CEDPO
CONFEDERATION OF EUROPEAN DATA PROTECTION ORGANISATIONS

Bonn, Bucharest, Dublin, Lisbon, Madrid, Milan, Paris, The Hague, Vienna, Warsaw

Generative AI: Implicazioni per la protezione dei dati

CEDPO AI Working Group
16 Ottobre 2023

Contatti:
https://cedpo.eu
info@cedpo.eu

Informazioni su questa guida

"L'intelligenza artificiale non è un concetto nuovo per i DPO (Data Protection Officer) e per i professionisti della protezione dei dati. L'IA generativa, tuttavia, lo è. Quando ChatGPT di OpenAI è stato lanciato nel novembre 2022, la maggior parte dei professionisti della protezione dei dati non aveva mai sentito parlare di IA generativa e di certo non si occupava di tali tecnologie nel proprio lavoro quotidiano. Ora, con ChatGPT nelle mani di oltre 100 milioni di utenti a livello globale e l'ingresso sul mercato di molti altri fornitori come Google Bard e Claude di Anthropic, è diventato una realtà operativa e una necessità per i professionisti della protezione dei dati gestire le conseguenze degli strumenti di IA generativa che vengono rapidamente utilizzati all'interno delle organizzazioni. Sia che questi strumenti vengano adottati semplicemente, sia che vengano messi a punto dalle organizzazioni utilizzando i propri set di dati, esistono implicazioni inedite e inesplorate sulla protezione dei dati con le quali i professionisti della protezione dei dati devono fare rapidamente i conti. L'obiettivo di questo documento è quello di guidare i professionisti della protezione dei dati attraverso il labirinto di problemi che si stanno sviluppando man mano che queste tecnologie vengono adottate rapidamente nelle organizzazioni. Tra le altre questioni chiave, questo documento esamina i rischi connessi alla condivisione dei dati, l'accuratezza dei dati personali, la conduzione di DPIA (Analisi dell'impatto sulla protezione dei dati) sugli strumenti di IA generativa, l'implementazione della protezione dei dati fin dalla progettazione, la scelta di una base giuridica adeguata alla formazione dei sistemi di IA generativa, l'ottimizzazione delle strutture organizzative, l'applicazione di tecniche di miglioramento sulla privacy e la gestione dei diritti degli interessati nel contesto di queste tecnologie. Non ci sarà futuro senza l'IA generativa, e con i dati che svolgono un ruolo cruciale nella formazione e nel funzionamento di questi sistemi, i DPO svolgeranno un ruolo centrale nel garantire che gli standard di protezione e governance dei dati siano al centro di queste tecnologie".

CEDPO
CONFEDERATION OF EUROPEAN DATA PROTECTION ORGANISATIONS

Indice dei contenuti

1. Esattezza dei dati personali

L'esattezza dei dati personali elaborati dagli strumenti di intelligenza artificiale generativa (AI) è una questione fondamentale per la protezione dei dati con tali tecnologie. L'articolo 5 paragrafo 1, lettera d) del GDPR stabilisce che "I dati personali sono esatti e, se necessario, aggiornati; devono essere adottate tutte le misure ragionevoli per cancellare o rettificare tempestivamente i dati inesatti rispetto alle finalità per le quali sono trattati".

È evidente, ed è una questione di buon senso, che il trattamento di dati personali inesatti può avere implicazioni reali per l'interessato, eppure gli strumenti di IA generativa, come ChatGPT, la chatbot basata sul testo ampiamente utilizzato da OpenAI, presentano intrinsecamente numerose inesattezze nei dati che elaborano. Per loro natura, questi strumenti ingeriscono vaste quantità di dati, provenienti da massicce operazioni di web scraping. Necessariamente, questi dati presentano tutte le loro imperfezioni e diventano parte del database su cui gli utenti di ChatGPT effettuano le loro richieste. Quando un utente riceve una risposta completamente o parzialmente inaccurata, questo genera ciò che i fornitori di IA chiamano "allucinazioni" o, nel linguaggio comune, "falsità".

OpenAI stessa, sul suo sito web, avverte gli utenti dei rischi che comporta e del fatto che non ci si possa fidare automaticamente dell'esattezza dei dati recuperati. In una sezione intitolata 'Limitazioni', si legge che "A volte ChatGPT scrive risposte che suonano plausibili ma sono errate o prive di senso"[1]. Come se non bastasse, OpenAI sottolinea che lo strumento spesso aggiunge inesattezze, indovinando cosa intende un utente incerto. Si afferma che: "Idealmente, il modello dovrebbe porre domande di chiarimento quando l'utente fornisce una richiesta ambigua. Invece, i nostri modelli attuali di solito indovinano cosa intendeva l'utente.[2] "

Se a ciò si aggiunge il fatto che i termini di trattamento dei dati di ChatGPT stabiliscono chiaramente che l'utente è il titolare del trattamento dei dati, mentre OpenAI è solo il responsabile del trattamento dei dati, dovrebbe essere chiaro, per gli utenti, che si tratta di un mercato "acquista con cautela". Perché? Perché, se una qualsiasi delle parti elabora ulteriormente dati personali inesatti,

[1] https://openai.com/blog/chatgpt
[2] Ibid.

diventerà responsabile di qualsiasi inosservanza all'articolo 5, paragrafo 1) lettera d) GDPR sopra citato. Nel contesto di ChatGPT, quindi, fare affidamento su dati personali inesatti forniti dallo strumento renderà l'utente responsabile di inosservanza del GDPR, specialmente quando tale riutilizzo influisce sui diritti fondamentali e sulle libertà degli interessati.

Le organizzazioni dovrebbero comprendere che non si tratta di una considerazione meramente teorica e che le autorità hanno già chiesto alle aziende di IA generativa di rendere conto dell'accuratezza dei loro dati. Nel marzo 2023, l'Autorità italiana per la Protezione dei Dati ha bloccato la distribuzione di ChatGPT in Italia, rilevando, tra le altre cose, che i dati non erano spesso esatti. Ha osservato, sulla base dei "test finora effettuati, che le informazioni rese disponibili da ChatGPT non sempre corrispondono alle circostanze reali, quindi vengono elaborati dati personali inesatti.[3] "

I Responsabili della Protezione dei Dati (RPD/ DPO) devono quindi essere consapevoli dei rischi legati al trattamento di dati inesatti. Gli utenti all'interno dell'organizzazione di un DPO dovrebbero ricevere linee guida chiare che li aiutino a capire che i risultati di qualsiasi strumento di IA generativa, come ChatGPT, sono accompagnati da un avviso, ovvero che l'utente è ancora responsabile della verifica dell'esattezza dei dati personali ottenuti. Questo è un punto critico.

Un ulteriore rischio correlato deriva dalla seconda clausola dell'articolo 5 paragrafo 1) lettera d), che stabilisce che i dati personali devono essere "mantenuti aggiornati". ChatGPT e strumenti simili, come Google Bard e Claude di Anthropic, si basano su attività di web scraping effettuate fino a un certo punto nel tempo, il che significa che il loro database diventa obsoleto e, alla fine, non rispondono necessariamente ad eventi aggiornati. Questo crea il rischio che gli utenti otterranno dati personali che non sono più pertinenti, o privi di contesto o semplicemente inesatti, dato che gli eventi sono cambiati o le informazioni si sono evolute nel periodo intercorso.

I DPO devono inoltre essere consapevoli dei modi in cui pregiudizi e discriminazioni non mitigati negli insiemi di formazione potrebbero indirettamente portare a dati imprecisi, aprendo nuovamente l'utente al rischio di un'ulteriore elaborazione di dati imprecisi. Un ultimo rischio

[3] https://www.gpdp.it/web/guest/home/docweb/-/docweb-display/docweb/9870847#english

relativo ai chatbot di IA generativa è il tono che utilizzano: un livello di certezza e autorità oracolare che potrebbe quasi essere definito un "modello oscuro", tanto è fuorviante il suo effetto sulla valutazione dei risultati di ricerca. Quando i chatbot di IA generativa sono palesemente sbagliati o inesatti, spesso lo sono in modo molto sicuro e definitivo, un atteggiamento che maschera il fatto che, come ammette ad esempio OpenAI, la risposta potrebbe semplicemente essere "senza senso". In qualsiasi risultato di ricerca, il tono della risposta dovrebbe essere ignorato, e ancora una volta, gli utenti dovrebbero rendersi conto che l'output di questi strumenti richiede una valutazione umana, soprattutto quando si tratta di questioni sull'esattezza dei dati personali coinvolti.

2. Condividere i dati personali con strumenti di intelligenza artificiale generativa

L'Intelligenza artificiale (IA) si è rapidamente evoluta da un concetto di fantascienza a una caratteristica relativamente comune della nostra vita. Un ramo emergente dell'AI è l'IA generativa, che può creare dati nuovi, precedentemente inesistenti, che imitano fedelmente i dati di input. I modelli di IA generativa possono, nelle giuste condizioni, generare testi, immagini, musica di alta qualità e altro ancora. Tuttavia, la comodità e il potenziale innovativo dell'IA generativa comportano un costo. Nonostante le sue promettenti capacità, la condivisione di dati personali con questi sistemi presenta notevoli rischi per la privacy, la riservatezza, l'integrità e la sicurezza dei dati. Comprendere questi rischi è essenziale per proteggere i diritti degli individui e mantenere un ambiente digitale sicuro.

