Rischi e opportunità dell'IA per migliorare equità e accessibilità

Percorso di co-creazione delle Raccomandazioni IA

Contesto

La definizione di Raccomandazioni sull'equità dei dati e degli algoritmi dell'Intelligenza Artificiale rientra tra le attività prioritarie dell'impegno 5 - Promozione dell'inclusività e dei diritti nell'accesso alle tecnologie e nell'utilizzo dell'Intelligenza Artificiale del 6° Piano d'Azione Nazionale per il governo aperto (6NAP) 2024-2026.

È realizzata attraverso un processo partecipativo articolato in 3 incontri di co-creazione che mira a coinvolgere sul tema dell'utilizzo dell'IA e il co-design dei sistemi di IA utilizzati dalle PA, gli stakeholder pubblici e delle organizzazioni della società civile interni ed esterni alla comunità di riferimento dell'iniziativa OGP Italia.

Temi degli incontri di co-creazione

Gli incontri di questo percorso partecipativo hanno come filo conduttore la difesa dello spazio civico e la promozione dei diritti nello sviluppo e nell'adozione dei sistemi di intelligenza artificiale. I **temi** su cui ruotano le principali domande che caratterizzeranno gli incontri sono: **consapevolezza, equità, governance e partecipazione, appropriatezza tecnologica**.

Questi temi definiscono il perimetro entro il quale formulare le raccomandazioni, che il gruppo di lavoro responsabile dell'impegno B5 sintetizzerà in un documento da inviare ai decisori nella pubblica amministrazione che dovranno garantire l'equità dei sistemi di intelligenza artificiale e l'applicazione dei diritti.

Riteniamo che l'insieme delle tematiche, durante i tre incontri previsti, possa contribuire a garantire l'adozione di IA più responsabile e attenta ai diritti fondamentali ed equità, e un'applicazione estensiva dei processi messi in campo.

Il primo incontro è dedicato al tema della **raccolta, preparazione e validazione dei dati che servono ad alimentare e addestrare il sistema IA**. La qualità dei dati determina la qualità delle decisioni prese dall'algoritmo. Eventuali *bias*, pregiudizi o discriminazioni presenti nei dati iniziali influenzeranno inevitabilmente l'intero sistema IA.

Il secondo incontro affronta il tema del**l'apertura del processo di progettazione, sviluppo e adozione degli algoritmi,** includendo scelte tecnologiche, modalità di funzionamento e trasparenza delle logiche decisionali adottate. In particolare, l'attenzione sarà rivolta all'adozione di soluzioni tecnologiche sul mercato e a discutere come le PA possono esercitare una funzione di committenza vigile e critica verso la scelta dei fornitori, indicando i requisiti rispetto al procurement per garantire il rispetto dei diritti fondamentali all'interno del loro "conformity assessment"(valutazione di conformità).

Il terzo incontro riguarda nello specifico **i sistemi di controllo e verifica dei sistemi di IA**, in particolare la fase di addestramento. In questa fase l'algoritmo viene effettivamente addestrato e affinato attraverso dati reali, e anche dati sintetici, testato in contesti concreti e poi costantemente monitorato e aggiornato nel tempo. Qui possono emergere chiaramente eventuali errori, discriminazioni o malfunzionamenti del sistema. L'addestramento richiede attenzione costante per garantire che il sistema funzioni in modo equo e affidabile durante tutto il suo ciclo di vita.

Calendario degli incontri

1° incontro: 15 luglio 2025, dalle 15.00 alle 17.00

2° incontro: 9 settembre 2025, dalle 15.00 alle 17.00

3° incontro: **7 ottobre 2025, dalle 15.30 alle 17.00**

Tutti gli incontri si svolgono online.

Partecipanti

Persone ed organizzazioni appartenenti al team dell'impegno B5 del 6NAP; attori-chiave individuati attraverso manifestazione di interesse; la generalità dei soggetti nella Community OGP IT che si sono iscritti all'evento.

Svolgimento degli incontri

Durata di ciascun incontro: 120 minuti

Fase di plenaria

5-10' - Saluti iniziali, breve presentazione dell'impegno B5 e presentazione modalità di lavoro: suddivisione in gruppi di lavoro dei/delle partecipanti.

