il processo decisionale, l'apprendimento nel contesto e la risposta in

[151] Amazon Web Services- Cosa sono i Trasformatori? - Spiegazione dei Trasformatori nell'intelligenza artificiale - AWS - Reperibile su: (https://aws.amazon.com/it/what-is/transformers-in-artificial-intelligence/#:~:text=I%20trasformatori%20rappresentano%20un%27architettura%20di%20rete%20neurale%20progettata,tracciamento%20delle%20relazioni%20tra%20i%20componenti%20della%20sequenza) Ultima visita al sito in data 15 febbraio 2025

[152] VASWANI A., *Attention Is All You Need*, in *arXiv preprint arXiv:1706.03762*, 2023

[153] *Ibidem*

contesti *zero-shot*[154]. Queste abilità costituiscono una conseguenza indiretta delle gigantesche dimensioni di questi modelli, e perciò sono presenti anche quando gli LLM non sono stati addestrati specificamente per possedere tali attributi[155][156]. Queste capacità emerse di riflesso hanno portato gli LLM ad essere ampiamente adottati in diversi contesti, tra cui quello multimodale, la robotica, la manipolazione di strumenti, la risposta a domande, etc.[157]

Alcuni esempi di modelli linguistici di grandi dimensioni includono la serie GPT, utilizzata in strumenti come *ChatGPT* di OpenAI e *Copilo*t di Microsoft; *Gemini*, sviluppato da Google (precedentemente noto come *Bard*); i modelli LLaMA di Meta; la serie *Grok* di X; e i modelli *Claude* di Anthropic.

Il funzionamento degli LLM si innesta da un lato sull'acquisizione della capacità di categorizzare e quindi di generare contenuti in linguaggio naturale spendibili per i più disparati utilizzi; dall'altro, su un processo di apprendimento da parte dell'algoritmo delle relazioni tra le parole. Detto processo di apprendimento ha natura puramente di analisi statistica.

Il vicolo cieco in cui giunge tutto l'entusiasmo prodotto dai risultati degli LLM deriva dal fatto che essi, dopotutto, operano sulla base di mere probabilità statistiche per generare del linguaggio grammaticalmente e semanticamente corretto, attenendosi meramente al contesto fornito.

Dunque, non sono in grado di considerare la veridicità del contenuto prodotto.

---

[154] Ovvero la situazione in cui il modello di IA è addestrato per riconoscere e categorizzare elementi o concetti senza averne prima conosciuto alcun esempio. Fonte: IBM- Bergmann Dave-Che cos'è lo zero-shot learning? | IBM.(2024)
Reperibile su: https://www.ibm.com/it-it/think/topics/zero-shot-learning#:~:text=Lo%20zero-shot%20learning%20%28ZSL%29%20%C3%A8%20uno%20scenario%20di,prima%20alcun%20esempio%20di%20tali%20categorie%20o%20concetti Ultima visita al sito in data 15 febbraio 2025

[155] J. Wei, Y. Tay, R. Bommasani, C. Raffel, B. Zoph, S. Borgeaud, D. Yo gatama, M. Bosma, D. Zhou, D. Metzler, et al., Emergent abilities of large language models, arXiv preprint arXiv:2206.07682 (2022)

[156] T. Webb, K. J. Holyoak, H. Lu, Emergent analogical reasoning in large language models, Nature Human Behaviour 7 (9) (2023) 1526–1541

[157] Amazon Web Services- Cosa sono i Trasformatori? - Spiegazione dei Trasformatori nell'intelligenza artificiale – AWS- Reperibile su: (https://aws.amazon.com/it/what-is/transformers-in-artificial-intelligence/#:~:text=I%20trasformatori%20rappresentano%20un%27architettura%20di%20rete%20neurale%20progettata,tracciamento%20delle%20relazioni%20tra%20i%20componenti%20della%20sequenza)
Ultima visita al sito in data 15 febbraio 2025

Allo stato attuale della tecnologia, e non di rado, gli LLM non sono immuni dal produrre inesattezze e finanche delle vere e proprie "allucinazioni".

Le "allucinazioni dell'IA" sono i casi in cui le reti neurali, ed i modelli generativi, producono risultati che non rispecchiano in nessun modo la realtà, e che non rientrano nei parametri prefissati, costituendo quindi materiale falso o erroneo[158]. Un esempio molto comune di ciò e di facilissima reperibilità nel *web*, sono alcune immagini e video generati tramite IA, i quali propongono persone fantasiosamente conformate e crasi impossibili tra oggetti diversi. Ci sono alcuni casi celebri però che lasciano trasparire la potenzialità lesiva della disinformazione che può propagarsi dalle allucinazioni dell'IA. Un caso noto al riguardo, è quello che coinvolge il *chatbot* di Google, *Bard*, al quale era stato chiesto di riferire sulle scoperte del telescopio spaziale *James Webb*. In quest'occasione *Bard* affermò che il telescopio aveva catturato le prime immagini di un pianeta situato al di fuori del sistema solare. L'affermazione infatti è completamente falsa in quanto i primi esopianeti sono stati fotografati dalla NASA nel 2004[159], mentre il telescopio *Webb* è stato lanciato nel 2021. È stato rilevato poi da Bruce Macintosh, che l'errore del *chatbot* è derivato da una errata categorizzazione e connessione logico-grammaticale delle informazioni reperite sul *web*[160].

Se si lascia momentaneamente da parte l'aspetto ludico e ricreativo della questione e si considerano gli LLM nella loro dimensione di strumenti utili e di supporto alla vita lavorativa e quotidiana di un individuo, emerge una effettiva fragilità nei risultati prodotti. Da quanto detto fin qui, tra i limiti in cui può incorrere l'operato di un LLM c'è dunque non solo la generazione di *output* fallaci dovuti errori logico-semantici, ma anche quella di risultati incongrui, dovuti alla scarsa qualità e alla composizione non equilibrata dei dati di addestramento; infatti, come anche accennato nei capitoli precedenti, ciò può accadere a causa della possibile e statisticamente verosimile presenza di dati personali,

---

[158] Intelligenza Artificiale Italia Blog- Cosa sono le allucinazioni dell'IA? AI hallucinations | Intelligenza Artificiale Italia Blog. Reperibile su: https://www.intelligenzaartificialeitalia.net/post/cosa-sono-le-allucinazioni-dell-ia-ai-hallucinations - Ultima visita al sito in data 15 febbraio 2025

[159] A giant planet candidate near a young brown dwarf - Direct VLT/NACO observations using IR wavefront sensing.- Reperibile su: https://www.aanda.org/articles/aa/pdf/2004/38/aagg222.pdf -Ultima visita al sito in data 17 febbraio 2025

[160] Bruce Macintosh su X: "@olgias @Google The nuance is that what was "revealed" in the above stories was the first photograph taken by JWST, not the first photograph taken by any telescope everywhere. It's partially a grammar-and-logic-of-words thing, but it seems like that's something you would like AI to get right." / X. Reperibile su: https://x.com/bmac_astro/status/1623456005320491008?s=20&t=TC0deTQ7hh2Y2j0jKyIt1A Ultima visita al sito in data 17 febbraio 2025

contenuti protetti da *copyright*, testi inappropriati, falsi o discriminatori tra il materiale di addestramento.

Nondimeno, i risultati prodotti possono essere semplicemente obsoleti alla stregua del fatto che i modelli elaborano unicamente i dati disponibili al momento dell'addestramento, risultando così incapaci di riflettere gli eventuali aggiornamenti successivi.

Un ulteriore fattore, che denota la ancora instabile affidabilità degli *output* prodotti dall'IA, ha che fare con la sostanziale imprevedibilità del contenuto di *output*: difatti nonostante un algoritmo venga alimentato con lo stesso *input* ripetutamente, esso può generare *output* diversi sia nel contenuto che nella forma.

1.1 LLMs: criticità e sviluppi futuri

L'evoluzione dei modelli, a partire da GPT-4, hanno comportato, come accennato nel paragrafo precedente, dei progressi straordinari in fatto di elaborazione del linguaggio naturale. Tuttavia la loro linea di sviluppo futura presenta delle sfide specifiche che sono state evidenziate in modo soddisfacente nel recentissimo studio panoramico e ricapitolativo di Humza N. et Al, da cui per l'appunto, è tratta la schematizzazione che segue[161]. Tra le questioni più calde ci sono sicuramente gli alti costi computazionali, la resilienza agli attacchi avversari, l'interpretabilità degli algoritmi, l'elaborazione simultanea dei dati ed il rispetto della normativa a tutela della riservatezza. Questi punti critici non solo sottolineano le complessità tecniche coinvolte, ma lasciano trapelare anche l'impatto più ampio e la traiettoria futura degli LLM nelle applicazioni nel mondo reale. Di seguito viene proposta una schematizzazione che ricapitola il livello attuale delle prestazioni degli LLM ed i potenziali sforzi per affrontarne le criticità.

Per quel che riguarda il costo computazionale, l'addestramento degli LLM richiede il dispendio di ampie risorse, che va a gravare sui costi di produzione e solleva, tra l'altro, preoccupazioni ambientali dovute al notevole consumo di energia che risulta dall'addestramento di modelli su larga scala. Il miglioramento delle prestazioni è direttamente proporzionale all'aumento delle risorse computazionali, ma, allo stesso

[161] HUMZA N., *A Comprehensive Overview of Large Language Models,* in *arXiv preprint arXiv:2307.06435,* 2024.

tempo, il tasso di miglioramento diminuisce gradualmente quando fattori come la dimensione del modello e quella del *set* di dati rimangono fisse, seguendo quindi la legge dei rendimenti decrescenti[162].

Bias ed equità: gli LLM possono ereditare e amplificare i pregiudizi sociali contenuti nei loro dati di formazione. Questi pregiudizi possono manifestarsi nei risultati del modello, portando a potenziali questioni etiche e di equità.

Overfitting: Sebbene gli LLM possiedano capacità di apprendimento sostanziali, sono suscettibili di restare intrappolati dai parametri dati dai dati di addestramento, portando ad una situazione in cui il modello non è più in grado di fare previsioni o trarre conclusioni accurate basate su dati diversi da quelli di addestramento. Questo può accadere in conseguenza di modelli molto complessi o quando viene effettuato un addestramento prolungato su dati campione. I casi di *overfitting*, quindi rendono completamente vano lo scopo di un modello di *machine learning*[163]. Una possibile soluzione a questo percorso sdrucciolevole viene proposta da Tänzer M[164] et al. nel perseguire la ricerca di un sempre migliore equilibrio tra memorizzazione e generalizzazione nello sviluppo del modello. Difatti, la memorizzazione consente al modello di ricordare dettagli specifici dai suoi dati di addestramento, assicurando che possa fornire risposte accurate a domande precise. Tuttavia, la generalizzazione consente al modello di fare inferenze e produrre risposte per *input* che non ha mai visto prima, il che è essenziale per far sì che il modello sia spendibile nella gestione di varie attività del quotidiano. Un'eccessiva impostazione tesa alla memorizzazione, infatti, può portare sì ad un'alta velocità di elaborazione, ma questo rende il modello poco flessibile e in difficoltà con nuovi *input*.

Disuguaglianze economiche e di ricerca: l'elevato costo della formazione e dell'implementazione degli LLM può far sì che il loro sviluppo diventi prerogativa solo di organizzazioni che abbiano a loro disposizione solide risorse finanziarie, peggiorando potenzialmente le disuguaglianze economiche e di ricerca nell'IA[165].

[162] E. Strubell, A. Ganesh, A. McCallum, Energy and policy considerations for deep learning in nlp, arXiv preprint arXiv:1906.02243 (2019).
[163] C. Zhang, S. Bengio, M. Hardt, B. Recht, O. Vinyals, Understanding deep learning (still) requires rethinking generalization, Communications of the ACM 64 (3) (2021) 107–115.
[164] M.Tänzer, S. Ruder, M. Rei, Memorisation versus generalisation in pre trained language models, arXiv preprint arXiv:2105.00828 (2021)
[165] S. M. West, M. Whittaker, K. Crawford, Discriminating systems, AI Now(2019) 1–33

Ragionamento e pianificazione: alcuni compiti di ragionamento e pianificazione, anche apparentemente semplici come la pianificazione di buon senso, che gli esseri umani trovano facili, rimangono ben oltre le attuali capacità degli LLM. Questo non è del tutto inaspettato, considerando che gli LLM generano principalmente completamenti di testo basati sulla probabilità e non offrono solide garanzie in termini di capacità di ragionamento[166].

Allucinazioni: oltre quanto detto precedentemente, è opportuno specificare che le allucinazioni possono essere classificate in tre categorie:

1) allucinazione in conflitto di *input,* in cui gli LLM producono contenuti che divergono dall'input fornito dagli utenti;

2) allucinazione in conflitto di contesto, in cui gli LLM generano contenuti che contraddicono le informazioni che hanno generato in precedenza;

3) allucinazione in conflitto con i fatti, la quale coinvolge la generazione di contenuti da parte dell'LLM che non si allineano con la conoscenza mondiale consolidata.

Prompt Engineering: i *prompt* fungono da *input* per gli LLM e la loro sintassi e semantica svolgono un ruolo cruciale nel determinare l'output del modello. Le variazioni rapide, a volte controintuitive per gli esseri umani, possono comportare cambiamenti significativi nell'*output* del modello e vengono affrontate attraverso l'ingegneria tempestiva, che prevede la progettazione di *query* in linguaggio naturale per guidare efficacemente le risposte degli LLM.

Conoscenza limitata: le informazioni acquisite durante la fase di addestramento sono limitate e possono diventare obsolete dopo breve tempo. Tuttavia, la ripetizione del *training* del modello, utilizzando dati aggiornati, è un'attività costosa.

Sicurezza e controllabilità: L'utilizzo di LLM comporta il rischio di generare contenuti dannosi, fuorvianti o inappropriati, sia per incidente che per quando vengono fornite richieste specifiche. Garantire che questi modelli siano utilizzati in modo sicuro richiede un impegno significativo.

Sicurezza e privacy: Come già detto, gli LLM sono inclini a far trapelare informazioni personali e a generare risposte false, non etiche e sregolate. Pertanto, la continua implementazione di misure di sicurezza è essenziale per garantire che gli LLM siano

[166] K. Valmeekam, A. Olmo, S. Sreedharan, S. Kambhampati, Large lan guage models still can't plan (a benchmark for llms on planning and reasoning about change), arXiv preprint arXiv:2206.10498 (2022).

sicuri e affidabili per le applicazioni in ambito di intelligenza artificiale particolarmente complesse.

Multimodalità: l'apprendimento multimodale, in cui gli LLM vengono addestrati su dati diversi come testo, immagini e video, mira a creare modelli con una comprensione più ricca. Tuttavia, ci sono degli aspetti che sono soliti presentare delle criticità, come l'allineamento dei dati e le esigenze computazionali più elevate.

Dimenticanza "catastrofica": gli LLM sono sempre più spesso (e ancor più spesso saranno) addestrati su set di dati di grandi dimensioni e poi ottimizzati su dati di domini specifici, riducendo così la necessità di risorse di addestramento. Tuttavia, problemi come "l'adattamento al dominio" e "l'oblio catastrofico", di fatto ostacolano la conservazione delle conoscenze pregresse quando arriva il momento di apprendere nuovi compiti.

Robustezza agli attacchi avversari: i modelli linguistici di grandi dimensioni hanno dimostrato grandi capacità in vari compiti, ma restano vulnerabili agli attacchi avversari, che con anche lievi alterazioni dell'*input* possono fuorviarli. Questo accade specialmente con modelli come BERT[167], l'ottimizzazione contro gli attacchi avversari può sicuramente migliorare l'aspetto della robustezza, ma può anche compromettere le capacità di generalizzazione. Man mano che gli LLM si integrano sempre di più in sistemi complessi, l'esame delle loro proprietà di sicurezza diventa cruciale, visto il pericoloso il sempre più frequente verificarsi di attacchi avversari ad LLM integrati all'interno di ML considerati affidabili. Questa vulnerabilità è particolarmente preoccupante nei domini critici per la sicurezza, che richiedono robusti strumenti di valutazione avversaria per garantire la resilienza degli LLM.

Interpretabilità e spiegabilità: la natura di "*Black Box*" degli LLM pone difficoltà nella comprensione dei loro processi decisionali, il che, soprattutto nei settori sociali più sensibili, costituisce un punto critico per stabilire accettazione e fiducia nei loro confronti. Nonostante le loro capacità avanzate, la mancanza di informazioni sul loro funzionamento

[167]BERT, acronimo di "Bidirectional Encoder Representations from Transformers", è un algoritmo di elaborazione del linguaggio naturale rilasciato da Google nel 2017. Il modello si è distinto per la sua capacità di comprendere le sfumature linguistiche, il contesto delle parole all'interno delle frasi, e per il rapido incremento nei suoi livelli di accuratezza. L'algoritmo è stato implementato con successo per migliorare la comprensione delle query da parte di Google e di conseguenza migliorare la restituzione dei risultati alle ricerche.
Fonte: Network360- Cos'è Bert, l'algoritmo che cambia il mondo del Natural Language Processing- Giugno 2020. Reperibile su: https://www.ai4business.it/intelligenza-artificiale/cose-bert-lalgoritmo-che-cambia-il-mondo-del-natural-language-processing/ Ultima visita al sito in data 17 febbraio 2025

ne limita l'efficacia e l'affidabilità. Comprendere la logica alla base delle risposte degli LLM è un tassello molto importante per garantire che operino in linea con i valori umani e gli *standard* legali.

Preoccupazioni sulla privacy: le questioni concernenti privacy e riservatezza nei modelli linguistici di grandi dimensioni sono aumentate di pari passo alla loro crescita in complessità e dimensioni, in particolare per quanto riguarda la condivisione dei dati e il potenziale uso improprio. Se i modelli vengono addestrati su dati contenenti anche dati personali, sorgono ulteriori preoccupazioni qualora tali modelli vengano resi disponibili pubblicamente, questo a causa del rischio di "rigurgitare" dati personali memorizzati durante l'addestramento.

Elaborazione in tempo reale: L'elaborazione in tempo reale nei modelli di lingua di grandi dimensioni è una caratteristica importante in varie applicazioni, soprattutto con la crescente popolarità delle *app* di intelligenza artificiale per dispositivi mobili, da un lato, e le preoccupazioni relative alla sicurezza delle informazioni e alla privacy, dall'altro. Tuttavia, gli LLM hanno spesso centinaia di livelli e milioni di parametri, che ne impediscono l'elaborazione in tempo reale a causa delle elevate esigenze computazionali e del peso limitato disponibile all'archiviazione su piattaforme *hardware*, in particolare negli ambienti di *edge computing*[168]. Sebbene alcuni sforzi come *MobileBERT*[169] mirino a ridurre i requisiti di memoria, devono ancora affrontare un notevole sovraccarico di esecuzione a causa dell'elevato numero di livelli del modello, che porta a un'elevata latenza di inferenza.

Dipendenze a lungo termine: i modelli linguistici di grandi dimensioni hanno mostrato notevoli progressi nella comprensione e nella generazione di testo, ma spesso hanno difficoltà a preservare il contesto e a gestire le dipendenze a lungo termine, in particolare

[168] *L'edge computing* è un modello di calcolo distribuito in cui l'elaborazione dei dati avviene in prossimità del luogo in cui sono generati. Questa prossimità migliora i tempi di risposta, come insight più rapidi, tempi di risposta migliori e maggiore disponibilità della larghezza di banda.
Fonte: IBM-Che cos'è l'edge computing? Reperibile su: https://www.ibm.com/it-it/topics/edge-computing Ultima visita al sito in data 17 febbraio 2025.

[169] Il mobileBERT è una versione compressa ed accelerata del modello base BERT. Tale modello è stato modificato per adattarsi anche a dispositivi mobili che avendo risorse computazionali più limitate, soffrirebbero di latenza nel funzionamento.
Fonte: Zhiqing S. et Al- MobileBERT: a Compact Task-Agnostic BERT for Resource-Limited Devices-arXiv:2004.02984. Aprile 2020
Reperibile su. https://arxiv.org/abs/2004.02984 Ultima visita al sito in data 17 febbraio 2025.

in conversazioni complesse a più turni o in documenti lunghi. Questa limitazione è il principale fattore responsabile di risposte incoerenti o irrilevanti.

Accelerazione *hardware*: la crescita degli LLM presenta, in questo caso, notevoli sfide, a causa delle crescenti esigenze computazionali e di memoria associate all'addestramento e all'implementazione di tali modelli. Le GPU[170] hanno svolto un ruolo cruciale nel soddisfare i requisiti *hardware* per l'addestramento degli LLM, nel solco dell'evoluzione nel settore del *networking*, volta ad ottimizzare gli elaboratori destinati ai carichi di lavoro di addestramento. Tuttavia, le dimensioni crescenti degli LLM, che hanno purtroppo superato la velocità del progresso dell'*hardware*, rendono il processo di addestramento del modello sempre più costoso. A tal proposito, la quantizzazione del modello potrebbe essere un approccio promettente[171] per colmare il divario crescente tra le dimensioni dell'LLM e la capacità dell'*hardware*. Sebbene l'accelerazione specializzata come GPU o TPU[172] possa ridurre significativamente il costo computazionale, rendendo più fattibili le applicazioni in tempo reale, rischierebbe comunque non risolvere completamente tutte le limitazioni, richiedendo ulteriori progressi nella tecnologia *hardware*.

Quadri normativi ed etici: i rapidi progressi nell'intelligenza artificiale hanno dato origine a sofisticati modelli linguistici di grandi dimensioni come GPT-4 di OpenAI e *Bard* di Google. Questi sviluppi sono un sonoro promemoria che sottolinea l'imperativo di coltivare la supervisione normativa, al fine di gestire le sfide etiche e sociali che accompagnano l'uso diffuso degli LLM. Proprio per fatto che gli LLM possono generare contenuti che sono suscettibili di essere utilizzati sia positivamente che negativamente, cornici etiche che esortano alla proattività, a ricercare misure politiche per guidarne l'uso responsabile, e assegnare la responsabilità per i loro risultati, sono punti cruciali che verranno discussi più avanti, insieme all'importanza delle valutazioni preventive, che

[170] Sono le unità di elaborazione grafica. Sono costituite da un circuito elettronico in grado di eseguire calcoli matematici ad alta velocità.

