Atti della settima conferenza AAAI/ACM su IA, etica e società (AIES 2024)

La correttezza della delega ai sensi del regolamento europeo sulla protezione dei dati e della legge sull'AI: Una prospettiva di sensibilità e necessità

Ioanna Papageorgiou
Università Leibniz di Hannover
Istituto di informatica giuridica
Hannover, Germania
ioanna.papageorgiou@iri.uni-hannover.de

Astratto

Il presente lavoro si occupa della convergenza tra il Regolamento europeo sulla protezione dei dati e la legge sull'IA all'interno del paradigma dei metodi computazionali che rendono operativa l'equità in presenza di dati demografici, in particolare attraverso l'uso di variabili proxy e tecniche inferenziali (*Proxy Fairness*). In particolare, esplora la natura giuridica dei dati coinvolti nella Proxy Fairness ai sensi del Regolamento europeo sulla protezione dei dati, concentrandosi sulla nozione giuridica di sensibilità. Inoltre, esamina la liceità del trattamento dei dati personali sensibili ai fini della Proxy Fairness ai sensi dell'AI Act, concentrandosi in particolare sul requisito giuridico della Necessità. Attraverso questa analisi, il documento mira a far luce su aspetti fondamentali della legittimità della Proxy Fairness nel contesto del diritto dell'UE, fornendo una base normativa a questa linea di approcci Fair-AI.

Introduzione

La crescente adozione di sistemi di IA in aree ad alto rischio della vita pubblica e gli studi approfonditi sul potenziale discriminatorio dell'IA (Mehrabi et al. 2021) hanno stimolato una proliferazione di metodi algoritmici che studiano e perseguono l'equità nei sistemi di IA (*Fair-AI*) (Ntoutsi et al. 2020; Schwartz et al. 2022; Mitchell et al. 2021). Questi metodi sono incentrati sull'individuazione, l'attenuazione e la valutazione dei pregiudizi nei gruppi protetti dalla legge e richiedono quasi sempre l'accesso ad attributi sensibili, come i dati demografici, che determinano l'appartenenza al gruppo. Tuttavia, questo spesso implica il trattamento di dati personali sensibili, che è in linea di principio vietato dalla legge sulla protezione dei dati dell'UE, ponendo delle sfide alla fattibilità degli approcci Fair-AI.

In risposta a questa sfida, un filone crescente di ricerca sull'IA (Ashurst e Weller 2023; Centre for Data Ethics and Innovation e Department for Science, Innovation and Technology 2023; Awasthi et al. 2021; Chen et al. 2019a; Yan, te Kao e Ferrara 2020; Zhu et al. 2023; Gupta et al. 2018) ha studiato metodi computazionali che consentono di operazionalizzare l'equità in assenza di dati demografici, in particolare attraverso l'uso di variabili proxy e tecniche inferenziali (*Proxy Fairness*). Oltre a essere sempre più studiati, i metodi di equità per procura sono già stati ampiamente impiegate in vari settori, tra cui quello fiscale, finanziario, della protezione dei consumatori (Elzayn et al. 2023; Baines e Courchane 2014; Consumer Financial Protection Bureau 2014; Elliott et al. 2009) e dei social media (Alao et al. 2021) (Belli et al. 2021), in particolare nel contesto statunitense. Tuttavia, finora è stata prestata scarsa attenzione all'interazione di questi metodi con le normative vigenti in materia di protezione dei dati, il che pone una significativa incertezza giuridica sulla legittimità di questi approcci.

Questa incertezza si intensifica di fronte agli sviluppi normativi in corso. In particolare, la prossima legge sull'IA ha affrontato anche la sfida della scarsità di dati nel contesto dell'equità, consentendo, per motivi di interesse pubblico, il trattamento dei dati personali sensibili ai fini dell'individuazione e della correzione degli errori nei sistemi di IA ad alto rischio. Precisamente, secondo l'articolo 10 (5) della legge sull'IA, il trattamento dei dati personali sensibili è consentito solo "nella misura in cui è *strettamente necessario* per garantire la protezione e la correzione degli errori in relazione ai sistemi di IA ad alto rischio...[enfasi aggiunta]". Sebbene la disposizione di abilitazione sembri essere indipendente dal metodo, il che significa che non è limitata a un particolare approccio di equità, il requisito di necessità stabilito influenza in modo significativo la scelta dei metodi di equità e, in misura maggiore, l'ambito di applicazione della Proxy Fairness.

Alla luce di quanto sopra, il presente lavoro si propone di esaminare le implicazioni della Proxy Fairness ai sensi del Regolamento generale sulla protezione dei dati e dell'AI Act, fornendo una base normativa a questa linea di approcci Fair-AI. In particolare, il documento fornisce i seguenti contributi:

- La prima sezione analizza la Proxy Fairness ai sensi del Regolamento generale sulla protezione dei dati, concentrandosi sulla nozione di "*sensibilità*" dei dati. Attingendo alla giurisprudenza consolidata della Corte di giustizia europea e a importanti studi giuridici, il documento esamina la natura dei dati coinvolti negli approcci di Proxy Fairness, facendo luce su aspetti sfumati della protezione dei dati che riguardano le *varianti di proxy* e le *inferenze sui dati*.
- La sezione successiva analizza la Proxy Fairness nell'ambito della legge sull'AI, in particolare rendendo operativa la nozione giuridica di "*necessità*". Basandosi sull'articolo 10 (5) della legge sull'AI, il documento esamina la necessità di acquisire dati personali sensibili nel contesto dei metodi di proxy fairness, facendo luce su aspetti fondamentali della loro

legittimità. Questo esame comporta una valutazione comparativa degli approcci di equità per delega rispetto alle alternative di default, considerando i criteri di necessità di *intrusività*, *efficacia* e *ragionevolezza*.

Sebbene l'analisi si basi su un quadro normativo europeo, la sua rilevanza va oltre i confini dell'Unione e si estende a iniziative Fair-AI extraterritoriali. In particolare, i fornitori di IA che immettono sistemi di IA ad alto rischio sul mercato dell'UE o che mettono in servizio tali sistemi nell'Unione rientrano nell'ambito di applicazione della legge, indipendentemente dal loro stabilimento o dalla loro ubicazione all'interno dell'Unione o di un Paese terzo.[1]Di conseguenza, sono anche soggetti agli obblighi di governance dei dati previsti dalla legge sull'IA, tra cui l'obbligo di esaminare i set di dati di formazione, test e valutazione per individuare eventuali distorsioni e adottare misure appropriate per individuarle, prevenirle e attenuarle.[2]Nella misura in cui i fornitori di IA non appartenenti all'UE trattano informazioni relative a persone situate nell'UE al fine di individuare o correggere le distorsioni nei loro set di dati, devono attenersi ai requisiti UE applicabili, compreso il requisito della necessità. [3]

Metodologia

Il documento adotta fondamentalmente una metodologia di ricerca giuridica dottrinale. Due atti legislativi dell'UE, il Regolamento generale sulla protezione dei dati e l'AI Act, fungono da quadro normativo dell'analisi, mentre le nozioni giuridiche di "sensibilità" dei dati e di "necessità" del trattamento ne costituiscono i punti centrali. Tuttavia, nella misura in cui esamina i fenomeni tecnici all'interno di questo quadro normativo e concettuale, come gli approcci di equità computazionale, l'analisi assume un approccio tecno-giuridico riguardo all'oggetto di indagine.

In particolare, il documento combina caratteristiche metodologiche descrittive, classificatorie e valutative (Kestemont 2018; Smits 2017), al fine di identificare il sistema giuridico sottostante applicabile nel contesto della Proxy Fairness e di indirizzarne l'applicazione legale.

Per descrivere la nozione giuridica di "sensibilità" dei dati, la prima sezione utilizza un'interpretazione giuridica grammaticale e sistemica, nonché un'interpretazione sulla base della giurisprudenza dell'UE, di fonti giuridiche non vincolanti e della dottrina giuridica. Successivamente, gli approfondimenti dell'analisi descrittiva, che seguono in gran parte un ragionamento deduttivo, informano la classificazione dei dati "per delega" e "per inferenza" nell'ambito della nozione descritta di dati personali sensibili, mentre gli approfondimenti della letteratura tecnica fungono da criterio esterno per questa classificazione.

La seconda sezione utilizza gli stessi strumenti internazionali e gli stessi criteri interni per descrivere la nozione giuridica di "necessità" e per classificare le caratteristiche della Proxy Fairness sotto le componenti della necessità. A causa dell'elemento comparativo intrinseco della Necessità, la sezione si basa in gran parte su una valutazione comparativa tra gli approcci di Proxy Fairness[4]e le loro controparti "Default"[5], secondo i criteri giuridici di in- trusività, efficacia e ragionevolezza. Gli spunti della letteratura informatica, STS (Science, Technol- gie, and Society) e dell'etica dell'IA sono incorporati come indicatori di valutazione, informando i criteri legali senza alterare il quadro di valutazione del documento.