Fase di introduzione (nel gruppo a cura di FT)

5-10' - Ingaggio da parte esperta/o progetto Opengov e facilitatore/trice

- perché il vostro parere è importante
- rilettura della guida alla discussione
- **lancio di 1 domanda** a cui rispondere: **Come possiamo assicurare equità nelle fasi di addestramento di un sistema di IA, al fine di ridurre discriminazioni e stereotipi?**

Fase di discussione (100')

90' dedicati a Tavoli online di discussione facilitata (figure esperte contribuiscono alla discussione e facilitatrici gestiscono il confronto)

- 10' di commento finale da parte dell'esperta/o del progetto Opengov sulla varietà dei contributi emersi. Riepilogo sui successivi passaggi del percorso partecipativo.
- chiusura dell'evento direttamente nelle stanze, senza tornare in plenaria.

Entro 1 settimana dall'incontro: condivisione con i/le partecipanti del report con i principali esiti delle sessioni di discussione.

Guida alla discussione 3° incontro

7 ottobre 2025

Introduzione

La sessione di lavoro di gruppo si svolge online, su piattaforma Zoom in stanze virtuali parallele, composte a partire dal numero dei/delle partecipanti presenti.

I/Le partecipanti sono affiancati/e da un/a facilitatrice per la gestione del confronto e un/a componente del team dell'Impegno B5 che contribuisce alla conoscenza del contesto. Dopo una breve presentazione del tema (5-10 minuti), i/le partecipanti sono invitati/e a contribuire con il proprio punto di vista alla discussione, a partire dalla domanda illustrata di seguito.

Lo staff presente ha il compito di raccogliere le opinioni espresse dai/dalle partecipanti in un confronto esplorativo che non ha come obiettivo la condivisione su una sola proposta ma l'arricchimento della riflessione intorno alla domanda. Al termine dell'incontro sarà prodotto un documento di restituzione che rende conto delle diverse sensibilità, percezioni e proposte emerse dal confronto.

Gli esiti di ciascun incontro saranno utilizzati per la definizione dei contenuti del testo delle Raccomandazioni sull'equità dei dati e degli algoritmi dell'IA.

Di che cosa discuteremo

In questo terzo e ultimo incontro partecipativo, ci concentriamo sulla fase di addestramento degli algoritmi. Questa è la fase in cui i dati servono a "insegnare" al modello: forniscono esempi, pattern, regole implicite su cui il modello costruisce la sua previsione. In pratica: il modello "impara" da un insieme di dati esemplificativi (le cosiddette *train data*) con etichette (in apprendimento supervisionato) o senza (in apprendimento non supervisionato).

Qui la qualità, la rappresentatività, la pulizia e la completezza dei dati sono critiche: dati distorti / parziali / sbilanciati producono modelli con *bias*, fragili o inefficaci. Esempio: dati addestrati solo su un gruppo demografico specifico porteranno a modelli che non generalizzano.

Studi mostrano che la qualità del dato (completezza, accuratezza, coerenza) ha un impatto diretto sulle performance del modello.
Se i dati sono chiusi, incompleti o privi di metadati, il modello è viziato in origine. Le pratiche FAIR (metadati ben strutturati, documentazione, interoperabilità) aiutano a mitigare questi problemi. Dati aperti ben documentati sono molto importanti in questa fase.

Ci interessa capire come garantire etica e abbattimento dei bias cognitivi, progettando momenti di monitoraggio e definendo la community per la valutazione e l'analisi dei risultati.
Rifletteremo su chi deve valutare le scelte tecnologiche, le modalità di funzionamento e la trasparenza delle logiche decisionali adottate. La discussione verterà in particolare sulla valutazione dei bias cognitivi: come vengono generati e come si possono prevenire e superare, grazie a quali interventi. Quanti e quali stakeholder possono (devono) essere coinvolti e come per garantire maggiore trasparenza e apertura.
Valuteremo se e come le Pubbliche Amministrazioni possono definire e verificare requisiti e dati su cui allenare gli algoritmi di IA, garantendo che, anche nelle gare, vengano inseriti vincoli e garanzie in tal senso.