[171] C. Guo, J. Tang, W. Hu, J. Leng, C. Zhang, F. Yang, Y. Liu, M. Guo, Y. Zhu, *Olive: Accelerating large language models via hardware friendly outlier-victim pair quantization, in: Proceedings of the 50th Annual International Symposium on Computer Architecture*, 2023, pp. 1–15

[172] Il TPU (*Tensor Processing Unit*) è un processore specializzato sviluppato da Google specificamente per accelerare i compiti di machine learning. Sono progettate per eseguire calcoli di bassa precisione ma su larga scala e sono ottimizzate per sopportare il carico di lavoro dei modelli di grandi dimensioni (LLM).
Fonte: Lavecchia Vito- Informatica e ingegneria online
Reperibile su: https://vitolavecchia.altervista.org/caratteristiche-e-differenza-tra-tpu-e-gpu-in-informatica/
Ultima visita al sito in data 17 febbraio 2025.

possono rappresentare un meccanismo di governance promettente per garantire che gli LLM, siano progettati e implementati in modo etico, legale e tecnicamente robusto.

2. IA e decisioni automatizzate: una sfida alla trasparenza

La rapida inflazione nello sviluppo dell'intelligenza artificiale, per quanto se ne voglia parlare, è stata innescata dai vantaggi pratici che gli algoritmi, tramite le loro decisioni automatizzate, hanno saputo offrire. I processi automatizzati infatti hanno ammaliato le nostre società con prestazioni economiche, veloci e foriere di una semplicità tale da aiutare nell'alleggerire gli apparati burocratici. Inoltre, vengono tuttora erroneamente considerati come capaci di restituire degli *output* prevedibili e privi di sfumature parziali. Dire che un algoritmo operi e prenda decisioni imparziali, è un'affermazione pericolosa perché può portare al fraintendimento del suo vero senso. Di per sé un algoritmo, in quanto cervello-motore operativo di una macchina, non può soffrire di preconcetti e non può essere portatore di credenze o valori, questo perché è privo di quell' "umanità" che gli consentirebbe di provare e veicolare delle emozioni. Detto ciò, sebbene, quindi, un algoritmo non venga progettato per perpetrare discriminazioni (o perlomeno non dovrebbe esserlo), in pratica può esserlo. Questo perché gli algoritmi sono influenzati dai dati che gli vengono somministrati in analisi. Dunque, per loro natura e variegata provenienza, i dati di addestramento riflettono le realtà da cui sono estrapolati, portandosi dietro anche pregiudizi ed errori. In ultimo, per completare questa impostazione, c'è da considerare l'ineliminabile tasso di fallibilità che caratterizza pressoché ogni cosa, e a maggior ragione le macchine e le tecnologie *data driven.*

L'IA, proprio per il fatto di essere una tecnologia *data driven,* si posiziona nell'occhio del ciclone per le conseguenze che i suoi limiti attuali possano comportare per la privacy delle persone. A tal proposito, si vedono traballare, nella loro efficacia, misure di sicurezza esplicitamente previste dal GDPR come l'anonimizzazione e la pseudonimizzazione, che sono state concepite come baluardi per impedire che i dati corressero il rischio di tornare ad essere personali. Questo è dovuto alla capacità degli algoritmi di re- identificare dati precedentemente resi anonimi e di dedurre dati personali

anche se addestrati con informazioni che non lo sono, tramite correlazioni statistiche operate su larga scala[173].

Questa evenienza, quand'anche remota, fornisce un'occasione per i terzi di avere indebito accesso alla sfera privata di persone inconsapevoli. E di inconsapevolezza degli utenti si parla non solo in casi di attacco doloso a sistemi, volto a sfruttarne le vulnerabilità, ma anche nei casi decisamente più rispondenti al quotidiano, come la persona che si trova ad acconsentire alla trasmissione dei propri dati per poter fare uso di un servizio specifico e che, quindi, non può immaginare eventuali trattamenti successivi divergenti da quanto acconsentito, anche solo per analisi diretta a profilazione su dati anonimizzati o pseudonimizzati, o ancor peggio per addestramento di IA.

La situazione attuale, dunque, restituisce un contesto in cui gli interessati hanno perso gran parte del controllo sui propri dati e non ne sono nemmeno pienamente consapevoli e l'IA, da strumento ausiliario alle attività umane, è diventato un ricettacolo meccanico di dati, pronto a minare in una miriade di modi i principi del GDPR.

2.1 Le decisioni automatizzate dell'articolo 22 GDPR

In questo contesto così delicato, M. Peluso[174] ha proposto un possibile quadro solutorio, basato sulle previsioni del GDPR, delle criticità sollevate da tecnologie che producono decisioni automatizzate.

Oggigiorno, e ancor di più che al momento di entrata in vigore del GDPR (nonostante sia un testo normativo che si può ritenere recente), le persone sono oggetto di processi decisionali automatizzati e profilazione di cui stentano a carpirne la portata. E spesso l'unico vero contatto con il processo decisorio automatizzato è il risultato da esso prodotto.

Su tal punto, la normativa di settore vede una specifica previsione nell'art. 22 GDPR rubricato "*Processo decisionale automatizzato relativo alle persone fisiche, compresa la*

[173] Peluso, Maria Grazia. *Intelligenza artificiale e tutela dei dati : prospettive critiche e possibili benefici per una governance efficace*. Giuffrè Francis Lefebvre, 2024.
[174] *Ibidem* (pp. 164 e ss.)

profilazione"[175]. In linea generale, l'articolo appronta una tutela generica per i casi in cui un individuo sia sottoposto a decisioni basate su di un trattamento automatizzato, ivi compresa quale sottoinsieme, la profilazione. La profilazione, dunque, è un tipo di trattamento dei dati automatizzato che può portare o meno ad una decisione capace di incidere giuridicamente sull'interessato.

Se della profilazione troviamo una definizione all'articolo 4 GDPR[176], invece delle decisioni automatizzate viene fatto solo un riferimento normativo all'articolo 22 GDPR, che può essere meglio compreso affiancandovi il chiarimento a riguardo, che si trova nelle apposite linee guida del Gruppo di Lavoro art 29[177]. Dalle parole di questo articolo si può tuttavia inferire quali siano le caratteristiche principali che denotano la peculiarità e la rischiosità connaturata a questo tipo di trattamento: difatti, "*una decisione basata unicamente sul trattamento automatizzato*"[178] è una decisione che viene presa da un sistema algoritmico che inferisce il proprio *output* sulla base di un set di dati senza la partecipazione del fattore umano. E nel momento in cui la decisione automatizzata è frutto di inferenza, tratta da un profilo della persona precedentemente creato, vi è anche profilazione; ecco che entra in gioco il senso dell'espressione "compresa la profilazione" all'art. 22.[179][180]

---

[175] Art. 22(1) GDPR: "*1. L'interessato ha il diritto di non essere sottoposto a una decisione basata unicamente sul trattamento automatizzato, compresa la profilazione, che produca effetti giuridici che lo riguardano o che incida in modo analogo significativamente sulla sua persona" […]*

[176] Art. 4 (4) GDPR: "*Profilazione: qualsiasi forma di trattamento automatizzato di dati personali consistente nell'utilizzo di tali dati personali per valutare determinati aspetti personali relativi a una persona fisica, in particolare per analizzare o prevedere aspetti riguardanti il rendimento professionale, la situazione economica, la salute, le preferenze personali, gli interessi, l'affidabilità, il comportamento, l'ubicazione o gli spostamenti di detta persona fisica*".

[177] Gruppo di Lavoro art. 29, Guidelines on automated individual decision-making and profiling for the purposes of regulation2016/679 (WP251), 2017 p. 8 : "Il processo decisionale automatizzato ha un ambito di applicazione diverso e può sovrapporsi parzialmente alla profilazione o derivare da essa. Il processo decisionale esclusivamente automatizzato è la capacità di prendere decisioni con mezzi tecnologici senza il coinvolgimento umano. Le decisioni automatizzate possono essere basate su qualsiasi tipo di dato"

[178] Art. 22 (1) GDPR.

[179] *Ibidem: "L'interessato ha il diritto di non essere sottoposto a una decisione basata unicamente sul trattamento automatizzato, compresa la profilazione, che produca effetti giuridici che lo riguardano o che incida in modo analogo significativamente sulla sua persona."*

[180] Gruppo di Lavoro art. 29, Guidelines on automated individual decision-making and profiling for the purposes of regulation2016/679 (WP251), 2017 p. 8: "*Le decisioni automatizzate possono essere prese con o senza profilazione; La profilazione può avvenire senza prendere decisioni automatizzate. Tuttavia, la profilazione e il processo decisionale automatizzato non sono necessariamente attività separate. Quello che inizia come un semplice processo decisionale automatizzato potrebbe diventare basato sulla profilazione, a seconda di come vengono utilizzati i dati."*

Riguardo quanto detto fin qui, ci sono vari esempi internazionali di decisioni algoritmiche erronee o discriminatorie, che hanno causato pesanti ripercussioni sugli interessati, una delle più note è il caso statunitense State v. Loomis[181], risalente al 2013, anno in cui il cittadino afroamericano Eric Loomis venne condannato a sei anni di reclusione, e cinque mesi di libertà vigilata, per oltraggio a pubblico ufficiale e appropriazione indebita di un veicolo usato per commettere una sparatoria. Per giungere alla sentenza, i giudici, hanno usato il *software* giudiziario COMPAS *(Correctional offender management profiling for alternative sanctions)*, il quale segnalava, per l'appunto, che l'imputato aveva un alto punteggio nella valutazione del rischio di recidiva. In questo contesto la società privata, proprietaria del *software*, negò al signor Loomis l'accesso al codice sorgente, e si rifiutò di fornire il dettaglio delle informazioni sulla base delle quali COMPAS prendeva le sue decisioni. Il rifiuto venne tenuto fermo, nonostante l'esplicito riferimento, nella sentenza, al punteggio fornito dal software, e fu inoltre motivato sulla base del segreto che copriva il codice sorgente. Nell'impugnazione della sentenza, Loomis sostenne che il diniego di accesso al codice sorgente ledeva il suo diritto costituzionalmente garantito ad un giusto processo, in quanto gli veniva precluso di conoscere la logica che soggiaceva alla previsione fatta dal *software*, e che costituiva una parte essenziale della motivazione della sentenza. Il caso si è concluso con la conferma da parte della Corte Suprema della sentenza di primo grado. Infatti, la Corte Suprema si è posta a sostegno delle argomentazioni dei giudici di primo grado, i quali, in sostanza, hanno affermato che la previsione di COMPAS non era altro che uno dei vari elementi, indipendenti l'uno dall'altro, presi in considerazione per giungere alla sentenza. Dunque, l'uso di COMPAS è stato ritenuto un fattore non determinante ai fini della decisione, bensì uno strumento di ausilio per i giudici. Di fatto, sono stati ritenuti prevalenti i diritti alla tutela della proprietà degli algoritmi, a scapito delle richieste di rivelazione dei codici da parte degli interessati. Nonostante la Corte Suprema abbia confermato la sentenza di primo grado, sono state poste alcune restrizioni all'uso di COMPAS, tra cui il divieto di utilizzarlo per decidere

[181] State v. Loomis, 881 N. W.2d 749, 767 (Wis. 2016).

se un imputato sia meritevole di incarcerazione o meno e quello di usarlo per stimare la durata della detenzione.[182]

Il caso Loomis è un importante punto di riflessione su come le macchine intelligenti, e gli algoritmi da cui sono mosse, abbiano bisogno di stringenti condizioni d'uso, specialmente quando le loro capacità predittive rischiano di produrre degli *output* potenzialmente lesivi dei diritti e delle libertà dell'interessato.

Potrebbe essere fatta una ulteriore considerazione conclusiva riflettendo, ancora una volta, sui potenziali rischi legati all'elaborazione dei dati da parte degli algoritmi, i quali, essendo per natura privi di capacità di giudizio critico, possono mal "collegare" le informazioni in *input* e restituire un *output* non solo affetto da fallacia logica, ma addirittura umanamente discriminatorio. Infatti, nel caso in oggetto il *software* COMPAS, utilizzato dai giudici, era in grado di fare previsioni utilizzando una complessa analisi statistica su dati raccolti tramite lunghi questionari (circa 137 quesiti) divisi in sezioni e informazioni aggiuntive estratte dai casellari giudiziari pubblici. Nonostante quest'algoritmo operasse sulla base di informazioni veritiere e neutrali, esso è risultato affetto da *bias* discriminatori di tipo razziale, che attribuivano agli imputati afroamericani punteggi di recidiva criminosa più alti di quelli attribuiti ad imputati di etnia caucasica.[183]

2.2. GDPR: l'*accountability* e la valutazione di impatto a tutela degli interessati soggetti a decisioni automatizzate

Alla luce delle considerazioni fatte fin qui, si può iniziare ad osservare un quadro più chiaro della posizione del GDPR in relazione allo sviluppo e all'utilizzo di tecnologie fortemente *data driven* quali la IA. Come si è già discusso preliminarmente nel capitolo primo, l'era dei *big data* ha dato inizio ad un *trend* in cui, di pari passo allo sviluppo di tecnologie sempre più tangenti la sfera privata delle persone, vi è il sempre crescente

[182] "Criminal Law — Sentencing Guidelines — Wisconsin Supreme Court Requires Warning Before Use of Algorithmic Risk Assessments in Sentencing. — 'State v. Loomis', 881 N.W.2d 749 (Wis. 2016)." *Harvard Law Review* 130, no. 5 (2017): 1530–37
Reperibile su: https://www.jstor.org/stable/44865547?seq=2

[183] How We Analyzed the COMPAS Recidivism Algorithm-*by Jeff Larson, Surya Mattu, Lauren Kirchner and Julia Angwin*- ProPublica- May 23, 2016
Reperibile su. https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm. Ultima visita al sito in data 17 febbraio 2025.

bisogno, da parte di dette tecnologie, di immense quantità di dati per poter performare. Di fronte a questo fatto oggettivo c'è poi la difficile ricerca di contemperare le esigenze di progresso tecnologico, con la soddisfacente applicazione dei principi del GDPR, primo fra tutti, (ma non l'unico) il consenso come base di legittimità del trattamento. Sul punto si è espressa in modo puntuale F. Faini la quale ha affermato che : "*Proprio tali criticità sostanziali, la mancata conoscenza preventiva delle finalità e le correlate difficoltà nel rispetto dei principi della normativa comportano difficoltà ad assicurare le informazioni da fornire da parte del titolare del trattamento*[184] *e il consenso libero, preventivo, specifico, inequivocabile e revocabile dell'interessato,*[185] *che rischiano di vanificarsi e di inficiare la stessa liceità del trattamento. Su cosa sarà informato e su cosa esprimerà il consenso l'interessato, se non si conoscono preventivamente le finalità di utilizzo dei Big Data?"* [186]
Su questo punto le visioni di M. Peluso[187], F.Faini[188] e E. Tosi[189] convergono nel ritenere che le difficoltà applicative della normativa sulla tutela dei dati sia principalmente dovuta all'opacità e alla chiusura sia sul quali dati vengono somministrati agli algoritmi sia sulla più generale gestione dei dati, e che, nell'era dei *big data*, non si possa più veramente parlare del consenso informato come efficace strumento di tutela.
A ben vedere, da questo scenario, deriva poi uno squilibrio di fatto tra potere del titolare e potere dell'interessato, che finisce per indebolire lo scudo di tutele normative azionabili dall'individuo.
Proseguendo il ragionamento in questo senso, si giunge ad una lettura del principio di *accountability* che valorizza le sue potenzialità soprattutto nella sua accezione di azioni preventive e precauzionali, le quali si estrinsecano nelle prescrizioni del GDPR tese spingere gli attori del trattamento ad agire nell'ottica di garantire una tutela tramite azioni di valutazione, gestione e prevenzione del rischio *ex ante*. Tale accezione del principio di *accountability* viene controbilanciata dal suo corrispettivo *ex post*, il quale si realizza in

[184] Artt. 12-14 GDPR
[185] Art. 7 GDPR
[186] F.Faini, *"Dati, Algoritmi e Regolamento europeo 2016/679"*, in *Regolare la tecnologia: il Reg. UE n. 2016/679 e la protezione dei dati personali. Un dialogo fra Italia e Spagna*, a cura di A. Mantelero e D. Poletti, 2018, pp. 343-344.
Reperibile su: https://www.academia.edu/67555128/Dati_algoritmi_e_Regolamento_europeo_2016_679
[187] Peluso, Maria Grazia. *Intelligenza artificiale e tutela dei dati : prospettive critiche e possibili benefici per una governance efficace*. Giuffrè Francis Lefebvre, 2024.
[188] Vedi nota 181 (F.Faini)
[189] E. Tosi, *Responsabilità civile per illecito trattamento dei dati personali e danno non patrimoniale*, Giuffrè, 2019

un duplice controllo: quello da parte dell'Autorità di protezione dati e quello dell'autorità giudiziaria, a cui spetta l'accertamento e la liquidazione dei danni, derivanti da illecito trattamento dei dati personali, e lesivi dei diritti alla riservatezza, alla tutela dei dati e all'identità dell'individuo.

Dunque, nell'impraticabilità più assoluta di limitare il progresso tecnologico, il quadro regolatorio delineato dal GDPR ha subito, nel contesto dello scenario dell'era dei *big data* un cambiamento di prospettiva[190]: è iniziata infatti ad emergere- come una delle poche soluzioni veramente praticabili per porre un freno alla crescente disparità di posizioni tra titolare ed interessato - il principio di *accountability* potenziato dalle sfumature dei principi di prevenzione e di precauzione della responsabilità civile, da perseguire, tenuto conto dello stato dell'arte, della tecnica e dei costi di attuazione, sino ai limiti di quanto sia considerato socialmente, economicamente e giuridicamente accettabile. Naturalmente escluso ogni sforzo sproporzionato al dovere precauzionale[191].

Aver spostato il punto focale sull'accezione preventiva del principio di *accountability*, comporta lo spostamento del fulcro dell'azione sul titolare del trattamento e sugli obblighi cui deve ottemperare. Spetta al titolare sia valutare le misure tecniche e organizzative più adatte alla natura dei dati, all'oggetto e alle finalità del trattamento, sia essere poi in grado di comprovarne l'effettività in relazione al rischio creato con il trattamento in questione. Al fine di rendere più agevole la prova - anche tenuto conto del fatto che la norma di riferimento, ovvero l'articolo 24 del GDPR, ha una portata alquanto ampia e non fornisce alcun elenco esaustivo di misure adeguate ma, piuttosto, caldeggia da parte del titolare un comportamento proattivo - vengono proposti agli artt. 40-43 del GDPR degli

---

[190] E. Tosi, *Responsabilità civile per illecito trattamento dei dati personali e danno non patrimoniale*, Giuffrè, 2019, p. 40: "*Tale impostazione segna il superamento della vecchia impostazione del trattamento dati come attività pericolosa tutelata, soprattutto, ex post sul piano del risarcimento del danno, all'impostazione ex ante di prevenzione e precauzione del danno stesso: analisi, gestione e controllo del rischio dell'attività di trattamento dei dati personali e responsabilizzazione del titolare sono i nuovi fattori che segnano il già segnalato cambio di prospettiva rispetto alla normativa previgente, permeando tutto l'impianto normativo del GDPR."*

[191] Art. 24, reg. UE n. 679/2016. *"1. Tenuto conto della natura, dell'ambito di applicazione, del contesto e delle finalità del trattamento, nonché dei rischi aventi probabilità e gravità diverse per i diritti e le libertà delle persone fisiche, il titolare del trattamento mette in atto misure tecniche e organizzative adeguate per garantire, ed essere in grado di dimostrare, che il trattamento è effettuato conformemente al presente regolamento. Dette misure sono riesaminate e aggiornate qualora necessario.*
2. Se ciò è proporzionato rispetto alle attività di trattamento, le misure di cui al paragrafo 1 includono l'attuazione di politiche adeguate in materia di protezione dei dati da parte del titolare del trattamento."

accountability tools[192], quali le certificazioni e i codici di condotta. L'intenzione con cui sono stati messi a punto questi strumenti non è solo quello dell'autoregolazione, ma vi è anche la prospettiva di ingenerare un senso di affidabilità negli utenti soggetti a trattamenti. Nondimeno, secondo una visione congiunta di Lucchini Guastalla[193] e Poletti-Causarano[194], codici di condotta e certificazioni sono strumenti di agevolazione probatoria flessibili ed adattabili sia rispetto agli obblighi del titolare, sia delle garanzie richieste in capo al responsabile del trattamento, e perciò efficaci nel supportare l'applicazione della normativa in materia di tutela dei dati[195].

Tuttavia, nemmeno quest'approccio è immune da osservazioni volte ad incrementare il coinvolgimento degli interessati. Sotto tale profilo, secondo D. Poletti e M. Causarano [196] risulta carente la logica dell' "*accountability with the public*", ovvero le modalità con cui si portano gli interessati a partecipare a tutte le fasi del trattamento: dall'adozione delle misure precauzionali all'attuazione passando per il monitoraggio.

Se le nuove prospettive applicative del principio di *accountability* sono mosse dall'esigenza di restituire rispetto e controllo sui dati agli interessati, ed incrementare la loro fiducia, da un lato, e, dall'altro, garantire una costante effettività nell'applicazione della normativa, allora sarebbe un atteggiamento auspicabile riuscire ad integrare maggiormente, ed in modo stabile, gli interessati, in un dialogo con titolari e responsabili del trattamento. Quest'evenienza però, al momento, resta affidata al personale apprezzamento dei titolari.

Tirando le fila del discorso, dunque, l'ampia portata applicativa dell'articolo 24 del GDPR, sta permettendo alla normativa di orientare la propria efficacia anche nell'attuale momento storico, che vive l'ascesa delle tecnologie *data driven* (tra cui l'IA) come un fattore di minaccia alla garanzia del diritto alla tutela dei dati.

