Rischi e opportunità dell'IA per migliorare equità e accessibilità

Report sessione di discussione gruppo 2

15 luglio 2025

Partecipanti presenti alla sessione

1. Glenda Gentili, AGID
2. Carmen Lancianese, Comune di Milano
3. Mirko Duradoni, Universitas Mercatorum
4. Mariella Pagliuca, PCM Dipartimento Pari Opportunità
5. Salvatore Condito (silente)

Staff Formez/Facilitazione

1. Iolanda Romano, Task force OGP
2. Francesca De Chiara, Esperta L1 Progetto OpenGov

Sviluppo della discussione

La facilitatrice introduce la sessione e dà lettura della domanda:

Come possiamo assicurare equità nei dati raccolti, in una fase antecedente allo sviluppo dell'algoritmo, per ridurre discriminazioni e stereotipi ?

Ciascun partecipante è invitato a contribuire alla discussione.

Segue una restituzione sintetica dei temi trasversali emersi dal confronto con attenzione alle diverse percezioni/priorità presentate, senza attribuzioni (ovvero senza riportare "chi ha detto cosa"). Laddove possibile sono segnalate le proposte che verificano un ampio livello di convergenza.

Restituzione del confronto

La discussione si avvia a partire dal "primo gradino" da considerare, che è la **raccolta dati disaggregata**, in modo da avere dati significativi per la categoria che vogliamo analizzare. In questo modo si evita la generalizzazione e si riesce ad essere più specifici sulle fasce di interesse: di genere, ma anche di età ed altre categorie. Il confronto prosegue affrontando il tema delle barriere che si incontrano nella raccolta dei dati: infatti queste dipendono dalle motivazioni per cui questi dati vengono raccolti. Si afferma infatti che generalmente le persone rilasciano i dati solo a fronte di certe motivazioni, mentre altre non lo fanno, per cui già in questa fase si verificano degli sbilanciamenti. Si verifica un fenomeno analogo a quello che avviene nel referendum, dove il problema è l'astensione. La domanda è quindi: "c'è un modo per massimizzare la partecipazione (nel rilascio dei dati) e contrastare l'astensione?" Ovviamente la risposta dipende dal tipo di dati. Le persone possono reagire ad un "incentivo estrinseco" (come sconti per servizi, "commisurato ma

verso il basso, perché dal punto di vista psicologico sorregge") ma ci sono anche incentivi di carattere psicosociale, che fanno sentire le persone come parte della risoluzione del problema. Viene fatto l'esempio di una piattaforma gamificata con gli incentivi estrinseci.

Il confronto prosegue sulla **qualità del dato e la sua conformità.** Alcuni partecipanti infatti affermano che il meccanismo incentivante può aiutare la raccolta, ma che permane il problema dei dati non strutturati. La domanda da porsi è: "quali metriche utilizziamo per garantire la *fairness* di queste informazioni"? Infatti, secondo diversi partecipanti, i dati raccolti, prima di essere riutilizzati, devono essere valutati e resi conformi alla *fairness,* concetto di cui si parla molto. La valutazione di conformità, trattata anche all'interno dell'AI Act deve infatti essere fatta all'origine. Si precisa che questo è il senso della regolamentazione, affermando, per esempio, che nelle Linee guida per l'adozione dell'IA nella PA, redatte da AgID, c'è una parte sulla valutazione di impatto che richiede valutare la conformità dei dati. Un altro punto evidenziato è il fatto che la fairness significa qualità di alto valore, non solo conformità: quest'ultima potrebbe essere anche legata solo a parametri tecnici ma non consentire la valutazione relativa all'equità (secondo la compliance con determinati parametri, come ad es. quello della leggibilità dei dati). Il requisito etico va quindi evidenziato. A questo proposto una partecipante ricorda che è stato rilasciato il 10 luglio il Code of practice redatto da esperti indipendenti che dà delle indicazioni a riguardo. Si articola su tre temi: trasparenza, copyright e sicurezza. Alcuni passaggi dell'AI Act già ad agosto entreranno in vigore si è ritenuto importante avere queste indicazioni. L'obiettivo di questo Codice è introdurre una "presunzione di conformità", che potrà essere messa in discussione, a favore delle aziende che lo adottano in vista dell'entrata in vigore dell'AI Act, prevista per il 2 agosto. Si tratta di uno strumento che invita le imprese del settore a valutare e migliorare autonomamente i propri sistemi, anticipando così le nuove regole. a questo proposito una partecipante ricorda che le linee guida di AgID sono state oggetto di consultazione e a breve saranno rilasciate.