In tal senso, è utile citare l'art. 5, comma 1, lett. d) — Trasparenza dell'addestramento nelle forniture alla PA dalla **legge delega che regolamenta l'uso dell'intelligenza artificiale nel lavoro** (Legge 23 settembre 2025, n. 132), appena approvata, che introduce l'obbligo di garantire trasparenza nelle modalità di addestramento delle applicazioni basate sull'IA generativa. La legge si applica agli acquisti della Pubblica Amministrazione (piattaforme di e-procurement) e dice in

sostanza che quando la PA sceglie fornitori di sistemi/modelli di IA, può privilegiare (non è un obbligo rigido, ma un criterio preferenziale) le soluzioni che mantengono i dati strategici in data center sul territorio nazionale, con disaster recovery e business continuity anch'essi in Italia; che **offrono elevati standard di sicurezza e trasparenza nelle modalità di addestramento e nello sviluppo di applicazioni di IA generativa**; il tutto nel rispetto della concorrenza e dei principi di non discriminazione e proporzionalità.

Le implicazioni pratiche sono le seguenti per i fornitori della PA: dovranno essere in grado di spiegare come addestrano il modello (procedure, fonti dei dati a livello di categorie/insiemi, controlli qualità, sicurezza del training pipeline) affinché la PA possa valutare la trasparenza dell'addestramento come fattore premiale. Se i fornitori trattano dati strategici, potrà essere un plus (o una richiesta di gara) l'hosting in data center nazionali con *disaster recovery (DR)* e *business continuity* (BC) in Italia. La legge enuncia anche un principio generale sulla qualità/trasparenza dei dati con cui si sviluppano (e quindi si addestrano) i sistemi e modelli di IA, da garantire e vigilare "secondo proporzionalità" rispetto ai settori d'impiego. Questo fa da cornice all'idea di trasparenza sull'addestramento.

L'art. 16 della stessa Legge, su dati, algoritmi e metodi per l'addestramento, delega il Governo a emanare, entro 12 mesi, uno o più decreti legislativi che definiscano una disciplina organica sull'uso di dati, algoritmi e metodi matematici per l'addestramento dei sistemi di IA. La delega vale senza introdurre obblighi ulteriori rispetto a quelli già previsti dall'AI Act della Unione Europea nei settori coperti da quel regolamento. I decreti dovranno prevedere di individuare i casi in cui serve definire il regime giuridico dell'uso di dati/algoritmi/metodi per l'addestramento, con i diritti e gli obblighi di chi li usa (es. condizioni d'uso dei dataset, doveri di documentazione, cautele); prevedere tutele (risarcitorie o inibitorie) e un apparato sanzionatorio in caso di violazioni; assegnare alle sezioni specializzate in materia di impresa la competenza sulle relative controversie.

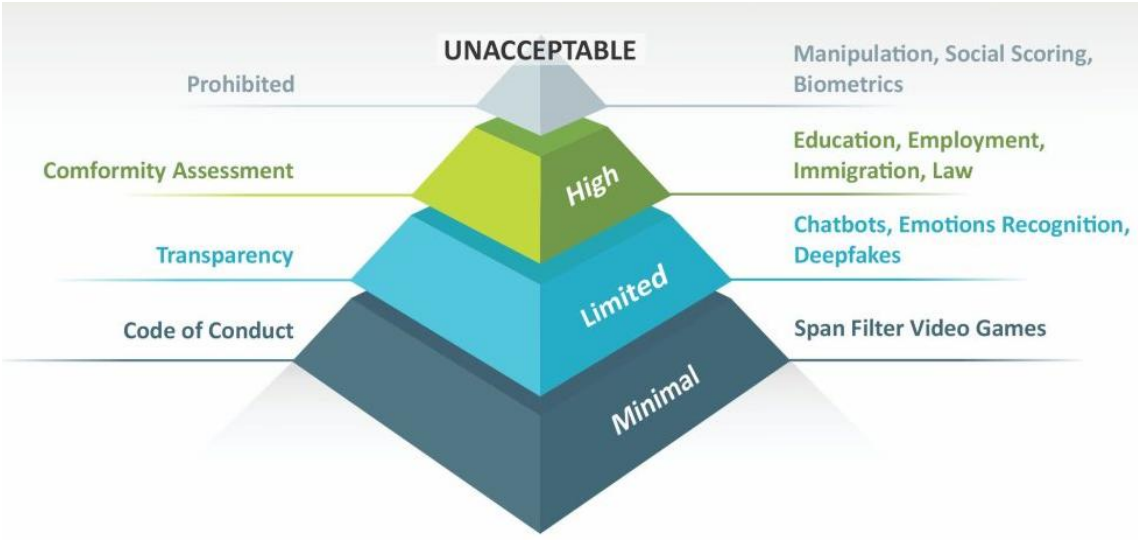
Oggi la legge stabilisce i principi e il perimetro; con i decreti delegati si avrà la disciplina di dettaglio su come usare dati/algoritmi/metodi nel training (chi può fare cosa, con quali tutele e sanzioni), mantenendo l'allineamento all'AI Act (no "over-compliance" nazionale nelle aree coperte dal regolamento UE). Dunque i fornitori dovranno documentare i processi di addestramento; garantire trasparenza su dati e procedure; e considerare hosting in data center nazionali.

