# **Bias & Fairness Mitigation Report – [Nome Progetto]**

**Data:** [gg/mm/aaaa]

**Versione:** [v1.0]

**Responsabile:** [Nome, Ruolo]

## **1. Sommario**

Sintesi delle attività svolte per identificare, valutare e mitigare bias e rischi di unfairness nel sistema AI.

## **2. Ambito e obiettivi**

* **Sistema analizzato:** [es. RAG+NER+GPT+Modulo adversariale]
* **Scopo:** Garantire che l’output del sistema sia il più possibile equo, non discriminatorio e trasparente.

## **3. Analisi dei potenziali bias**

| **Fase/processo** | **Possibili bias** | **Cause potenziali** | **Evidenze/Riscontri** |
| --- | --- | --- | --- |
| Raccolta dati | Bias di selezione | Campioni non rappresentativi | [esempi] |
| Pre-processing (NER) | Bias nei dati grezzi | Errori anonimizzazione/parole chiave mancanti | [esempi] |
| GPT output | Bias algoritmico | Addestramento/prejudice modello | [esempi, citazioni log] |
| Interazione utente | Feedback loop | Bias da auto-rinforzo | [descrizione, log] |

## **4. Metriche di fairness e risultati**

| **Metrica utilizzata** | **Valore rilevato** | **Soglia/target** | **Note/commenti** |
| --- | --- | --- | --- |
| Demographic Parity | [valore] | [target] |  |
| Equalized Odds | [valore] | [target] |  |
| Analisi linguaggio/lexical | [valore] | [target] |  |
| Audit umano (campione risp.) | [valore] | [target] |  |

## **5. Strategie e misure di mitigazione adottate**

* **Data balancing:** [oversampling/undersampling/reweighting applicati, su quali dati]
* **Prompt engineering:** [esempi di prompt per ridurre bias]
* **Modulo adversariale:** [come e quando interviene, log delle segnalazioni]
* **Revisione manuale:** [casi critici/escalation]
* **Monitoraggio continuo:** [dashboard, alert, periodicità]

## **6. Audit, test e revisioni**

* **Test periodici:** [cadenza, esiti principali]
* **Risultati audit interni/esterni:** [sintesi, eventuali raccomandazioni]
* **Incidenti o casi di bias rilevati:** [data, descrizione, risoluzione]
* **Azioni correttive intraprese:** [cosa, quando, da chi]

## **7. Trasparenza e comunicazione**

* **Comunicazione agli stakeholder:** [come vengono informati utenti/management sui rischi di bias]
* **Accessibilità della documentazione:** [come e dove viene archiviata/condivisa]

## **8. Piano di miglioramento continuo**

* **Obiettivi futuri:** [nuove metriche, test, formazione personale, aggiornamento modelli]
* **Prossima revisione prevista:** [data]
* **Responsabile revisione:** [Nome, ruolo]

## **9. Allegati**

* Report metriche dettagliate
* Log audit fairness
* Prompt utilizzati per test fairness
* Casi studio e simulazioni

**[Firma, data, eventuale timbro]**