Pur attenendosi a un approccio normativo di tipo dottrinale, il pa- gamento adotta una prospettiva interdisciplinare, esponendo le complesse interdipendenze tra il diritto (della protezione dei dati), l'informatica e l'etica nel contesto della Fair-AI e creando un crocevia discorsivo da sfruttare e arricchire ulteriormente con le rispettive discipline e le loro metodologie.

Equità della delega ai sensi del GDPR: una prospettiva di sensibilità

La nozione di dati sensibili, ben consolidata nella legislazione europea sulla protezione dei dati, è utilizzata nel Regolamento generale sulla protezione dei dati per indicare "categorie particolari di dati personali", che per la loro natura sono particolarmente sensibili in relazione ai diritti fondamentali e richiedono quindi una protezione più estesa rispetto ai dati personali "ordinari" (considerando 51 del GDPR). L'articolo 9 del GDPR definisce in modo restrittivo i dati sensibili come quelli che rivelano l'origine razziale ed etnica, le opinioni politiche, le convinzioni religiose o filosofiche, l'appartenenza sindacale, nonché i dati genetici, biometrici e sanitari, e ne consente il trattamento solo in presenza di rigorose eccezioni.

La legge sull'IA prevede tale eccezione all'articolo 10 (5), consentendo il trattamento di dati personali sensibili per il rilevamento e la correzione di errori nei sistemi di IA ad alto rischio, facendo esplicito riferimento all'articolo 9 del GDPR e al suo concetto di dati sensibili.[6]Per valutare quindi la correttezza di Proxy in base a questa eccezione della legge sull'IA, è necessario innanzitutto indagare in che misura essa comporti il trattamento di dati sensibili ai sensi del GDPR.[7]A tal fine,

---

[1]Si veda l'art. 2 ([1])della legge sull'AI. 2 (1) Legge AI.

[2]Si veda l'art. 10 (2) (g), (f) della legge sull'AI. 10 (2) (g), (f) Legge AI

[3] [1]fornitori di IA non sono obbligati a debi- care i loro set di dati con informazioni sensibili relative a persone nell'UE. Tuttavia, ai sensi dell'articolo 10 (3) e (4) della legge sull'IA, essi sono tenuti ad assicurarsi che gli insiemi di dati per la formazione, la convalida e il test possiedano proprietà statistiche adeguate per quanto riguarda gli individui a cui è destinato il sistema di IA ad alto rischio, considerando anche le caratteristiche specifiche dell'ambiente geografico in cui il sistema di IA ad alto rischio opererà. Questo requisito può portare alla personalizzazione degli insiemi di dati utilizzati a fini di equità per adattarli al contesto locale dell'Unione, incentivando l'utilizzo di insiemi di dati relativi all'UE.

[4]L'analisi giuridica prende in considerazione un paradigma di base di Proxy Fairness, che prevede l'inferenza dell'attributo sensibile pertinente da caratteristiche correlate (variabili proxy) attraverso l'uso di un classificatore at- tributo, noto anche come modello proxy, al fine di valutare o controllare l'equità. Sebbene le varianti degli approcci alla correttezza per procura possano influenzare alcuni aspetti dell'analisi, la logica sottostante si applica *mutatis mutandis* a qualsiasi metodo di correttezza che si occupi in qualche misura di procure e inferenze di dati sensibili.

[5]Nel contesto di questa analisi, il termine "Default" si riferisce ad approcci situati in un regime non demograficamente scarso, che si avvalgono direttamente degli attributi sensibili "reali" ottenuti direttamente o indirettamente dagli interessati.

[6]I termini "categorie particolari di dati" e "dati sensibili" sono utilizzati in modo intercambiabile all'interno del GDPR. Si veda il considerando 10 del GDPR.

[7]Il "trattamento" dei dati comporta, ai sensi dell'articolo 4, paragrafo 2, del RGPD, un'ampia gamma di operazioni eseguite sui dati personali, tra cui la raccolta, l'organizzazione, la strutturazione, la conservazione, la modifica, la rivalutazione, l'utilizzo o la divulgazione dei dati.

Il capitolo che segue distingue tra i due principali pilastri dei dati coinvolti nella Proxy Fairness, ovvero i dati *proxy* e quelli *inferiti*, e li valuta in base alla nozione di sensitività.[8]L'analisi giuridica considera un paradigma di base della Proxy Fairness, che prevede l'inferenza dell'attributo sensitivo pertinente da caratteristiche correlate (variabili proxy) attraverso l'uso di un classificatore di attributi - noto anche come modello proxy -, al fine di valutare o controllare l'equità. Oltre a fornire le basi necessarie per la successiva analisi, questa sezione rappresenta un esercizio giuridico critico. Infatti, gli approcci inferenziali, a causa delle loro intrinseche complessità e dell'interferenza indiretta con i dati sensibili, rischiano spesso di sfuggire alle maglie del Regolamento europeo sulla protezione dei dati e al suo rigoroso regime di protezione dei dati sensibili.

Dati proxy

Gli approcci basati su proxy sono concettualmente costruiti attorno all'utilità e all'uso di attributi proxy, ovvero dati che sono associati e che potrebbero essere utilizzati come sostituti o sostituzioni di tratti demografici reali non disponibili o inaccessibili. Il cognome o l'indirizzo di un individuo, ad esempio, potrebbero fungere da proxy per la sua etnia ed essere utilizzati per calcolare la probabilità di appartenenza a una specifica razza o etnia. Nel contesto della Proxy Fairness, i dati di origine ordinaria vengono elaborati per dedurre gli attributi sensibili pertinenti, che (inferenze) vengono poi utilizzati insieme ai metodi standard di fairness per la valutazione o l'ottimizzazione della fairness. Dal punto di vista della protezione dei dati, ciò solleva la questione non banale della natura del primo blocco di dati coinvolto nella Proxy Fairness, ovvero i dati proxy, chiamati anche dati che producono inferenze.

Per rispondere a questa domanda è necessario innanzitutto basarsi su una lettura testuale del GDPR e in particolare dell'articolo 9 (1), che definisce i dati sensibili come "dati *che rivelano l'*origine razziale o etnica...[enfasi aggiunta]". La formulazione di questa disposizione consente esplicitamente un'interpretazione ampia, comprendendo non solo i dati intrinsecamente sensibili - cioè quelli che per loro natura contengono informazioni sensibili - ma anche i dati da cui si possono dedurre informazioni sensibili relative a un individuo. Ciò è dovuto al termine chiave "rivelare", che descrive la natura del legame tra i dati personali considerati sensibili e il contenuto informativo di tali dati (Spiecker et al. 2023).

Questo approccio interpretativo espansivo è stato esplicitamente approvato dal Gruppo di lavoro Articolo 29 in diverse linee guida e infine affermato dalla Corte di giustizia europea (CGE) nella sua recente giurisprudenza (Nowak 2017). In questa specifica sentenza, la Corte di giustizia europea ha esplicitamente ampliato l'ambito di applicazione del GDPR dai dati *intrinsecamente* sensibili alle informazioni che consentono *direttamente* di dedurre informazioni sensibili da attraverso operazioni intellettuali come la deduzione o il riferimento incrociato. In particolare, secondo la CGUE, è sufficiente che i dati siano semplicemente in grado di rivelare indirettamente informazioni sensibili.

Di conseguenza, anche i dati di origine ordinaria possono essere trattati come dati sensibili, quando si può dimostrare che consentono di inferire attributi sensibili (cfr. (Wachter e Mittelstadt 2018; Quinn e Malgieri 2020; Malgieri e Comande` 2017)). Questa interpretazione è particolarmente rilevante nel contesto della Proxy Fairness, che si basa concettualmente sulla capacità dei modelli proxy di inferire gli attributi sensibili mancanti da dati apparentemente non sensibili. Per esempio, il codice postale di un individuo può non essere intrinsecamente sensibile, ma se combinato con altri dati nel contesto della Proxy Fairness, può portare a dedurre l'effettiva etnia dell'individuo. Non considerare questi dati come sensibili rischierebbe, secondo la giurisprudenza consolidata della Corte di giustizia europea, di compromettere l'efficacia della protezione speciale offerta dal GDPR ai dati sensibili e, in misura maggiore, i diritti e le libertà fondamentali ad essi associati.

Tuttavia, la CGUE ha implicitamente liquidato con questa giurisprudenza un'ampia mole di studi giuridici che hanno messo in dubbio la rilevanza e l'efficacia contemporanea del concetto di dati sensibili del GDPR e hanno di conseguenza cercato di restringerne l'ambito di applicazione. Particolarmente preoccupati dal "rischio di inflazione" dei dati sensibili a fronte degli sviluppi tecnologici, molti studiosi hanno proposto approcci moderati che introducono ulteriori fattori soggettivi o oggettivi nella definizione di dati sensibili (Quinn e Malgieri 2020; Wachter e Mittelstadt 2018; Solove 2024; Schiff, Ehmann e Selmayr 2017). I para- grafi successivi discutono il caso della Proxy Fairness alla luce di questi approcci restrittivi.