---

[192] D. Poletti, M. Causarano, "Autoregolamentazione privata e tutela dei dati personali: tra codici di condotta e meccanismi di certificazione", in Privacy digitale, a cura di E. TOSI, Giuffrè, 2019, p. 378

[193] E. Lucchini Guastalla, "Privacy e data protection: principi generali", in Tosi, Emilio, et al. *Privacy digitale : riservatezza e protezione dei dati personali tra GDPR e nuovo Codice privacy*. Giuffrè Francis Lefebvre, 2019. P. 83

[194] Poletti, M. Causarano, "Autoregolamentazione privata e tutela dei dati personali: tra codici di condotta e meccanismi di certificazione", in Privacy digitale, a cura di E. TOSI, Giuffrè, 2019,

[195] E. Lucchini Guastalla, "Privacy e data protection: principi generali", in Tosi, Emilio, et al. *Privacy digitale : riservatezza e protezione dei dati personali tra GDPR e nuovo Codice privacy*. Giuffrè Francis Lefebvre, 2019. P.380

[196] Ibidem

In particolare, è proprio l'assenza di un elenco esaustivo di misure ritenute adeguate, o obblighi per il titolare, che riflette la concessione di una libertà che implica una maggiore responsabilizzazione. Responsabilizzazione che si declina, in modo concreto, indipendentemente dallo specifico trattamento posto in essere[197].

Lo stesso Gruppo di Lavoro art. 29, a suo tempo, nell'Opinione 3/2010 si espresse in modo favorevole a trasformare i principi generali di protezione dei dati, da materia normativa rigida a politiche e procedure concrete, definite a livello di titolare del trattamento, e calibrate per le specifiche esigenze nascenti dai diversi trattamenti, chiaramente nel rispetto delle leggi e dei regolamenti applicabili[198].

Prendendo in considerazione sul punto anche quanto sostenuto da Comandè[199], quest'approccio, nel complesso rispettoso dell'architettura del GDPR, e che posiziona la *compliance* normativa come livello base di tutela, si riflette positivamente anche nella disciplina dell'intelligenza artificiale, consistendo allo stesso tempo anche in uno strumento a sostegno dell'attribuzione della responsabilità civile e teso alla diminuzione dei rischi di danno.

2.2.1. La valutazione preventiva di impatto

Allo stesso tipo di logica precauzionale delle previsioni di responsabilizzazione del titolare del trattamento, risponde l'articolo 35 del GDPR, dedicato alla valutazione di impatto sula protezione dei dati. In sunto, la disposizione predispone che venga condotta una valutazione di impatto quando il trattamento possa comportare, in conseguenza

---

[197] G. Finocchiaro, "Il principio di accountability- GDPR tra novità e discontinuità", in Giurisprudenza italiana, n. 12, Utet, 2019, p. 2777. *"In questo ambito, il principio di accountability può costituire un approccio quanto mai appropriato al problema, dal momento che alloca il rischio presso il soggetto, cioè il titolare del trattamento dei dati, che meglio è in grado di esaminare il contesto e di valutare come affrontarlo e che sarà chiamato a dimostrare l'adeguatezza delle scelte adottate."*

[198] Gruppo di Lavoro art. 29, Parere 3/2010, p. 9, punto 27.

[199] G. Comandè, "*Intelligenza Artificiale e responsabilità tra liability e accountability. Il carattere trasformativo della IA e il problema della responsabilità*", in Analisi giuridica delle economia, n.1, 2019. P.171.

"*la centralità del ruolo dei dati nel ciclo di vita delle AI suggerisce di prendere in attenta considerazione la collocazione del principio di accountability al centro delle regole di responsabilità [...]. Resta il fatto che la qualità, la quantità dei dati, il contesto di raccolta, le modalità di selezione rimangono centrali nella definizione di molti profili anche nuovi della responsabilità civile connessa alle IA."*

dell'uso di nuove tecnologie, rischi elevati per i diritti e le libertà degli interessati[200]. Tuttavia, solo a rischi considerati elevati[201] consegue l'obbligo di procedere ad una valutazione di impatto, la quale dovrà tenere in considerazione le modalità di svolgimento del trattamento, la base giuridica su cui poggia, le finalità perseguite e anche le misura che il titolare ha previsto di usare per prevenire ricadute sugli interessati[202].

Quest'approccio, basato sulla limitazione modulata del rischio a seconda della sua intensità, ben si adatta all'enorme circolazione di dati trattati da tecnologie *data driven* che caratterizza l'era dei *big data*.

Se si riflette sul fatto che la valutazione in merito alla rischiosità del trattamento rimane ancorata all'autonomia del titolare (che dovrà poi essere in grado di dimostrare l'adeguatezza e la lungimiranza delle misure prese), ben si percepisce la linea di continuità con il principio di *accountability*. Da questo punto di vista, si può constatare come gli oneri procedurali in capo al titolare siano stati in parte stornati [203] [204], così come è stata esclusa, per ragioni di onerosità e velocità del mercato digitale, una valutazione preventiva da parte delle Autorità di controllo.[205] Anche se, lasciare che il titolare del trattamento, sulla base di una sua previa autovalutazione, agisca da filtro rispetto al controllo dell'Autorità garante, è una scelta che potrebbe in astratto portare a sottostime del rischio da parte del titolare.[206]

---

[200] Art. 35(1)(2) GDPR: "*1. Quando un tipo di trattamento, allorché prevede in particolare l'uso di nuove tecnologie, considerati la natura, l'oggetto, il contesto e le finalità del trattamento, può presentare un rischio elevato per i diritti e le libertà delle persone fisiche, il titolare del trattamento effettua, prima di procedere al trattamento, una valutazione dell'impatto dei trattamenti previsti sulla protezione dei dati personali. Una singola valutazione può esaminare un insieme di trattamenti simili che presentano rischi elevati analoghi.*

2. Il titolare del trattamento, allorquando svolge una valutazione d'impatto sulla protezione dei dati, si consulta con il responsabile della protezione dei dati, qualora ne sia designato uno." […]

[201] Al riguardo il Gruppo di Lavoro art. 29 ha promulgato nel 2017 delle linee guida in materia di valutazione di impatto che poi sono state adottate anche dal EDPB.

WP 248-rev.01. Linee guida in materia di valutazione d'impatto sulla protezione dei dati e determinazione della possibilità che il trattamento possa presentare un rischio elevato ai fini del regolamento (UE) 2016/679.

[202] D'acquisto G. , Naldi M., *Big data e Privacy by Design*, Giappichelli editore, 2019. P. 30 ss

[203] F. Mollo, *"Gli obblighi previsti in funzione di protezione dei dati personali", Cap X in "Persona e mercato dei dati. Riflessioni sul GDPR"*, a cura di Zorzi Galgano N., Cedam, 2019. P. 290 ss

[204] Nella Direttiva del 1995, all' art. 18 era previsto un generalizzato obbligo di notificare alle Autorità di controllo i trattamenti.

[205] R. Torino, *"La valutazione di impatto (Data Protection Impact Assessment)*", in *"I dati personali nel diritto europeo"*, pp. 855 ss

[206] E. Lucchini Guastalla, "Privacy e data protection: principi generali", in Tosi, Emilio, et al. *Privacy digitale : riservatezza e protezione dei dati personali tra GDPR e nuovo Codice privacy*. Giuffrè Francis Lefebvre, 2019

Il legislatore europeo però fa una interessante presunzione di alto rischio, al punto 3 lett. (a) dell'art. 35, in cui introduce l'obbligo di valutazione di impatto per tutti i trattamenti che prevedono *"una valutazione sistematica e globale di aspetti personali relativi a persone fisiche, basata su un trattamento automatizzato, compresa la profilazione*"[207]. Secondo l'avviso di M. Peluso, che qui si condivide, la previsione è stata concepita per far fronte alle criticità legate all'utilizzo delle tecnologie *data driven*, che, grazie all'obbligo di valutazione di impatto, sono possono restituire un più alto livello di trasparenza sull'intero processo automatizzato[208].

In ultimo, resta da precisare che il vaglio dell'Autorità di controllo viene ulteriormente imposto, qualora le misure previste dalla valutazione di impatto non fossero sufficienti a limare i rischi derivanti dal trattamento; in questo caso, l'art. 36 GDPR obbliga il titolare a consultare preventivamente l'Autorità di controllo, la quale dispone di poteri: di indagine, correttivi, autorizzativi e consultivi, a norma dell'art. 58 GDPR. In questo modo, il legislatore ha dato ulteriore risonanza al principio di precauzione che, in un contesto di rapido sviluppo tecnologico come quello attuale, aiuta a tutelare gli utenti *ex ante* il verificarsi di eventuali danni[209].

3. La figura del *Data Protection Officer* e l'inizio della convivenza con l'IA

Con l'entrata in vigore del GDPR, nel maggio 2018, è stata introdotta la figura del Responsabile della protezione dei dati, meglio conosciuto nella sua versione inglese di *Data Protection Officer* (DPO). L'istituzione di questa figura può essere vista come un corollario del principio di *accountability* così come chiosato dall'art. 5(2) GDPR[210]. Principio che, lungo tutto il Regolamento, si protende a segnare il passaggio ad una

[207] Art. 35(3) GDPR:
"3. La valutazione d'impatto sulla protezione dei dati di cui al paragrafo 1 è richiesta in particolare nei casi seguenti:
a) una valutazione sistematica e globale di aspetti personali relativi a persone fisiche, basata su un trattamento automatizzato, compresa la profilazione, e sulla quale si fondano decisioni che hanno effetti giuridici o incidono in modo analogo significativamente su dette persone fisiche;"

[208] Peluso, Maria Grazia. *Intelligenza artificiale e tutela dei dati : prospettive critiche e possibili benefici per una governance efficace*. Giuffrè Francis Lefebvre, 2024. P. 242.

[209] Ibidem p. 245

[210] Ai sensi dell'art. 5(2) del GDPR il titolare del trattamento è competente per il rispetto dei principi di liceità, correttezza, trasparenza, limitazione delle finalità, minimizzazione, esattezza, limitazione della conservazione, integrità e riservatezza; dovendo in ogni caso essere in grado di dimostrarlo.

concezione più dinamica della protezione dati: fatta di scelte da e soppesare azioni da intraprendere[211][212].

Nell'era dei *big data* e delle tecnologie alimentate dalle grandi quantità di dati come lo è l'IA, il DPO è la figura giuridica ed aziendale creata dal GDPR per supportare la protezione dei dati personali in modo specializzato e professionale, senza però sacrificare quella flessibilità tanto importante per scalare le evoluzioni tecnologiche e sociali del nostro tempo[213].

Sulle promesse di questa nuova figura, già si era espresso l'allora Presidente dell'Autorità Garante per la Protezione dei Dati Personali Antonello Soro, il quale, il quale si riferì al DPO come uno "*strumento "nuovo" di governance della dimensione digitale che, se adeguatamente valorizzato, consentirà anche una corretta ed efficace reingegnerizzazione dello "stato digitale"*[214].

Il vento di riforma, portato dall'entrata in vigore del GDPR, ha attribuito ai soggetti che trattano dati personali, *ipso facto*, la responsabilità sulla loro tutela. Mutuando la metafora utilizzata da R. Panetta: *"Dopo la riforma chi tratta dati personali è posto in un ruolo quasi genitoriale rispetto ai trattamenti che predispone; infatti, anche per il titolare la posizione di garanzia deriva da un elemento puramente fattuale […]"*[215]. Ai titolari, dunque, è richiesto in modo permanente e a prescindere dalla presenza o meno del DPO, di garantire forme effettive e concrete di tutela attraverso un costante processo di adeguamento delle misure adottate.

All'interno dello scenario, fin qui brevemente tratteggiato, il DPO agisce come possibile filtro intermedio tra il soggetto a cui fa capo la gestione del trattamento (titolare o responsabile che sia) ed il complesso di azioni volte a rendere questi soggetti GDPR *compliant*[216].

---

[211] R. Panetta (a cura di), *Circolazione e protezione dei dati personali, tra libertà e regole del mercato*, Giuffrè Francis Lefebvre, 2019

[212] In proposito, si veda anche European Data Protection Supervisor – EDPS, Additional EDPS comments on the data protection reform package, Brussels, 2013. P. 8, punto 31.

[213] R. Panetta et al. *Il Data Protection Officer tra regole e prassi*. Seconda edizione., Giuffrè Francis Lefebvre, 2023.

[214] *Tra privacy e open data intesa possibile* – Intervento di Antonello Soro, 13 ottobre 2014. Reperibile sul sito del Garante: https://www.garanteprivacy.it/home/docweb/-/docweb-display/content/id/3467378

[215] *Ibidem*

[216] L. Ferola, *La «nuova» figura del responsabile della protezione dei dati personali e le sue caratteristiche*, in R. Panetta (a cura di), , pp.347-365

A questo punto, si può proseguire il ragionamento iniziato nel paragrafo 2 di questo capitolo, nel quale si è approfondita la ricerca di un equilibrio tra la necessità di trasparenza nel trattamento e processo decisionale automatizzato nel contesto della IA, indagando sui profili evolutivi del DPO che, in questi sei anni di maturazione applicativa del GDPR, ha visto il proprio ruolo arricchirsi di sfaccettature e nuove competenze da acquisire.

Prima di procedere però, è opportuno ricapitolare brevemente la cornice normativa da cui la figura del DPO prende le mosse.

3.1 La nomina del DPO

La sezione 4 del GDPR, che racchiude gli artt. dal 37 al 39, condensa tutti gli aspetti sia teorici che pratici riguardanti questo soggetto. Un primo aspetto che vale la pena rilevare è il fatto che la sua designazione è un atto che riguarda tanto i titolari quanto i responsabili del trattamento, indice questo non solo della propensione del Regolamento responsabilizzare i soggetti che, a vario titolo, gestiscono i trattamenti, ma anche della consapevolezza che, rispetto al secolo scorso, l'era dei *big data* avrebbe portato sempre più di frequente e con proporzioni sempre più grandi, all'esternalizzazione dei trattamenti di dati[217]. Di conseguenza, i responsabili del trattamento vengono considerati delle vere e proprie ramificazioni operative dei titolari[218][219].

L'articolo 37 (1)GDPR[220] introduce tre ipotesi tassative in cui la nomina di un DPO è obbligatoria:

---

[217] *Ibidem*

[218] E. Pelino , M. E. Carpenelli , Artt. 37-39 GDPR, in L. BOLOGNINI, E. PELINO (a cura di), *Codice della disciplina Privacy*, Giuffrè Francis Lefebvre, 2019, p. 269

[219] R. Panetta et al. *Il Data Protection Officer tra regole e prassi*. Seconda edizione., Giuffrè Francis Lefebvre, 2023. P. 7

[220]Articolo 37 (1) GDPR: "*1. Il titolare del trattamento e il responsabile del trattamento designano sistematicamente un responsabile della protezione dei dati ogniqualvolta: a) il trattamento è effettuato da un'autorità pubblica o da un organismo pubblico, eccettuate le autorità giurisdizionali quando esercitano le loro funzioni giurisdizionali; b) le attività principali del titolare del trattamento o del responsabile del trattamento consistono in trattamenti che, per loro natura, ambito di applicazione e/o finalità, richiedono il monitoraggio regolare e sistematico degli interessati su larga scala; oppure c) 2. le attività principali del titolare del trattamento o del responsabile del trattamento consistono nel trattamento, su larga scala, di categorie particolari di dati personali di cui all'articolo 9 o di dati relativi a condanne penali e a reati di cui all'articolo 10."*

a) In caso di trattamento svolto da un'autorità od organismo pubblico[221].

b) Quando le attività principali del titolare del trattamento o del responsabile del trattamento consistono in trattamenti che richiedono il monitoraggio regolare e sistematico di interessati su larga scala.

c) Quando le attività principali del titolare del trattamento o del responsabile del trattamento consistono nel trattamento su larga scala di categorie particolari di dati o di dati personali relativi a condanne penali e reati.

Al di fuori di questi tre casi, la nomina di un DPO è facoltativa. C'è da precisare poi che, qualora un'organizzazione opti per la nomina volontaria, al DPO si applicheranno tutti i requisiti previsti dalla disciplina come nel caso di nomina obbligatoria; la disciplina prevista dal GDPR non è in alcun modo rimodulabile[222].

3.2. L'indipendenza del DPO

Dal combinato disposto dell'art. 38 (3) GDPR[223] e delle ultime righe del Considerando 97[224] si ricavano le linee generali che tratteggiano l'indipendenza dell'agire del DPO. Per indipendenza del DPO si intende, principalmente, quel grado di autonomia che egli deve conservare riguardo l'approccio da tenere e gli obiettivi da conseguire. La natura mista di questo soggetto, legato contrattualmente alle organizzazioni presso cui presta servizio, fa sì che, per un effettivo svolgimento delle sue mansioni, che possono anche sfociare in censure alle misure scelte dal titolare o responsabile, abbia bisogno di totale indipendenza funzionale. L'art. 38 GDPR, in conclusione del comma (3), si preoccupa di fornire un'ultima indicazione riguardo: il posizionamento del DPO rispetto alle altre figure che fanno parte dell'organizzazione per cui lavora. Egli è tenuto a riportare le proprie

[221] L'art. 37(1)(a) nello specifico prevede che siano escluse le autorità giurisdizionali quando esercitano le loro funzioni giurisdizione.

[222] WP 243 rev.01-Linee guida sui responsabili della protezione dei dati ('DPO') p. 5-6: *"When an organisation designates a DPO on a voluntary basis, the requirements under Articles 37 to 39 will apply to his or her designation, position and tasks as if the designation had been mandatory."*

[223] Articolo 38(3)GDPR: 3. "*Il titolare del trattamento e il responsabile del trattamento si assicurano che il responsabile della protezione dei dati non riceva alcuna istruzione per quanto riguarda l'esecuzione di tali compiti. Il responsabile della protezione dei dati non è rimosso o penalizzato dal titolare del trattamento o dal responsabile del trattamento per l'adempimento dei propri compiti. Il responsabile della protezione dei dati riferisce direttamente al vertice gerarchico del titolare del trattamento o del responsabile del trattamento."*

[224] Considerando 97 GDPR:*"[..] Tali responsabili della protezione dei dati, dipendenti o meno del titolare del trattamento, dovrebbero poter adempiere alle funzioni e ai compiti loro incombenti in maniera indipendente"*

determinazioni solamente *"al vertice gerarchico del titolare del trattamento o del responsabile"*. Se così non fosse, dover riportare le proprie determinazioni anche ai livelli intermedi dell'organizzazione, potrebbe risultare controproducente per la sua possibilità di contrastare le misure dei vertici[225].

3.3. Poteri, compiti e funzioni del DPO

Il DPO, nell'esecuzione dei suoi compiti deve avere pieno accesso alle strutture dell'organizzazione, e deve poter ricevere tutte le risorse necessarie al sostegno della sua opera da parte del titolare e del responsabile[226].

Il potere di contrastare le scelte di trattamento, poste in essere dal titolare ed in esecuzione dal responsabile, trova un limite nel fatto che egli non dispone di un vero e proprio potere di veto, non essendo un'estensione dell'Autorità di controllo presso le organizzazioni.

Tuttavia, egli ben può formalizzare la sua opinione contraria in uno scritto[227]. Su questo punto, il Gruppo Art. 29 si è espresso anche riguardo il limite negativo dei poteri del DPO, chiosando che: *"L'autonomia dei DPO non significa, tuttavia, che essi dispongono di poteri decisionali che vanno oltre i loro compiti ai sensi dell'articolo 39"*[228]. Dunque, in ogni caso, i poteri affidati al DPO sono coperti da un duplice ordine di "tetti" sovrapposti: il primo e più grande è il dettato del GDPR, il secondo e chiaramente più piccolo, è quanto concordato con i vertici dell'organizzazione destinataria della sua opera.

Compiti e funzioni del DPO sono contenuti nel raggruppamento non esaustivo dell'Art. 39[229], l'elenco prevede 5 gruppi di compiti: 1) informazione e consulenza; 2) sorveglianza

---

[225] R. Panetta et al. *Il Data Protection Officer tra regole e prassi*. Seconda edizione., Giuffrè Francis Lefebvre, 2023. P. 47-64

[226] Art 38 (2)GDPR: *"Il titolare e del trattamento e il responsabile del trattamento sostengono il responsabile della protezione dei dati nell'esecuzione dei compiti di cui all'articolo 39 fornendogli le risorse necessarie per assolvere tali compiti e accedere ai dati personali e ai trattamenti e per mantenere la propria conoscenza specialistica."*

[227] C. Zhang, S. Bengio, M. Hardt, B. Recht, O. Vinyals, Understanding deep learning (still) requires rethinking generalization, Communications of the ACM 64 (3) (2021) 107–115

[228] WP 243 rev.01-Linee guida sui responsabili della protezione dei dati ('DPO') p. 15

[229] Articolo 39 GDPR: "*1. a) Il responsabile della protezione dei dati è incaricato almeno dei seguenti compiti: informare e fornire consulenza al titolare del trattamento o al responsabile del trattamento nonché ai dipendenti che eseguono il trattamento in merito agli obblighi derivanti dal presente regolamento nonché da altre disposizioni dell'Unione o degli Stati membri relative alla protezione dei dati; b) sorvegliare l'osservanza del presente regolamento, di altre disposizioni dell'Unione o degli Stati membri relative alla protezione dei dati nonché delle politiche del titolare del trattamento o del responsabile del trattamento in*

sulla corretta osservazione del Regolamento; 3) parere, ove richiesto, sulla valutazione di impatto; 4) cooperazione con l'Autorità di controllo; 5) fungere da punto di raccordo con l'Autorità di controllo.

3.4. I primi esiti della convivenza

Giungendo alle conclusioni, è opinione di chi scrive che la figura del DPO, anche, e forse soprattutto, nella sua accezione facoltativa, abbia una notevole importanza, non solo in qualità di esperto, ma anche come polo di riferimento e di indirizzo per i vertici del trattamento (come anche dipendenti), verso la conformità al GDPR e l'adozione di *best practices* volte a dimostrare una maggiore responsabilizzazione.

Tuttavia, si deve pur ammettere che non si possa pretendere, da un solo soggetto, *l'expertise* necessaria per comprendere questioni di natura più tecnica o, semplicemente, campi in cui l'esperienza e la professionalità acquisita potrebbero essere non dominanti. Per questa ragione, il DPO ha la possibilità di usare le risorse che l'organizzazione ha messo a disposizione per le sue attività, per ricorrere a consulenze specifiche.

A questo punto, una domanda sorge quasi da sé: in un orizzonte in cui l'utilizzo della IA ha conseguenze sempre più intense nello svolgimento della vita umana[230], quali sono le prospettive evolutive del ruolo di DPO?