Un altro tema affrontato riguarda il **linguaggio da utilizzare**. Una partecipante afferma che un processo di rivoluzione inizia "dal basso", ovvero, che per addestrare un sistema (di AI) che garantisca una risposta affidabile occorre avere dati il più possibile neutri, quindi che non contengono bias all'origine. A suo avviso per fare ciò il linguaggio deve essere il più possibile inclusivo, per esempio, non usare il maschile sovraesteso (nel linguaggio naturale un gruppo viene sempre indirizzato con il plurale maschile, invece occorre eliminare questi bias anche nel linguaggio). Il gruppo discute del tema del linguaggio considerandolo importante. Un altro approccio proposto è quello di utilizzare l'intelligenza collettiva. Se vogliamo che l'AI non usi le "master narratives" (sul genere, ma non solo) dobbiamo anzitutto correggerle e poi con il contributo degli esseri umani **creare delle counter narratives,** e, in più sessioni di addestramento, insegnare all'AI a considerare altre banche dati per adottare nuove narrative per contribuire all'inclusività. Anche un'altra partecipante si esprime sull'idea di linguaggio neutro sostenendo che annulla le differenze di cui invece dobbiamo tenere conto. Una raccomandazione del Consiglio d'Europa, legata alla Convenzione Quadro del COE sull'IA (generica sulle discriminazioni di genere) include la richiesta di evitare i bias prodotti dall'umano. Uno dei temi su cui si insiste è di introdurre quante più figure e professionalità femminili nel processo di sviluppo del

software, che introducono una visione delle donne, che apportano il proprio apparato valoriale. Dal dato di input all'output finale integrare una visione altra rispetto all' "ingegnere uomo maschio etero bianco" (the "default male").

Si prosegue affrontando il tema dell'**addestramento dell'AI**. Secondo un altro partecipante si può fallire solo se consideriamo che i dati non possono essere rimaneggiati. Cercare dati completamente *fair* rischia di essere una missione impossibile, perché nel processo comunque qualcosa può andare storto. Se partiamo con un modello teorico, creiamo i dati: ma in realtà dobbiamo considerare che si può comunque sempre tornare sui dati e correggere il tiro. Su questo anche un altro partecipante si dice d'accordo: immettere nel sistema delle indicazioni il più possibile neutre (inteso come senza vizi iniziali) ci permette di avere una maggiore inclusività. Si avanza il dubbio che non si riesca a prendere in considerazione tutti i punti di vista, ma il punto è che "non si riesce mai a considerare tutti". Ecco perché è importante **i**l dato disaggregato, perché è quello che fornisce l'informazione per il sesso, l'età, la disabilità o abilità e mi dà le informazioni dettagliate per quella categoria. Poi nel processo lavorativo vanno inserite più categorie possibili. Occorre iniziare dalle differenze più evidenti, poi quando ci si trova nell'algoritmo di fronte a delle discriminazioni evidenti si deve tornare indietro e "correggere il tiro". Per questo si parla di *addestramento* dell'AI.