Per cominciare, i fattori soggettivi sono incentrati sull'"intenzionalità" e guardano essenzialmente alle intenzioni e agli scopi dichiarati dei responsabili del trattamento dei dati rispetto alla generazione di inferenze, escludendo così i dati che accidentalmente o casualmente rivelano informazioni sensibili. Dato che l'uso di dati proxy nel contesto della Proxy Fairness comporta concettualmente l'inferenza intenzionale o affermativa di informazioni sensibili rilevanti, su cui si baserebbe l'individuazione e la correzione dei pregiudizi, i dati proxy sembrano superare il test di sen- sibilità secondo gli approcci propositivi suggeriti. Questo è plausibile, dato che i dati sensibili sono tipicamente il risultato desiderato del processo analitico del classificatore nei metodi proxy standard.[9]

Inoltre, la Proxy Fairness sembra soddisfare la maggior parte dei criteri oggettivi di sensibilità, che considerano il contesto di elaborazione e includono, tra l'altro, i costi e la quantità di tempo necessari per l'inferenza, la tecnologia disponibile al momento dell'elaborazione (Malgieri e Comande` 2017) e la facilità o l'affidabilità dell'inferenza (Malgieri e Co- mande` 2017; Wachter e Mittelstadt 2018). In primo luogo, i modelli proxy sostengono di essere oggi una soluzione relativamente semplice e facilmente de-

[8]Poiché la classificazione dei dati come "sensibili" ai sensi del GDPR implica essenzialmente la loro classificazione come "personali", gli aspetti del legame personale dei dati - in particolare quelli relativi al contesto della correttezza della delega - sono inevitabilmente inclusi. Tuttavia, l'analisi presuppone principalmente l'esistenza di dati personali ai sensi del GDPR, mentre le questioni relative all'identificabilità e all'anonimato esulano dal campo di applicazione del presente documento.

[9]Questo fattore può variare negli approcci in cui i dati proxy non sono trattati come tali, ma piuttosto come un segnale molto approssimativo sulla potenziale iniquità, senza trarre conclusioni o dedurre l'attributo sensibile stesso (Andrus et al. 2021).

soluzione utilizzabile, senza implicare complessità significative o costi proibitivi (Centre for Data Ethics and Innovation e Department for Science, Innovation and Technology 2023; Ashurst e Weller 2023). In particolare, un'ampia gamma di strumenti proxy è attualmente disponibile come open source [ad esempio pre- dictrace (Kaplan 2023), ZRP (ZestAI 2024)] o prodotti commerciali (Namsor 2024), con metodologie proxy per la razza e l'etnia come la Bayesian Improved Surname Geocoding (BISG) già ampiamente utilizzata nel contesto statunitense per scopi di equità (Elliott et al. 2009; Consumer Finan- cial Protection Bureau 2014). Allo stesso modo, la ricerca accademica ha suggerito con enfasi che i dati sensibili, e in particolare quelli relativi all'etnia, possono essere *facilmente* dedotti da una serie di altri dati apparentemente innocui, come l'ubicazione e il cognome, le note dei medici o persino i like di Facebook (Solove 2024; Kosinski, Stillwell e Graepel 2013).

Tuttavia, la questione se i metodi di correttezza proxy stiano anche stimando *in modo affidabile* l'attributo sensibile pertinente appare più controversa. Diversi commentatori (Wachter e Mittelstadt 2018; Gola e Heckmann 2022; Finck 2021) sottolineano che i dati proxy ordinari dovrebbero essere riclassificati solo se forniscono una base affidabile o statisticamente significativa per la generazione di inferenze, sebbene la stessa Corte di giustizia europea non specifichi alcun livello di soglia richiesto per la "capacità" di rivelazione dei dati proxy.[10]Nel contesto esaminato della Proxy Fairness, sebbene i metodi proxy attualmente disponibili o impiegati spesso (auto)segnalino un alto livello di accuratezza nel pre-determinare l'attributo sensibile in questione, i tassi di accuratezza riportati si rivelano spesso inaffidabili e incoerenti tra i gruppi de- mografici, soprattutto in presenza di deriva concettuale e del modello (Centre for Data Ethics and Innovation e Department for Science, Innovation and Technology 2023; LeClair, Parker, and Young 2023). Nel complesso, la valutazione della Proxy Fairness rispetto al fattore dell'affidabilità condizionerebbe la "sensibilità" dei dati proxy al livello di accuratezza del classificatore di attributi: maggiore è l'accuratezza predittiva del classificatore di attributi, migliore è l'indicazione della sensibilità dei dati.

Dati desunti

Anche il secondo blocco di dati coinvolto nella Proxy Fairness, ovvero le inferenze generate, è stato oggetto di un intenso dibattito accademico, soprattutto a causa della loro natura *artificiale* e *probabilistica*. Dato che le inferenze sono sottoprodotti algoritmici di un processo analitico su altri dati (proxy), piuttosto che direttamente osservati o raccolti dal soggetto interessato, sollevano la seguente questione preliminare: devono essere considerati dati personali ai sensi del GDPR, distinti dai dati da cui sono stati dedotti?

Il GDPR definisce i dati personali come "*qualsiasi informazione*. . . " e le linee guida disponibili del Gruppo di lavoro articolo 29 (Gruppo di lavoro articolo 29 per la protezione dei dati 2007, 2017, 2016) richiedono un'interpretazione ampia che non è legata a un tipo o a una forma specifica di informazione e può quindi includere non solo i dati raccolti ma anche quelli desunti.[11]

Anche la Corte di giustizia europea ha affrontato indirettamente la questione (Nowak 2017), estendendo l'ambito di applicazione dei dati personali alle informazioni "oggettive" e "soggettive", come opinioni o valutazioni, purché si riferiscano a soggetti interessati. Attraverso l'analogia giuridica e assimilando le inferenze a opinioni o tipi di valutazioni soggettive, diversi studiosi (Hallinan e Zuiderveen Borgesius 2020; Wachter e Mittelstadt 2018) hanno utilizzato questa giurisprudenza per includere le inferenze generate da algoritmi nell'ambito di applicazione del GDPR. Di conseguenza, le inferenze generate nel contesto della Proxy Fairness possono essere viste come "opinioni" probabilistiche, che emergono sulla base di serie di dati fattuali (Proxy Data) sottoposti a un quadro analitico e interpretativo (l'algoritmo di classificazione) al fine di generare nuove conclusioni probabili sugli individui rappresentati nelle serie di dati (ad esempio, la loro etnia). In quanto tali, rientrerebbero nell'ambito di applicazione del GDPR dei dati personali in base alla giurisprudenza citata. Inoltre, secondo l'approccio seguito dalla CGUE e dal Gruppo di lavoro articolo 29, le inferenze generate nel contesto della Proxy Fairness si riferirebbero agli interessati in virtù del loro "contenuto"[12], rappresentando informazioni *sui* soggetti interessati (ad esempio la loro origine etnica) e descrivendoli direttamente.[13]

È importante notare che le deduzioni generate dovrebbero essere classificate come dati personali sensibili a prescindere dalla loro accuratezza o validità, ossia dal fatto che prevedano accuratamente l'effettiva origine etnica dell'interessato. Come affermato dal Gruppo di lavoro articolo 29 (Gruppo di lavoro articolo 29 sulla protezione dei dati 2007), non è necessario che le informazioni siano vere o provate per essere considerate dati personali ai sensi del regolamento sulla protezione dei dati, evidenziando che le norme sulla protezione dei dati prevedono già la possibilità di informazioni errate. Ciò è particolarmente rilevante nel contesto della Proxy Fairness, in quanto le inferenze previste si basano su correlazioni probabilistiche piuttosto che sulla causalità, mancando spesso di validità scientifica.[14]Infine, includendo esplicitamente le informazioni soggettive sotto forma di opinioni o valutazioni ai sensi del GDPR, la Corte di giustizia europea ha anche indicato che i dati personali non si riferiscono solo a fatti veri o verificabili. Di conseguenza, l'utilizzo di dati desunti da modelli proxy per valutare o correggere pregiudizi costituirebbe un trattamento di dati sensibili, indipendentemente dall'accuratezza del modello proxy.

[10]Cfr. (Wachter e Mittelstadt 2018), che interpreta la decisione del Tribunale nella causa (Egan e v. Parlamento europeo 2012) come un'affermazione che l'affidabilità è un attributo essenziale dei dati sensibili. [11]Si veda anche la dichiarazione del Gruppo di lavoro "Articolo 29" nel contesto dei dati biometrici: "*l'estrazione di informazioni dai campioni è una raccolta di dati personali, a cui si applicano le norme della direttiva* (art. 29 WP 2012)".