Indipendentemente dal fatto che i DPO assumano la piena responsabilità della conformità dell'IA all'interno di un'organizzazione, è chiaro che devono far parte della discussione ogni volta che i sistemi di IA trattano (o potrebbero) dati personali[231].

---

materia di protezione dei dati personali, compresi l'attribuzione delle responsabilità, la sensibilizzazione e la formazione del personale che partecipa ai trattamenti e alle connesse attività di controllo; c) fornire, se richiesto, un parere in merito alla valutazione d'impatto sulla protezione dei dati e sorvegliarne lo svolgimento ai sensi dell'articolo 35; d) cooperare con l'autorità di controllo; e e) 2. fungere da punto di contatto per l'autorità di controllo per questioni connesse al trattamento, tra cui la consultazione preventiva di cui all'articolo 36, ed effettuare, se del caso, consultazioni relativamente a qualunque altra questione. Nell'eseguire i propri compiti il responsabile della protezione dei dati considera debitamente i rischi inerenti al trattamento, tenuto conto della natura, dell'ambito di applicazione, del contesto e delle finalità del medesimo."

[230] Si veda da ultimo il recente caso della CGUE C-634/21, SCHUFA Holding (Scoring): Judgment of the Court (First Chamber) of 7 December 2023 (request for a preliminary ruling from the Verwaltungsgericht Wiesbaden — Germany) — OQ v Land Hessen

[231] AI and Personal Data A Guide for DPOs "Frequently Asked Questions", CEDPO AI Working Group, June 2023. P. 13.

Per quel che riguarda l'aspetto pratico, il GDPR, quindi, ammette che i DPO regolino l'uso dell'IA nel trattamento dei dati personali. Nel fare ciò, è essenziale che il DPO lavori basandosi anche sulle implicazioni e i limiti del sistema di intelligenza artificiale, soprattutto per quanto riguarda eventuali pregiudizi che potrebbero potenzialmente verificarsi.

Il limite del ruolo di garanzia del DPO sta nel fatto che, come evidenziato dalla CEDPO[232], porre sulle sue spalle poteri decisionali sulla governance dell'IA, potrebbe sia comprometterne l'indipendenza (costringendolo a rendere conto del proprio operato non più solo ai vertici responsabili del trattamento), sia ingenerare un conflitto di interessi ai sensi dell'art. 38 (6)[233] (attribuendogli, oltre al compito di far rispettare la normativa sulla protezione dati, anche parzialmente la responsabilità dell'attuazione delle attività di trattamento).

[232] Is the DPO the right person to be the AI Officer? CEDPO AI and Data Working Group Micro-Insights Series, July 2024. P. 5.

[233] Articolo 38 (6)GDPR. *"Il responsabile della protezione dei dati può svolgere altri compiti e funzioni. Il titolare del trattamento o il responsabile del trattamento si assicura che tali compiti e funzioni non diano adito a un conflitto di interessi."*

CAPITOLO QUARTO

SI POSSONO VERAMENTE MITIGARE GLI EFFETTI DELLO SCRAPING?

L'elaborato si è aperto dando, nel capitolo primo, un ampio contesto sullo *scraping* e sugli attriti che l'uso di questo strumento provoca al rispetto della disciplina sulla protezione dei dati personali. Nel secondo capitolo, è stata tracciata una cornice giuridica di riferimento che, poi, nel capitolo terzo è stata ulteriormente ampliata nei suoi aspetti pratici, andando ad evidenziare quali aspetti strutturali e funzionali della IA, possano essere i più inclini a rappresentare un rischio per l'integrità del diritto alla protezione dei dati personali, proprio per il fatto di coinvolgere grandi quantità di dati derivanti dallo *scraping*.

Giungiamo ora alla fase conclusiva, concentrando il ragionamento sulle possibili attività che, seppur non del tutto solutorie del problema, possono essere considerate alla stregua di cautele attuative del principio di *accountability* e della tutela dati *by design*, volte a ridurre, per quanto consentito dallo stato dell'arte e della tecnica, l'uso non autorizzato dei dati personali pubblicamente disponibili.

Negli ultimi anni, è cresciuta infatti l'esigenza di implementare approcci e tecniche che permettessero alle aziende, di ogni dimensione, di ispirarsi a tali cautele e adattarne l'uso al proprio contesto. A questa necessità ha sicuramente contribuito il rapido diffondersi dell'utilizzo dei modelli di IA generativa, i quali, essendo sempre più spesso addestrati con dati accessibili sul *web*, e predisposti all'apprendimento e aggiornamento continuo sulla base delle interazioni con l'utente, ben può accadere che "rigurgitino" risultati relativi a persone fisiche, delle quali abbiano assorbito i dati personali in fase di addestramento.

Questa attenzione è stata anche sostenuta dalla (giustificata) risonanza che hanno avuto casi come quello, già discusso, riguardante Clearview AI e, temporalmente antecedente, quello che vide LinkedIn difendere il proprio diritto a adottare misure per impedire lo

scraping dei dati pubblici dei propri utenti con specifico riferimento all'attività della società di analisi dati "HiQ"[234].

A questo punto quindi, si può notare uno scollamento tra la correttezza dei principi, che regolano il trattamento dei dati personali, e la difficoltà nel raggiungere un sufficiente grado di conformità a tale normativa. A partire dal rispetto del principio della trasparenza ex art. 5 (1) (a) del GDPR[235], che pure come sottolineato al Considerando 58 del GDPR, risulta essere un aspetto chiave: "[…] *in situazioni in cui la molteplicità degli operatori coinvolti e la complessità tecnologica dell'operazione fanno sì che sia difficile per l'interessato comprendere se, da chi e per quali finalità sono raccolti dati personali che lo riguardano, quali la pubblicità online.*"[236] Trasparenza che svolge, inoltre, un ruolo cruciale verso l'effettività della protezione dei dati *by design* e nel garantire l'*accountability* (come stabilito dal Considerando 78[237]) all'interno del contesto dei sistemi di intelligenza artificiale.

Tuttavia, garantire la trasparenza dei sistemi di intelligenza artificiale con riferimento alle loro pratiche utilizzate in materia di dati, tramite spiegazioni chiare sul loro funzionamento e sulle decisioni che prendono, può essere difficile a causa della complessità dei sistemi stessi. Ma, ed è opinione di chi scrive, questa complessità non può, e non deve, trovare una giustificazione in sé stessa, perché allora il percorso evolutivo della IA generativa sarebbe uno scontrarsi continuo con la normativa sulla tutela dei dati personali, e di ciò, due esempi molto recenti sono a portata di mano. Il primo caso è quello che ha coinvolto OpenAI, ovvero la società titolare di ChatGPT, alla quale è stato destinato il provvedimento correttivo e sanzionatorio del Garante adottato a

[234] hiQ Labs, Inc. v LinkedIn Corporation, Case 17-cv-03301-EMC

[235] Art. 5 GDPR: "1. *I dati personali sono:*
a) trattati in modo lecito, corretto e trasparente nei confronti dell'interessato. […]"

[236] Considerando 58 del GDPR.

[237] Considerando 78 del GDPR:"[…] *offrire trasparenza per quanto riguarda le funzioni e il trattamento di dati personali, consentire all'interessato di controllare il trattamento dei dati e consentire al titolare del trattamento di creare e migliorare le caratteristiche di sicurezza. In fase di sviluppo, progettazione, selezione e utilizzo di applicazioni, servizi e prodotti basati sul trattamento di dati personali o che trattano dati personali per svolgere le loro funzioni, i produttori dei prodotti, dei servizi e delle applicazioni dovrebbero essere incoraggiati a tenere conto del diritto alla protezione dei dati allorché sviluppano e progettano tali prodotti, servizi e applicazioni e, tenuto debito conto dello stato dell'arte, a far sì che i titolari del trattamento e i responsabili del trattamento possano adempiere ai loro obblighi di protezione dei dati.* […]"

novembre 2024[238] all'esito di un'istruttoria[239] che ha richiesto più di un anno di lavori. Alla società è stata contestata la mancata notifica all'Autorità riguardo la violazione dei dati subìta a marzo 2023, la mancanza di una base giuridica per il trattamento dei dati personali dei propri utenti (fatto, questo, che ha comportato anche la violazione del principio di trasparenza e lasciati insoddisfatti gli obblighi informativi nei confronti degli utenti). Come parte del provvedimento correttivo, ed a norma dell'art. 166 comma 7 del Codice Privacy[240], il Garante ha richiesto ad OpenAI di realizzare una campagna informativa della durata di sei mesi da trasmettere sui principali media. Quest'aspetto del provvedimento, anche se non direttamente legato allo *scraping* di dati, è tuttavia particolarmente rilevante, se si considera l'importanza che ha la creazione di consapevolezza nel pubblico sul funzionamento delle chatbots basati sulla IA generativa, e sulla raccolta dei dati degli utenti e non-utenti per l'addestramento dei modelli stessi.

Creare consapevolezza e comprensione nel pubblico, infatti, non solo rende edotti gli interessati riguardo i diritti da loro esercitabili (tra cui opposizione, rettifica e cancellazione), ma, al netto delle misure che possono essere messe in atto per contrastare lo *scraping*, un pubblico adeguatamente informato è un pubblico in grado di moderare con consapevolezza sia la propria presenza nel *web* che l'utilizzo dell'intelligenza artificiale generativa.

Il caso più recente in linea temporale, il quale peraltro è solo al principio del suo sviluppo, tratta dell'istruttoria aperta dal Garante nei confronti delle società che gestiscono la chatbot R1 di DeepSeek uscito a Gennaio 2025. In data 28 gennaio 2025 il Garante ha comunicato di aver inviato una richiesta di informazioni[241] alle società che gestiscono DeepSeek. Le questioni che DeepSeek avrebbe dovuto chiarire vertevano sui tipi di dati

[238] Garante per la Protezione dei Dati Personali-Provvedimento del 2 novembre 2024 - [doc. web n. 10085455]

[239] Il 31 marzo 2024, contestualmente alla limitazione immediata del trattamento dei dati degli utenti italiani disposta nei confronti di OpenAI, il Garante ha contestualmente aperto un'istruttoria sul caso. Garante per la Protezione dei Dati Personali- Intelligenza artificiale: il Garante blocca ChatGPT. Raccolta illecita di dati personali. Assenza di sistemi per la verifica dell'età dei minori-31 marzo 2024 [Doc-Web 9870847].

[240] Articolo 166 comma 7 del Codice Privacy: "*Nell'adozione dei provvedimenti sanzionatori […] può essere applicata la sanzione[…] dell'ingiunzione a realizzare campagne di comunicazione istituzionale volte alla promozione della consapevolezza del diritto alla protezione dei dati personali, sulla base di progetti previamente approvati dal Garante e che tengano conto della gravità della violazione. […]*"

[241] Garante per la Protezione dei Dati Personali- COMUNICATO STAMPA - IA: il Garante privacy chiede informazioni a DeepSeek. Possibile rischio per i dati di milioni di persone in Italia- 28 gennaio 2025 [Doc-Web n.10096856]

raccolti, la loro fonte, per quali finalità, la base giuridica del trattamento ed il luogo di conservazione dei dati. Il Garante ha, inoltre, specificamente chiesto, nel caso in cui i dati siano stati raccolti attraverso *web scraping*, di spiegare come gli utenti siano stati mesi al corrente sul trattamento dei loro dati. A seguito della risposta, ritenuta insufficiente nel contenuto, il Garante ha disposto, in data 30 gennaio 2025, la limitazione urgente ed immediata del trattamento dei dati degli utenti[242]. Nello specifico, il riscontro fornito da Deepseek in data 29 Gennaio 2025 ha declinato ogni dovere di risposta e collaborazione delle proprie società nei confronti del Garante, asserendo di non essere mai entrate nel mercato italiano (e non aver pianificato di farlo), e dunque, sostenendo l'inapplicabilità del GDPR in relazione alle attività di trattamento di dati personali da loro effettuate. Tuttavia, nello stesso riscontro pervenuto il 29 gennaio, le società hanno anche affermato di aver provveduto a rimuovere l'applicazione "DeepSeek" dagli *app store* italiani. La limitazione ordinata dal Garante è stata disposta, quindi, in conseguenza di un clima di totale assenza di collaborazione da parte di DeepSeek, in violazione dell'articolo 31 del GDPR[243], e stanti le affermazioni contraddittorie delle società riguardo la loro effettiva entrata nel mercato italiano, in violazione dell'Art. 3 (2) lett. a) del GDPR che prevede lo specifico ambito di applicazione territoriale della normativa[244].

Se, quindi, è auspicabile che questa complessità trovi un modo per diventare spiegabile (con riferimento alla comprensibilità degli algoritmi) e rispettosa dei principi che governano il trattamento dei personali, parallelamente viene in risalto l'importanza che assumono le azioni a contrasto dello *scraping*, le quali, seppure non decisive, fungono da "tampone" in una situazione che attualmente è *in itinere*.

Dunque, è all'interno di una cornice ancora piuttosto sguarnita a livello di casi pratici che, nel 2022, la società americana di indagini di mercato *business to business* "NewtonX"[245], ha condotto una ricerca al fine di indagare sulle *best practices* "anti-*scraping*" adottate

[242] Garante per la Protezione dei Dati Personali-Provvedimento del 30 gennaio 2025- [doc. web n. 10098477]

[243] Articolo 31 GDPR: "Il titolare del trattamento, il responsabile del trattamento e, ove applicabile, il loro rappresentante cooperano, su richiesta, con l'autorità di controllo nell'esecuzione dei suoi compiti.".

[244] Per approfondire il concetto di territorialità si veda il Capitolo I, paragrafo 3.1.1

[245] NewtonX è una società multinazionale che opera nel commercio interaziendale di servizi di ricerche di mercato. Ha sede a New York. La società è stata fondata nel 2016 con l'obiettivo di sfruttare l'IA per creare una piattaforma in grado di fornire alle aziende clienti, dati utili sulla base dei quali prendere decisioni business-critical.
Fonte: AI Magazine- NewtonX Using AI to power business decisions- by Tilly Kenyon, 2021.
Reperibile su: https://aimagazine.com/ai-strategy/newtonx-using-ai-power-business-decisions. Ultima visita al sito in data 20 Gennaio 2025

dalle imprese operanti in ambiti in cui è connaturata la pubblicazione di ampie quantità di dati, anche personali. Proprio da questa ricerca, sono seguite, come in una reazione a catena, da parte di enti differenti e a livello internazionale una serie di *white papers* e dichiarazioni congiunte (tra cui la dichiarazione congiunta delle Autorità Garanti internazionali [246] di cui si è parlato nel capitolo primo[247], i *papers* della MUSA[248], la nota informativa rilasciata dal Garante a maggio 2024[249] o, l'ultimo in ordine di pubblicazione, è il *Concluding Statement* dell'Autorità Garante canadese OPC[250]) miranti sia a responsabilizzare sia ad offrire strumenti alle SMCs (*social media companies*) e alle piattaforme *web* che per loro natura ospitano in forma pubblicamente accessibile dati personali.

1. Che gli effetti dello *scraping* siano acuiti dallo sfumare della distinzione tra dati personali e non?

Prima di svolgere alcune considerazioni sulle cautele che possono contribuire a limitare lo *scraping* dei dati dal *web*, è utile fare una riflessione su quello che costituisce, per chi scrive, il fulcro della questione: ovvero la necessità di una tutela dei dati personali, giuridica e pratica, che si concretizzi anche, per quanto possibile, in un effettivo controllo, da parte degli interessati, sulla circolazione dei propri dati.

[246] Joint statement on data scraping and the protection of privacy -August 24, 2023- disponibile sul sito della Information Commissioner's Office (ICO):
Reperibile su: https://ico.org.uk/media/about-the-ico/documents/4026232/joint-statement-data-scraping-202308.pdf

[247] Si veda il Capitolo I, Paragrafo 3.

[248] La Mitigating Unauthorized Scraping Alliance (MUSA) è un'organizzazione indipendente che, attraverso la collaborazione tra esperti di settore, accademici, policymakes ed aziende, redige dossier in cui delinea alcune possibili azioni di contrasto non vincolanti e volontarie volte a rilevare, prevenire, mitigare lo scraping di dati non autorizzato su piattaforme e siti web pubblicamente accessibili. L'obiettivo dell'organizzazione è quello di produrre uno standard di settore generalizzato, e dunque mutuabile indipendentemente dalla collocazione geografica, in grado di far allineare le aziende nell'ostacolare lo scraping.

[249] Garante per la Protezione dei Dati Personali-Web scraping ed intelligenza artificiale generativa: nota informativa e possibili azioni di contrasto- Provvedimento del 20 maggio 2024- [doc. web n. 10020316]

[250] Concluding joint statement on data scraping and the protection of privacy- Ottobre 2024. Consultabile sul sito dell'OPC: https://www.priv.gc.ca/en/opc-news/speeches-and-statements/2024/js-dc_20241028/#fn3

Perché, forse, la complessità del fenomeno e l'importanza dell'impegno nel limitarlo con ogni strumento disponibile, sta proprio in una riflessione sulla personalità dei dati o, meglio, della loro suscettibilità di ricondurre ad una persona determinata.

Partendo dalla definizione di dato personale fornita dal GDPR all'articolo 4(1)[251], si considerano dati personali quelle informazioni che riferiscono direttamente o indirettamente ad una persona fisica, sia essa identificata o identificabile. Questa distinzione, tuttavia, è tanto fondamentale quanto ricca di sfaccettature, che il recente sviluppo di tecnologie *data driven* come la IA, ha ulteriormente arricchito di complessità. Lasciando da parte la questione della base giuridica del trattamento, ci si intende ora concentrare su di un altro aspetto: l'attuale livello di efficacia dello strumento dell'anonimizzazione sui *dataset* utilizzati per addestrare le IA generative. Come si è già avuto modo spiegare[252], lo *scraping* è un'attività di raccolta dati automatizzata ed indiscriminata, ciò implica che dai *bot web crawler* non viene effettuata nessuna valutazione sulla riconducibilità o meno a una persona fisica determinata dei dati raccolti, questa è un'operazione che verrà (auspicabilmente) effettuata in un secondo momento sul *dataset*, una volta composto e prima di essere utilizzato. Dunque, prima di utilizzare un *dataset*, ed indipendentemente dal fatto che presenti un rischio sistemico o meno, il fornitore deve provvedere a rendere i dati che utilizza il più possibile sicuri, in modo da scongiurare lesioni di diritti e libertà individuali. Questo avviene per mezzo di una serie di azioni come la pulizia, il filtraggio, l'eliminazione di distorsioni nel *set* di dati raccolti, ed anche attraverso le varie tecniche di anonimizzazione[253]. Tutte queste operazioni sono, nell'insieme, fondamentali per evitare che l'algoritmo, in un determinato ed imprevedibile momento del suo funzionamento, "rigurgiti" dei dati personali. Nonostante però l'importanza, a monte, della distinzione tra dati personali e non e, a valle, l'importanza di renderli anonimi quando utilizzati come materiale di sviluppo per la IA, lo stesso funzionamento algoritmico può rendere molto difficile distinguere in pratica tra le due categorie.

---

[251] Art 4(1) GDPR: *"«dato personale»: qualsiasi informazione riguardante una persona fisica identificata o identificabile («interessato»); si considera identificabile la persona fisica che può essere identificata, direttamente o indirettamente, con particolare riferimento a un identificativo come il nome, un numero di identificazione, dati relativi all'ubicazione, un identificativo online o a uno o più elementi caratteristici della sua identità fisica, fisiologica, genetica, psichica, economica, culturale o sociale;"*

[252] Si veda il Capitolo I, paragrafo 2.

[253] AI Act-Allegato XI- Sezioni 1 e 2

Il GDPR, nel Considerando 26[254], pone un criterio distintivo per valutare l'applicabilità dei principi di protezione dei dati, ovvero l'identificabilità di una persona attraverso quei dati. Dal Considerando 26 viene poi tracciata una linea di separazione tra dati pseudonimizzati (i quali, comunque, potrebbero essere attribuiti ad una persona fisica determinata tramite l'incrocio con altre informazioni) e quelli anonimi, i quali, a fine Considerando, vengono esclusi dal raggio d'azione della normativa. Proseguendo nella lettura del Considerando 26, per valutare l'identificabilità di una persona fisica, è necessario prendere in considerazione tutti i mezzi ragionevolmente suscettibili di essere utilizzati. Ciò include fattori oggettivi, come i costi ed il tempo necessari per l'identificazione, ma anche delle tecnologie che il momento storico mette a disposizione. Come accennato sopra, per espressa previsione del Considerando 26, i dati anonimi, o meglio - utilizzando le parole della normativa- "*i dati personali resi sufficientemente anonimi*" restano fuori dall'ambito applicativo del GDPR.

Dati i limiti tecnologici, insiti nelle varie tecniche di anonimizzazione, e date le crescenti abilità dei modelli di IA generativa, quanto possono essere ritenuti veramente resistenti alla re-identificazione i dati anonimizzati e, dunque, non più personali?[255]

Si noti che "identificato" non significa necessariamente "denominato". Può anche solo essere sufficiente poter stabilire una connessione affidabile tra dati specifici e un individuo noto[256].

---

[254]Considerando 26 GDPR: "*È auspicabile applicare i principi di protezione dei dati a tutte le informazioni relative a una persona fisica identificata o identificabile. I dati personali sottoposti a pseudonimizzazione, i quali potrebbero essere attribuiti a una persona fisica mediante l'utilizzo di ulteriori informazioni, dovrebbero essere considerati informazioni su una persona fisica identificabile. Per stabilire l'identificabilità di una persona è opportuno considerare tutti i mezzi, come l'individuazione, di cui il titolare del trattamento o un terzo può ragionevolmente avvalersi per identificare detta persona fisica direttamente o indirettamente. Per accertare la ragionevole probabilità di utilizzo dei mezzi per identificare la persona fisica, si dovrebbe prendere in considerazione l'insieme dei fattori obiettivi, tra cui i costi e il tempo necessario per l'identificazione, tenendo conto sia delle tecnologie disponibili al momento del trattamento, sia degli sviluppi tecnologici. I principi di protezione dei dati non dovrebbero pertanto applicarsi a informazioni anonime, vale a dire informazioni che non si riferiscono a una persona fisica identificata o identificabile o a dati personali resi sufficientemente anonimi da impedire o da non consentire più l'identificazione dell'interessato. Il presente regolamento non si applica pertanto al trattamento di tali informazioni anonime, anche per finalità statistiche o di ricerca."*

[255] Il Gruppo di Lavoro Art.29 nell'opinione WP 216 - 05/2014 on Anonymisation Techniques, fa un'analisi tecnica e della robustezza delle più diffuse tecniche di anonimizzazione quali: noise addition, permutation, differential privacy, aggregation, k-anonymity, l-diversity and t-closeness.