Come la maggior parte dei sistemi di intelligenza artificiale, l'IA generativa è basata sui dati. La formazione tradizionale dell'IA prevede la somministrazione di grandi insiemi di dati nei modelli di IA, che possono quindi apprendere modelli e caratteristiche da questi dati. Una volta completata la formazione, il sistema di IA è in grado di generare output basati sui modelli e le caratteristiche apprese. Ciò significa che una volta che i dati personali fanno parte del set di apprendimento dell'IA, contribuiscono alla formazione del modello interno dell'IA e ne influenzano inevitabilmente il comportamento ed i risultati. In effetti, i dati diventano "parte" dell'IA nel senso che contribuiscono

alla comprensione e alla conoscenza del sistema. Ciò comporta notevoli problemi di protezione dei dati quando i dati personali sono utilizzati come dati di apprendimento.

I modelli di IA generativa formati a partire da dati personali possono potenzialmente estrarre informazioni sensibili come nomi, indirizzi, informazioni sulla salute o addirittura dati finanziari, per poi ripubblicare tali dati nei risultati di ricerca per diversi utenti. Inoltre, i modelli di IA generativa possono amplificare l'esposizione generando più dati simili all'input originale. Le terze parti possono quindi sfruttare questi dati per attività illegali, tra cui pubblicità invasive, truffe di phishing o, in casi più gravi, frodi o furti d'identità. Ciò evidenzia le complessità del controllo sull'utilizzo dei dati personali da parte dei modelli di IA generativa. Una volta che i dati personali sono stati condivisi con i modelli di IA generativa, gestirne e tracciarne l'uso diventa un compito intricato (se non impossibile), a causa della natura del modo in cui i sistemi di IA elaborano le informazioni e memorizzano e replicano i dati in diversi sistemi. Pertanto, ritirare i dati personali condivisi con i modelli di IA generativa potrebbe essere estremamente difficile o irrealizzabile. La lezione per i DPO è che gli utenti devono comprendere esattamente quali tipi di informazioni possono e non possono essere condivise con gli strumenti di IA generativa, perché una volta condivisi i dati personali, si è superato il punto di non ritorno, e sarà molto difficile annullare ciò che è stato fatto.

Uno dei rischi più allarmanti associati alla condivisione di dati personali con l'IA generativa è la creazione e la proliferazione dei "deepfakes". Per deepfakes si intende l'applicazione dell'IA al fine di creare, alterare o manipolare contenuti, come immagini, audio e video, in modo tale da creare contenuti iperrealistici ma completamente falsi. Addestrandosi sui dati personali, l'IA generativa può generare media sintetici che impersonano in modo convincente persone naturali o giuridiche. Questi deepfake possono quindi essere utilizzati in modo malevolo, ad esempio per campagne di disinformazione, frodi o molestie. A ciò si aggiunge il fatto che l'accuratezza delle decisioni dell'IA generativa dipende fortemente dalla qualità e dalla diversità dei dati di addestramento in ingresso. Se questi dati personali sono distorti, anche i risultati dell'IA possono diventare anch'essi distorti, portando a conseguenze ingiuste.

L'IA generativa è molto promettente per numerose applicazioni, ma l'uso dei dati personali deve essere gestito con attenzione per mitigare i potenziali rischi. Utilizzando rigorosi controlli di

protezione dei dati, pratiche etiche nell'AI e solide protezioni legali, potrebbe essere possibile sfruttare il potenziale dell'IA generativa preservando al contempo i diritti di protezione dei dati e promuovendo un ambiente digitale sicuro e protetto.

3. Qual è la base giuridica appropriata?

La base giuridica che si applica correttamente alla formazione dei sistemi di Intelligenza Artificiale con dati personali è una considerazione fondamentale. A prima vista, non esiste un candidato che legittimi chiaramente questa attività di trattamento e che sostenga anche i diritti di protezione dei dati delle persone interessate. Si tratta di una considerazione critica perché il volume di dati di addestramento utilizzato per le applicazioni AI generative è enorme e in continua crescita. Se si vuole che tali attività di formazione continuino e che l'IA mantenga le sue promesse, non ci si può basare su una base giuridica incerta per quanto riguarda i dati personali. Inoltre, il Regolamento sull'Intelligenza Artificiale (AI Act) non è particolarmente istruttivo su questo punto, dato che l'Articolo 10 (che tratta la governance dei dati e la governance dei dati di formazione per i sistemi AI) non stabilisce una base giuridica specifica per l'uso dei dati personali per la formazione dei sistemi di AI. Pertanto, occorre rivolgersi al GDPR per trovare una base giuridica adeguata per questa attività. In primo luogo, esamineremo brevemente come i dati vengono utilizzati per formare i sistemi AI generativi. Questo avviene in quattro modi diversi:

1. Quando i dati personali sono stati estrapolati da Internet;
2. Quando i dati personali sono stati forniti dagli utenti del sistema AI, ad esempio quando questi ultimi inviano richieste agli strumenti AI generativi;
3. Quando i dati personali sono stati raccolti da terzi, come ad esempio broker di dati o aziende che dispongono di database rilevanti per la fase di formazione dell'AI (ad esempio, un database di sentenze giudiziarie per uno strumento AI predittivo in ambito legale);
4. Quando gli sviluppatori/operatori AI utilizzano i dati personali contenuti nei loro database per formare il sistema AI.

In questi casi, in sensi dell'Articolo 6 del GDPR, le basi giuridiche più rilevanti sono tre: contratto, legittimo interesse e consenso.

1. Contratto

L'articolo 6, paragrafo 1, lettera b) del GDPR stabilisce che il contratto può costituire una base giuridica per il trattamento dei dati personali quando "il trattamento è necessario all'esecuzione di un contratto di cui l'interessato è parte o all'esecuzione di misure precontrattuali adottate su richiesta dell'interessato stesso".

L'applicazione del primo aspetto della base giuridica del contratto (ossia l'esecuzione del contratto stesso) richiederebbe la dimostrazione che la formazione del sistema di intelligenza artificiale (e non l'uso del sistema una volta addestrato) sia strettamente necessaria all'esecuzione di un contratto con l'interessato. Questo requisito di necessità è interpretato in modo molto restrittivo dalle autorità per la protezione dei dati. Secondo l'European Data Protection Board (EDPB), non dovrebbe essere possibile eseguire l'oggetto principale del contratto specifico con l'interessato se non viene effettuato il trattamento dei dati personali in questione. In altre parole, il trattamento dei dati personali dovrebbe essere una condizione necessaria per l'esecuzione del contratto.

Considerando questa interpretazione restrittiva, c'è poco spazio per l'utilizzo del contratto come base giuridica durante la formazione di un sistema di intelligenza artificiale. Questa base potrebbe teoricamente essere applicata quando l'uso del sistema di intelligenza artificiale è l'oggetto del contratto stipulato tra l'operatore dell'IA e l'utente, e quando non vi è altra modalità di esecuzione di tale contratto se non addestrando l'IA con i dati degli utenti.

Per quanto riguarda il secondo aspetto di questa base giuridica, ossia le misure precontrattuali, la sua applicazione richiederebbe la dimostrazione che un interessato abbia fatto una richiesta nel contesto della potenziale stipula di un contratto e che non vi sia altra modalità per soddisfare le sue richieste se non formare (e non solo utilizzare una volta formato) l'IA. Si tratta di un'opzione ancora più restrittiva e limitata rispetto alla prima parte di questa base giuridica.

Nel complesso, le circostanze in cui la base giuridica del contratto potrebbe essere utilizzata per giustificare la formazione di sistemi di intelligenza artificiale con dati personali sono molto limitate e, in termini pratici, questa base non sarà un'opzione valida per giustificare tali attività di trattamento. Nel caso dell'IA generativa, il contratto come base giuridica è, in ogni caso, particolarmente inadatto, visto che di solito non esiste un contratto tra gli interessati i cui dati vengono utilizzati e le organizzazioni responsabili della formazione di tali sistemi con quei dati.

2. Legittimo interesse

Il legittimo interesse potrebbe essere applicato solo a condizione che il titolare del trattamento completi una valutazione del legittimo interesse per garantire che tali interessi non prevalgano sugli interessi o sui diritti e le libertà fondamentali dell'interessato.

Ciò può tuttavia essere difficile, soprattutto perché, nella maggior parte dei casi, l'organizzazione che si occupa della formazione di strumenti di IA generativa, come OpenAI, non è in contatto diretto con gli interessati, né ha alcuna forma di relazione con questi ultimi. A questo proposito, vanno ricordate le recenti azioni del Garante italiano per la Protezione dei Dati contro ChatGPT. Nel marzo 2023, l'autorità ha bloccato ChatGPT nel territorio italiano fino a quando OpenAI non fosse stata in grado di rispondere in modo soddisfacente ad alcune domande, una delle quali prevedeva che OpenAI dovesse specificare la base giuridica sulla quale formare ChatGPT con dati personali. Nella sua risposta a questo punto, OpenAI ha identificato a identificato nel legittimo interesse la base giuridica. Si tratta di un impegno e di una dichiarazione estremamente significativi da parte di OpenAI, in quanto vincola di fatto l'enorme compito di addestramento dei sistemi di IA generativa a una base giuridica che è intrinsecamente incerta, dato l'esplicito diritto degli interessati, ai sensi dell'articolo 21 del GDPR, di opporsi a tale trattamento.

Per potersi effettivamente avvalere del Legittimo interesse come base giuridica sarebbe necessario in particolare:

- uno studio caso per caso del contesto di formazione e dell'utilizzo dell'AI, nonché della raccolta dei dati personali utilizzati per verificare che il trattamento dei dati soddisferà le ragionevoli aspettative degli interessati;
- una dimostrazione della rigorosa necessità di questo trattamento e del fatto che l'IA non può funzionare in modo efficiente senza essere addestrata con i dati personali in questione;
- una maggiore trasparenza del trattamento dei dati nei confronti degli interessati. Tutte le informazioni richieste dal GDPR devono essere fornite agli interessati in modo adeguato;
- un sistema di opt-out efficace reso noto agli interessati entro un periodo ragionevole prima che i loro dati vengano forniti al sistema AI[4];
- in generale, un sistema efficiente per garantire il rispetto dei diritti degli interessati, il che potrebbe essere difficile da implementare date le particolarità del funzionamento dell'IA generativa.