[12]Secondo il triplice approccio seguito dal Gruppo di lavoro "Articolo 29" (Gruppo di lavoro "Articolo 29" per la protezione dei dati personali 2007) e dalla giurisprudenza della Corte di giustizia europea, le informazioni possono "riferirsi" a un individuo "in ragione del loro contenuto, della loro finalità o del loro effetto". A questo proposito, gli at- tributi dedotti nel contesto della Proxy Fairness differiscono da altri tipi di inferenze, ad esempio quelle generate nel contesto della profilazione, che si riferiscono agli interessati in virtù delle loro "finalità" o "effetti" (Wachter e Mittelstadt 2018).

[13]Questo legame può essere contestato negli approcci in cui le inferenze sono generate a livello aggregato di gruppo (ad esempio, "l'X per cento di questo set di dati è costituito da donne/nativi") e non sono collegate a un individuo (Centre for Data Ethics and Innovation e Department for Science, Inno- vation and Technology 2023; Andrus et al. 2021).

[14]Si veda la sezione "Ragionevolezza".

modello in questione.

Infine, in termini di sensibilità, sembra chiaro che le inferenze tratte per valutazioni o interventi di equità sarebbero di natura sensibile quando dis- chiudono direttamente o rappresentano intrinsecamente una categoria di dati che, in quanto tale, è altamente protetta dal GDPR, come nel caso dell'"etnia".

Equità della delega ai sensi della legge sull'AI: una prospettiva di necessità

La sezione precedente ha analizzato il modo in cui la Proxy Fairness entra in relazione con il Regolamento europeo sulla protezione dei dati, soffermandosi in particolare sul regime speciale del Regolamento per i dati sensibili. Nella misura in cui comportano il trattamento di dati personali sensibili per l'individuazione e la correzione di errori, gli approcci di proxy fairness rientrano nell'ambito di applicazione dell'articolo 10 del Regolamento.

(5) della legge sull'AI.

Come già menzionato, questa disposizione delinea un'eccezione al divieto di trattamento dei dati personali sensibili previsto dal GDPR, aprendo questa possibilità ai fini della rilevazione e della correzione dei pregiudizi. Sebbene questa eccezione si applichi a qualsiasi tipo di metodo di correttezza che comporti il trattamento di dati sensibili, sia esso Default o Proxy, il requisito di necessità prescritto influisce in modo significativo sulla scelta del metodo di correttezza in un caso specifico. In particolare, ai sensi dell'art. 10 (5) della legge sull'AI, il metodo proxy ha un impatto significativo sulla scelta del metodo di correttezza in un caso specifico. 10 (5) della legge sull'IA, il trattamento di dati sensibili è consentito solo "nella misura in cui è strettamente *necessario* per garantire l'individuazione e la correzione dei pregiudizi negativi in relazione ai sistemi di IA ad alto rischio [enfasi aggiunta]", ossia solo in base al requisito della necessità.

Il principio di necessità, che è stato una con- dizione ricorrente al trattamento dei dati personali, stabilisce essenzialmente che il trattamento dei dati è consentito solo nella misura in cui non sia disponibile un'alterna- tiva *meno intrusiva* ma *altrettanto efficace*, che *possa ragionevolmente* raggiungere l'obiettivo in questione (Garante europeo della protezione dei dati 2023; Schantz e Wolff 2017). Secondo la giurisprudenza della Corte di giustizia europea (Meta Platforms e altri 2023; TK contro Asocia¸tia de Proprietari bloc M5A-ScaraA 2018), deve essere interpretato in modo da riflettere pienamente l'obiettivo del regolamento sulla protezione dei dati e, cosa importante, in combinazione con il principio *di minimizzazione dei dati* sancito dall'articolo 5, paragrafo 1, lettera c), del GDPR.

A questo proposito, i fornitori di IA che intendono avvalersi dell'eccezione prevista dalla legge sull'IA e trattare dati personali sensibili per l'individuazione e la correzione di errori devono effettuare un test di necessità, che prevede il confronto delle alternative disponibili in base ai livelli di a) *intrusività*, b) *efficacia* e c) *ragionevolezza*. In particolare, ai sensi dell'articolo 10 (5) (f) della versione finale della legge sull'IA, i fornitori di IA sono esplicitamente tenuti a redigere una giustificazione specifica nell'ambito della conservazione dei dati, in cui spiegano che l'operazione di trattamento era conforme al principio di necessità. Di conseguenza, la sola utilità di un metodo di equità non può giustificare il trattamento di dati personali sensibili.

Condurre questo test di necessità è un compito estremamente complesso per i fornitori di IA, in quanto richiede non solo una comprensione completa dello stato dell'arte della Fair-AI, ma anche l'applicazione di nozioni giuridiche vaghe e aperte. Questa complessità è intensificato dal fatto che i metodi per l'individuazione e la correzione dei bias sono ancora agli albori, con applicazioni limitate nel mondo reale, ma in rapida evoluzione. Di conseguenza, ciò pone delle sfide per la conduzione di una valutazione rigorosa e basata sull'evidenza della disponibilità, dell'efficacia e dei compromessi dei diversi metodi di equità per un caso specifico.

D'altra parte, dato che il requisito della necessità è intrinsecamente legato alla legittimità del trattamento dei dati, un'interpretazione e un'applicazione errate rischiano non solo di ledere il diritto fondamentale degli interessati alla protezione dei dati, ma anche di esporre i fornitori di IA a gravi conseguenze, che vanno dal danno alla reputazione a pesanti sanzioni finanziarie.

I paragrafi seguenti esaminano gli approcci di correttezza per procura in base al requisito della necessità, in particolare confrontandoli con gli approcci predefiniti che col- legano e utilizzano direttamente gli attributi sensibili "reali"[15], lungo gli assi di necessità dell'intrusività, dell'efficacia e della ragionevolezza.

Dato che l'applicazione del requisito di necessità è intrinsecamente dipendente dal contesto e comporta un livello di discrezionalità da parte dei fornitori di IA, questo contributo non può fornire risposte convincenti in merito alla legittimità complessiva dei metodi esaminati. Invece, facendo luce su aspetti sfumati del requisito della necessità, mira a sostenere il trattamento lecito di dati personali sensibili nel contesto dell'equità e a fornire a professionisti, ricercatori e fornitori di IA gli strumenti necessari per interpretare e applicare la necessità in vari contesti di equità.

Intrusività

Per aderire al principio di necessità, i fornitori di IA devono innanzitutto valutare se l'obiettivo di equità desiderato possa essere raggiunto con mezzi meno intrusivi, ossia con mezzi che interferiscono meno con i diritti di protezione dei dati (Gola e Heckmann 2022). Alcuni dei criteri ritenuti rilevanti per valutare la gravità dell'interferenza includono, tra l'altro, il *volume* e il *tipo di* dati trattati, le *possibilità di collegamento* e i *rischi di abuso dei dati* (Schantz e Wolff 2017).

Poiché i metodi di proxy fairness richiedono l'elaborazione delle caratteristiche "correlate" degli individui (ad esempio, l'indirizzo o il cognome) per dedurre l'attributo mancante di interesse (ad esempio, l'etnia), essi comportano *de facto* l'elaborazione di un volume maggiore di dati rispetto agli approcci predefiniti, aggiungendo di fatto il "blocco di dati" extra delle informazioni proxy. Inoltre, nella misura in cui i dati proxy e le inferenze generate sono classificati come dati sensibili ai sensi del GDPR[16], la Proxy Fairness implica *de jure* il trattamento di un maggior numero di dati di natura sensibile, aumentando così la severità in base al criterio del tipo di dati.

Risultati simili emergono quando si considera il principio della minimizzazione dei dati, che implica, tra l'altro, la minimizzazione del numero di dati e del loro utilizzo nella misura più ampia possibile (Gola e Heckmann 2022). In particolare, questo principio mira a ridurre non solo la quantità di dati, ma piuttosto la relazione con i dati stessi.

[15]Per attributi sensibili "reali" il documento si riferisce a quelli ottenuti direttamente o indirettamente dagli interessati, in contrapposizione a nuovi dati desunti da dati già disponibili.

[16]Si veda la sottosezione "Dati non ufficiali".

dei dati a una persona fisica, ossia la facilità con cui i dati possono essere collegati a un individuo (Spiecker et al. 2023; Panel for the Future of Science and Technology, EPRS - European Parliamentary Research Service, Scientific Fore- sight Unit (STOA) 2020). Di conseguenza, si può affermare che la raccolta e l'elaborazione di varie categorie di dati relativi a un soggetto nel contesto della Proxy Fairness aumentano non solo le possibilità di collegamento dei dati, ma anche, in misura maggiore, la facilità di identificazione del soggetto, con alcuni tipi di dati proxy comuni, come i cognomi, particolarmente identificativi.