Per garantire compliance sul training, dunque attendere e prepararsi ai futuri decreti; mappare l'uso di dati, algoritmi e metodi; definire le policy interne su diritti/obblighi e gestione delle

controversie.

Tutte le tematiche sopra citate andranno lette anche con l'obiettivo di capire come la società civile, le OSC e altri possibili stakeholder potranno essere inclusi.

Per orientarsi ulteriormente ai fini della discussione, trovate qui la piramide di rischio in cui sono classificati gli usi dei sistemi di IA, così come definiti nell'AI ACT approvato dal Parlamento europeo nel marzo 2024.



Nella tabella seguente invece, trovate riferimenti su obiettivi, dataset, rischi, obblighi normativi che differenziano i modelli di base (GPAI - General Purpose AI) e i modelli AI specializzati.

	GPAI (General-Purpose AI) Modello di base (foundation model) Esempi: GPT, LLaMA, Mistral, Gemini.	**Altri modelli AI specializzati** Modello addestrato (o fine-tuned) per un compito preciso
Obiettivo iniziale	Versatilità, uso generale	Task specifico (es. visione medica, traduzione)
Dataset	Ampio, eterogeneo, non limitato a un dominio	Dati mirati e curati per un ambito
Scala	Molto grande, alto costo computazionale	Più contenuto, spesso gestibile in-house
Riuso	Fine-tuning o adattamento per casi d'uso specifici	Già pronto per lo scopo previsto
Rischi	Bias amplificati e propagati in tanti contesti	Bias più circoscritti al dominio specifico
Obblighi normativi	Trasparenza sui dati di training, documentazione, test per modelli "sistemici" (AI Act + L. 132/2025)	Obblighi legati al settore d'uso (es. sanità, lavoro, PA)

La domanda di fondo che vi proponiamo per la discussione è la seguente:

Come possiamo assicurare equità nelle fasi di addestramento di un sistema di IA, al fine di ridurre discriminazioni e stereotipi?

A partire da questa domanda, vi chiediamo di discutere in particolare:

- Quali elementi di trasparenza nell'addestramento possiamo già garantire oggi?
- Come documentiamo i dati usati per l'addestramento?
- Quali rischi possono essere collegati alla localizzazione dei data center?
- Quali strumenti di tutela servirebbero in caso di violazioni nell'uso dei dati di training?
- Come prepararci ai futuri decreti senza sovraccaricare i processi attuali?
- Quali best practice aziendali potrebbero diventare standard legali?
- Come si correggono eventuali distorsioni?
- Che tipo di formazione esiste per un addestramento responsabile ed etico dei sistemi IA?
- Quali meccanismi partecipativi si possono adottare per controllare e verificare la correttezza dell'addestramento e dei risultati prodotti dall'IA (es. test aperti, segnalazioni pubbliche, osservatori permanenti)?
- Come dev'essere strutturata la governance per consentire la segnalazione e correzione tempestiva di eventuali distorsioni e problematiche emerse nella fase di addestramento e nell'utilizzo reale?