3. Consenso

Anche la base del consenso potrebbe essere applicata, ma solo in circostanze chiaramente circoscritte. Sebbene in casi estremi potrebbe essere l'unica base giuridica possibile (ad esempio, nel caso di trattamento di categorie particolari di dati o dati relativi a minori), come regola generale ha uno spazio molto limitato nella formazione dei sistemi di intelligenza artificiale generativa, come attualmente concepiti. L'intero apparato utilizzato per la formazione dei sistemi di intelligenza artificiale rende quasi impossibile ottenere il consenso. Ciò è dovuto al fatto che, in primo luogo, la maggior parte dei dati utilizzati per formare tali sistemi viene acquistata da intermediari che hanno ottenuto questi dati attraverso lo scraping su Internet, un'attività che non comporta l'ottenimento del consenso da parte dei soggetti dei dati sottostanti. In effetti, la liceità stessa del data scraping come attività commerciale è tutt'altro che certa, come evidenziato dalla recente comunicazione congiunta di dodici autorità globali per la protezione dei dati, tra cui l'Information Commissioner's

[4] Se il sistema di opt-out viene portato a conoscenza dell'interessato dopo l'addestramento dell'IA, l'opposizione al trattamento sarebbe nella maggior parte dei casi, da un lato, irrilevante in quanto il trattamento dei dati sarebbe già avvenuto e, dall'altro, molto difficile da fermare quando una grande quantità di dati personali relativi a numerosi interessati viene fornita a un'IA (si veda la parte XXX relativa all'esercizio dei diritti dell'interessato).

Office del Regno Unito[5]. Per utilizzare il consenso come base giuridica, sarebbe necessario soddisfare tutti i requisiti per un consenso valido ai sensi del GDPR, il che significa che dovrebbe derivare da un'azione chiara e affermativa, dovrebbe essere dato liberamente, specifico, informato e inequivocabile. Si tratta di un'asticella molto alta da raggiungere nel mondo della formazione dei sistemi di intelligenza artificiale.

Se il fornitore di intelligenza artificiale non è in contatto con gli interessati, come avviene generalmente, il consenso dovrà essere raccolto dall'utente del sistema di intelligenza artificiale, cioè dall'organizzazione con la quale l'interessato ha una relazione. Tuttavia, questo avviene di solito a posteriori, quando il sistema di IA è già stato addestrato, per cui l'obiezione al trattamento sarebbe nella maggior parte dei casi irrilevante, poiché il trattamento dei dati è già avvenuto. Inoltre, sarebbe molto difficile invertire la rotta quando il sistema di IA avrà già assorbito grandi quantità di dati personali relativi a numerosi interessati.

In conclusione, il legittimo interesse è probabilmente la base più adatta per formare i sistemi di intelligenza artificiale con i dati personali; tuttavia, come detto sopra, non fornisce una base certa, data la necessità di effettuare una valutazione degli interessi legittimi, nonché il fatto che gli interessati possono opporsi a tale trattamento in qualsiasi momento.

4. Rischi del jailbreak e tutele per la protezione dei dati

Poco dopo il rilascio di ChatGPT, gli hacker hanno iniziato a cercare di "eludere" la chatbot dell'AI, cercando di aggirare le sue protezioni e di fargli dire cose inappropriate o irrazionali. Questi tentativi intricati, che mirano ad eludere le restrizioni imposte ai programmi di intelligenza artificiale, sono diventati noti come "Jailbreak". Questo termine era originariamente utilizzato nel contesto della tecnologia digitale per indicare il tentativo di ottenere accesso al sistema operativo di uno

[5] https://ico.org.uk/about-the-ico/media-centre/news-and-blogs/2023/08/joint-statement-on-data-scraping-and-data-protection/

smartphone o di un tablet, specialmente quelli prodotti da Apple, al fine di eseguire software modificati o non autorizzati.

Nel contesto dei modelli di intelligenza artificiale generativa, il termine ora si riferisce alla progettazione di richieste che inducono le chatbot a eludere le regole sulla produzione di contenuti odiosi o sulla scrittura di atti illegali. Questi attacchi comportano la manipolazione dei sistemi di intelligenza artificiale generativa per produrre contenuti contrari alle loro regole previste, come la generazione di materiale illegale. Un altro uso di questi attacchi potrebbe essere la diffamazione e l'attacco personale a un individuo una volta che i suoi dati personali sono stati divulgati.

Una società di sicurezza specializzata in AI è stata in grado di violare GPT-4, l'ultima chatbot generatore di testo di OpenAI, in poche ore dal rilascio iniziale del sistema. Utilizzando suggerimenti accuratamente elaborati, il CEO della società di sicurezza ha eluso i sistemi di sicurezza di OpenAI e ha fatto sì che GPT-4 generasse dichiarazioni omofobe, creasse e-mail di phishing ed appoggiasse la violenza. Questo comportamento deviante rappresenta un grave rischio in quanto ha il potenziale di esporre i dati personali che sono stati involontariamente, o forse addirittura intenzionalmente, inseriti nel sistema e, quindi, può essere manipolato da attori malintenzionati.

Un attacco strettamente correlato è l'attacco prompt injection, che può inserire silenziosamente dati o istruzioni dannose nei modelli di intelligenza artificiale. Un attacco di tipo prompt injection mira a ottenere una risposta non prevista da parte degli strumenti basati su LLM. Per poi ottenere un accesso non autorizzato, manipolare le risposte o aggirare le misure di sicurezza. Le tecniche e le conseguenze specifiche degli attacchi di tipo prompt injection variano a seconda del sistema.

I jailbreaks e gli attacchi di tipo prompt injection sono una forma di hacking non convenzionale, che utilizza frasi ben formulate al posto del codice per sfruttare le debolezze dei sistemi di intelligenza artificiale. Sebbene questi attacchi si concentrino attualmente sull'elusione dei filtri dei contenuti, i ricercatori della sicurezza mettono in guardia dal rischio di furto di dati e di attività criminali informatiche diffuse, man mano che i sistemi di intelligenza artificiale generativa diventano più diffusi.

Numerosi servizi online e prodotti popolari si basano su grandi insiemi di dati per formare e migliorare i loro algoritmi di intelligenza artificiale. I flussi di dati provenienti dalle reti, dalle piattaforme di social media, dai dispositivi mobili e da varie altre fonti contribuiscono alla vasta quantità di informazioni che le aziende utilizzano per addestrare i loro sistemi di apprendimento automatico. È quindi importante notare che alcuni dei dati contenuti in questi set di dati potrebbero essere considerati dati personali, anche da parte di utenti meno attenti alla protezione dei dati. Purtroppo, a causa dell'uso improprio e della cattiva gestione dei dati personali da parte di alcune aziende, la protezione dei dati è diventata una questione politica globale urgente.

In modo analogo, gran parte dei nostri dati sensibili viene raccolta anche per migliorare i processi abilitati dall'intelligenza artificiale. Questi dati svolgono un ruolo cruciale nel promuovere l'adozione del machine learning, poiché gli algoritmi sofisticati si basano su tali dati per prendere decisioni in tempo reale. Algoritmi di ricerca, assistenti vocali, sistemi di consulenza e altre soluzioni AI sfruttano ampie serie di dati degli utenti del mondo reale per fornire risultati personalizzati e rilevanti.

All'inizio del 2023, è stato lanciato un sito web chiamato "Jailbreak Chat" in cui vengono raccolti e condivisi i suggerimenti per le chatbots AI come ChatGPT provenienti da forum online. I visitatori del sito possono contribuire con i propri jailbreak, provare i suggerimenti inviati da altri e votare la loro efficacia. Gli utenti malintenzionati potrebbero sfruttare questi jailbreak per raccogliere dati personali contenuti nei sistemi al fine di commettere crimini come il furto di identità e la creazione di deepfake per impersonare gli individui.

Le implicazioni di jailbreak e attacchi prompt injection diventano più significative quando questi sistemi ottengono l'accesso a dati personali e sensibili. Ad esempio, se un attacco di prompt injection riuscito istruisce un assistente personale AI a ignorare le istruzioni precedenti e inviare un'e-mail a tutti i contatti, potrebbe portare non solo all'imbarazzo dell'individuo ma anche a problemi diffusi per le persone coinvolte, nonché alla rapida diffusione di contenuti dannosi tra le reti personali e lavorative dell'individuo.

Garantire la sicurezza dei modelli di base come ChatGPT è di primaria importanza, dato che il loro uso diventa sempre più diffuso. Tuttavia, gli hacker non si arrenderanno facilmente. Con l'evoluzione

dei sistemi AI, i jailbreak sono diventati più complessi. Alcuni coinvolgono più personaggi, storie intricate, traduzioni e persino elementi di codifica per generare output specifici.

Alcuni "red team" autorizzati effettuano attacchi di inserimento di prompt su modelli AI per scoprirne le vulnerabilità. Un red team nella sicurezza informatica rappresenta il team di sicurezza offensiva, responsabile di scoprire vulnerabilità di sicurezza attraverso penetration test. Con l'IA generativa, questi team cercano attacchi che includono vulnerabilità reali, influenzano il comportamento del sistema o ingannao gli utenti per aggirare la sicurezza del sistema. Altri tentativi provengono da appassionati che amano mostrare risultati divertenti o inquietanti sui social media. Questo approccio alla sicurezza è subottimale perché è frammentato e si affida all'esposizione virale e a individui influenti per suggerire le correzioni.