Finora risulta evidente che il risultato della valutazione dell'intrusività in un caso specifico dipende fortemente dal tipo di modello utilizzato e in particolare dal numero e dal tipo di caratteristiche di input che richiede per la stima degli attributi e/o dell'equità. A questo proposito, gli approcci che studiano l'equità con l'uso di una piccola quantità di proxy e soprattutto con proxy "deboli" (cioè imprecisi) (Zhu et al. 2023) diventano sempre più rilevanti.

Infine, spostando l'attenzione sul criterio dei *rischi di abuso* si ottengono risultati in qualche modo contraddittori. L'uso improprio dei dati descrive tipicamente i casi in cui i dati raccolti vengono utilizzati illegalmente per altri scopi, in particolare in modi che potrebbero danneggiare le persone. Un rischio, invece, descrive l'esistenza della possibilità che si verifichi un evento con conseguenze dannose (van Dijk, Gellert e Rommetveit 2016). In particolare, ai sensi del GDPR, la nozione di rischio comprende due dimensioni, la *gravità* del danno ai diritti degli interessati e la *probabilità* che l'evento dannoso e il danno si verifichino (Considerando 75 GDPR, (DSK Datenschutzkonferenz 2018)). Confrontare i metodi Proxy e De- fault in termini di rischi di danno non è un esercizio facile, dato che ogni approccio comporta compromessi e rischi unici per i diritti fondamentali.

Innanzitutto, il considerando 75 del GDPR identifica i casi in cui agli interessati viene impedito di esercitare il controllo sui propri dati personali come potenziali casi di uso dannoso dei dati che comportano rischi per i loro diritti e libertà. Allo stesso modo, il considerando 7 del GDPR afferma che di fronte ai rapidi sviluppi tecnologici "le persone fisiche dovrebbero avere il controllo dei propri dati personali". Il trattamento dei dati nel contesto della Fairness basata sull'inferenza è intrinsecamente più soggetto a tali rischi dannosi, in quanto nasconde la partecipazione e il controllo degli interessati sulla generazione di informazioni relative ad aspetti sensibili della loro identità. In particolare, il trattamento di caratteristiche identitarie sensibili come la razza, la religione o l'orientamento sessuale come qualità che possono essere previste dall'esterno, produce nuove forme di controllo sulla capacità di un individuo di definire se stesso e, in misura maggiore, sulla sua autonomia (cfr. Keyes 2018). Di conseguenza, l'applicazione della correttezza per delega invade maggiormente il diritto alla privacy e alla protezione dei dati, in particolare i loro razionali di protezione dell'*autonomia* e dell'*autodeterminazione informativa*, rispettivamente. In effetti, sebbene gli studiosi di diritto siano in disaccordo su molti aspetti della privacy, vi è un maggiore consenso sul fatto che la privacy svolga un ruolo importante nella protezione dell'identità e dell'autonomia di un individuo e che l'autodeterminazione informativa costituisca una motivazione fondamentale del GDPR (Puri 2021; Thouvenin 2021).

D'altra parte, il considerando 75 del GDPR descrive anche i casi di discriminazioni come potenziali usi dannosi dei dati sensibili, mentre il GDPR stesso, e in particolare il suo regime di dati sensibili, è volto a proteggere il diritto fondamentale di non discriminazione (Spiecker et al. 2023). È evidente che gli abusi dei principi di non discriminazione rappresentano un rischio significativo di uso improprio legato al trattamento dei dati sensibili. Numerosi esempi testimoniano il modo in cui la raccolta e l'archiviazione dei dati sulla popolazione hanno facilitato le violazioni dei diritti umani e il perseguimento di vari gruppi sulla base di classificazioni razziali o religiose (Seltzer e Anderson 2001). A questo proposito, e facendo riferimento alla dimensione della "probabilità" contenuta nella nozione di rischio del GDPR, potremmo sostenere che gli approcci predefiniti comportano maggiori rischi per i diritti di non discriminazione. In particolare, dato che un rapido riferimento ai dati di appartenenza al gruppo facilita l'uso improprio e che i dati intrinsecamente sensibili consentono di fatto un riferimento più rapido rispetto ai proxy, si può ragionare sul fatto che la raccolta e la conservazione diretta dei dati intrinsecamente sensibili comportano un rischio più elevato di uso improprio rispetto a quello dei dati proxy. Per approfondire, anche se i dati proxy sono in grado di rivelare attributi sensibili e come tali devono essere altamente protetti, non va trascurato il fatto che è comunque necessario un ulteriore livello di deduzione. Ciò implica un certo livello di sforzo inferenziale e di infrastruttura in caso di accesso non autorizzato o di fuga di dati, che riduce la probabilità di un loro uso improprio a fini di discriminazione razziale.[17]

Centrare la valutazione della necessità su rischi e danni concreti piuttosto che sui tipi di dati è in linea con le voci sempre più numerose degli studiosi che sostengono un approccio alla protezione dei dati sensibili orientato al rischio e contestuale (Solove 2024; Ohm 2015; Simitis, Hornung e Spiecker 2019). Daniel Solove (Solove 2024), ad esempio, sostiene che *per essere efficace, la legge sulla privacy deve concentrarsi sull'uso, sul danno e sul rischio piuttosto che sulla natura dei dati personali*, mentre il *danno e il rischio dipendono dalla situazione* e *raramente possono essere determinati al di fuori di un contesto.*[18]

Mentre questa ricerca ha finora esplorato il rischio e il testo come mezzi per valutare vari gradi di sensibilità dei dati, il presente contributo si concentra sulla nozione di rischio come mezzo per valutare l'intrusività del trattamento dei dati e, in misura maggiore, la sua necessità. Grazie alla sua natura intrinsecamente flessibile e contestuale, il concetto di necessità fornisce un quadro adeguato per incorporare fattori circostanziali e legati al rischio, senza interferire con l'approccio consolidato della Corte di giustizia europea relativo alla sensibilità "naturale" di determinate categorie di dati.

Infine, sebbene i rischi e i danni associati alla perdita di autonomia non siano trascurabili, si potrebbe sostenere che concentrarsi sui rischi di discriminazione sia più in linea con le motivazioni alla base degli articoli 10 (5) della legge sull'AI e 9 (2) del GDPR, nonché con la sistematica della necessità del GDPR. In particolare, anche

[17]A questo proposito è fondamentale il tipo di misure tecniche e organizzative adottate per salvaguardare la sicurezza dei dati, tra cui, ad esempio, la segregazione degli insiemi di dati proxy da quelli desunti.

[18]Analogamente, secondo (Spiecker et al. 2023), la natura intrusiva di alcuni dati personali è situazionale e la definizione di un insieme di casi che hanno diritto a un livello di protezione più elevato è una dimensione della realtà sociale.

Sebbene l'accresciuta protezione dei dati sensibili di cui agli articoli 9 del GDPR e 10 (5) dell'AI Act miri a colpire i rischi per tutti i diritti fondamentali, tali rischi sono prevalentemente percepiti in termini di un'elevata probabilità di discriminazione.[19]Inoltre, mentre la maggior parte degli studiosi concorda sul fatto che l'autodeterminazione informativa è una logica del GDPR, essa non si riflette di per sé in modo prominente nel concetto di necessità. In particolare, la necessità funge da requisito in questi motivi di trattamento dei dati che si basano meno sull'esercizio dell'autodeterminazione informativa, come l'interesse pubblico o il legittimo interesse del responsabile del trattamento, rispetto a quelli basati sul consenso o sulla pubblicazione dei dati da parte dell'interessato stesso.[20]

Efficacia

Il rispetto del requisito della necessità non richiede di dare priorità a qualsiasi tipo di alternativa più blanda, ma solo a quelle alternative più blande che possono raggiungere l'obiettivo perseguito in modo comparabilmente efficace.

In una seconda fase, i fornitori di IA devono confrontare le alternative identificate rispetto alla loro efficacia nel rilevare e correggere i bias, basandosi su prove teoriche e/o em- piriche relative all'utilità e ai limiti dei metodi di equità presi in considerazione. Ciò include argomentazioni qualitative e quantitative sul modo in cui i gruppi de- mografici rilevanti sarebbero meglio serviti dall'in- tervento pianificato, come le metriche di performance e di equità, l'accuratezza delle stime di equità e i relativi compromessi. Le considerazioni basate esclusivamente sulla convenienza o sull'efficacia in termini di costi operativi o di risorse organizzative non soddisfano il criterio di efficacia della necessità. Nel contesto della presente analisi, la considerazione dell'efficacia induce a un confronto di alto livello tra gli approcci De- fault e Proxy Fairness sulla base delle evidenze teoriche ed empiriche presentate in letteratura in merito alla loro efficacia nel rilevare e correggere i pregiudizi (razziali).