[256] Information Commissioner's Office, 'Anonymisation: Managing Data Protection Risk Code of Practice' (2012). Pg. 21

Su questo punto, il Gruppo di Lavoro Art. 29 si era espresso nel 2013, ancor prima di rilasciare l'opinione sulle tecniche di anonimizzazione[257]. In questa prima occasione aveva preconizzato il sempre più difficile raggiungimento di un soddisfacente grado di anonimizzazione dovuto ai progressi tecnologici nell'utilizzo dei dati e all'aumentare della loro quantità in modo spropositato. Aveva, inoltre, aggiunto che la re-identificazione consistesse in un'area già, allora, piuttosto grigia, all'interno della quale un titolare del trattamento potesse facilmente cadere nell'errore di ritenere che un *set* di dati fosse sufficientemente anonimizzato, quando, in realtà, una terza parte motivata sarebbe stata comunque in grado di identificare alcuni individui dalle informazioni rilasciate. Oggi, a più di un decennio di distanza "la terza parte motivata" cui faceva riferimento il Gruppo di Lavoro, possono essere considerati anche gli algoritmi, che attraverso la combinazione di vari *set* di dati, possono agevolmente identificare gli interessati sulla base di dati presumibilmente anonimizzati.

Come recente esempio che riguarda un modello generativo si può citare l'esito di una ricerca del 2019[258] in cui un modello generativo è stato in grado di stimare con precisione la probabilità che una persona specifica venisse re-identificata correttamente, anche in un *set* di dati fortemente incompleto. Utilizzando il modello, è emerso che il 99,98% degli interessati i cui dati erano stato anonimizzati, sono stati correttamente re-identificati utilizzando *set* con soli 15 attributi demografici. I risultati suggeriscono che sia irrealistico ritenere che *set* di dati anonimizzati, e anche pesantemente campionati, riescano a soddisfare i moderni *standard* di anonimizzazione cui fa riferimento il GDPR.

Come accennato, nel 2014 il Gruppo di lavoro Art. 29 si espresse su come valutare opportunamente il livello di resistenza delle tecniche di anonimizzazione usate, considerando tre criteri:

a) se è ancora possibile individuare una persona fisica;
b) se è ancora possibile effettuare un collegamento tra dati relativi ad una stessa persona fisica o gruppo di persone fisiche;
c) se è ancora possibile inferire delle informazioni relative ad una persona fisica.

[257] Gruppo di Lavoro Art.29 -WP 203- Opinion 03/2013 on purpose limitation

[258] Rocher, Luc, Julien M. Hendrickx, and Yves-Alexandre de Montjoye. "Estimating the Success of Re-Identifications in Incomplete Datasets using Generative Models." Nature Communications, vol. 10, no. 1, 2019

In caso di mancanza di riscontro per tutti e tre i punti, i dati possono essere considerati sufficientemente anonimizzati. In caso di riscontro, i criteri forniti individuano il titolare del trattamento, e quindi un'idea del livello di garanzia ottenibile con la tecnica in questione, messa alla prova con lo stato attuale della tecnologia.

Va notato che, mentre il Considerando 26 fa esplicito riferimento all' "individuazione", l'inferenza e la collegabilità sono approcci per giungere all'individuazione che sono stati presi in considerazione dal Gruppo di Lavoro Art. 29, ma non esplicitamente menzionati nel GDPR[259].

I criteri proposti dal Gruppo di Lavoro, rispecchiano, dunque, livelli di efficacia diversa dell'anonimizzazione.

Specificamente: l'individuazione si riferisce alla possibilità di isolare alcuni o tutti i dati che identificano un individuo nell'insieme di dati considerato[260].

La collegabilità denota il rischio generato dalla possibilità di ricondurre alla stessa persona fisica, o ad un gruppo di persone fisiche, due o più tipi di informazioni. Perciò, se è possibile stabilire (ad esempio mediante l'analisi di correlazione) che delle informazioni sono assegnabili ad uno stesso gruppo di individui, ma non è possibile procedere con la singola identificazione, allora la tecnica utilizzata fornisce solo resistenza contro l'individuazione ma non contro la collegabilità[261].

Infine, l'inferenza è realizzabile anche quando l'individuazione e la collegabilità non sono più possibili. L'inferenza è stata definita dal Gruppo di Lavoro come "*la possibilità di dedurre, con significativa probabilità, il valore di un attributo dai valori di un insieme di altri attributi*", dove, per chiarezza di vocabolario: un *set* di dati è composto da diversi record relativi ad individui, ovvero gli interessati. A sua volta, ogni *record* correlato ad un interessato è composto da un insieme di valori o "voci", che rispondono a diversi attributi (ad esempio, un valore può essere la cifra "2013", la quale risponde all'attributo "anno")[262].

Il Gruppo di Lavoro ha sottolineato che soddisfare i tre criteri sopra descritti è molto difficile, per via del fatto che l'anonimizzazione e la re-identificazione sono campi di

[259] Finck, Michèle, and Frank Pallas. "They Who must Not be identified—distinguishing Personal from Non-Personal Data Under the GDPR." International Data Privacy Law, vol. 10, no. 1, 2020. Pg 16.
[260] Gruppo di Lavoro Art.29 -WP 216 - Opinion 05/2014 on Anonymisation Techniques.-Pg 11.
[261] *Ibidem*
[262] *Ibidem*-Pg 12

ricerca attivi ed in continuo aggiornamento. Ciò viene poi confermato dall'analisi fatta delle tecniche più comunemente utilizzate, e dalla quale è emerso che ogni metodo lascia comunque un rischio residuo, e che quindi, è possibile trarne la maggior efficienza possibile, solo se la loro applicazione è progettata in modo appropriato e diversificato, basandosi, caso per caso, sul contesto e gli obiettivi tecnici.

Tornado a riallacciare il discorso all'iniziale questione dello sfumare della distinzione tra dati personali e non, con l'avvento di tecniche di analisi dei dati e *hardware* sempre più performanti, nonché della crescente produzione di dati, sta diventando sempre più semplice mettere in relazione i dati con le persone fisiche a cui appartengono[263].

Alcuni hanno osservato che la legge sulla protezione dei dati personali potrebbe di fatto diventare una legge a tutela di tutti i dati, poiché in un futuro prossimo, tutti i dati potrebbero essere considerati dati personali e quindi soggetti al GDPR[264]. Ciò può essere vero nella misura in cui il concetto di dati personali, che determina l'ambito di applicazione materiale della protezione dei dati, continua ad espandersi e ad applicarsi ad una gamma di situazioni sempre più vasta. E questo assume ancor più senso se si include nel ragionamento il concetto di "*datafication*", ossia il processo di trasformazione della vita umana in una fonte continua di dati[265]. Attraverso questo processo, ampi domini della vita umana sono diventati suscettibili di essere elaborati attraverso forme di analisi automatizzate su larga scala[266].

Alla luce dei progressi tecnologici che stiamo vivendo, e soprattutto alle crescenti capacità elaborative della IA, stabilire il rischio di re-identificazione risulta una sfida ancora aperta.

[263] Finck, Michèle, and Frank Pallas. "They Who must Not be identified—distinguishing Personal from Non-Personal Data Under the GDPR." International Data Privacy Law, vol. 10, no. 1, 2020, pp. 11-36

[264] Purtova, Nadezhda. "The Law of Everything. Broad Concept of Personal Data and Future of EU Data Protection Law." Law, Innovation and Technology., vol. 10, no. 1, 2018, pp. 40–81

[265] Mejias, Ulises A., and Nick Couldry. -"Datafication."- Internet Policy Review, vol. 8, no. 4, 2019, pp. 1-10

[266] Mayer-Schönberger, V., & Cukier, K.- Big Data: una rivoluzione che trasformerà il modo in cui viviamo, lavoriamo e pensiamo.- 2013- Pp. 78-94.

2. Il primo report sulle azioni a contrasto dello *scraping*: l'indagine condotta dalla società americana NewtonX

Nell'Ottobre 2022 la NewtonX ha condotto una ricerca[267] su un campione di 1300 esperti in gestione ed analisi di dati provenienti da imprese nei settori: IT, *social media*, servizi finanziari, *gaming*, *e-commerce*. Per far parte del campione sono state selezionate solo imprese con più di 50 dipendenti e aventi sede tra Stati Uniti, Regno Unito ed Unione Europea.

Nel *whitepaper*[268] in cui sono riportati in forma riassuntiva i risultati dell'indagine, è altresì specificato che il progetto è stato finanziato da Meta Inc. senza che questo però abbia avuto influenza sull'esito del lavoro.

L'obiettivo di questa ricerca è offrire una più chiara comprensione dei fattori da tenere in considerazione per predisporre efficacemente delle azioni a contrasto della raccolta dei dati pubblicamente accessibili sul *web*, compresa una panoramica sui differenti approcci adottati dalle varie categorie di imprese considerate, per favorire la consapevolezza e prevenzione degli agenti del settore.

L'indagine pone l'accento sulle piattaforme *web* che, nonostante la natura della loro attività implichi la pubblica fruibilità dei dati degli utenti, si occupano attivamente di realizzare un insieme di misure appositamente combinate e aggiornate per ostacolarne l'indebita raccolta.

Sebbene la ricerca di NewtonX prenda in esame anche quanto riportato da esperti appartenenti ad ordinamenti giuridici diversi da quello europeo, e ancor più da quello italiano, il suo contenuto ha una portata volutamente ampia e generalizzata tale da poter rispecchiare una realtà e delle azioni pratiche che vadano oltre le differenze nelle normative sulla tutela dei dati applicabili. Su questo punto, si precisa che nell'illustrare

---

[267] NewtonX Case Study- NewtonX Data Extraction Prevention Best Practices Study- Novembre 2022. Reperibile su: https://6114340.fs1.hubspotusercontent-na1.net/hubfs/6114340/Product%20Marketing/NewtonX%20Data%20Extraction%20Prevention%20Best%20Practices%20Study.pdf?utm_campaign=NP-EM-Data_Extraction_Prevention_Case_Study-BOFU-11-14-22&utm_source=data%20extraction%20prevention&utm_medium=pdf&utm_content=whitepaper . Ultima visita al sito in data 19 febbraio 2025.

[268] NewtonX Reports- NewtonX Data Extraction Prevention Whitepaper.pdf- dicembre 2022. Reperibile su:https://6114340.fs1.hubspotusercontent-na1.net/hubfs/6114340/Product%20Marketing/NewtonX%20Data%20Extraction%20Prevention%20Whitepaper.pdf. Ultima visita al sito in data 19 febbraio2025

gli accorgimenti e gli approcci proposti nel *report*, essi verranno anche opportunamente contestualizzati all'interno della cornice normativa del GDPR.

Il *report* si apre suggerendo, sulla base di quanto emerso dall'indagine condotta sul campione, che, per contrastare efficacemente lo *scraping*, è opportuno elaborare una strategia, un piano d'azione, che sia naturalmente commisurato ai rischi insiti nel trattamento e ai tipi di dati esposti. Questo passa come prima cosa, attraverso delle operazioni preliminari di natura organizzativa che abbiano lo scopo di mettere in luce i rischi che lo *scraping* su quei dati comporterebbe alla tutela dei diritti e delle libertà degli interessati. Da qui, l'importanza di sviluppare consapevolezza della qualità e quantità dei dati che una piattaforma rende pubblici; dopodiché, è utile quantificare l'allocazione di risorse che il giusto equilibrio di tecniche preventive e di rilevamento richiede; ed infine, disporre di procedure di revisione e condivisione contribuisce alla robustezza dell'impalcatura formata dalle misure tecniche approntate.

Riguardo la consapevolezza: per una piattaforma o sito *web* (*rectius* per il titolare del trattamento dei dati che vengono ivi pubblicati), è un passo primario e fondamentale essere consapevoli di quanti e quali tipi dati personali (foto, video, geolocalizzazione, email, *post* e commenti) vengono esposti attraverso i propri canali; questo permette, infatti, al titolare, di determinare, nel modo più appropriato, delle misure che tengano conto del livello di tutela adatto per dimostrare uno sforzo nel senso dell'*accountability*, come anche della necessità di evitare di aggravare l'esperienza dell'utente, ed operare un eccessivo trattamento di dati (ad esempio tramite oneri di registrazione ingiustificati).

Pare opportuno precisare che, la *user experience* è un aspetto che il titolare non dovrebbe porre sulla bilancia delle valutazioni da fare al momento di determinare delle tutele a favore di un equo trattamento e circolazione dei dati, perché equivarrebbe, in pratica, a declassare l'importanza del rispetto dei principi della normativa sulla tutela dei dati. Tuttavia, è emerso dallo studio[269] che una parte delle aziende campione usa come scriminante nella scelta delle cautele a contrasto dello *scraping*, l'arrecare il minimo disturbo possibile ad una scorrevole fruizione delle piattaforme/servizi.

Il report suggerisce un passo ulteriore nella consapevolezza dei dati trattati, ovvero quello di classificarli in base alla possibilità o meno di inferire, tramite la loro rielaborazione, informazioni aggiuntive sull'interessato. È più difficile (e probabilmente anche

[269] NewtonX Case Study- NewtonX Data Extraction Prevention Best Practices Study-Novembre 2022-Pg.5.

infruttuoso) voler prevenire dei rischi senza però conoscere bene l'oggetto da proteggere e le sue vulnerabilità. A tal proposito, come esemplificazione, può essere richiamato il caso di Clearview AI, di cui si è detto nel capitolo primo[270]. Nel caso richiamato, come affermato nel provvedimento del Garante[271], il trattamento posto in essere da Clearview consisteva nello *scraping* di immagini dal *web* e nella loro successiva rielaborazione ed analisi tramite *software*, per creare delle rappresentazioni vettoriali dei volti, al fine di renderli suscettibili di essere indicizzati e poi comparabili ad altre immagini. Tale processo di rielaborazione tecnica, è, quindi, in grado di rendere la fotografia di una persona, a tutti gli effetti, un dato biometrico ad essa relativo. C'è altresì da aggiungere che di ogni fotografia possono essere estratti i metadati associati, i quali dischiudono informazioni su data, ora e geolocalizzazione del luogo dove è stata scattata la foto.

Il ragionamento riguardo l'allocazione di risorse da destinare al giusto equilibrio di tecniche preventive e di rilevamento andrà condotto dal titolare muovendo dalle considerazioni precedentemente fatte in merito alle caratteristiche dei dati trattati ma anche alle dimensioni dell'impresa.

Tuttavia, nonostante alcune tecniche siano largamente diffuse e di relativa speditezza applicativa (come il blocco degli IP e l'utilizzo di *CAPCHA*), e possano dunque essere messe in campo senza far parte di una specifica strategia anti-*scraping* (quindi come misure tecniche isolate), è consigliabile elaborare un piano tecnico-organizzativo con l'ausilio di esperti e avvalersi di fornitori di servizi esterni specializzati per quelle operazioni che richiedono competenze particolari (come possono essere i sistemi di *age* o *identity verification*). Il 64% degli esperti partecipanti allo studio[272] ha riferito che, nel contesto della loro impresa di appartenenza, l'uso di risorse esterne specializzate, ha contribuito ad irrobustire a livello organizzativo le basi del piano.

Un aspetto altrettanto importante per l'efficacia organizzativa del piano è l'utilità di combinare fattore tecnico e fattore umano, in modo tale da migliorare e adattarlo ai cambiamenti, se necessario. In questo senso, le tecniche basate sull'automazione (ad esempio, le più diffuse ma non le più efficaci, sono il blocco dell'IP e degli *User-Agent* aggressivi o sconosciuti) risultano fondamentali per la prevenzione ed il rilevamento dello

[270] Capitolo I, paragrafo 3.1.1

[271] Ordinanza ingiunzione nei confronti di Clearview AI-Registro dei provvedimenti n. 50 del 10 febbraio 2022- [doc. web n. 9751362] - Pg. 14

[272] NewtonX Case Study- NewtonX Data Extraction Prevention Best Practices Study-Novembre 2022-Pg.7.

scraping, specialmente su larga scala. Il coinvolgimento del giudizio umano resta, però imprescindibile per controllare l'efficacia delle misure messe in atto, specialmente quelle di natura tecnica, nei confronti delle quali ben ci si potrebbe approcciare impostandole e, poi, dimenticando ogni manutenzione e aggiornamenti necessari a tenere il piano al passo con i rapidi sviluppi tecnologici.

L'ultimo punto chiave che un titolare del trattamento dovrebbe prendere in considerazione per concepire e strutturare una efficace strategia anti-*scraping*, è il mantenimento di documentazione aggiornata riguardo le misure prese ed il loro ruolo all'interno della coerenza del piano. Inoltre, risulta essere importante per il 54% [273] degli esperti intervistati, anche il fatto che le informazioni, e la documentazione relativa al piano, vengano condivise sia all'interno dell'impresa stessa, che anche con altre imprese, in un'ottica di scambio di idee per il perseguimento di un obiettivo comune.

3. Le azioni di contrasto a livello internazionale

Nel 2023 anche la *Mitigating Unauthorized Scraping Alliance* (MUSA) e l'ICO insieme ad altre undici autorità per la protezione dei dati di tutto il mondo, nella loro dichiarazione congiunta sullo *scraping* e la tutela dei dati, hanno contribuito ad ampliare i suggerimenti elaborati a partire dalla ricerca condotta dalla NewtonX.

Questi documenti sono stati elaborati in modo indipendente, e diretti a soggetti differenti. Infatti, il *Joint Statement* delle Autorità garanti internazionali[274] è specificamente diretto, da un lato, alle *social media companies* (SMC) per esortarle a tenere un comportamento più tutelativo dei dati degli utenti, dall'altro destina a questi ultimi un breve *vade-mecum* su come gestire in modo più consapevole le impostazioni privacy delle piattaforme che utilizzano. Il paper pubblicato dalla MUSA, invece, è rivolto alla generalità delle imprese che lavorino ed operino esponendo al pubblico dei dati personali[275].

[273] Ibidem. Pg. 8

[274] Joint statement on data scraping and the protection of privacy -August 24, 2023- disponibile sul sito della Information Commissioner's Office (ICO): https://ico.org.uk/media/about-the-ico/documents/4026232/joint-statement-data-scraping-202308.pdf

[275] *Mitigating Unauthorized Scraping Alliance* (MUSA)- Industry Practices to Mitigate Unauthorized Data Scraping- 30 Marzo 2023. Reperibile sul sito: https://antiscrapingalliance.org/industry-practices-to-mitigate-unauthorized-data-scraping/. Ultima visita al sito in data 19 febbraio 2025.

Nel 2024, a livello internazionale si è aggiunto il contributo dell'Autorità Garante canadese OPC[276] approvato da quindici membri dell'*International Enforcement Cooperation Working Group* (IEWG) della *Global Privacy Assembly*. La dichiarazione dell'OPC si è basata sugli approcci suggeriti dal *Joint Statement* delle Autorità garanti internazionali alle SMC e agli utenti, per poter selezionare delle tecniche adatte alle piccole e medie imprese (PMI). Le PMI, infatti, difficilmente dispongono delle stesse risorse finanziarie o possibilità tecniche delle s*ocial media companies* di rilevanza globale. Ciononostante, le PMI non possono considerarsi esonerate dalla responsabilità di proteggere i dati pubblicati sui propri siti *web* o piattaforme. Proprio per questo, è opportuno che anche le azioni di contrasto adottate dalle PMI (e non solo quelle delle SMC) consistano in combinazioni multilivello di controlli tecnici e procedurali; approccio che garantirebbe un grado di protezione rafforzato proprio in virtù della differenziazione nelle tecniche. Invero, strumenti come il rilevamento dei *bot*, il *rate limiting* e i CAPTCHA, sono tra le soluzioni più diffuse ed allo stesso tempo accessibili al *budget* generalmente più modesto delle PMI.

4. (Segue) E a livello nazionale

A livello nazionale, la nota informativa emanata dal Garante nel maggio 2024[277] si colloca, rispetto ai documenti poc'anzi illustrati, sia come un'importante contestualizzazione dello *scraping* all'interno della cornice giuridica tracciata dal GDPR, sia come approfondimento sulle potenzialità, e sui limiti, delle misure tecniche più utilizzate ed efficaci.

Tutti e quattro i documenti presentati, condividono lo scopo di delineare e promuovere l'adozione di pratiche di settore non vincolanti e volontarie, da parte di imprese che ospitino sulle proprie piattaforme online dati personali pubblicamente accessibili, volte a rilevare, prevenire e mitigare lo *scraping* non autorizzato su tali dati. È bene, tuttavia, tenere in mente che le procedure descritte, hanno la ragionevole capacità di mitigare il fenomeno, ma non di impedirlo del tutto. Questo, specialmente se si ragiona in ottica pro-

[276] Office of the privacy Commissioner of Canada.

[277] Garante per la Protezione dei Dati Personali-Web scraping ed intelligenza artificiale generativa: nota informativa e possibili azioni di contrasto- Provvedimento del 20 maggio 2024 - [doc. web n. 10020316]

futuro, è dovuto alla continua evoluzione delle tecnologie di *scraping* (a cui fa in ogni caso da contraltare lo sviluppo delle azioni di contrasto) e all'ineliminabile presenza dei dati personali dal mondo *web*.

Nelle righe che seguono verranno esaminate le misure selezionate dal Garante, e ricondotte ad una logica di attuazione del principio di *accountability*.

4.1. Creazione di aree riservate

Le aree riservate, all'interno di un sito *web*, sono sezioni di un sito a cui l'accesso è possibile solo tramite delle credenziali. Questo tipo di cautela tecnico - organizzativa offre il vantaggio di essere una misura implementabile con sforzi economici e capacità tecniche contenute. Difatti, in alternativa all'installazione di un *plugin* o all'utilizzo di piattaforme dedicate, è possibile optare per un *editor* di siti *web* che disponga di tale servizio.[278]

Stante il presupposto che l'addestramento della IA generativa necessita di vaste quantità di dati, e che, nell'attuale momento storico, l'attività di *scraping* garantisce una varietà ed una realisticità dei dati difficilmente ottenibile altrimenti, sottrarre i dati alla pubblica disponibilità, può indirettamente contribuire a schermarli dalla raccolta automatizzata.