Mentre aziende come OpenAI, Google e Microsoft hanno intrapreso misure per affrontare gli attacchi di jailbreaks e di prompt injection, i ricercatori dietro questi attacchi continuano a trovare nuovi modi per sfruttare le vulnerabilità. Lo sviluppo di sistemi AI generativi richiede approcci che vanno oltre i tradizionali metodi di red teaming, come l'utilizzo di un secondo modello di IA per analizzare i suggerimenti o la netta separazione tra i suggerimenti del sistema e quelli dell'utente.

L'automazione e le tecniche avanzate sono necessarie per identificare e mitigare i jailbreak e gli attacchi di prompt injection su scala. Automatizzando il processo di identificazione delle vulnerabilità e dei comportamenti indesiderati, i ricercatori intendono scoprire e affrontare un numero maggiore di questi rischi per la sicurezza.

Questi tipi di tecniche automatizzate possono essere considerate il punto di partenza per un impegno più profondo da parte degli sviluppatori di AI nel valutare la sicurezza dei loro sistemi. Coinvolgendo una gamma diversificata di partecipanti e dando priorità alla trasparenza e alla responsabilità, l'obiettivo è migliorare la sicurezza, la affidabilità e l'uso etico della tecnologia AI generativa. Le valutazioni di terze parti, la mitigazione automatizzata dei jailbreak e l'uso del red-teaming svolgeranno un ruolo fondamentale nel raggiungere questo obiettivo e nel migliorare le pratiche legate allo sviluppo dell'IA per soddisfare i requisiti sia del GDPR che del prossimo AI Act.

5. Come vengono tutelati i diritti degli interessati attraverso gli strumenti di intelligenza artificiale generativa?

La tecnologia dell'AI generativa, o GenAI, consiste in sistemi di intelligenza artificiale in grado di generare testi, immagini o altri media in risposta a stimoli. I modelli generativi apprendono gli schemi e la struttura dei dati di input, generando successivamente nuovi contenuti simili ai dati di addestramento, ma con un grado di novità, invece di limitarsi a classificare o prevedere dati. Questi sistemi di intelligenza artificiale si basano spesso su trasformatori generativi preaddestrati (GPT), reti neurali artificiali costruite sull'architettura dei trasformatori, preaddestrati su ampi insiemi di dati di testo non etichettati e capaci di generare testi simili a quelli umani. Utilizzano modelli linguistici di grandi dimensioni (LLM) per produrre dati basati sul set di dati di addestramento utilizzato per crearli.

Comprendere la tecnologia alla base dell'AI generativa è fondamentale per capire che questi strumenti comprendono diverse fasi, e che i dati personali possono essere trattati in ciascuna fase. Tuttavia, il trattamento dei dati personali in una fase non implica necessariamente l'elaborazione dei dati in un'altra.

Le fasi in cui potrebbero applicarsi i diritti degli interessati secondo la legge sulla protezione dei dati nel contesto dell'AI generativa includono:

1. La fase di formazione dei dati, quando vengono incorporati dati personali.
2. La fase di implementazione, in cui i dati personali vengono utilizzati per generare contenuti e i risultati dei contenuti stesso.
3. Il modello stesso, che potrebbe contenere dati personali.

È inoltre essenziale sottolineare che i software di AI generativa possono elaborare indirettamente i dati, in particolare quelli relativi all'utente, come i dati dell'account o i metadati relativi all'uso della soluzione.

Nei modelli comuni di machine learning, identificare gli individui a cui si riferiscono i dati di addestramento è una sfida potenziale per garantire i loro diritti. Di solito, questi dati includono solo

le informazioni pertinenti alle previsioni, senza identificatori univoci degli interessati. Vengono sottoposti a varie misure di prelaborazione per renderli adatti agli algoritmi di machine learning, spesso trasformando i dati personali in una forma che è più difficile (ma non impossibile) da collegare ad individui specifici. Le leggi sulla protezione dei dati potrebbero quindi applicarsi a questi dati trasformati, poiché potrebbero ancora essere utilizzati per identificare individui. Questo processo richiede considerazione nella risposta alle richieste dei diritti degli interessati.

Questo processo è diverso per i modelli di AI generativa rispetto a quelli comuni di machine learning, come spiegato nel paragrafo precedente. I modelli di AI generativa sono spesso formati con dati accessibili sul web e il loro valore spesso nel generare risultati relativi a persone fisiche. Questo implica una notevole quantità di dati personali nei dati di formazione di questi modelli. Di conseguenza, questi sei di dati potrebbero essere oggetto di richieste dei soggetti interessati.

Nei modelli di AI generativa, l'"apprendimento continuo" pone anche sfide uniche in termini di conformità al GDPR. Questi modelli vengono regolarmente aggiornati in base alle interazioni degli utenti, il che significa che i dati personali vengono continuamente elaborati. Questi dati provengono principalmente dalle interazioni e dagli stimoli degli utenti, e va notato che i soggetti interessati e i fornitori di dati non sono necessariamente la stessa entità nel contesto dei modelli AI ad apprendimento continuo.

Alla luce di queste considerazioni, la gestione dei diritti dei dati ai sensi del Regolamento generale sulla protezione dei dati (GDPR) nel contesto dei modelli di AI generativa presenta sfide uniche, in particolare per i diritti di Cancellazione, Rettifica, Accesso e Obiezione.

Il primo problema condiviso è la non-reperibilità dei dati nei modelli di AI generativa. Come accennato in precedenza, questi modelli attingono dati da una vasta gamma di fonti, come il web scraping e le interazioni degli utenti. Questo approccio multiforme alla raccolta dei dati rende più difficile tracciare i contribuiti dei singoli utenti. Inoltre, a differenza dei tradizionali sistemi di archiviazione dei dati, nei sistemi GenAI i dati personali sono profondamente incorporati in algoritmi complessi, rendendo più difficile l'isolamento di dati specifici. Questo rende complicato l'adempimento dei diritti del GDPR, dal momento che è necessario identificare se e dove i dati personali vengono elaborati all'interno del sistema.

Un ulteriore livello di complessità è costituito dalla questione dei "dati personali dedotti". Si tratta di conclusioni che il modello può trarre in base alla sua formazione. Ad esempio, un modello di AI generativa potrebbe dedurre l'appartenenza politica di un utente in base alle interazioni con i dati. L'opinione prevalente tende a includere queste deduzioni nelle risposte alle richieste dei diritti degli interessati, poiché potrebbero rivelare indirettamente informazioni personali. Anche il concetto di "dati di gruppo dedotti" merita attenzione. Questo tipo di dati viene generato sulla base di modelli più ampi riconosciuti durante l'addestramento. Il fatto che questi dati di gruppo siano considerati personali dipende dal loro successivo trattamento e utilizzo.

Oltre alle sfide comuni, ve ne sono anche di specifiche legate ai diritti individuali che richiedono la modifica o la cancellazione dei dati. Modificare o rimuovere i dati dal set di formazione dopo una richiesta del soggetto interessato (DSR) potrebbe influire sulla validità e correttezza del modello. Spesso, i dati originali servono come base per tali processi di convalida. Inoltre, la cancellazione o la modifica di dati già incorporati nel modello spesso implicano la rimozione o la modifica di tali dati per riqualificare il modello, un'operazione costosa e dispendiosa in termini di tempo.

In sintesi, l'intersezione tra i diritti del GDPR e i modelli di AI generativa presenta un labirinto di sfide, ognuna con le proprie complessità. La stessa natura di questi modelli, dal modo in cui incorporano ed elaborano i dati alle difficoltà nel tracciare i contributi individuali, aggiunge livelli di complessità all'adempimento del GDPR. Sebbene non esista una soluzione universale per affrontare in modo fluido queste sfide, il panorama in evoluzione offre alcune soluzioni emergenti che potrebbero fungere da punti di partenza per la compliance.

Per cominciare, nonostante l'assenza di una soluzione universale, è possibile adottare misure proattive. L'implementazione del principio della "privacy by design e by default" durante le fasi di creazione e implementazione del modello GenAI fornisce un livello fondamentale di protezione dei dati che viene integrato fin dall'inizio.

Nell'affrontare il complesso mondo della protezione dei dati, si potrebbe prendere in considerazione una strategia preventiva che limiti la portata dei dati e delle loro caratteristiche identificative. In questo modo, sarebbe possibile alleviare potenzialmente molte delle complessità che potrebbero emergere nel ciclo di elaborazione dei dati. La minimizzazione dei dati potrebbe essere una parte

essenziale di questa pianificazione iniziale, guidando il responsabile del trattamento dei dati a raccogliere solo ciò che è veramente necessario. Basandosi su questo, le tecniche di anonimizzazione dei dati personali o l'uso di tecnologie avanzate per la privacy (PET), come i dati sintetici, potrebbero consentire una ulteriore riduzione dell'ambito potenzialmente interessato dal DSR.

Inoltre, investire in misure proattive come il data mapping e l'etichettatura dei dati è cruciale. Tali misure offrono chiarezza sull'origine e sulle caratteristiche dei dati di formazione, semplificando la gestione delle richieste dei diritti nelle fasi successive.

Nel momento in cui i modelli di AI generativa passano dalla fase di sviluppo a quella dell'implementazione, l'attenzione si sposta verso l'ottimizzazione dell'adattabilità e della tracciabilità. In questa fase, mantenere un registro meticoloso dell'elaborazione del trattamento dei dati dei dati non è solo una buona prassi, ma diventa indispensabile per facilitare la risposta alle richieste dei diritti degli interessati. Ciò è tanto più importante se si considera la maggiore malleabilità dei dati in questa fase. Inoltre, le sfide dell'"apprendimento continuo" nei modelli implementati possono essere affrontate in modo efficace attraverso tecniche di dimensionamento. Ciò consente un efficiente ripristino a uno stato precedente del modello senza la necessità di ricominciare da zero. Questo collegamento garantisce che siano soddisfatte sia l'adattabilità che la tracciabilità, fornendo un quadro solido per la conformità.