Da un lato, semplici metodi proxy sono già stati impiegati in diversi ambiti[21], dimostrando così, in termini pratici, un livello di efficacia nella valutazione e persino nella contabilizzazione delle disparità razziali, senza raccogliere o basarsi direttamente su dati etnici reali (Bogen, Rieke e Ahmed 2020; Andriotis e Ensign 2015). Ricerche scientifiche come quella di (Diana et al. 2022) hanno anche dimostrato che è possibile addestrare in modo efficiente dei proxy che possono sostituire le caratteristiche sensibili mancanti per addestrare in modo efficace i classificatori a valle soggetti a una varietà di condizioni di equità demografica.

[19]Si veda il Considerando 71 del GDPR; le Linee guida per la regolamentazione degli archivi personali computerizzati (Joinet, on Prevention of Discrimi- nation, and of Minorities. Special Rapporteur on the Study of the Relevant Guidelines in the Field of Computerized Personal Files 1988), che giustificano un'ulteriore protezione per i dati sensibili sulla base della premessa che tali dati sono "suscettibili di dare luogo a discriminazioni illegali o arbitrarie"; anche (Ku¨hling e Buchner 2020), che vede l'articolo 9 del GDPR come una specificazione normativa dell'articolo 21 della Carta dei diritti fondamentali, cioè il diritto fondamentale alla non discriminazione, in contrapposizione a (Albers e Veit 2020), che trovano troppo stretta questa riduzione teleologica. Per un approfondimento sulle motivazioni della protezione dei dati sensibili si rimanda a (Quinn e Malgieri 2020).

[20]Cfr. artt. 6 (1) e 9 (2) GDPR e (Thouvenin 2021).

[21]Cfr. l'introduzione del documento.

tensioni. Inoltre, è stato sostenuto (Andrus et al. 2021), che i dati demografici desunti potrebbero talvolta offrire approfondimenti più obiettivi e accurati rispetto ai dati demografici auto-riportati o etichettati, rendendoli più efficaci per compiti specifici di rilevamento dei pregiudizi, come nei casi di pregiudizi legati alla razza percepita. Tuttavia, l'accuratezza e l'utilità dei metodi proxy comuni, che si basano su una serie di ipotesi, è stata fortemente contestata nella ricerca accademica a causa della loro tendenza a sovrastimare le disparità demografiche e a introdurre errori e pregiudizi sistematici (Ashurst e Weller 2023; Chen et al. 2019a; LeClair, Parker e Young 2023; Imai et al. 2023).

D'altra parte, l'uso diretto di dati demografici reali, purché di alta qualità, offre evidentemente le stime più accurate delle metriche di equità di gruppo, consentendo un'analisi più approfondita delle disparità di gruppo (Ashurst e Weller 2023).[22]La ricerca contemporanea suggerisce che i metodi di equità che si basano su proxy, anche se ben formati, dovrebbero comunque essere considerati una soluzione secondaria in termini di efficienza (Ashurst e Weller 2023; Chen et al. 2019b). Ciò sembra plausibile poiché i metodi proxy, per definizione, si sforzano di riprodurre l'alternativa predefinita in contesti vincolati in cui i dati rilevanti sono inaccessibili o non disponibili in tutte le fasi della pipeline di sviluppo e implementazione.

Per concludere, il confronto dell'efficacia, come l'intero test di necessità, deve essere condotto caso per caso e l'esito varierà inevitabilmente a seconda della metrica o del concetto di equità tar- getato, nonché del tipo di modello e del settore in esame.

Ragionevolezza

Infine, la necessaria valutazione delle alternative è condizionata da un terzo elemento, quello della ragionevolezza per il responsabile del trattamento dei dati. Questo aspetto della necessità è esplicitamente previsto dal considerando 39 del GDPR, che afferma che "i dati personali devono essere trattati solo se la finalità del trattamento non può essere *ragionevolmente* soddisfatta con altri mezzi [enfasi aggiunta]".

Per approfondire, secondo l'opinione prevalente (Schantz e Wolff 2017; Information Commissioner's Office Ac- cessed 2023-11-11), la necessità non implica che il trattamento dei dati in questione sia assolutamente indispensabile a causa di ragioni tecniche o di altro tipo per raggiungere lo scopo in questione, né suggerisce che esso diventi irraggiungibile senza di esso. Implica "unicamente" che non sia disponibile un'alternativa efficace e più mite che sia *ragionevole* in termini di fattibilità personale, operativa o finanziaria (Gola e Heckmann 2022). In particolare, nulla di praticamente impossibile, proibitivo o illegale può essere richiesto ai fornitori di IA per soddisfare la necessità. Il criterio della ragionevolezza sembra in gran parte adattarsi a considerazioni di utilità, tra cui, nel caso della Proxy Fairness, la facilità o la fattibilità della sua implementazione insieme ai costi e alle risorse organizzative associate.

[22]Tuttavia, i dati autodichiarati sono anche potenzialmente soggetti a inac- curia (Andrus et al. 2021), limitando così l'efficacia della Default Fairness.

In primo luogo, i metodi proxy sono spesso rappresentati come l'unica risorsa per i professionisti che cercano di offrire informazioni quantitative per un'analisi approfondita, data l'indisponibilità e l'irrecuperabilità degli attributi sensibili "reali" all'interno dei set di dati esistenti. La correttezza dei proxy, invece, si basa potenzialmente sui dati già in possesso dei fornitori di IA, eliminando così il tempo e le risorse necessarie per raccogliere direttamente dati demo- grafici sensibili. A questo proposito, i proxy offrono un'alternativa relativamente semplice e implementabile, rendendo la Proxy Fairness una strategia immediatamente fattibile, praticamente utilizzabile e a basso costo che può facilitare vari tipi di equità algoritmica (Andrus e Villeneuve 2022; Centre for Data Ethics and In- novation e Department for Science, Innovation and Tech- nology 2023).[23]Ciò pone la questione critica se
- a condizione che sia meno invasivo - sarebbe anche ragionevole imporre ai fornitori di IA di raccogliere dati demografici "da zero", direttamente dagli interessati, nonostante i possibili disagi, il consumo di tempo o di risorse e la potenziale conseguente obsolescenza di un ampio corpus di set di dati esistenti.

Tuttavia, richiedendo una "*stretta*" necessità, l'Articolo 10
(Il comma 5 della legge sull'AI innalza la soglia legale di necessità nel caso di rilevamento e correzione di errori, superando lo standard del GDPR per altri casi di interesse pubblico (cfr. art. 9, comma 1, lettera g)).
). Questa condizione aggiuntiva potrebbe influire sull'approccio e sul rigore della valutazione della necessità, suggerendo una necessità imperativa ab- soluto per il trattamento di dati sensibili a fini di equità - in un senso di "conditio sine qua non" - o uno standard meno rigoroso quando si esplorano opzioni alternative. Per soddisfare la necessità nel primo scenario sarebbe necessario che sia assolutamente impossibile, sulla base dello stato dell'arte, garantire l'individuazione e la correzione dei pregiudizi senza il trattamento di dati personali sensibili, sia per inferenza che per uso diretto.

Tuttavia, la valutazione della ragionevolezza non deve essere ridotta a un calcolo di utilità netta. Come già detto, le alternative illegali ovviamente non soddisfano il criterio di ragionevolezza. Secondo l'autore, lo stesso dovrebbe valere*, mutatis mutandis*, per le alternative non etiche. In particolare, gli approcci basati sull'inferenza hanno raccolto critiche significative da parte degli etici dell'IA a causa delle preoccupazioni sulla loro mancanza di validità scientifica e delle implicazioni etiche associate alla previsione di qualità dell'identità intrinsecamente sensibili. Questi tipi di approcci sono stati criticati, tra l'altro, perché assomigliano alla fisiognomica (Engelmann et al. 2022), riecheggiano pratiche coloniali (Scheuerman, Pape e Hanna 2021) e riducono notevolmente l'agenzia e l'identità degli individui. Tali preoccupazioni etiche sono sostenute dalla maggior parte delle teorie filosofiche contemporanee sull'identità personale, che sostengono l'idea che "*la libertà di interpretazione di sé è un elemento costitutivo dei confini concettuali dell'identità personale*" (Engelmann e Grossklags 2019).

In base alla valutazione della necessità, questa linea di ricerca sull'etica critica può assumere una dimensione normativa, dove la generazione di inferenze moralmente discutibili, come la deduzione dell'orientamento sessuale dai tratti del viso, potrebbe essere considerata un trattamento irragionevole di dati sensibili ai sensi del GDPR e dell'AI Act. Allo stesso modo, gli approcci predefiniti che raccolgono dati personali sensibili sulla base dell'articolo 10 (5) dell'AI ACT, anche se con mezzi non etici, non soddisfano il criterio di ragionevolezza.

[23]Tuttavia, la correzione dei bias con l'uso di metodi proxy più complessi o con l'ascrizione manuale implica in genere competenze aggiuntive e risorse eccessive, soprattutto nel caso di grandi insiemi di dati (Andrus e Villeneuve 2022; Ashurst e Weller 2023).