Un esempio dell'utilità immediata di questo tipo di misura possono essere i *forum* o le *chat* per lo scambio tra utenti ospitate su siti o piattaforme ed accessibili solo previa registrazione. Questi, infatti, sono luoghi virtuali in cui è molto più probabile che le persone dischiudano, attraverso le proprie opinioni, informazioni personali e stati d'animo.

La creazione di aree riservate, come del resto una qualsiasi altra misura, per ottemperare al proprio scopo, richiede di essere manutenuta nel tempo: in questo caso è necessario un elevato impegno gestionale e robustezza nei sistemi di sicurezza.

Di contro, una cautela di tipo volontario, volta a perfezionare l'attuazione del principio di *accountability* da parte di un titolare del trattamento, non può comportare un trattamento

[278] IONOS- Come creare un sito web con un'area riservata o protetta da una password- Aprile 2022 Articolo reperibile su: https://www.ionos.it/digitalguide/siti-web/creare-siti/area-riservata-per-gli-utenti/. Ultimo accesso 19 febbraio 2025.

di dati eccessivo[279], in violazione del principio di minimizzazione di cui all'articolo 5(1)(c) del GDPR[280].

Seguendo il principio di minimizzazione dei dati, in combinazione con il principio di limitazione della finalità, i titolari del trattamento non sono quindi *ipso facto* autorizzati a trattare più dati personali di quanto non sia strettamente necessario per raggiungere la finalità dichiarata[281]. Ciò non significa, tuttavia, che il principio di minimizzazione dei dati preveda l'obbligo di ridurre al minimo assoluto il trattamento dei dati. Bensì, si riferisce a un obbligo di ridurre al minimo la raccolta dei dati a un livello adeguato alle finalità del trattamento[282]. Le tre dimensioni di cui art 5(1)(c), ovvero: adeguatezza, pertinenza e limitazione delle finalità dovrebbero quindi essere prese in considerazione nel loro connubio.

Sul punto della minimizzazione, è intervenuto anche lo studio condotto dal Parlamento Europeo riguardo l'impatto del GDPR sullo sviluppo della IA[283], tale studio ha collegato il concetto di minimizzazione a quello di proporzionalità. Ciò in virtù del fatto che la minimizzazione non esclude l'inclusione di ulteriori dati personali in un trattamento; questo, infatti, ben può accadere nella misura in cui l'aggiunta di tali dati apporti un beneficio, rispetto alle finalità del trattamento che superi i rischi aggiuntivi per gli interessati.

Tornando alla creazione di aree riservate, come si diceva, ciò non può concretizzarsi in una violazione del principio di minimizzazione, sottoponendo gli interessati ad oneri di

[279] Garante per la Protezione dei Dati Personali-Web scraping ed intelligenza artificiale generativa: nota informativa e possibili azioni di contrasto- Provvedimento del 20 maggio 2024- [doc. web n. 10020316]. Pg. 4

[280] Articolo 5(1)(c) GDPR: "1. I dati personali sono: […]
c) adeguati, pertinenti e limitati a quanto necessario rispetto alle finalità per le quali sono trattati. […]"

[281] Witt, Cornelius, and Jan De Bruyne. "The Interplay between Machine Learning and Data Minimization under the GDPR: The Case of Google's Topics API." International Data Privacy Law., vol. 13, no. 4, 2023. PG 288-289

[282] Voigt, Paul e Axel von dem Bussche. Il Regolamento generale sulla protezione dei dati (GDPR) dell'UE: una guida pratica. 2a ed., Springer, 2024. Pg 138-139

[283] European Parliament, 'The impact of the General Data Protection Regulation (GDPR) on artificial intelligence' (European Parliamentary Research Service, Scientific Foresight Unit-PE 641.530 – June 2020) Pg:47-48

registrazione ingiustificati[284]. In tal senso, si può menzionare una recente decisione[285] dell'Ufficio del *Data Protection Ombudsman* finlandese, il quale ha comminato una sanzione amministrativa di 856.000 euro ad una società di vendita al dettaglio di articoli di elettronica ed elettrodomestici, per non aver specificato il periodo di conservazione degli *account* dei clienti del proprio negozio online e per aver imposto la creazione di *account* per effettuare acquisti sul sito. Su questo aspetto, l'Autorità finlandese ha ritenuto che rendere la creazione di un *account* cliente un requisito per effettuare acquisti *online* e non aver definito in alcun modo il periodo di conservazione dei dati personali raccolti, non potesse essere giustificato dalla possibilità, per gli interessati, di richiedere la cancellazione dei loro dati in un secondo momento. Al punto 55 della decisione[286] l'autorità chiarisce che: "*Va notato che il principio di minimizzazione dei dati di cui all'articolo 5, paragrafo 1, lettera c) del GDPR presuppone inoltre che il titolare del trattamento garantisca che i dati personali siano conservati in una forma che consenta l'identificazione degli interessati per un arco di tempo non superiore a quello necessario per le finalità per le quali i dati personali sono trattati".*

Dati questi presupposti, la creazione di aree riservate è una misura che ha bisogno di essere ben calibrata nella sua realizzazione pratica e giustificata da una preventiva valutazione dei costi - benefici in termini di conformità al Regolamento sulla tutela dei dati personali.

---

[284]Garante per la Protezione dei Dati Personali-Web scraping ed intelligenza artificiale generativa: nota informativa e possibili azioni di contrasto- Provvedimento del 20 maggio 2024- [doc. web n. 10020316]. Pg. 4

[285]Office of the Data Protection Ombudsman- Administrative fine imposed on Verkkokauppa.com for failing to define storage period of customer data – requiring customers to register was also illegal.
Sommario della decisione (in Suomi) disponibile in inglese sul sito dell'Autorità e consultabile a: https://tietosuoja.fi/en/-/administrative-fine-imposed-on-verkkokauppa.com-for-failing-to-define-storage-period-of-customer-data-requiring-customers-to-register-was-also-illegal

[286] Tietosuojavaltuutetun toimiston -Tietosuojavaltuutetun ja seuraamuskollegion päätökset- 6.3.2024 TSV/26/2020. Pg. 10
Reperibile su:
https://tietosuoja.fi/documents/6927448/204092115/P%C3%A4%C3%A4t%C3%B6s+TSV.26.2020.pdf/cc31f8b8-a4ec-e622-501d-6b0e2e1a53ca/P%C3%A4%C3%A4t%C3%B6s+TSV.26.2020.pdf?t=1710776065426s

4.2 Inserimento di clausole *ad hoc* nei termini di servizio

Il *web scraping* è un'attività capace di ripercuotersi ed innescare conseguenze anche in ambiti giudici ulteriori a quello della tutela dati, quale ad esempio il diritto d'autore, o, in relazione agli eventuali Termini di Servizio (ToS) adottati dalla piattaforma oggetto di *scraping*, anche il diritto contrattuale.

In principio, occorre chiarire che i ToS non si presentano tutti alla stessa maniera, ma la loro presentazione può essere di due tipi principali:

a) I ToS "*Clickwrap*"[287] richiedono l'accettazione esplicita da parte dell'utente, in genere tramite un *clic* su un pulsante o una casella da spuntare per confermare l'accettazione. Il *clickwrap* è un modo di stipulare il contratto dei termini di servizio sempre più utilizzato ed in genere preferito per il ruolo attivo che l'utente riveste accettando esplicitamente il contratto che gli viene presentato nella finestra *pop-up* sullo schermo.

b) I ToS *"Browsewrap"* invece sono contratti passivi e non richiedono il consenso esplicito dell'utente. In questo caso il consenso rimane implicito nel comportamento e nella permanenza dell'utente sul sito. Dunque, la vincolatività del contratto si basa sul presupposto che il prosieguo della navigazione sul sito costituisca accettazione implicita dei termini[288]. In presenza di Tos *browsewrap* generalmente viene visualizzato un *banner* in cui si avvisa l'utente che proseguendo nella navigazione, si intendono accettati i termini di servizio. In caso di necessità di consultazione, nella norma, il testo dei ToS è consultabile attraverso il menù o nel piè di pagina.

Nel senso della piena rispettabilità delle clausole contenute nei termini di servizio di una piattaforma, si espresse la CJEU nel caso Ryanair vs. PR Aviation del 2015[289], a seguito di rinvio pregiudiziale da parte della Corte Suprema dei Paesi Bassi. In quell'occasione, tra le varie questioni, alla CJEU fu posta anche quella di stabilire se il divieto contenuto

---

[287] Jordan Yerman- Is Web Scraping Legal? Navigating Terms of Service and Best Practices- *Ethical Web Data Collection Initiative (EWDCI)*, 7 Jan 2025- Reperibile su: https://ethicalwebdata.com/is-web-scraping-legal-navigating-terms-of-service-and-best-practices/. Ultimo accesso 18 febbraio 2025.

[288] Minucci G. e Lombardi G.- Web scraping: accordi browse wrap e tutele disponibili- Lexia- ottobre 2024.- Reperibile su: https://www.lexia.it/2024/10/31/web-scraping/. Ultimo accesso 1 Febbraio 2025.

[289] CJEU-(Case C-30/14)
Judgment of the Court (Second Chamber) of 15 January 2015 (request for a preliminary ruling from the Hoge Raad der Nederlanden — Netherlands) — Ryanair Ltd v PR Aviation BV
(Reference for a preliminary ruling — Directive 96/9/EC — Legal protection of databases — Database not protected by copyright or the sui generis right — Contractual limitation on the rights of users of the database)

nei ToS del sito di *Ryanair*[290] di utilizzare qualsiasi *software* o sistema automatico per estrarre (*scraping*) dati dal sito e dalle banche dati in esso contenute a fini commerciali, fosse legittimo ai sensi della Direttiva 96/9/CE a tutela dei database[291]. La risposta della Corte si basò sulla premessa che l'insieme di dati disponibili sul sito di Ryanair non costituissero una banca dati, né protetta dal diritto d'autore né dal diritto sui generis e che, quindi, la Direttiva[292] non potesse essere invocata. Pertanto, secondo la Corte, l'autore di una banca dati non ricompresa tra quelle nello scopo della Direttiva 96/9/CE, può legittimamente tutelare i dati presenti sulla propria piattaforma, invocando una protezione su base contrattuale[293].

L'inserimento di apposite clausole che vietano lo *scraping* nei termini di servizio della piattaforma o del sito, dunque, costituisce una cautela di natura puramente giuridica, operante *ex post* per quel che riguarda le conseguenze della loro violazione: difatti, in caso di mancato rispetto, legittima i gestori delle piattaforme ad agire in giudizio per far dichiarare l'inadempimento contrattuale della controparte; ma operante anche *ex ante* fungendo da deterrente.

4.3 Monitoraggio del traffico di rete

Tra gli accorgimenti di tipo puramente tecnico, implementabili per rilevare flussi anomali di dati su un sito od una piattaforma, c'è il monitoraggio del traffico di rete. Quest'operazione viene condotta utilizzando un *software* apposito in grado di controllare lo stato di una rete e generare delle mappe topologiche che permettono una localizzazione esatta dei problemi sulla base delle informazioni raccolte ed analizzate.

---

[290] Termini d'uso | Ryanair.com
Reperibili su sito dell'azienda: https://www.ryanair.com/it/it/azienda/termini-d-uso. Ultima visita al sito in data 19 febbraio 2025.

[291] Direttiva 96/9/Ce Del Parlamento Europeo E Del Consiglio dell'11 marzo 1996 relativa alla tutela giuridica delle banche di dati

[292] La questione verteva sul combinato disposto degli articoli 6(1),8,15 della Direttiva 96/9/CE

[293] Dispositivo della sentenza caso -(Case C-30/14): "La direttiva 96/9/CE del Parlamento europeo e del Consiglio, dell'11 marzo 1996, relativa alla tutela giuridica delle banche di dati, dev'essere interpretata nel senso che essa non è applicabile a una banca dati non tutelata né dal diritto d'autore né dal diritto sui generis ai sensi di tale direttiva, con la conseguenza che gli articoli 6, paragrafo 1, 8 e 15 della direttiva medesima non ostano a che il creatore di una banca dati siffatta stabilisca limitazioni contrattuali all'utilizzo della stessa da parte dei terzi, fatto salvo il diritto nazionale applicabile."
Reperibile su: https://curia.europa.eu/juris/document/document.jsf?docid=162299&doclang=IT. Ultimo accesso al sito in data 19 febbraio 2025.

Grazie alla mappatura generata dal *software*, è possibile avere una panoramica esaustiva, sullo stato dei componenti di rete, sulle prestazioni, sulle risorse sovraccariche ed è in grado, di conseguenza, di ottimizzare il flusso dei dati. Uno dei metodi impiegati per monitorare e risolvere eventuali anomalie consiste nell'impostazione di soglie specifiche di fruizione di quella reta da parte da parte di un singolo utente. L'utente riceve, quindi, avvisi istantanei ogni volta che tale soglia viene superata. Alcune soglie sono tipo statico, altre, tra cui i moderni sistemi di monitoraggio, utilizzano il *Machine Learning* per determinare la normale prestazione di tutte le metriche di una rete in base all'orario e al giorno della settimana[294].

Il vantaggio principale che gli strumenti di monitoraggio danno è l'avere una visione diretta dei dispositivi connessi ad una rete, e del modo in cui alterano il flusso dei dati.

Una tutela che può, logicamente, accompagnare il monitoraggio della rete è il *Rate Limiting*, ovvero la limitazione del traffico di rete anomalo. La limitazione della frequenza, anche se non risolutiva nella gestione dell'attività dei bot, può tuttavia, aiutare a ridurre alcuni loro tipi di attività dannose, tra cui anche lo *scraping* di contenuti.

La limitazione della velocità viene eseguita all'interno di un'applicazione, anziché sul *server web* stesso. In genere, la limitazione della velocità si basa sul rilevamento degli indirizzi IP (l'indirizzo IP è il modo principale in cui un'applicazione identifica chi o cosa sta effettuando la richiesta) da cui provengono le richieste e del tempo trascorso tra ogni richiesta da loro effettuata. Se ci sono troppe richieste da un singolo IP entro un determinato periodo di tempo, la soluzione di limitazione della velocità non soddisferà le richieste dell'indirizzo IP per un certo periodo di tempo.[295]

Monitoraggio e limitazione della velocità sono tuttavia delle azioni efficaci contro determinati tipi di *bot* specifici (ad esempio quelli responsabili di attacchi *DDoS*), perciò, nei restanti casi, rischiano di rimanere nella sfera del solo rilevamento della problematica, mancando di risolutività. Proprio per questa ragione, si rendono necessari anche sistemi di gestione dei *bot*, i quali siano in grado approcciare in modo completo la loro attività,

[294] IBM- Cos'è il monitoraggio della rete? (2024)
Reperibile su: https://www.ibm.com/it-it/topics/network-monitoring. Ultimo accesso in data 3 febbraio 2025

[295] Cloudflare- What is Rate Limiting. Reperibile su: https://www.cloudflare.com/learning/bots/what-is-rate-limiting/. Ultimo accesso al sito in data 3 febbraio 2025.

arrivando a bloccarla, finanche identificando i *bot* probabilmente dannosi (questo è possibile per i *software* di ultima generazione che utilizzano il *machine learning*).
Da quanto detto fin qui, risulta chiaro come le misure descritte, se prese singolarmente, abbiano un impatto irrisorio in termini di efficacia di contrasto e che la loro forza stia nell'azione congiunta di più misure attive contemporaneamente. Perciò, al problema dei *bot* che, a gran velocità, navigano nelle diverse aree di siti e piattaforme per raccogliere qualsiasi tipo di contenuto, si possono opporre delle misure che agendo di concerto sono in grado di colpire le diverse dimensioni del problema: rilevamento, rallentamento, blocco del traffico anomalo ed individuazione dei possibili agenti dannosi.[296]

4.4 Intervento sui bot

Avendo a questo punto compreso che gli esecutori materiali dello *scraping* sono i *bot*, la maggior parte delle misure tecniche che riescano ad ostacolare il loro accesso a siti e piattaforme, si rivela essere un passo in più verso una protezione maggiormente completa.
I mezzi per intervenire sui *bot* sono molteplici ed in continuo aggiornamento, così come del resto lo sono i modi per aggirare le protezioni. Di seguito verranno menzionate, a titolo esemplificativo, alcune soluzioni tra le più diffuse e ritenute utili.

a) Le verifiche CAPTCHA e reCAPTCHA

Le verifiche CAPTCHA sono test progettati per determinare se un utente online è davvero un essere umano oppure un *bot*. La parola CAPTCHA è un acronimo che sta per "*Completely Automated Public Turing test to tell Computers and Humans Apart*".
I CAPTCHA classici, che sono tuttora in uso, prevedono l'identificazione da parte degli utenti di alcune lettere. Le lettere in questione sono appositamente distorte in modo tale che i *bot* non siano in grado di riconoscerle. Per superare il test, gli utenti devono interpretare il testo distorto, digitando le lettere corrette in un campo del modulo e poi inviare. Se quanto digitato è errato, agli utenti viene chiesto di riprovare. Tali test sono comuni, ad esempio, nei moduli di accesso, in quelli di registrazione dell'*account* o nei sondaggi online. Tuttavia, il progresso tecnologico ha portato sul mercato *bot* che usano

[296] *Ibidem.*

il *Machine Learning* per riconoscere i testi distorti; questo ha determinato la necessità di rendere più complessi i test esistenti.

La tecnologia reCAPTCHA è stata sviluppata dai ricercatori della *Carnegie Mellon University* di Pittsburgh, poi acquisita da Google nel 2009. Oggi è un servizio gratuito che Google offre in sostituzione dei CAPTCHA classici. Nel corso del tempo, Google ha ampliato le funzionalità dei test reCAPTCHA inserendo il test di *Turing* avanzato, basato sull'intelligenza artificiale per distinguere il comportamento umano da quello di un *bot*[297]. Le funzionalità e i tipi di questi test si sono quindi ampliate, venendo a ricomprendere anche: il riconoscimento delle immagini, la spunta di caselle di controllo e la valutazione generale del comportamento dell'utente durante la navigazione del sito. In quest'ultimo caso, l'interazione del programma con l'utente è del tutto assente: l'algoritmo sul quale è basato il *bot*, è programmato per interagire con il *back-end* e la pagina *web* del *client* per attivare una sequenza di eventi di autenticazione del comportamento dell'utente sul sito (tra cui i movimenti del cursore del *mouse* ed il tempo trascorso sulla singola sezione del sito), per poi giungere all'assegnazione, a ciascun visitatore, di un punteggio[298] da 0,0 a 1,0. Dove: 0,0 indica che l'interazione è ad alto rischio e potrebbe celare un *bot*, mentre 1,0 indica che l'interazione è a basso rischio e dunque, molto probabilmente umana[299].

b) Modifica periodica del markup HTML

Modificare gli elementi del markup può contribuire a rendere meno agevole lo *scraping* da parte dei *bot.*

c) Incorporazione di dati

Incorporare dei dati solitamente costituiti da breve testo, come numeri di telefono od *e-mail,* all'interno di oggetti multimediali, può rendere più complesso il rilevamento dei dati.

---

[297] De santis Luca- ReCAPTCHA: scopri cos'è, come funziona e implementarlo sul tuo sito -*Doctor web Agency*- ottobre 2024. Reperibile su: https://www.doctor-web.it/recaptcha-cose-come-funziona/ Ultima visita al sito in data 3 Febbraio 2025.

[298] Google for Developers- reCAPTCHA V3- ottobre 2024. Reperibile su: https://developers.google.com/recaptcha/docs/v3?hl=it#:~:text=reCAPTCHA%20v3%20restituisce%20un%20punteggio%20per%20ogni%20richiesta,chiavi%20reCAPTCHA%20v3%20nella%20Console%20di%20amministrazione%20reCAPTCHA. Ultima visita al sito in data 19 febbraio 2025

[299] Cloudflare-How CAPTCHAs work | What does CAPTCHA mean? | Cloudflare. Reperibile su: https://www.cloudflare.com/learning/bots/how-captchas-work/ -Ultimo accesso al sito in data 3 febbraio 2025.

d) Rendering della pagina web

Il *rendering* è il processo attraverso il quale un *browser* interpreta e visualizza il codice sorgente di una pagina *web* allo scopo di generare la versione visuale e interattiva che gli utenti vedono e con cui interagiscono[300]. In altre parole, è la resa del codice sorgente in una rappresentazione grafica e funzionale.

Il *rendering* dinamico dei contenuti, che genera e carica dinamicamente tramite *JavaScript* il contenuto delle pagine *web*, aggiunge complessità e maggiore lentezza nella fruizione del sito da parte dei *bot* e rende obsoleti i *bot* tradizionali basati sull'analisi HTML statica[301].

e) Intervento sul file robot.txt.

Il *robots.txt* è un file di testo che contiene delle direttive volte a gestire il traffico dei *crawler* dei motori di ricerca. Sostanzialmente, comunica ai *bot* quali contenuti di un sito dovrebbero essere sottratti all'indicizzazione[302]. Il file *robots.txt.* fa parte di un protocollo basato sulla conformità volontaria chiamato *Robot Exclusion Protocol* (abbreviato in REP). Il protocollo è costituito da regole, tra le quali quelle volte a consentire "*allow*" o non consentire "*disallow*" l'accesso di determinati *bot* ad un URI[303]. L'efficacia di questo protocollo è tuttavia limitata ai *bot* che contengano specificamente all'interno del loro "programma", la finalità di *scraping* per sviluppo della IA generativa. Restando quindi esclusi i *bot* di cui gli sviluppatori non hanno condiviso i dettagli tecnici. Infine, è importante ribadire che la volontarietà di adesione degli *scraper* a detto protocollo, che non costituisce quindi uno *standard* riconosciuto, ne limita ulteriormente l'efficacia.

---

[300] Sacheli Giovanni-Come analizzare il rendering di un sito web da parte di Google- EVE Milano- Gennaio 2025. Reperibile su: https://www.evemilano.com/blog/analisi-rendering/- Ultima visita al sito in data 4 febbraio 2025.