6. Privacy by Design: come costruire strumenti di intelligenza artificiale generativa in conformità con il GDPR

La protezione dei dati sin dalla fase di progettazione svolge un ruolo cruciale nel garantire la conformità al Regolamento Generale sulla Protezione dei Dati (GDPR). Essa implica la salvaguardia dei dati personali fin dalle prime fasi di progettazione e durante l'intero ciclo di vita del sistema. L'idea della protezione dei dati sin dalla progettazione ha avuto origine da un insieme più generale di principi sulla privacy denominati Privacy by Design, sviluppati inizialmente in Canada nei primi anni 2000.

La Privacy by Design è un approccio all'ingegneria dei sistemi che è stato inizialmente sviluppato da Ann Cavoukian e formalizzato in un rapporto sulle tecnologie di miglioramento della privacy da un team congiunto del Commissario per l'Informazione e la Privacy dell'Ontario (Canada), dell'Autorità Olandese per la Protezione dei Dati e dell'Organizzazione Olandese per la Ricerca Scientifica Applicata nel 1995. Il framework della Privacy by Design è stato pubblicato nel 2009 ed è stato adottato dall'Assemblea Internazionale dei Commissari per la Privacy e dalle Autorità di Protezione dei Dati nel 2010. Nello stesso anno, la Conferenza internazionale delle autorità di protezione dei dati e dei commissari per la privacy ha approvato all'unanimità una risoluzione che riconosce la privacy by design come una componente essenziale della tutela della privacy. A ciò ha fatto seguito la decisione della Federal Trade Commission degli Stati Uniti di includere la Privacy by Design tra le tre pratiche consigliate per la protezione della privacy online.

Poco dopo il 2010, l'Europa ha iniziato a lavorare alla revisione delle proprie leggi sulla protezione dei dati. Ispirandosi alla Privacy by Design e ai suoi principi, l'Europa ha messo a punto dei criteri di privacy by design per la protezione dei dati introdotti nel 2018 con l'articolo 25 del Regolamento generale sulla protezione dei dati (GDPR).

Negli ultimi anni, il rapido sviluppo dell'IA generativa ha portato a una maggiore consapevolezza dei potenziali rischi e delle questioni etiche legate alla progettazione di sistemi che elaborano dati personali. Queste considerazioni riguardano non solo rischi complessi legati alla protezione dei dati, come la divulgazione di informazioni sensibili e la conservazione delle cronologie delle chat, ma anche una serie di minacce ai diritti degli interessati e dei cittadini dell'UE, inclusi il "diritto all'oblio" e gli algoritmi tendenziosi. Questo diritto consente agli interessati di richiedere la cancellazione dei propri dati personali da parte di un'azienda. Se la cancellazione dei dati dai database è relativamente semplice, la rimozione dei dati dai modelli di machine learning è un compito più complesso. Tecniche di anonimizzazione e pratiche di minimizzazione dei dati possono aiutare a trovare un equilibrio tra il rispetto dei diritti degli interessati e la salvaguardia della complessiva funzionalità del modello di IA generativa.

Un aspetto da considerare dal punto di vista umano è che, a causa della complessità dei moderni sistemi di IA, le persone coinvolte nella costruzione e nell'implementazione dei sistemi di IA spesso

hanno una gamma più ampia di competenze e background rispetto ai normali sviluppatori di sistemi, tra cui l'ingegneria del software tradizionale, l'amministrazione dei sistemi, i data scientist, gli statistici e gli esperti di dominio.

A causa di questa vasta gamma di competenze, è possibile che vi sia una minore comprensione dei requisiti di conformità alla sicurezza, nonché di quelli della legge sulla protezione dei dati. Per queste persone, la sicurezza dei dati personali potrebbe non essere sempre stata una priorità, in particolare se in passato hanno realizzato applicazioni di IA con dati non personali o in un ambito di ricerca in cui i dati personali erano protetti in sandbox.

Gli algoritmi distorti sono un ulteriore problema per la protezione dei dati. I sistemi di IA generativa apprendono da grandi quantità di dati e se questi dati sono distorti, gli algoritmi possono perpetuare e amplificare questi pregiudizi nei loro risultati. Ciò solleva questioni etiche relative all'equità, alla discriminazione e ai potenziali danni causati da un contenuto distorto generato dall'IA se utilizzato per prendere decisioni importanti, che possono cambiare la vita dei soggetti interessati.

Le allucinazioni dell'IA si riferiscono a casi in cui i sistemi generativi di IA producono risultati non basati su informazioni reali o accurate. Queste allucinazioni possono trarre in inganno gli utenti e avere potenziali implicazioni per la sicurezza degli interessati. I sistemi di IA generativa devono fornire risultati affidabili e attendibili, soprattutto per quanto riguarda i cittadini europei i cui dati personali e la cui accuratezza sono protetti dal GDPR.

La crescita dei deepfakes, ossia contenuti audio o video realistici ma manipolati, è stata associata anche alla tecnologia dell'intelligenza artificiale generativa. I deepfakes hanno il potenziale necessario per manipolare l'opinione pubblica, diffondere disinformazione e rappresentare un rischio per la sicurezza pubblica. Le implicazioni etiche dei deepfakes evidenziano la necessità di misure solide per prevenirne la creazione e per individuarne e contrastarne la diffusione.

L'aspetto fondamentale della protezione dei dati sin dalla progettazione è la trasparenza. Questa svolge un ruolo cruciale nella protezione dei dati da parte della progettazione e garantisce la responsabilità all'interno dei sistemi di IA. Le organizzazioni devono essere trasparenti sulle loro

pratiche in materia di dati, fornendo chiare spiegazioni sul funzionamento dei sistemi di IA e sulle decisioni che prendono. Tuttavia, il raggiungimento della trasparenza nei sistemi di IA può risultare difficile a causa della loro complessità. È essenziale sviluppare metodi e strumenti che consentano di spiegare le previsioni algoritmiche agli utenti finali in modo significativo e comprensibile.

Inoltre, è importante adottare pratiche di sicurezza e protezione dei dati di ultima generazione nello sviluppo di sistemi di IA e nell'uso dei dati personali in contesti di IA. La formazione sulla GDPR è essenziale per garantire che il personale abbia le competenze necessarie per affrontare i rischi legati alla protezione dei dati e alla sicurezza.

L'efficacia dei modelli di IA dipende fortemente dalla qualità dei dati che ricevono, rendendo la protezione dei dati un aspetto integrante della loro progettazione. L'utilizzo di dati sensibili durante la formazione di algoritmi di IA generativa può comportare la comparsa di informazioni personali nelle uscite delle chatbot o compromettere la sicurezza dei dati durante gli attacchi informatici. Ulteriori complicazioni derivano dal fatto che le pratiche comuni su come trattare i dati personali in modo sicuro nell'ambito della scienza dei dati e dell'ingegneria dell'IA sono ancora in fase di sviluppo. Nell'ambito della conformità al principio di sicurezza del GDPR, un'organizzazione deve assicurarsi di monitorare attivamente e tenere conto delle pratiche di sicurezza più avanzate durante lo sviluppo di sistemi di IA e l'utilizzo di dati personali in un contesto di IA.

Non è possibile elencare tutti i rischi di sicurezza che potrebbero aggravarsi con l'uso dell'IA per il trattamento dei dati personali. A prescindere dal rischio, tuttavia, le aziende devono assicurarsi che il personale abbia le competenze e le conoscenze adeguate per affrontare non solo i rischi per la sicurezza, ma anche quelli per la protezione dei dati. È qui che entra in gioco l'importanza della formazione sul GDPR.

L'efficacia dei modelli di IA dipende dalla qualità dei dati che ricevono, rendendo la protezione dei dati un aspetto integrante della loro progettazione. L'utilizzo di dati sensibili durante la formazione di algoritmi generativi di IA può far emergere informazioni personali negli output delle chatbot o compromettere la sicurezza dei dati durante i cyberattacchi.

Per questo motivo, quando si progettano prodotti di IA, è fondamentale dissociare i dati personali dai singoli utenti attraverso l'uso di set di dati sintetici con anonimizzazione completa e identificatori non reversibili per l'addestramento algoritmico, la verifica e il controllo di qualità. L'implementazione di controlli rigorosi sull'accesso ai dati all'interno dell'azienda e la conduzione di audit regolari possono aiutare a prevenire le violazioni dei dati personali.

Inoltre, è importante sottolineare che un maggior numero di dati non equivale necessariamente a soluzioni migliori. Testare gli algoritmi utilizzando la minimizzazione dei dati può aiutare a determinare la quantità minima di dati necessaria per un caso d'uso valido. Inoltre, è fondamentale fornire agli utenti un processo semplificato per consentire agli utenti di richiedere la rimozione dei propri dati personali.

L'adozione di tecniche di apprendimento conflittuale, che prevedono la combinazione di set di dati contrastanti durante il processo di apprendimento automatico, può aiutare a identificare difetti e distorsioni nei risultati degli algoritmi di IA. Inoltre, l'esplorazione dell'uso di set di dati sintetici che non contengono dati personali effettivi è un approccio potenziale, anche se sono necessarie ulteriori ricerche per valutarne l'efficacia.

Le organizzazioni devono allineare l'uso responsabile dell'IA con i principi esistenti di protezione dei dati delineati nel GDPR. Queste linee guida dovrebbero includere vari aspetti come l'accountability, l'intervento umano, l'accuratezza, la sicurezza, la prevenzione dei pregiudizi e la comprensibilità del processo decisionale automatizzato.