Conclusione

A fronte della crescente popolarità degli approcci di proxy fairness nell'ambito della Fair-AI e della mancanza di un quadro giuridico completo corrispondente, il presente lavoro ha esplorato gli aspetti della Proxy Fairness nell'ambito del General Data Protection Regu- lation e dell'AI Act. Le nozioni giuridiche di "sensibilità" e "necessità" hanno fornito il quadro concettuale per l'analisi.

Basandosi sull'articolo 9 (1) del GDPR, il documento ha analizzato la natura dei dati coinvolti nella Proxy Fairness, facendo luce sugli aspetti più sfumati della protezione dei dati che riguardano le variabili proxy e le inferenze sui dati. In questo modo, ha dimostrato che i metodi inferenziali non sono in linea di principio esenti dalla portata del GDPR e del suo ampio regime per i dati sensibili. Successivamente, il documento ha esaminato la liceità del trattamento dei dati sensibili per la Proxy Fairness ai sensi dell'articolo 10 (5) dell'AI Act. Applicando il requisito della necessità e confrontando gli approcci di Proxy Fairness con quelli di Default Fairness in relazione alla loro intrusività, efficacia e ragionevolezza, il documento ha approfondito gli aspetti fondamentali della legittimità degli approcci di rilevamento e correzione dei pregiudizi che trattano dati sensibili.

Garantire la conformità legale e al contempo navigare nel panorama dell'equità, in particolare quando sono in gioco dati personali sensibili, rappresenta evidentemente un compito impegnativo per i fornitori e gli operatori dell'IA. Tuttavia, rimane di estrema importanza, data la minaccia incombente di sanzioni normative e, in particolare, i rischi per i diritti fondamentali delle persone. Facendo luce sulle sfumature normative coinvolte nella Proxy Fairness e guidando la liceità del trattamento dei dati sensibili in questo contesto, il presente documento ha cercato di assistere i fornitori di IA nella conformità normativa e di salvaguardare i diritti di protezione dei dati degli interessati. Infine, ma non per questo meno importante, l'analisi condotta ha gettato le basi per ulteriori ricerche scientifiche sull'intersezione tra legge sulla protezione dei dati, etica e Fair-AI, che vanno oltre una concettualizzazione avversaria della correttezza rispetto alla privacy.

Ringraziamenti

Questo lavoro è stato finanziato dal programma di ricerca e innovazione Horizon 2020 dell'Unione Europea nell'ambito delle azioni Marie Skłodowska-Curie (accordo di sovvenzione numero 860630) per il progetto: NoBIAS - Artificial Intelligence without Bias. Inoltre, questo lavoro riflette solo il punto di vista dell'autore e l'Agenzia esecutiva per la ricerca europea (REA) non è responsabile dell'uso che può essere fatto delle informazioni in esso contenute.

Riferimenti

Alao, R.; Bogen, M.; Miao, J.; Mironov, I.; e Tannen, J. 2021. Come Meta sta lavorando per valutare l'equità in relazione alla razza negli Stati Uniti attraverso i suoi prodotti e sistemi.

Albers, W. e Veit, B. 2020. *BeckOK DatenschutzR*. C.H.BECK Mu¨nchen 2020, 37 edizione. DS-GVO Art. 9 Rn. 4, beck-online.

Andriotis, A. e Ensign, R. L. 2015. Il governo degli Stati Uniti utilizza il test della razza per pagamenti da 80 milioni di dollari.

Andrus, M.; Spitzer, E.; Brown, J.; e Xiang, A. 2021. Ciò che non possiamo misurare, non possiamo capire: Sfide all'acquisizione di dati demografici nel perseguimento dell'equità. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, 249-260. New York, NY, USA: Association for Computing Machinery. ISBN 9781450383097.

Andrus, M. e Villeneuve, S. 2022. Equità algoritmica basata sulla demografia: Caratterizzare i rischi della raccolta di dati demografici nel perseguimento dell'equità. In *Pro- ceedings of the ACM Conference on Fairness, Accountabil- ity, and Transparency*. ACM. ISBN 978-1-4503-9352-2.

Art29WP, A. . W. P. 2012. Parere 3/2012 del Gruppo di lavoro sulla protezione dei dati dell'articolo 29 sugli sviluppi delle tecnologie biometriche.

Gruppo di lavoro articolo 29 sulla protezione dei dati. 2007. Parere sul concetto di dati personali.

Gruppo di lavoro articolo 29 sulla protezione dei dati. 2016. Linee guida sul diritto alla portabilità dei dati. *Gruppo di lavoro sulla protezione dei dati*, (16/EN): 9-11. In archivio presso la Columbia Business Law Review.

Gruppo di lavoro articolo 29 sulla protezione dei dati. 2017. Linee guida sul processo decisionale individuale automatizzato e sulla profilazione ai fini del Regolamento 2016/679. Documento n. 17/EN, WP251rev.01.

Ashurst, C. e Weller, A. 2023. Equità senza dati de- mografici: A Survey of Approaches. In *Proceeds of the 3rd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*. New York, NY, USA: Association for Computing Machinery. ISBN 9798400703812.

Awasthi, R.; Beutel, A.; Kleindessner, M.; Morgenstern, J.; e Wang, X. 2021. Valutazione della correttezza dei modelli di apprendimento automatico in presenza di informazioni incerte e incomplete. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 206-214.

Baines, A. P. e Courchane, M. J. 2014. Prestito equo: Implications for the indirect auto finance market. Studio, Associazione americana dei servizi finanziari.

Belli, L.; Yee, K.; Tantipongpipat, U.; Gonzales, A.; Lum, K.; e Hardt, M. 2021. Verifica algoritmica a livello di contea dei pregiudizi razziali nella home timeline di Twitter. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, 368-378.

Bogen, M.; Rieke, A.; e Ahmed, S. 2020. Consapevolezza nella pratica: tensioni nell'accesso ai dati sensibili degli attributi per l'an- tidiscriminazione. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, FAT* '20, 492-500. New York, NY, USA: Association for Computing Machinery. ISBN 9781450369367.

Centro per l'etica e l'innovazione dei dati e Dipartimento per la scienza, l'innovazione e la tecnologia. 2023. Consentire un accesso responsabile ai dati demografici per rendere più equi i sistemi di IA. Rapporto di ricerca e analisi. Pubblicato il 14 giugno 2023.

Chen, J.; Kallus, N.; Mao, X.; Svacha, G.; Udell, M. 2019a. Equità senza consapevolezza: Valutare la disparità quando la classe protetta non è osservata. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, 339-348. New York, NY, USA: Association for Computing Machinery. ISBN 9781450361255.

Chen, J.; Kallus, N.; Mao, X.; Svacha, G.; Udell, M. 2019b. Equità senza consapevolezza: Valutare la disparità quando la classe protetta non è osservata. In *FAT*, 339-348. ACM.

Ufficio di protezione finanziaria dei consumatori. 2014. Utilizzo di informazioni pubblicamente disponibili per fornire un'approssimazione alla razza e all'etnia non identificate: Una metodologia e una valutazione. Rapporto tecnico, Consumer Financial Protection Bureau.

Diana, E.; Gill, W.; Kearns, M.; Kenthapadi, K.; Roth, A.; e Sharifi-Malvajerdi, S. 2022. Proxy multiaccurati per la correttezza a valle. In *Proceedings of the ACM Conference on Fairness, Accountability, and Transparency (FAccT)*. Association for Computing Machinery. ISBN 978- 1-4503-9352-2.

DSK Datenschutzkonferenz. 2018. Risiko fu¨r die Rechte und Freiheiten natu¨rlicher Personen.

Egan, K.; e contro il Parlamento europeo, M. H. 2012. Egan e Hackett contro il Parlamento. ECLI:EU:C:2019:1064. Sentenza del Tribunale (Quinta Sezione) del 28 marzo 2012.

Elliott, M. N.; Morrison, b. A.; Fremont, A.; McCaffrey, D. F.; Pantoja, P.; e Lurie, N. 2009. Utilizzo dell'elenco dei cognomi del Census Bureau per migliorare le stime di razza/etnia e le disparità associate. *Health Services and Outcomes Research Methodology*, 9(2): 69-83.

Elzayn, H.; Smith, E.; Hertz, T.; Ramesh, A.; Goldin, J.; Ho, D. E.; e Fisher, R. 2023. Misurare e attenuare le disparità razziali nelle verifiche fiscali. Rapporto tecnico, Stanford Institute for Economic Policy Research (SIEPR).

Engelmann, S. e Grossklags, J. 2019. Preparare il terreno: Towards Principles for Reasonable Image Inferences. In *Adjunct Publication of the 27th Conference on User Mod- eling, Adaptation and Personalization*, UMAP'19 Adjunct, 301-307. New York, NY, USA: Association for Computing Machinery. ISBN 9781450367110.