[301]Abbas Assad- Keeping Data Safe: How to Counter Web Scraping Attacks- Techopedia- Agosto 2023. Reperibile su: https://www.techopedia.com/keeping-data-safe-how-to-counter-web-scraping-attacks - Ultima visita al sito in data 4 febbraio 2025

[302] Koster M. et Al, Internet Engineering Task Force (IETF)- RFC 9309 Robots Exclusion Protocol- *RFC Editor*- settembre 2022.
Reperibile su: RFC 9309: Robots Exclusion Protocol. Ultima visita al sito in data 19 febbraio 2025

[303] Lo *Uniform Resource Identifier* è una sequenza di caratteri che identifica una risorsa presente nel web (sia essa una pagina web, immagini o documenti).
myPOS-Cos'è l'URI: definizione, esempi e differenze con URL e URN | myPOS-Ottobre 2024. Reperibile su: https://www.mypos.com/it-it/blog/suggerimenti/cos-e-uri-definizione-esempi. Ultimo accesso al sito in data 4 febbraio 2025.

4.5 Conclusioni sull'agire volontario nella dimostrazione dell'*accountability*

Le misure, appena discusse, si collocano nell'alveo delle varie possibili azioni considerabili da un titolare per garantire ed essere in grado di dimostrare che il trattamento da lui effettuato sia conforme al GDPR in relazione alle attività di *scraping*.
A parere di chi scrive, tali misure costituiscono, per ora, una solo "fotografia" dello sforzo massimo attualmente possibile di dimostrare un'*accountability* robusta da parte del titolare.
In quest'ottica, il principio di *accountability* prende la forma di segnali luminosi che tracciano la strada verso l'evoluzione di un diritto alla tutela dei dati che non subisca gli effetti del progresso tecnologico, bensì che ne sia componente attiva.

CONCLUSIONI

L'idea di questo elaborato muove dall'interesse nel mettere in luce che, al netto degli innegabili benefici per la collettività, di cui è foriera l'intelligenza artificiale, la questione del suo sviluppo merita un'attenzione particolare per l'impatto che questa tecnologia ha sul rispetto del diritto alla tutela dei dati personali degli interessati. Nello specifico, attraverso le considerazioni fatte ed i documenti esaminati, si è inteso dimostrare che, il proteggere dal *web scraping* i dati resi pubblici, non consta meramente di una serie di azioni "tecniche" volte ad ostacolarlo, bensì rientra nel più ampio tracciato della conformità al principio di *accountability*, che dovrebbe guidare l'azione del titolare del trattamento.
Per lo sviluppo del lavoro di tesi, si è scelto di incentrare la discussione attorno ad un tipo di intelligenza artificiale che, negli ultimi anni ha avuto gran diffusione ed è riuscita a farsi strada nella quotidianità delle persone grazie alla versatilità delle sue prestazioni, ovvero: l'intelligenza artificiale generativa.
A questo proposito, si è reso opportuno, nel Capitolo I, analizzare il fenomeno del *web scraping* e le sue immediate interazioni con la disciplina sulla tutela dei dati personali di cui al GDPR.

Dalle riflessioni svolte su questa normativa[304] è emerso il ruolo centrale ricoperto dal titolare del trattamento. Per l'appunto, il titolare è un soggetto che, partendo dalle linee essenziali del suo ruolo tracciate dal GDPR[305], è poi anche il fulcro dell'attuazione del principio di *accountability.* Infatti, grazie alle scelte da lui operate, in virtù della discrezione consentita dalla normativa[306], tale principio ha potenzialità di "estendersi" e concretizzarsi in misure di tutela fortemente legate alle necessità del caso. La considerazione di questo "spiraglio di libertà" è proprio quel che ha permesso di giungere, nel Capitolo IV, ad apprezzare l'opportunità del contenuto della nota informativa emanata dal Garante nel maggio 2024[307] ed a coglierne, aldilà delle misure pratiche consigliate a contrasto dello *scraping*, anche la portata esortativa diretta nei confronti dei titolari del trattamento, ad approcciare il proprio dovere di responsabilizzazione in modo proattivo ed aggiornato. Solo in questo modo potrà costituire un agire veramente orientato all'efficacia, teso alla massima realizzazione del principio di *accountability* ed in grado di tenere il passo all'evoluzione tecnologica che permette lo sviluppo di *software* per la raccolta dati sempre più sofisticati (e potenziati da IA).

Per meglio carpire il valore dell'approccio olistico in attuazione del principio di *accountability* suggerito, a più riprese, dal GDPR, è opportuno richiamare quanto affrontato nei Capitoli II e III.

Infatti, nonostante le due normative principali considerate (GDPR e AI Act) predispongano, per differenti scopi, vari singoli strumenti (sia obbligatori che facoltativi), tra cui ad esempio: da un lato la DPIA ex art. 35 GDPR e, dall'altro la valutazione d'impatto sui diritti fondamentali ex art. 27 AI Act al fine di individuare trattamenti e sistemi dall'impatto potenzialmente rischioso[308]; i codici di condotta ex Art. 56 AI Act[309] con cui i fornitori di GPAI fissano volontariamente i propri *standard* di sicurezza, trasparenza e gestione del rischio; Il DPO ex artt. 37-39 GDPR, il cui atto di nomina

[304] Capitolo I, paragrafo 3.

[305] Art. 4 (7) GDPR: *"«titolare del trattamento»: la persona fisica o giuridica, l'autorità pubblica, il servizio o altro organismo che, singolarmente o insieme ad altri, determina le finalità e i mezzi del trattamento di dati personali. quando le finalità e i mezzi di tale trattamento sono determinati dal diritto dell'Unione o degli Stati membri, il titolare del trattamento o i criteri specifici applicabili alla sua designazione possono essere stabiliti dal diritto dell'Unione o degli Stati membri*".

[306] Capitolo I, paragrafo 3.2.1.

[307] Garante per la Protezione dei Dati Personali-Web scraping ed intelligenza artificiale generativa: nota informativa e possibili azioni di contrasto- Provvedimento del 20 maggio 2024- [doc. web n. 10020316]

[308] Capitolo II, paragrafi 2.1 . Capitolo III, paragrafo 2.2.

[309] Capitolo II, paragrafo 3.2.

facoltativa da parte del titolare del trattamento (o anche da parte del responsabile)[310] è espressione del suo impegno nel cercare conformità sia con le disposizioni del GDPR che con la normativa nazionale sulla tutela dati (D.lgs. 30 giugno 2003, n. 196); e considerando anche gli strumenti di origine non-normativa illustrati nel Capitolo IV[311], non c'è (per ora) una unica soluzione adatta od un unico strumento in grado di contrastare il *web scraping* finalizzato all'addestramento di sistemi di IA generativa.

Questa consapevolezza, dell'insufficienza del singolo strumento, lascia tuttavia spazio per apprezzare l'effetto di un approccio di tipo olistico[312], in cui il risultato, dato dall'unione di misure orientate a contrastare aspetti differenti del problema, restituisce un valore aggiunto in termini di risultato, rispetto alla semplice somma delle singole misure. È opportuno quindi, che i gestori di siti *web* e piattaforme che rivestono anche la posizione di titolari del trattamento di dati personali, resi pubblici *online* per finalità diverse, e sulla base di differenti condizioni di legittimità, valutino, in relazione al proprio caso, un pacchetto di misure che seppur non totalmente esaustive, riescano a contenere gli effetti dello *scraping* proprio in virtù della loro azione combinata.

[310] Capitolo III, paragrafo 3
[311] Capitolo IV, Paragrafi 2,3, e 4
[312] Enciclopedia treccani Online- "Olismo
Tesi secondo cui il tutto è più della somma delle parti di cui è composto […]".

BIBLIOGRAFIA

Legislazione

- Carta Dei Diritti Fondamentali Dell'unione Europea del Dicembre 2000 (GU C 202 del 7 Giugno 2016)
- Commissione Europea- Proposta Di Regolamento Del Parlamento Europeo E Del Consiglio Che Stabilisce Regole Armonizzate Sull'intelligenza Artificiale (Legge Sull'intelligenza Artificiale) E Modifica Alcuni Atti Legislativi Dell'unione.- Bruxelles, 21.4.2021 COM(2021) 206 final 2021/0106 (COD)
- D.lgs. 30 giugno 2003, n. 196 recante "Codice in materia di protezione dei dati personali, disposizioni per l'adeguamento dell'ordinamento nazionale al Regolamento (UE) 2016/679 del Parlamento europeo e del Consiglio, del 27 aprile 2016, relativo alla protezione delle persone fisiche con riguardo al trattamento dei dati personali, nonché alla libera circolazione di tali dati e che abroga la Direttiva 95/46/CE
- Direttiva (Ue) 2016/680 Del Parlamento Europeo E Del Consiglio del 27 aprile 2016 relativa alla protezione delle persone fisiche con riguardo al trattamento dei dati personali da parte delle autorità competenti a fini di prevenzione, indagine, accertamento e perseguimento di reati o esecuzione di sanzioni penali, nonché alla libera circolazione di tali dati e che abroga la decisione quadro 2008/977/GAI del Consiglio
- Direttiva 96/9/Ce Del Parlamento Europeo E Del Consiglio dell'11 marzo 1996 relativa alla tutela giuridica delle banche di dati
- Parlamento europeo- (Reperibile su) Testi approvati - Implicazioni dei Big Data in termini di diritti fondamentali: privacy, protezione dei dati, non discriminazione, sicurezza e attività di contrasto (2016/2225(INI)) , Martedì 14 marzo 2017
- Regolamento (UE) 2016/679 del Parlamento europeo e del Consiglio, del 27 aprile 2016, relativo alla protezione delle persone fisiche con riguardo al trattamento dei dati personali, nonché alla libera circolazione di tali dati e che abroga la direttiva 95/46/CE (regolamento generale sulla protezione dei dati)
- Regolamento (UE) 2024/1689 del Parlamento europeo e del Consiglio, del 13 giugno 2024, che stabilisce regole armonizzate sull'intelligenza artificiale e modifica i regolamenti (CE) n, 300/2008, (UE) n, 167/2013, (UE) n, 168/2013, (UE) 2018/858, (UE) 2018/1139 e (UE) 2019/2144 e le direttive 2014/90/UE, (UE) 2016/797 e (UE) 2020/1828 (regolamento sull'intelligenza artificiale)
- Senato della Repubblica XIX Legislatura -Fascicolo Iter DDL S. 1146 -Disposizioni e delega al Governo in materia di intelligenza artificiale
- Trattato sul funzionamento dell'Unione europea del 13 dicembre 2007 - versione consolidata (GU C 202 del 7.6.2016)

Soft Law

- Commissione Europea Bruxelles, - Libro Bianco sull'Intelligenza Artificiale - Un approccio europeo all'eccellenza e alla fiducia1-9.2.2020 COM(2020) 65 final
- Commissione Europea- European AI Alliance-A Practical Organizational Framework for AI Accountability- Ottobre 2023
- Commissione Europea-Comunicazione Della Commissione Al Parlamento Europeo, Al Consiglio, Al Comitato Economico E Sociale Europeo E Al Comitato Delle Regioni. L'intelligenza artificiale per l'Europa-COM(2018) 237 final. Bruxelles 25 aprile 2018
- EDPB- Linee guida 8/2020 sul targeting degli utenti di social media- Versione 2.0 - Adottate il 13 aprile 2021
- EDPB, Report of the work undertaken by the ChatGPT Taskforce, 23 May 2024.
- EDPB-EDPS Parere congiunto 5/2021 sulla proposta di regolamento del Parlamento europeo e del Consiglio che stabilisce regole armonizzate sull'intelligenza artificiale (legge sull'intelligenza artificiale) -18 giugno 2021
- EDPB-Guidelines 1/2024 on processing of personal data based on Article 6(1)(f) GDPR -Version 1.0 -Adopted on 8 October 2024
- EDPB-Linee guida 3/2019 sul trattamento dei dati personali attraverso dispositivi video Versione 2.0 Adottate il 29 gennaio 2020
- EDPB-Linee guida 4/2019 sull'articolo 25 Protezione dei dati fin dalla progettazione e per impostazione predefinita Versione 2.0 Adottate il 20 ottobre 2020
- EDPB-Opinion 28/2024 on certain data protection aspects related to the processing of personal data in the context of AI models -Adopted on 17 December 2024
- EDPS-Generative AI and the EUDPR. First EDPS Orientations for ensuring data protection compliance when using Generative AI (2024).
- European Parliament, 'The impact of the General Data Protection Regulation (GDPR) on artificial intelligence' (European Parliamentary Research Service, Scientific Foresight Unit)-PE 641.530 – June 2020
- European Union Agency for Fundamental Rights (FRA),-Your Rights Matter: Data Protection And Privacy- Fundamental Rights Survey (2020)
- Gruppo Di Lavoro Articolo 29 Per La Protezione Dei Dati - 00569/13/EN WP 203 - Opinion 03/2013 on purpose limitation- Adopted on 2 April 2013
- Gruppo Di Lavoro Articolo 29 Per La Protezione Dei Dati - 02356/09/EN WP 168- The Future of Privacy contribution to the Consultation of the European Commission on the legal framework for the fundamental right to protection of personal data - Adopted on 01 December 2009
- Gruppo Di Lavoro Articolo 29 Per La Protezione Dei Dati - 0829/14/EN WP216- Opinion 05/2014 on Anonymisation Techniques -Adopted on 10 April 2014
- Gruppo Di Lavoro Articolo 29 Per La Protezione Dei Dati -00062/10/EN WP 173 - Opinion 3/2010 on the principle of accountability- Adopted on 13 July 2010

- Gruppo Di Lavoro Articolo 29 Per La Protezione Dei Dati -16/EN WP 243 rev.01 - Guidelines on Data Protection Officers (‘DPOs’) -Adopted on 13 December 2016, As last Revised and Adopted on 5 April 2017
- Gruppo Di Lavoro Articolo 29 Per La Protezione Dei Dati -17/EN WP 251rev.01- Guidelines on Automated individual decision-making and Profiling for the purposes of Regulation 2016/679 -Adopted on 3 October 2017, As last Revised and Adopted on 6 February 2018
- Gruppo Di Lavoro Articolo 29 Per La Protezione Dei Dati -17/IT WP 248 rev.01- Linee guida in materia di valutazione d'impatto sulla protezione dei dati e determinazione della possibilità che il trattamento "possa presentare un rischio elevato" ai fini del regolamento (UE) 2016/679- adottate il 4 aprile 2017
- Gruppo di Lavoro Articolo 29 Per La Protezione Dei Dati- 17/IT WP260 rev.01- Linee guida sulla trasparenza ai sensi del regolamento 2016/679 adottate il 29 novembre 2017,Versione emendata adottata l’11 aprile 2018
- Gruppo Di Lavoro Articolo 29 Per La Protezione Dei Dati- 844/14/IT WP 217 -Parere 6/2014 sul concetto di interesse legittimo del responsabile del trattamento ai sensi dell’articolo 7 della direttiva 95/46/CE- adottato il 9 aprile 2014.
- Information Commissioner’s Office (ICO) -big data, artificial intelligence, machine learning and data protection- 2017
- Information Commissioner’s Office (ICO): Joint statement on data scraping and the protection of privacy -August 24, (2023)
- Parlamento Europeo- “Intelligenza artificiale: questioni relative all'interpretazione e applicazione del diritto internazionale”. Risoluzione del Parlamento europeo del 20 gennaio 2021 sull'intelligenza artificiale: questioni relative all'interpretazione e applicazione del diritto internazionale nella misura in cui l'UE è interessata relativamente agli impieghi civili e militari e all'autorità dello Stato al di fuori dell'ambito della giustizia penale (2020/2013(INI). G.U. (2021/C 456/04)

Bibliografia e Sitografia

- A. C. Amato Mangiameli -"Intelligenza Artificiale, Big Data e Nuovi Diritti." Rivista Italiana Di Informatica e Diritto, vol. 4, no. 1, 2022
- Abbas Assad- Keeping Data Safe: How to Counter Web Scraping Attacks- Techopedia- Agosto 2023. Ultima visita al sito in data 4 febbraio 2025
- AI Magazine- *NewtonX Using AI to power business decisions- by* Tilly Kenyon, 2021.
- Amazon Web Services- Cosa sono i dati strutturati? - Spiegazione dei dati strutturati - AWS- Sito visitato in data 10 gennaio 2025
- Amazon Web Services- Cosa sono i Trasformatori? - Spiegazione dei Trasformatori nell'intelligenza artificiale - AWS- Ultima visita al sito in data 15 febbraio 2025
- Amazon Web Services- Dati strutturati e dati non strutturati: differenza tra dati collezionabili - AWS- Sito consultato in data 10 gennaio 2025
- Barbierato D., Trattamento dei dati personali e “nuova” e responsabilità civile, in Responsabilità civile e previdenza, n.6/2019

- Baxter K. e Schlesinger Y.- "Managing the Risks of Generative AI"- Harvard Business Review- 2023
- Boolean (reperibile su): Boolean | HTML: cos'è, come funziona e a cosa serve | Boolean Blog . Ultima visita al sito in data 15 febbraio 2025.
- Brown, M., Gruen, A., Maldoff, G., Messing S. & Sanderson Z., Zimmer M.-Web Scraping for Research: Legal, Ethical, Institutional, and Scientific Considerations- 10.48550/arXiv.2410.23432.- (2024)
- Campbell F. -Data scraping - what are the privacy implications? Privacy & data protection. 20(1) -2019
- Capuzzo G., Minority Report. Uno studio su intelligenza artificiale e comparazione giuridica tra UE, USA e Cina- Rivista Critica del Diritto Privato- Anno XL - 4 Dicembre 2022 Trimestrale - ISSN 1123-1025
- CEDPO- AI and Personal Data A Guide for DPOs "Frequently Asked Questions", CEDPO AI Working Group, June 2023. Reperibile su: Microsoft Word - CEDPO AI and Data FAQ 12 June 2023_F.docx
- CEDPO AI Working Group- Generative AI: The Data Protection Implications, 16 Ottobre 2023. Reperibile su: generative-ai-the-data-protection-implications-16-10-2023.pdf
- CEDPO- Is the DPO the right person to be the AI Officer? CEDPO AI and Data Working Group Micro-Insights Series, July 2024. Reperibile su: The-DPO-and-the-AI-Officer.pdf
- Cloudflare-How CAPTCHAs work | What does CAPTCHA mean? | Cloudflare- Ultimo accesso al sito in data 3 febbraio 2025
- Cloudflare-What is rate limiting? | Rate limiting and bots | Cloudflare- Ultimo accesso al sito in data 3 febbraio 2025.
- Cocuccio M., Dimensione "patrimoniale" del dato personale e tutele risarcitorie, in Diritto di Famiglia e delle Persone (II), fasc. 1, 1 marzo 2022
- Comandè G., "*Intelligenza Artificiale e responsabilità tra liability e accountability. Il carattere trasformativo della IA e il problema della responsabilità*", in Analisi giuridica delle economia, n.1, 2019.
- Consiglio d'Europa (Ad Hoc Committee On Artificial Intelligence)-CAHAI (2020)23- Studio di fattibilità su un quadro giuridico per la progettazione, lo sviluppo e l'applicazione dell'IA basato sulle norme del Consiglio d'Europa. Strasburgo 17 dicembre 2020
- Contissa G. , Galli F., Gordano F., Sartor G., Il Regolamento europeo sull'intelligenza artificiale, in "i-lex. Scienze Giuridiche, Scienze Cognitive e Intelligenza Artificiale", Rivista semestrale online. Fascicolo 2. Dicembre 2021. ISSN 1825-1927
- CorCom – (reperibile su) Big data, il mercato italiano vale 3,4 miliardi. Vercellis: "Ora però le aziende devono darsi una strategia" Nov. 2024. Sito consultato in data 18 Dicembre 2024
- Criminal Law — Sentencing Guidelines — Wisconsin Supreme Court Requires Warning Before Use of Algorithmic Risk Assessments in Sentencing. — 'State v. Loomis', 881 N.W.2d 749 (Wis. 2016)." Harvard Law Review 130, no. 5 (2017):

1530–37. Reperibile su: Criminal Law — Sentencing Guidelines — Wisconsin Supreme Court Requires Warning Before Use of Algorithmic Risk Assessments in Sentencing. — "State v. Loomis", 881 N.W.2d 749 (Wis. 2016) on JSTOR

- D'acquisto G. , Naldi M., -Big data e Privacy by Design-, Giappichelli editore, 2019.
- David Kaye, -Promotion and protection of the right to freedom of opinion and expression, 29 agosto 2018,- A/73/348
- De santis Luca- ReCAPTCHA: scopri cos'è, come funziona e implementarlo sul tuo sito -*Doctor web Agency*- ottobre 2024. Consultabile a: Recaptcha: cos'è, come funziona e come implementarlo. Ultima visita al sito in data 3 Febbraio 2025.
- Di Matteo F. "La riservatezza dei dati biometrici nello Spazio europeo dei diritti fondamentali: sui limiti all'utilizzo delle tecnologie di riconoscimento facciale." in Freedom, Security & Justice: European Legal Studies: 2023 n.1. ISSN 2532-2079
- Durante, M., Floridi, L. - A Legal Principles Based Framework for AI Liability Regulation. In: Mökander, J., Ziosi, M. (eds) The 2021 Yearbook of the Digital Ethics Lab. Digital Ethics Lab Yearbook. Springer(2022)
- E. Pelino , M. E. Carpenelli , Artt. 37-39 GDPR, In L. BOLOGNINI, E. PELINO (a cura di), *Codice della disciplina Privacy*, Giuffrè Francis Lefebvre, 2019.
- European Union Agency for Fundamental Rights (FRA), Facial recognition technology: fundamental rights considerations in the context of law enforcement, del 27 novembre 2019
- Faini F., "Dati, Algoritmi e Regolamento europeo 2016/679", in Regolare la tecnologia: il Reg. UE n. 2016/679 e la protezione dei dati personali. Un dialogo fra Italia e Spagna, a cura di A. Mantelero e D. Poletti, 2018. Reperibile su: Dati, algoritmi e Regolamento europeo 2016/679
- Fei, Lanfang. "A Comparative Study on Public Interest Considerations in Data Scraping Dispute." International Journal of Law in Context, vol. 20, no. 4, 2024
- Ferola L. , *La «nuova» figura del responsabile della protezione dei dati personali e le sue caratteristiche*, In R. Panetta (a cura di), Circolazione e protezione dei dati personali, tra libertà e regole del mercato, Giuffrè Francis Lefebvre, 2019
- Finck, M., and Frank P. "They Who must Not be identified—distinguishing Personal from Non-Personal Data Under the GDPR." International Data Privacy Law, vol. 10, no. 1, 2020
- Finocchiaro G., "Il Principio Di Accountability- Gdpr Tra Novità E Discontinuità", In Giurisprudenza italiana, n. 12, Utet, 2019
- G. Chauvin et Al.- A giant planet candidate near a young brown dwarf - Direct VLT/NACO observations using IR wavefront sensing- Astronomy &Astrophysics. (2004)
- Garante per la Protezione dei Dati Personali- COMUNICATO STAMPA - IA: il Garante privacy chiede informazioni a DeepSeek. Possibile rischio per i dati di milioni di persone in Italia- 28 gennaio 2025 [Doc-Web n.10096856]
- Garante per la protezione dei Dati Personali- Intervento di Guido Scorza in "Dati pescati a strascico dall'intelligenza artificiale, perché la nostra indagine"-Agenda Digitale " novembre 2023. Reperibile su: Scorza: "Dati pescati a strascico dall'intelligenza artificiale, perché la... - Garante Privacy