Investimenti continui in misure di privacy, formazione nell'audit algoritmico e l'adozione di metodologie di etica, sicurezza e protezione dei dati nella progettazione sono necessari per affrontare efficacemente le opportunità e i rischi legati all'IA generativa. Tecnologie quali la differential privacy offrono tecniche di conservazione della privacy che possono essere incorporate nei sistemi di IA generativa. Tecnologie quali la privacy differenziale offrono tecniche di tutela della privacy che possono essere incorporate nei sistemi di intelligenza artificiale generativa. Metodi modulari per la pulizia degli insiemi di dati, tra cui la deduplicazione e i requisiti di divulgazione dei dati di addestramento, contribuiscono ad affrontare le sfide legate alla privacy.

Gli sforzi congiunti della comunità di protezione dei dati e ingegneristica, uniti all'impegno delle singole organizzazioni e dei professionisti della privacy, svolgono un ruolo indispensabile nell'affrontare le preoccupazioni sulla protezione dei dati legate all'IA generativa. Aderendo ai principi di privacy by design e integrando valutazioni complete sulla protezione dei dati e dei diritti fondamentali, le organizzazioni possono lavorare per una realizzazione affidabile dell'IA generativa mantenendo la conformità al GDPR. È essenziale continuare a investire nella formazione sulla protezione dei dati, nella verifica degli algoritmi e nell'integrazione di metodologie di etica, sicurezza e protezione dei dati nella progettazione per garantire l'uso responsabile ed etico dell'IA generativa.

7. Tecnologie di miglioramento privacy e Dati Sintetici

Gli strumenti di Intelligenza Artificiale generativa sono complessi, e come tutte le tecnologie simili, presentano molte sfide legali significative. La IA generativa richiede una grande quantità di dati, ma tali dati (soprattutto dati di alta qualità) possono essere difficili da reperire o possono essere legalmente protetti, sia dal punto di vista della proprietà intellettuale che della legislazione sulla protezione dei dati.

Dal punto di vista della protezione dei dati, le tecnologie per il miglioramento della privacy (PETs) possono rappresentare una soluzione valida per affrontare le problematiche legate alla protezione dei dati, in termini di minimizzazione dei dati, integrità, confidenzialità e data protection by design. L'Agenzia dell'Unione Europea per la Cybersicurezza (ENISA) definisce le PETs come "soluzioni software e hardware (ad esempio, sistemi che comprendono processi, metodi o conoscenze tecniche) per ottenere funzionalità specifiche di protezione dei dati o per proteggere dai rischi di privacy un individuo o un gruppo di persone".

Tra le varie PETs che potrebbero essere impiegate nel contesto della IA generativa, gli algoritmi di sintesi dei dati che generano dati "artificiali", meglio conosciuti come dati sintetici, possono svolgere un ruolo fondamentale.

Secondo il Garante europeo della protezione dei dati (GEPD), "I dati sintetici sono dati artificiali generati a partire da dati originali a partire da un modello addestrato a riprodurre le caratteristiche e la struttura dei dati originali (...). Il processo di generazione, chiamato anche sintesi, può essere realizzato utilizzando diverse tecniche, come gli alberi decisionali o algoritmi di apprendimento profondo. I dati sintetici possono essere classificati in base al tipo di dati di origine: il primo tipo utilizza dati reali, il secondo utilizza conoscenze raccolte dagli analisti e il terzo tipo è una combinazione di questi due".

In sostanza, i dati sintetici sono dati generati dal computer e derivati da dati reali esistenti o da algoritmi e modelli che replicano, completamente o parzialmente, caratteristiche, schemi e proprietà dei dati del mondo reale. L'uso dei dati sintetici può portare molti vantaggi nella formazione degli strumenti di IA generativa, in particolare perché:

a) riduce la necessità di raccogliere grandi quantità di dati personali reali. Nella fase di addestramento del modello di IA, questo aspetto è particolarmente importante poiché consente agli ingegneri di generare insiemi di dati molto più grandi a partire da quantità relativamente ridotte di dati personali;
b) consente un'etichettatura quasi perfetta (ad esempio, definita con precisione per lo sviluppo di un modello di IA specifico) e dati di migliore qualità, integrando o sostituendo i set di dati del mondo reale. Uno studio di Gartner ha previsto che "entro il 2024, il 60% dei dati utilizzati per lo sviluppo di progetti di IA e di analisi sarà generato in modo sintetico";
c) se opportunamente individuati e corretti, riducono potenzialmente il pregiudizio o lo squilibrio statistico dei set di dati originali, aumentando così l'equità del processo decisionale che si basa sui dati;
d) rafforza la privacy e riduce la superficie di attacco alla sicurezza informatica limitando il rischio di perdita di confidenzialità, integrità o disponibilità di informazioni personali reali;
e) riduce i costi in tutte le fasi della catena del valore dei dati limitando la necessità di un'eccessiva raccolta, pulizia, preparazione e archiviazione dei dati.

Tuttavia, ciò non significa che i dati sintetici siano la soluzione completa per tutti i problemi legati alla protezione dei dati. Ci sono ancora alcuni problemi legali che devono essere prese in considerazione dai responsabili della protezione dei dati (DPO).

In primo luogo, i dati sintetici non corrispondono necessariamente a dati anonimi, il che significa che il rischio di re-identificazione, in una misura o nell'altra, rimarrà. In pratica, i dati sintetici mirano a replicare i dati del mondo reale e quanto più si tratta di un proxy accurato, che mantiene tutte le caratteristiche e gli schemi dei dati originali, tanto più sarà efficiente per il modello generativo di intelligenza artificiale addestrato su tali dati; d'altro canto, però, l'aspetto negativo è che tale efficienza aumenterà, in modo direttamente proporzionale, il rischio di ri-identificazione. Ciò significa che il rischio di dedurre dati relativi a un individuo specifico dal set di dati sintetici o dal modello di IA stesso non si estinguerà.

Come ha osservato l'Ufficio del Commissario per l'Informazione del Regno Unito (ICO): "Dovreste concentrarvi sulla misura in cui le persone sono identificate o identificabili nei dati sintetici, e quali informazioni su di loro verrebbero rivelate se l'identificazione avesse successo. Alcuni metodi di generazione di dati sintetici hanno dimostrato di essere vulnerabili agli attacchi di "Model Inversion", di "Membership Inference" e di "Attribute Disclosure Risk". Questi possono aumentare il rischio di dedurre l'identità di una persona...".

L'uso di altre PET (come la privacy differenziale) o la soppressione degli outlier (punti di dati con caratteristiche uniche di identificazione), può contribuire a ridurre il rischio di ri-identificazione dei dati personali, ma non a eliminarlo completamente.

Inoltre, la fase di generazione dei dati sintetici può comportare il trattamento dei dati personali, soprattutto durante la raccolta e l'analisi di set di dati reali, il che implica la necessità di rispettare il GDPR e gli obblighi ad esso correlati. Occorre inoltre menzionare l'obbligo di fornire tutte le informazioni previste dall'Art. 13 del GDPR ai soggetti i cui dati vengono raccolti e successivamente utilizzati per scopi di formazione dell'IA, nonché di individuare una base giuridica per il trattamento ai sensi dell'Art. 6 del GDPR.

Infine, l'obbligo di rispettare rigorosamente i principi previsti dall'Art. 5 del GDPR sussiste quando si tratta di dati personali. In particolare, alcuni dei seguenti principi dell'Art. 5 sono degni di nota nel caso dell'IA generativa:

a) trasparenza: non si limita alle informazioni da fornire ai soggetti interessati ai sensi dell'Art. 13 del GDPR, come accennato in precedenza, ma anche agli utenti, con riferimento agli output sintetici generati dai modelli di IA, al fine di evitare il rischio di deep fake e/o manipolazione sociale;

b) limitazione della finalità: poiché i dati sintetici possono derivare da dati reali, che possono contenere informazioni personali, è necessario sottolineare che tali dati sono stati raccolti per scopi specifici, esplicite e legittime e che l'ulteriore trattamento (ad esempio, per la sintesi dei dati e l'addestramento successivo del modello di IA) non è incompatibile con gli scopi iniziali.

   Un principio simile è stato stabilito in relazione al processo di anonimizzazione da parte del WP Art. 29 (opinione 5/2014) secondo cui: "il processo di anonimizzazione, ossia il trattamento di dati personali per ottenere la loro anonimizzazione, è un esempio di "ulteriore trattamento". Pertanto, tale trattamento deve essere conforme al test di conformità secondo le linee guida fornite dal Gruppo di lavoro nel parare 03/2013 sulla limitazione delle finalità". In particolare, per quanto riguarda la fase di addestramento dei modelli di IA, il riferimento alle "finalità statistiche" non è in linea di principio incompatibile con gli scopi iniziali ai sensi della lett. b) dell'art. 5, ss.1.[6];

c) accuratezza e correttezza: occorre prestare attenzione per evitare il rischio di "allucinazioni" o di duplicazione di pregiudizi, errori o inesattezze contenuti nel set di dati originale. Questo è particolarmente importante se il modello di IA formato dai dati sintetici verrà poi utilizzato per prendere decisioni che potrebbero influenzare i diritti o gli interessi delle persone. Di fondamentale importanza per questo specifico scopo sarà lo sviluppo di tecniche che

[6] Su questo tema, si veda lo studio del Gruppo di esperti per il futuro della scienza e della tecnologia (Servizio di ricerca del Parlamento europeo) "The Impact of the GDPR on artificial intelligence", giugno 2020.

consentano di spiegare gli output generati dai sistemi di IA addestrati utilizzando dati sintetici.