Engelmann, S.; Ullstein, C.; Papakyriakopoulos, O.; e Grossklags, J. 2022. Cosa pensano le persone che l'intelligenza artificiale dovrebbe dedurre dai volti. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*. Seoul, Re- pubblica di Corea: ACM.

Garante europeo della protezione dei dati. 2023. Valutazione della necessità di misure che limitano il diritto fondamentale alla protezione dei dati personali: Un kit di strumenti.

Finck, M. 2021. *Insight personali nascosti e invischiati nel modello algoritmico: The Limits of the GDPR in the Personalisation Context*, 95-107. Cambridge University Press.

Gola, P.; e Heckmann, D. 2022. *Datenschutz- Grundverordnung, Bundesdatenschutzgesetz: DS-GVO / BDSG*. C.H. Beck, 3 edizione.

Gupta, M. R.; Cotter, A.; Fard, M. M.; e Wang, S. L. 2018. Equità dei proxy. *CoRR*, abs/1806.11212.

Hallinan, D. e Zuiderveen Borgesius, F. 2020. Le opinioni possono essere sbagliate! La nostra opinione. Sul principio di accuratezza nella legge sulla protezione dei dati. *Diritto internazionale della privacy dei dati*, ipz025.

Imai, K.; McCartan, C.; Goldin, J.; e Ho, D. E. 2023. Esaminare le disparità razziali quando la razza non è osservata.

Ufficio del Commissario per l'Informazione. Accesso al 2023-11-11. Guida alla base giuridica.

Joinet, L.; sulla prevenzione della discriminazione, Stati Uniti; e delle minoranze. Relatore speciale per lo studio delle linee guida pertinenti nel campo degli archivi personali computerizzati, P. 1988. Linee guida per la regolamentazione degli archivi personali computerizzati: Rapporto finale.

Kaplan, J. 2023. predictrace: Predictrace: Predict the Race and Gender of a Given Name Using Census and Social Security Administration Data. https://github.com/jacobkap/predictrace. https://jacobkap.github.io/predictrace/.

Kestemont, L. 2018. *Manuale di metodologia giuridica: Dall'obiettivo al metodo*. Intersentia Ltd. Pubblicato online da Cambridge University Press.

Keyes, O. 2018. Le macchine che sbagliano: Implicazioni Trans/HCI del riconoscimento automatico di genere. *Proc. ACM Hum.-Comput. Interact.*, 2(CSCW).

Kosinski, M.; Stillwell, D.; Graepel, T. 2013. I tratti e gli attributi privati sono prevedibili dalle registrazioni digitali del comportamento umano. *Atti dell'Accademia nazionale delle scienze*, 110(15): 5802-5805.

Ku¨hling, J.; e Buchner, B. 2020. *Datenschutz-Grundverordnung, Bundesdatenschutzgesetz: DS-GVO / BDSG Kommentar*. C.H. Beck, 3 edizione.

LeClair, B.; Parker, W.; e Young, A. 2023. Rapporto Frazer Nash - Algorithmic Bias: uno studio tecnico sulla fattibilità dell'uso di metodi proxy per il monitoraggio dei bias algoritmici in modo da preservare la privacy. Relazione tecnica, Frazer Nash Consultancy.

Malgieri, G. e Comande`, G. 2017. Sensitive-by-distance: i dati quasi sanitari nell'era degli algoritmi. *Information & Communications Technology Law*, 26(3): 229-249.

Mehrabi, N.; Morstatter, F.; Saxena, N.; Lerman, K.; Galstyan, A. 2021. Un'indagine su bias e correttezza nell'apprendimento manuale. *ACM Comput. Surv.*, 54(6).

Piattaforme Meta e altri. 2023. Sentenza della Corte (Grande Sezione) del 4 luglio 2023 (richiesta di pronuncia pregiudiziale dell'Oberlandesgericht Du sseldorf - Germania) - Meta Platforms Inc., già Facebook Inc. ECLI:EU:C:2023:537, par. 109.

Mitchell, S.; Potash, E.; Barocas, S.; D'Amour, A.; e Lum, K. 2021. Equità algoritmica: Scelte, ipotesi e definizioni. *Annual Review of Statistics and Its Application*, 8(1): 141-163.

Namsor. 2024. NamSor: Origine dei nomi. https://namsor. app/. Accesso: 2024-12-04.

Nowak. 2017. Sentenza della Corte (seconda sezione) del 20 dicembre 2017, Corte di giustizia dell'Unione europea C-434/16.

Ntoutsi, E.; et al. 2020. Bias in data-driven artificial intelli- gence systems - An introductory survey. *WIREs Data Min- ing Knowl. Discov.*, 10(3).

Ohm, P. 2015. Informazioni sensibili. *Southern California Law Review*, 88.

Panel per il futuro della scienza e della tecnologia, EPRS - Servizio europeo di ricerca parlamentare, Unità di previsione scientifica (STOA). 2020. L'impatto del Regolamento generale sulla protezione dei dati (GDPR) sull'intelligenza artificiale: Studio.

Puri, A. 2021. Una teoria della privacy di gruppo. *Cornell Journal of Law and Public Policy*, 30: 477-538.

Quinn, P. e Malgieri, G. 2020. La difficoltà di definire i dati sensibili - Il concetto di dati sensibili nel quadro di protezione dei dati dell'UE. *Rivista di diritto tedesco*. (In arrivo).

Schantz, P. e Wolff, H. A. 2017. *Il nuovo Daten- schutzrecht: Datenschutz-Grundverordnung e Bundes- datenschutzgesetz in der Praxis*. C.H.BECK.

Scheuerman, M. K.; Pape, M.; e Hanna, A. 2021. Autoessenzializzazione: Il genere nell'analisi facciale automatizzata come progetto coloniale esteso. *Big Data & Society*, 8(2): 20539517211053712.

Schiff, A.; Ehmann; e Selmayr. 2017. *Datenschutz-Grundverordnung DS-GVO Kommentar*. C.H.BECK, 3. edizione au- flage.

Schwartz, R.; Vassilev, A.; Greene, K.; Perine, L.; Burt, A.; e Hall, P. 2022. Verso uno standard per l'identificazione e la gestione dei pregiudizi nell'intelligenza artificiale. Rapporto tecnico 1270, Pubblicazione speciale del NIST.

Seltzer, W. e Anderson, M. 2001. Il lato oscuro dei numeri: Il ruolo dei sistemi di dati demografici nelle violazioni dei diritti umani. *Ricerca sociale*, 68(2): 481-513.

Simitis, Hornung e Spiecker. 2019. *Nomos Kommentar Datenschutzrecht DSGVO mit BDSG*. Nomos.

Smits, J. M. 2017. Che cos'è la dottrina giuridica? Sugli obiettivi e i metodi della ricerca giuridico-dogmatica. 207-228. Documento di lavoro dell'Istituto europeo di diritto privato di Maastricht n. 2015/06.

Solove, D. J. 2024. I dati sono ciò che i dati fanno: regolamentare in base al danno e al rischio invece che ai dati sensibili. *North Western University Law Review*, 118: 1081. Documento di ricerca sugli studi giuridici della GWU n. 2023-22, Documento di ricerca sul diritto pubblico della GWU Law School n. 2023-22.

Spiecker; Papakonstantinou; Hornung; e Hert, D. 2023. *Regolamento generale sulla protezione dei dati: Commento al GDPR articolo per articolo*. C.H.BECK. ISBN 978-3-406-74386-.

Thouvenin, F. 2021. Autodeterminazione informativa: Una motivazione convincente per la legge sulla protezione dei dati? *JIPITEC*, 12(4): 2021.

TK contro Asocia¸tia de Proprietari bloc M5A-ScaraA. 2018. Sentenza della Corte (Terza Sezione) dell'11 dicembre 2019. ECLI:EU:C:2019:1064, par. 47.

van Dijk, N.; Gellert, R.; e Rommetveit, K. 2016. Un rischio per un diritto? Oltre le valutazioni del rischio di protezione dei dati. *Computer Law and Security Review*, 32(2): 286-306.

Wachter, S. e Mittelstadt, B. 2018. Un diritto alle inferenze ragionevoli: Re-Thinking Data Protection Law in the Age of Big Data and AI. *Columbia Business Law Review*, 2019(2). 5 ottobre 2018.

Yan, S.; te Kao, H.; e Ferrara, E. 2020. Bilanciamento di classe equo: Enhancing Model Fairness Without Observing Sensitive Attributes. In *Proceedings of the 29th ACM*.

ZestAI. 2024. zrp: Zest Race Predictor. https://github.com/zestai/zrp. Accesso: 28 marzo 2024.

Zhu, Z.; Yao, Y.; Sun, J.; Li, H.; e Liu, Y. 2023. Le deleghe deboli sono sufficienti e preferibili per l'equità con attributi sensibili mancanti. In *Proceedings of the 40th In- ternational Conference on Machine Learning*, ICML'23. JMLR.org.