- Gartner: (reperibile su) Definition of Big Data - IT Glossary | Gartner. Ultima visita al sito 10 gennaio 2025.
- Geeksforgeeks- Introduction to Web Scraping - GeeksforGeeks Novembre 2024. Ultima visita al sito in data 10 febbraio 2025
- Genna I. , Artificial Intelligence Act, con il voto del Coreper il tempo dei cambiamenti è finito - Federprivacy. Febbraio 2024
- Google for Developers- reCAPTCHA V3- ottobre 2024- consultabile a: reCAPTCHA v3 | Google for Developers. Ultimo accesso al sito in data 4 febbraio 2025
- Google Machine Learning Education, "Adversarial Testing for Generative AI"- Google Developers- Dicembre 2024. Reperibile su Adversarial Testing for Generative AI | Machine Learning | Google for Developers. Ultima visita al sito in data 15 febbraio 2025.
- Google Search Central-(Reperibile su) In-Depth Guide to How Google Search Works | Google Search Central | Documentation | Google for Developers . Ultima visita al sito in data 15 febbraio 2025.
- Guo C., J. Tang, W. Hu, J. Leng, C. Zhang, F. Yang, Y. Liu, M. Guo, Y. Zhu, Olive: Accelerating large language models via hardware friendly outlier-victim pair quantization, in: Proceedings of the 50th Annual International Symposium on Computer Architecture, 2023
- Hacker, Philipp. "A Legal Framework for AI Training Data-from First Principles to the Artificial Intelligence Act." Law, Innovation and Technology, vol. 13, no. 2, 2021
- How We Analyzed the COMPAS Recidivism Algorithm-*by Jeff Larson, Surya Mattu, Lauren Kirchner and Julia Angwin*- ProPublica- May 23, 2016. Reperibile su: How We Analyzed the COMPAS Recidivism Algorithm — ProPublica
- Humza N.-A Comprehensive Overview of Large Language Models, in arXiv preprint arXiv:2307.06435-(2024)
- Iaselli M. - Il diritto alla portabilità dei dati- Altalex - 2017. Reperibile su: Il diritto alla portabilità dei dati personali. Ultima visita al sito in data 15 gennaio 2025
- Iaselli M. -Web scraping: un'analisi del provvedimento del Garante Privacy- Federprivacy, Giugno 2024
- IBM- Bergmann Dave-Che cos'è lo zero-shot learning? | IBM.(2024) Ultima visita al sito in data 15 febbraio 2025
- IBM- Che cos'è l'edge computing? Ultima visita al sito in data 17 febbraio 2025.
- IBM- Cos'è il monitoraggio della rete? (2024). Consultabile a: Che cos'è il monitoraggio della rete? | IBM. Ultimo accesso in data 3 febbraio 2025
- IBM -Cosa sono i Big Data? | Ultima visita al sito in data 11 gennaio 2025
- IBM- Murphy Mike-What are foundation models? - IBM Research (2022). Ultima visita al sito in data 15 febbraio 2025.
- Intelligenza Artificiale Italia Blog- Cosa sono le allucinazioni dell'IA? AI hallucinations | Intelligenza Artificiale Italia Blog. Ultima visita al sito in data 15 febbraio 2025.
- International Data Corporation -(Reperibile su) IDC Study_infographic_2017_USv2- Sito consultato in data 10 gennaio 2025.

- IONOS- Come creare un sito web con un'area riservata o protetta da una password- Aprile 2022. Articolo accessibile a: Creare siti web con area riservata: contenuti solo per i membri - IONOS. Ultimo accesso 1 febbraio 2025.
- Karjalainen, Tuulia. "All Talk, No Action? The Effect of the GDPR Accountability Principle on the EU Data Protection Paradigm." European Data Protection Law Review (EDPL), vol. 8, no. 1, 2022
- King J., Meinhardt C. -Rethinking Privacy in the AI Era Policy Provocations for a DataCentric World- Standford University HAI – 2024
- Kokoulina, Olga. "Towards Future-Proof, Rights-Respecting Automated Data Collection: An Examination of European Jurisprudence." Vanderbilt Journal of Entertainment and Technology Law., vol. 26, no. 4, 2024
- Koster M. et Al, Internet Engineering Task Force (IETF)- RFC 9309 Robots Exclusion Protocol- RFC Editor- settembre 2022. Consultabile a: RFC 9309: Robots Exclusion Protocol
- Lala Fabrizio -Data collection via web scraping: privacy and facial recognition after Clearview- ilex– Rivista di Scienze Giuridiche, Scienze Cognitive ed Intelligenza Artificiale. Vol. 16 n. 2 (2023)
- Lavecchia Vito- Informatica e ingegneria online- INFORMATICA E INGEGNERIA ONLINE | Informatica e Ingegneria Online- Ultima visita al sito in data 17 febbraio 2025
- Liu, Bing-Sentiment Analysis: Mining Opinions, Sentiments, and Emotions - Cambridge University Press, 2020
- Lo Sapio G., Intelligenza artificiale: rischi, modelli regolatori, metafore, in federalismi.it – ISSN 1826-3534, n. 27/2022. P. 249
- Lucchini Guastalla E., "Privacy e data protection: principi generali", in Tosi, Emilio, et al. Privacy digitale : riservatezza e protezione dei dati personali tra GDPR e nuovo Codice privacy. Giuffrè Francis Lefebvre, 2019.
- Mancosu, Moreno, and Federico Vegetti. "What You can Scrape and what is Right to Scrape: A Proposal for a Tool to Collect Public Facebook Data." Social Media + Society, vol. 6, no. 3, 2020.
- Mayer-Schönberger, V., & Cukier, K.- Big Data: una rivoluzione che trasformerà il modo in cui viviamo, lavoriamo e pensiamo.- 2013
- Mejias, Ulises A., and Nick Couldry.-"Datafication"- Internet Policy Review, vol. 8, no. 4, 2019
- Minucci G. e Lombardi G.- Web scraping: accordi browse wrap e tutele disponibili- Lexia- ottobre 2024.- Web scraping: accordi browse wrap e tutele disponibili - LEXIA
- Mitigating Unauthorized Scraping Alliance (MUSA)- Industry Practices to Mitigate Unauthorized Data Scraping- 30 Marzo 2023
- Mollo F., "Gli obblighi previsti in funzione di protezione dei dati personali", Cap X in "Persona e mercato dei dati. Riflessioni sul GDPR", a cura di Zorzi Galgano N., Cedam, 2019

- MyPOS-Cos'è l'URI: definizione, esempi e differenze con URL e URN | myPOS- Ottobre 2024. Ultimo accesso al sito in data 4 febbraio 2025.
- Network360- Cos'è Bert, l'algoritmo che cambia il mondo del Natural Language Processing- Giugno 2020. Cos'è Bert, l'algoritmo che cambia il mondo del Natural Language Processing - AI4Business
- NewtonX Case Study- NewtonX Data Extraction Prevention Best Practices Study- Novembre 2022. Reperibile su: NewtonX Data Extraction Prevention Best Practices Study. Ultima visita al sito in data 15 febbraio 2025.
- NewtonX Reports- NewtonX Data Extraction Prevention Whitepaper- Dicembre 2022. Reperibile su: Data Extraction Prevention: Best practices to combat data scraping. Ultima visita al sito in data 10 gennaio 2025.
- NewtonX: Using AI to power business decisions | AI Magazine. Sito consultato in data 20 Gennaio 2025
- Novelli, Claudio et al. "Generative AI in EU Law: Liability, Privacy, Intellectual Property, and Cybersecurity." ArXiv.org. (2024)
- Novelli, Claudio, Mariarosaria Taddeo, and Luciano Floridi. "Accountability in Artificial Intelligence: What it is and how it Works." AI & Society, vol. 39, no. 4, 2024
- Oracle Cloud Italia- Tipi di dati strutturati e non strutturati | Oracle Italia- Sito consultato in data 10 gennaio 2025
- Osservatori.net Digital Innovation del Politecnico di Milano- (reperibile su) Data Strategy per la valorizzazione dei Dati: mercato e maturità delle aziende italiane nel 2024 . Ultima visita al sito in data 15 febbraio 2025.
- Osservatorio Internet Media del Politecnico di Milano: "Cosa sono i Cookie, a cosa servono e come funzionano"- (Febbraio 2025). Reperibile su: Cosa sono i Cookie, a cosa servono e come funzionano. Ultimo visita al sito in data 15 febbraio 2025.
- Pagallo, Ugo, and Jacopo Ciani Sciolla. "Anatomy of Web Data Scraping: Ethics, Standards, and the Troubles of the Law." European Journal of Privacy Law & Technologies, no. 2, 2023
- Panetta R. (a cura di), Circolazione e protezione dei dati personali, tra libertà e regole del mercato, Giuffrè Francis Lefebvre, 2019
- Panetta R. et al. Il Data Protection Officer tra regole e prassi. Seconda edizione., Giuffrè Francis Lefebvre, 2023.
- Parasol, Max. Data Protection but Not Data Privacy-Cambridge University Press, 2021.
- Peluso, Maria Grazia. Intelligenza artificiale e tutela dei dati : prospettive critiche e possibili benefici per una governance efficace. Giuffrè Francis Lefebvre, 2024.
- Poletti D., M. Causarano, "Autoregolamentazione privata e tutela dei dati personali: tra codici di condotta e meccanismi di certificazione", in Privacy digitale, a cura di E. Tosi, Giuffrè, 2019
- Purtova, Nadezhda. "The Law of Everything. Broad Concept of Personal Data and Future of EU Data Protection Law." Law, Innovation and Technology., vol. 10, no. 1, 2018

- Raposo, Vera L. "Ex Machina: Preliminary Critical Assessment of the European Draft Act on Artificial Intelligence." International Journal of Law and Information Technology, vol. 30, no. 1, 2022
- Reilly, Casey. “The implications of data scraping: it benefits big business, but what does it mean for you?” Journal of high technology law : a student publication of Suffolk University Law School. 24.1 (2023).
- Rezende, Isadora N. "Facial Recognition in Police Hands: Assessing the ‘Clearview Case’ from a European Perspective." New Journal of European Criminal Law, vol. 11, no. 3, 2020
- Rocher, Luc, Julien M. Hendrickx, and Yves-Alexandre de Montjoye. "Estimating the Success of Re-Identifications in Incomplete Datasets using Generative Models." Nature Communications, vol. 10, no. 1, 2019
- Sacheli Giovanni-Come analizzare il rendering di un sito web da parte di Google-EVE Milano- Gennaio 2025. Ultima visita al sito in data 4 febbraio 2025.
- Sartor G. Sartor, The Impact Of The General Data Protection Regulation (Gdpr) On Artificial Intelligence,- European Parliamentary research service (EPRS), panel for the future of science and technology, June 2020. Reperibile su: EPRS_STU(2020)641530_EN.pdf
- Schepisi C., Le “dimensioni della regolazione dell’intelligenza artificiale nella proposta di regolamento della Commissione, in Quaderni AISDUE, ISSN 2723-9969, Sezione “Atti convegni AISDUE”, n. 16/2022
- Scialdone, M. & Vittoria La Rosa, M. -Exploring the Questions and Challenges of Artificial Intelligence Generative. Models in Europe. International Journal of Knowledge Processing Studies (KPS) (2023)
- Scialdone, Marco. “My Data Is Mine : AI, Data Protection, and a Digital Society”. Milano: Giuffrè Francis Lefebvre, (2024).
- Sciolla J. -The normative challenges of data scraping: legal hurdles and steps forward-ilex-Rivista di Scienze Giuridiche, Scienze Cognitive ed Intelligenza Artificiale. -Vol. 16 n. 2 (2023)
- SCM.Sirisuriya, De S. Importance of Web Scraping as a Data Source for Machine Learning Algorithms-Review, 2023 IEEE 17th International Conference on Industrial and Information Systems (ICIIS). IEEE, 2023
- Scraping Robot-“Web Scraping History: The Origins of Web Scraping” (2022). Reperibile su: Storia del web scraping: le origini del web scraping. Ultima visita al sito in data 10 febbraio 2025
- Sito consultato in data 18 Dicembre 2024
- Solove, Daniel J., Artificial Intelligence and Privacy (February 1, 2024). 77 Florida Law Review (forthcoming Jan 2025), GWU Legal Studies Research Paper No. 2024-36
- Soro Antonello.-Tra privacy e open data intesa possibile – Intervento di Antonello Soro, 13 ottobre 2014. Consultabile sul sito del Garante: Tra privacy e open data intesa possibile - Intervento di Antonello Soro,... - Garante Privacy.

- Strubell E., A. Ganesh, A. McCallum, Energy and policy considerations for deep learning in nlp, arXiv preprint arXiv:1906.02243 (2019).
- T. Webb, K. J. Holyoak, H. Lu, Emergent analogical reasoning in large language models, Nature Human Behaviour 7, (9), (2023). Pp. 1526–1541
- Tänzer M., S. Ruder, M. Rei, Memorisation versus generalisation in pre trained language models, arXiv preprint arXiv:2105.00828 (2021)
- Torino R., "La valutazione di impatto (Data Protection Impact Assessment)", in "I dati personali nel diritto europeo".
- Tosi E., Responsabilità civile per illecito trattamento dei dati personali e danno non patrimoniale, Giuffrè, 2019
- Tupay, Paloma Krõõt et al. "Is European Data Protection Toxic for Innovative AI? An Estonian Perspective." Juridica international. 30 (2021)
- Valmeekam K., A. Olmo, S. Sreedharan, S. Kambhampati, Large lan guage models still can't plan (a benchmark for llms on planning and reasoning about change), arXiv preprint arXiv:2206.10498 (2022).
- Voigt, Paul e Axel von dem Bussche. Il Regolamento generale sulla protezione dei dati (GDPR) dell'UE: una guida pratica. 2a ed., Springer, 2024.
- Wandhöfer Ruth, Hazem Danny Nakib- "The Regulatory Dimension of Data Protection, Privacy and AI", in "Redecentralisation Building the Digital Financial Ecosystem" - Cap 7. Pp 159-178. Cham : Springer International Publishing. (2023)
- Wei J., Y. Tay, R. Bommasani, C. Raffel, B. Zoph, S. Borgeaud, D. Yo Gatama, M. Bosma, D. Zhou, D. Metzler, et al., Emergent abilities of large language models, arXiv preprint arXiv:2206.07682 (2022)
- West S. M., M. Whittaker, K. Crawford, -Discriminating systems -AI Now- (2019)
- Witt, Cornelius, and Jan De Bruyne. "The Interplay between Machine Learning and Data Minimization under the GDPR: The Case of Google's Topics API." International Data Privacy Law., vol. 13, no. 4, 2023.
- Yerman Jordan - Is Web Scraping Legal? Navigating Terms of Service and Best Practices- Ethical Web Data Collection Initiative (EWDCI), 7 Jan 2025- Ultimo accesso 1 febbraio 2025.
- Z. Gold, and M. Latonero, Robots welcome: Ethical and legal considerations for web crawling and scraping, Wash. JL Tech. &Arts 13 (2018)
- Zhang C., S. Bengio, M. Hardt, B. Recht, O. Vinyals, Understanding deep learning (still) requires rethinking generalization, Communications of the ACM 64 (3) (2021)
- Zhao, B.-Web Scraping. In: Schintler, L.A., McNeely, C.L. (eds) Encyclopedia of Big Data. Springer-(2022).
- Zhiqing S. et Al- MobileBERT: a Compact Task-Agnostic BERT for Resource-Limited Devices- arXiv:2004.02984.Ultima visita al sito in data 17 febbraio 2025. Aprile 2020

Delibere e provvedimenti

- AGCM-GPDP-AGCOM.“Indagine conoscitiva sui big data”. (Febbraio 2020). Reperibile su: Microsoft Word - IC_Big data_imp.docx
- AGCOM- "Big data" -Interim report nell’ambito dell’indagine conoscitiva di cui alla delibera n. 217/17/CONS. (giugno 2018). Reperibile su: interim ita.pdf
- AGCOM n. 36/02/CONS.- Regole e modalità organizzative per la realizzazione e l'offerta di un servizio di elenco telefonico generale e adeguamento del servizio universale
- AGCOM-Delibera 180/02/CONS- Regole e modalità organizzative per la realizzazione e l'offerta di un servizio di elenco telefonico generale: modalità attuative
- CGUE C-40/17, Fashion ID GmbH & Co. KG v Verbraucherzentrale NRW e V
- CGUE C-634/21, SCHUFA Holding (Scoring): Judgment of the Court (First Chamber) of 7 December 2023 (request for a preliminary ruling from the Verwaltungsgericht Wiesbaden — Germany) — OQ v Land Hessen
- CJEU, judgment of 13 May 2014, Case C 131/12, Google Spain SL and Google Inc.
- CJEU, judgment of 19 October 2016, Case C 582/14, Breyer
- CJEU, judgment of 4 May 2017, Case C 13/16, Rīgas satiksme
- CJEU, judgment of 7 December 2023, Joined Cases C-26/22 and C-64/22, SCHUFA Holding (Libération de reliquat de dette)
- CNIL- Decisione n° MED 2021-134 of 1st November 2021 issuing an order to comply to the company Clearview AI. Reperibile su: Decision n° MED 2021-134 of 1st November 2021 issuing an order to comply to the company CLEARVIEW AI
- CNIL, Délibération n° 2022-125 du 15 décembre 2022 portant avis sur le projet d'arrêté relatif à la création au sein de la direction générale de la concurrence, de la consommation et de la répression des fraudes (DGCCRF) d'un traitement de données à caractère personnel dénommé « Polygraphe » (demande d'avis n° 22014966).
- Data Protection Commission in the matter of Meta Platforms Ireland Ltd- Decision of the Data Protection Commission made pursuant to Section 111 of the Data Protection Act 2018 and Article 60 of the General Data Protection Regulation. - Reperibile su: Final Decision_IN-21-4-2_Redacted.pdf
- Decisione dell’Autorità di controllo tedesca del Land di Amburgo (HmbBfDI) (decisione 545/2020; 32.02-102). Reperibile su: 545_2020_Anhörung_CVAI_DE_Redacted.pdf
- Garante della Protezione dei Dati Personali- Registro dei provvedimenti n. 114 dell'11 aprile 2023 [doc. web n. 9874702] (sospensione della limitazione provvisoria ad OpenAI)
- Garante della Protezione dei Dati Personali- Registro dei provvedimenti n. 621 del 21 dicembre 2023 [doc. web n. 9972593]
- Garante per la Protezione dei Dati Personali - Ordinanza ingiunzione nei confronti di Clearview AI-Registro dei provvedimenti n. 50 del 10 febbraio 2022- [doc. web n. 9751362]
- Garante per la Protezione dei Dati Personali- Elenchi telefonici on line e "ricerca inversa": illegittimi se la fonte non è il d.b.u- Registro dei provvedimenti n. 4 del 14 gennaio 2016 [doc. web n. 6053915]
- Garante per la Protezione dei Dati Personali- Registro dei provvedimenti n. 201 del 17 maggio 2023- [doc. web n. 9903067]

- Garante per la Protezione dei Dati Personali- Registro dei provvedimenti n. 112 del 30 marzo 2023 [doc. web n. 9870832]. (limitazione provvisoria al trattamento per OpenAI
- Garante per la Protezione dei Dati Personali-Web scraping ed intelligenza artificiale generativa: nota informativa e possibili azioni di contrasto- Provvedimento del 20 maggio 2024- [doc. web n. 10020316]
- IMY-DI-2020-2719:A126.614/2020 del 10 febbraio 2021. Disponibile sul sito dell'Autorità. Reperibile su: IMY (Sweden) - DI-2020-2719 - GDPRhub
- Parlamento europeo (Tematiche) -Big data: definizione, benefici e sfide (infografica) -Marzo 2023
- State v. Loomis, 881 N. W.2d 749, 767 (Wis. 2016).
- hiQ Labs, Inc. v LinkedIn Corporation, Case 17-cv-03301-EMC
- Garante per la Protezione dei Dati Personali-Registro dei provvedimenti n. 755 del 2 novembre 2024- [doc. web n. 10085455]
- Garante per la Protezione dei Dati Personali-Provvedimento del 30 gennaio 2025- [doc. web n. 10098477]
- Office of the Data Protection Ombudsman- Administrative fine imposed on Verkkokauppa.com for failing to define storage period of customer data – requiring customers to register was also illegal.
- Tietosuojavaltuutetun toimiston -Tietosuojavaltuutetun ja seuraamuskollegion päätökset- 6.3.2024 TSV/26/2020. Pg. 10
- CJEU-(Case C-30/14)- Judgment of the Court (Second Chamber) of 15 January 2015 (request for a preliminary ruling from the Hoge Raad der Nederlanden — Netherlands) — Ryanair Ltd v PR Aviation BV(Reference for a preliminary ruling — Directive 96/9/EC — Legal protection of databases — Database not protected by copyright or the sui generis right — Contractual limitation on the rights of users of the database
- Garante per la Protezione dei Dati Personali- Intelligenza artificiale: il Garante blocca ChatGPT. Raccolta illecita di dati personali. Assenza di sistemi per la verifica dell'età dei minori-31 marzo 2024 [Doc-Web 9870847].