8. Problemi specifici dell'IA generativa basata sulle immagini e sull'audio

Nel caso di applicazioni AI generative non basate su testo, come gli strumenti per la generazione di immagini, audio e video, esistono chiare implicazioni sulla protezione dei dati. Applicazioni molto diffuse, come Midjourney e Stable Diffusion, che consentono agli utenti di generare rapidamente immagini e video inserendo richieste di testo, si basano su grandi volumi di immagini e contenuti video. Questi dati sottostanti comprendono numerose categorie di dati personali sufficienti a identificare gli interessati, tra cui l'immagine e le sembianze dell'interessato spesso rappresentate nei risultati.

In particolare, i DPO possono aspettarsi che le seguenti categorie di dati personali siano coinvolte in tali strumenti:

- immagini degli interessati;
- rappresentazioni artistiche degli interessati
- filmati di soggetti interessati; e
- dati audio e vocali

Le organizzazioni dovranno comprendere che ulteriori elaborazioni di tali dati rientrano nel campo di applicazione del GDPR. Ad esempio, se un dipartimento marketing desidera creare materiale promozionale e utilizza immagini di soggetti dei dati ottenute dall' AI generativa, dovrà elaborare tali immagini nel rispetto delle leggi sulla protezione dei dati e dei principi fondamentali come la trasparenza, la liceità e la correttezza.

Inoltre, bisogna considerare la questione della combinazione dei dati provenienti da fonti di AI generativa con dati provenienti da altre fonti. Sebbene i dati ricevuti dallo strumento di AI generativa

potrebbero non identificare il soggetto dei dati, l'atto di combinarli con dati alternativi potrebbe farlo, e, ancora una volta, far emergere i requisiti del GDPR. Ciò potrebbe essere particolarmente rilevante quando, ad esempio, l'unione di immagini da diverse fonti porta all'identificazione degli individui.

Nei casi di utilizzo più creativi, in cui le organizzazioni desiderano modificare, alterare o cambiare significativamente la presentazione di immagini, video o contenuti audio, ciò dovrebbe essere fatto nel rispetto dei diritti fondamentali e delle libertà dei soggetti dei dati. I rischi, ad esempio, di diffamare o danneggiare i soggetti dei dati, dovrebbero sempre essere presi in considerazione e, se si ritiene che l'elaborazione possa comportare un rischio elevato, dovrebbe essere condotta una Valutazione di Impatto sulla Protezione dei Dati (DPIA).

Infine, nel caso in cui le organizzazioni desiderino creare contenuti "deepfake" legittimi, come ad esempio video aziendali ufficiali, le questioni del consenso degli interessati e della trasparenza nell'elaborazione dovrebbero essere considerate fondamentali.

9. Gestione del rischio di protezione dei dati

Eseguire una valutazione dell'impatto sulla protezione dei dati (DPIA) durante l'implementazione o l'uso di un sistema AI generativo diventa ancora più cruciale quando, come spesso accade, questi strumenti non sono stati ancora compresi adeguatamente, sia dal punto di vista della strategia aziendale che della gestione del rischio. La comprensione dei rischi per i dati personali derivanti dall'elaborazione AI generativa è ancora in evoluzione e tutti i DPO devono essere pronti a minacce e sfide ancora non previste. Per gestire questi rischi emergenti, dovrebbero essere presi in considerazione i seguenti fattori.

1. **Rischi per gli interessati**

La relazione tra l'utente e l'IA, così come gli impatti che il trattamento avrà sugli individui, devono essere al centro dell'analisi. I potenziali rischi per i soggetti dei dati includono:

- Impatti derivanti da una decisione parzialmente o completamente automatizzata prodotta dall'AI generativa. Le conseguenze di tali decisioni possono consistere in perdite di opportunità finanziarie o addirittura in restrizioni dei diritti fondamentali.
- Rischio di rafforzare la discriminazione e i pregiudizi nei confronti di alcuni utenti.
- Rischi derivanti dal trattamento di dati di categorie particolari di dati come delineato nell'Art. 9 del GDPR. Ad esempio, un tool AI generativo potrebbe dedurre da determinati dati personali della persona interessata (dalle modalità di espressione o dall'uso di determinate parole) l'origine etnica, le posizioni politiche o filosofiche o persino l'orientamento sessuale ed applicare un trattamento differenziato su questa base. Per identificare tali rischi, l'azienda che implementa lo strumento AI generativo dovrebbe effettuare una verifica regolare della qualità dei risultati generati.
- In termini di sicurezza informatica, le informazioni disponibili all'attaccante nel sistema AI possono costituire un vettore di minaccia. Uno scenario denominato "white box", in cui l'attaccante può dedurre/trovare molte informazioni tecniche per preparare il suo attacco, comporta una maggiore esposizione rispetto a un sistema "black box" in cui l'attaccante può accedere solo alle informazioni prodotte dal sistema come output. In particolare, i seguenti attacchi sono specifici per le fasi definite del progetto AI:

<table>
<tr><td rowspan="5">Fase di apprendimento</td><td rowspan="5">tipo di attacco</td><td rowspan="2">virus</td><td>attacchi backdoor</td></tr>
<tr><td>attacchi di virus</td></tr>
<tr><td rowspan="3">esfiltrazione</td><td>attacchi di membership inference</td></tr>
<tr><td>attacchi di model inversion</td></tr>
<tr><td>attacchi di model extracion</td></tr>
</table>

Fase di produzione	tipo di attacco	manipolazione	attacchi di elusione
			attacchi di riprogrammazione
			rifiuto del dispositivo
		esfiltrazione	attacchi di membership inference
			attacchi di model inversion
			attacchi di model extracion

2. Identificazione delle strategie di mitigazione

La Valutazione dell'impatto sulla protezione dei dati (DPIA), come sempre, dovrebbe essere condotta prima dell'avvio del progetto e dovrebbe poi, tramite la attraverso la protezione dei dati fin dalla progettazione, informare e guidare la fase di progettazione di qualsiasi strumento AI generativo. Nel caso dell'AI generativa, dovrebbero essere adottati i seguenti piani di mitigazione dei rischi:

- Messa a punto controllata tramite conversazioni esemplari, in cui un LLM viene addestrato a riprodurre un corpus di conversazioni che illustrano quello che si ritiene essere un comportamento desiderato.
- Messa a punto di un modello di valori in cui gli operatori premiano i risultati più soddisfacenti.
- Inoltre, le misure organizzative dovrebbero mirare a garantire una valutazione costante dei risultati forniti dallo strumento AI generativo, sia a livello dell'operatore umano che lo utilizza, sia di un'entità organizzativa che analizza i risultati su larga scala al fine di garantire un elevato livello di qualità dei risultati nel tempo.
- Analogamente, dovremmo cercare il più possibile di ottenere una situazione di comprensione delle decisioni prese dal modello AI generativo per consentire un vero controllo umano. In questo modo si può evitare un'eccessiva fiducia nei risultati prodotti

dagli strumenti di IA generativa. Tale fiducia eccessiva porterebbe, in assenza di controlli umani efficaci, alla produzione di decisioni completamente automatizzate.

Un'ulteriore considerazione per i DPO è l'emergente requisito di governance dell'AI di condurre Valutazioni di impatto sui diritti fondamentali (FRIA). Nel testo provvisorio del Regolamento sull'AI, che, alla data di pubblicazione di questo documento, è ancora in fase di discussione all'interno della legislatura dell'UE, è stato inserito l'obbligo di effettuare le valutazioni d'impatto sui diritti fondamentali. L'intenzione è che tale valutazione debba essere completata da un fornitore o utente di un sistema AI, nel caso in cui vi siano rischi per i diritti fondamentali e le libertà degli interessati.

Dato che le FRIA sono, di fatto, simili alle DPIA nel mondo dell'AI, con particolari sovrapposizioni nella comprensione di come le attività di elaborazione impattino sui diritti fondamentali, i DPO dovrebbero aspettarsi che questo lavoro venga loro assegnato una volta che il Regolamento sull'AI entrerà in vigore. Sebbene, per certi versi, i DPO siano in una posizione unica e qualificata per svolgere questo lavoro, non è detto che abbiano una conoscenza naturale dei nuovi rischi tecnologici che vengono rapidamente creati dalle tecnologie dell'IA. Per questo motivo, i DPO dovrebbero già essere impegnati nella ricerca e nella comprensione dei rischi specifici legati all'AI per i dati personali.

Dal punto di vista pratico, potrebbe essere possibile condurre FRIA e DPIA come un unico esercizio, ma qualunque metodo sia alla fine scelto, i DPO devono iniziare ora a sviluppare la conoscenza del rischio legato all'AI, in previsione del Regolamento sull'AI.

10. Trasparenza e IA generativa

Quando si raccolgono e si forniscono dati, compresi quelli personali, a un'intelligenza artificiale (AI) ai fini della sua formazione, e quando questo trattamento dati è regolato dal GDPR, l'entità che gestisce questa formazione (l'operatore dell'AI) deve garantire la trasparenza di tale trattamento dati ai sensi dell'articolo 5 paragrafo 1 lettera a) e dell'articolo 12 e seguenti del suddetto regolamento.

È possibile identificare tre diverse fonti di dati:

- Lo scraping dei dati da siti web con l'ausilio di robot o sistemi AI (Caso 1);
- La fornitura di dati da parte degli utenti del sistema o dei fornitori di dati riguardanti altri individui (Caso 2);
- La fornitura di dati che riguardano la propria persona da parte degli utenti dell'IA (Caso 3).

Per ognuno di questi casi d'uso, i modi per garantire la trasparenza del trattamento dei dati variano in base al tipo di formazione richiesto dall'AI.