Il gruppo si interroga quindi su **come procedere nel concreto,** sostenendo che "immaginarsi tutto a monte è difficile" e che ci vuole un processo ricorsivo e di valutazione dell'output. Qui è fondamentale includere le tante diversità che esistono nel processo tecnico di sviluppo dell'algoritmo, ma anche nella valutazione dell'output. Si veda esempio della storia della buona notte raccontata da una partecipante: la storia riguarda un esperimento in cui si è chiesto a Chat Gpt di proporre una storia della buona notte per un bambino e una bambina e il sistema, nonostante i diversi tentativi, ha continuato a proporre narrazioni stereotipate in cui il bambino sviluppava competenze di tipo tecnico e alto, come l'astronauta, mentre la bambina era relegata a ruoli nel mondo dell'arte o dell'insegnamento. È importante quindi il ruolo delle persone, soprattutto in fase di valutazione, perché diventano degli *early detector* che ci permettono di tornare indietro e capire dove sono generati i bias. Questo è vero anche per le categorie psicologiche stigmatizzate (come per la salute mentale). Occorre cercare una sinergia tra AI e persone, come "un sistema bipede a due gambe".

Viene rilanciato da una partecipanti il tema della consapevolezza della PA, ma su questo non ci sono interventi.

In ultimo, si affronta il tema del **monitoraggio periodico** e il gruppo discute del fatto che l'utilità del monitoraggio dipende dal punto in cui lo si immagina (a valle o a monte?). Si è già detto di coinvolgere persone con percezioni diverse durante il processo, sicuramente anche in fase di valutazione di output, ma si discute anche dell'opzione di un organo terzo (advisory board) che per conto delle PA possa valutare man mano i risultati e proporre delle correzioni. Occorre però evitare che gli organi di valutazione lavorino direttamente per le aziende, così da garantire la terzietà del loro lavoro.

Il lavoro si conclude con una veloce rilettura del report per verificarne l'adeguatezza, che è confermata dai partecipanti. In merito alle prossime date, la facilitatrice le comunica e chiede se ci sono eventuali obiezioni: una partecipante sarebbe lieta se si potesse spostare il prossimo incontro dal 9 settembre ad altra data, dato che è impegnata in una commissione e le piacerebbe partecipare.

Appendice - risorse citate dalle partecipanti

- The General-Purpose AI Code of Practice
- Convenzione quadro sull'intelligenza artificiale, i diritti umani e lo Stato di diritto sull'IA
- Study on the impact of artificial intelligence systems, their potential for promoting equality, including gender equality, and the risks they may cause in relation to non-discrimination

Conclusioni ed esiti della sessione

La sessione ha prodotto numerosi spunti concreti, tra cui, in sintesi:

1. raccolta e qualità dei dati: è essenziale la raccolta di dati disaggregati (per genere, età, ecc.) per evitare generalizzazioni e garantire analisi specifiche; la conformità tecnica non basta: è necessario integrare un requisito etico nella valutazione della qualità del dato;

2. linguaggio e neutralità. Il linguaggio inclusivo è cruciale: occorre evitare il maschile sovraesteso e bias linguistici; al contrario alcuni sostengono che annulli le differenze invece di valorizzarle;

3. addestramento e inclusività dell'AI. Deve basarsi su dati il più possibile neutrali e privi di bias all'origine ma è anche auspicabile correggere i dati e il modello in itinere (approccio ricorsivo) e creare delle counter-narratives coinvolgendo l'intelligenza collettiva;

4. inclusione di punti di vista diversi. Non si può rappresentare ogni punto di vista, ma i **dati disaggregati** aiutano a rendere visibili differenze rilevanti (genere, età, abilità). Occorre integrare progressivamente più categorie sociali nel processo di sviluppo e di revisione degli algoritmi.

5. valutazione degli output. La valutazione dell'output è fondamentale per individuare e correggere stereotipi e bias: serve un sistema ricorsivo, che preveda correzioni in corso d'opera grazie anche al contributo umano come early detector di bias.

6. sinergia tra AI e persone. È necessario un modello di collaborazione simbiotica tra AI e persone: un "sistema bipede a due gambe".

7. governance e valutazione esterna. Si propone la creazione di un organo terzo (advisory board) per valutare gli output e suggerire modifiche, garantendo indipendenza dalle aziende.