Caso 1

La trasparenza è una questione delicata e forse impegnativa quando si considera lo scraping di dati online, principalmente a causa del fatto che qualsiasi dato personale raccolto in questo modo non viene raccolto direttamente dall'interessato. Di conseguenza, l'Articolo 14 del GDPR dovrebbe applicarsi a tali dati, vale a dire i dati personali che non sono stati raccolti direttamente dall'interessato. Questo articolo concede al soggetto dei dati il diritto di ottenere dal responsabile del trattamento la conferma la conferma che sia in corso un trattamento dei suoi dati personali e, in caso affermativo, l'accesso ai suoi dati personali, insieme ad altre informazioni essenziali quali le finalità del trattamento, le categorie di dati trattati e così via. Inoltre, dovrebbe applicarsi l'Articolo 15 del GDPR relativo al diritto di accesso dell'interessato ai propri dati personali.

In questo scenario, si presentano diverse difficoltà all'operatore AI. In particolare:

- Identificare i dati personali tra i dati recuperati automaticamente dall'AI, che di solito sono costituiti da grandi quantità di dati
- Identificazione diretta di ogni singolo soggetto interessato;
- Ottenere informazioni di contatto sufficienti per informare ciascun interessato del trattamento dei suoi dati.

Alla luce di queste difficoltà, potrebbe essere applicato l'Articolo 14 paragrafo 5 lettera b) del GDPR. Questa sezione dell'articolo stabilisce che un titolare del trattamento non è tenuto a fornire le informazioni richieste a ogni interessato quando "la fornitura di tali informazioni si rivela impossibile o comporterebbe uno sforzo sproporzionato". La giurisprudenza di varie autorità per la protezione dei dati dimostra che questa eccezione dovrebbe essere interpretata in modo molto rigoroso. Detto

ciò, date le difficoltà individuate in precedenza riguardo ai modelli AI generativi, tale eccezione potrebbe essere applicata. In tal caso, l'operatore AI sarebbe comunque tenuto e rispettare i requisiti di trasparenza nei confronti dell'interessato.

Ai sensi dell'Articolo 14 paragrafo 5 lettera b) GDPR , il Titolare del trattamento deve adottare misure adeguate per proteggere i diritti e le libertà degli interessati. Tali misure includono la pubblicazione dell'informativa privacy del titolare del trattamento sul proprio sito web, ma anche, eventualmente, adottare misure più rigorose come l'esempio fornito dall'Autorità di Protezione dei Dati italiana nella regolamentazione di ChatGPT all'inizio del 2023. In definitiva, OpenAI ha accettato di effettuare una campagna informativa, di natura non promozionale, su tutti i principali mass media italiani (radio, televisione, giornali e Internet) per informare le persone della probabile raccolta dei loro dati personali per la formazione ChatGPT. Hanno inoltre concordato di mettere a disposizione sul sito web del titolare del trattamento uno strumento attraverso il quale tutte gli interessati possano esercitare il loro diritto di accesso ai propri dati personali.

D'altra parte, per quanto riguarda il diritto di accesso, può essere applicato anche l'articolo 11 del GDPR, che stabilisce che:

"1. Se le finalità per le quali un titolare del trattamento tratta dati personali non richiedono l'identificazione di un interessato da parte del titolare del trattamento, quest'ultimo non è obbligato a mantenere, acquisire o trattare informazioni aggiuntive per identificare l'interessato al solo scopo di conformarsi al presente regolamento".

2.Laddove, nei casi di cui al paragrafo 1 del presente articolo, il titolare del trattamento sia in grado di dimostrare di non poter identificare l'interessato, il titolare del trattamento, se possibile, informa l'interessato stesso. In tali casi, gli articoli da 15 a 20 GDPR non si applicano, salvo che l'interessato, al fine di esercitare i propri diritti di cui a tali articoli, fornisca ulteriori informazioni che ne consentano l'identificazione".

Inoltre, come viene ricordato nel considerando 4 del GDPR "il diritto alla protezione dei dati personali non è un diritto assoluto; deve essere considerato in relazione alla sua funzione nella società e deve essere bilanciato con altri diritti fondamentali, in conformità al principio di proporzionalità". Di

conseguenza, si potrebbe sostenere che non possono essere imposti sforzi sproporzionati all'operatore AI per identificare l'interessato e individuare i suoi dati personali nei dati di addestramento dell'AI.

Alla luce di quanto sopra, l'operatore AI che si trova ad affrontare una richiesta di accesso dovrebbe:

1. Verificare se i dati personali relativi all'interessato possono essere identificati;
2. Fornire all'interessato tutti i dati personali identificati;
3. Informare l'interessato che potrebbero esserci dati personali che lo riguardano che l'operatore AI non è in grado di individuare/fornire date le caratteristiche del trattamento dei dati in corso.

Inoltre, per ottemperare all'Articolo 25 del GDPR e al principio di protezione dei dati by design, l'operatore AI potrebbe essere obbligato a dimostrare che può anticipare tali richieste di accesso e che abbia esaminato tutte le possibilità tecniche che potrebbe ragionevolmente utilizzare per individuare i dati personali relativi a ciascun interessato (e di rivalutare regolarmente tali possibilità).

Caso 2

Poiché i dati vengono solitamente forniti agli operatori di IA lungo la catena di fornitura da altre terze parti (un utente o un fornitore di dati), queste terze parti potrebbero assistere l'operatore di intelligenza artificiale nel garantire la trasparenza nel trattamento dei dati, fornendo strumenti e indicazioni su come estrarre al meglio i dati personali dall'insieme di dati, dato che sono proprio queste terze parti a fornire inizialmente i set di dati. Queste terze parti potrebbero anche aiutare l'operatore di intelligenza artificiale a gestire le richieste di accesso da parte degli interessati per gli stessi motivi.

Caso 3

Quando i dati personali vengono raccolti direttamente dagli utenti, si applica l'Articolo 13 del GDPR. Il titolare del trattamento dei dati deve fornire informazioni specifiche all'interessato al momento della raccolta dei dati, come ad esempio l'identità e i dati di contatto del titolare del trattamento dei dati; i dati di contatto del loro responsabile della protezione dei dati; le finalità del trattamento per

cui sono raccolti i dati personali, nonché la base giuridica per il trattamento; insieme ad altre informazioni specifiche.

11. Ottimizzazione delle strutture organizzative

All'interno di qualsiasi organizzazione, dal punto di vista della struttura di gestione, il tema dell'Intelligenza Artificiale Generativa dovrà essere affrontato in modo multidimensionale, come riflesso della complessità della tecnologia e dei suoi impatti. Non sarà possibile per le aziende far lavorare ogni divisione da sola e senza farle interagire tra loro.

L'impatto dell'IA è una questione aziendale; pertanto, richiede un approccio integrato. Tale approccio integrato è essenziale per evitare la dispersione degli sforzi, ma soprattutto per garantire che le decisioni chiave vengano prese in modo multidisciplinare.

Per raggiungere questo obiettivo, le organizzazioni dovrebbero istituire una task force di lavoro sull'IA, focalizzato sull'IA e sulla sua governance. La creazione di tale gruppo di lavoro potrebbe essere un'iniziativa guidata dal DPO, una delle funzioni più esposte a questo tema, dato che deve gestire alcune considerazioni sull'IA in un contesto in cui sono coinvolti dati personali. In alternativa, potrebbe essere avviata e guidata da una funzione IT, come un Chief Data Officer o un Chief Technology Officer.

Questo gruppo di lavoro coinvolgerà significativamente l'ufficio legale, le funzioni di compliance e, in particolare, la protezione dei dati. Per gli aspetti tecnici, dovrebbe essere coinvolto il dipartimento di sicurezza informatica. Il gruppo di lavoro potrebbe coinvolgere anche il personale addetto alle comunicazioni e delle pubbliche relazioni, poiché sarà necessario comunicare internamente, e potenzialmente anche esternamente, le decisioni prese dal gruppo di lavoro. Il leader del gruppo di lavoro potrebbe istituire dei gruppi di lavoro focalizzati, in cui membri selezionati si concentrano su questioni specifiche e riportano i loro risultati al gruppo di lavoro. Il diagramma sopra fornisce un'idea indicativa della composizione di questi gruppi di lavoro e di come si relazionerebbero con la task per la Governance dell'IA.

La missione del gruppo di lavoro è quella di rispondere all'immediata esigenza di una governance responsabile dell'IA all'interno dell'organizzazione e di esaminare e gestire i rischi nell'uso dell'IA generativa, in particolare per quanto riguarda i dati personali, i pregiudizi, le questioni etiche, la regolamentazione emergente sull'IA e numerose questioni legali come i diritti di proprietà intellettuale e l'esposizione alla responsabilità.

Il principale obiettivo di questo gruppo di lavoro sarà definire un piano d'azione. Un aspetto critico di questo piano d'azione sarà condurre un elenco dei sistemi di IA utilizzati dall'azienda, compresa l'IA generativa. Un altro aspetto critico è la definizione di ruoli e responsabilità per tutte le funzioni del gruppo.

Il ruolo di questo gruppo di lavoro sull'IA è anche quello di sensibilizzare sull'IA a tutti i livelli dell'azienda. Questo punto è importante, poiché il rischio proviene naturalmente dai dipendenti che sono gli utenti quotidiani della tecnologia, ma deve essere collegato al più alto livello decisionale, perché decidere come utilizzare (o non utilizzare) l'IA generativa è una strategia aziendale.

Come compito iniziale, il gruppo di lavoro dovrebbe preparare linee guida preliminari per l'organizzazione riguardo all'uso responsabile dell'IA generativa, che dovrebbero, ad esempio, includere la raccomandazione di non inserire dati personali nelle richieste di strumenti rilevanti come ChatGPT, né di caricare immagini con persone identificabili.

Indipendentemente dalla complessità della tecnologia e della sua implementazione, il ruolo del DPO in questa task force è in ultima analisi quello di garantire che tutti i dati personali elaborati attraverso le tecnologie AI siano conformi al GDPR.