Progetto di ricerca—Progetto di ricerca "*Computable models of fairness and their applications in legal procedures*"—"Modelli computabili di equità e loro applicazione nelle procedure giuridiche"

Ottobre 2024-Settembre 2025

Il progetto EquAl (Equitable Algorithms) affronta valutazioni, decisioni e previsioni algoritmiche, per promuovere l'equità e contrastare la discriminazione verso singoli individui e gruppi. Obiettivi di EquAl:
- Fornire una comprensione dei concetti di ingiustizia e iniquità algoritmica e discriminazione, collegando le nozioni adottate nelle scienze sociali, nel diritto, nella statistica e nell'intelligenza artificiale.
- Esaminare il modo in cui le tecnologie possono promuovere l'equità e supportare l'individuazione e il contrasto dell'ingiustizia e della discriminazione algoritmica, in particolare per quanto riguarda la valutazione delle richieste di asilo.
- Sviluppare un prototipo di sistema di supporto alle decisioni per supportare le decisioni sulle richieste di asilo e testarne l'equità.
In questo contesto, la ricerca proposta mira a sviluppare approcci algoritmici per valutare la fattibilità del processo decisionale automatizzato e metodi per identificare e porre rimedio a iniquità e discriminazioni, in particolare per quanto riguarda le domande di asilo. Ciò comporterà due approcci: utilizzare metodi algoritmici ex post, per valutare il processo decisionale umano, e utilizzarli ex ante, per suggerire nuove decisioni, o ad interim, come componente di un processo decisionale in corso. L'uso ex-post sarà testato attraverso l'analisi algoritmica di una serie di decisioni umane, gli approcci ex-ante e provvisori saranno analizzati abbinando le possibili tecnologie agli scenari rilevanti. Verrà definita una metodologia per valutare l'equità nelle decisioni umane e automatizzate, generalizzando l'approccio adottato nel caso studio sull'asilo.

Piano delle attività — Piano di attività

La ricerca sarà organizzata secondo le seguenti fasi e attività.

Mese 1-6

La prima fase si concentrerà su:
(i) sviluppo e applicazione di metodi di anonimizzazione dei dati dei casi oggetto di analisi
(ii) una revisione completa dei criteri statistici utilizzati nell'apprendimento automatico per valutare l'equità delle previsioni/decisioni automatizzate e un confronto con le nozioni filosofiche, etiche e sociali di (in)equità.
(iii) l'estrazione e l'identificazione automatizzata delle caratteristiche di input rilevanti (tratti, caratteristiche e contesti) a disposizione dei decisori nella valutazione delle richieste di asilo, come emergono dai documenti raccolti
(iv) la creazione di un dataset contenente i documenti originali, le caratteristiche estratte e i risultati associati. L'insieme dei dati sarà bilanciato in relazione alle caratteristiche dei richiedenti (es.: paesi di origine e transito, genere, età, motivi di vulnerabilità) nonché all'esito delle decisioni (tutela sovvenzione/tutela rifiuto/domanda manifestamente infondata).
(v) l'impiego di metodi statistici e analitici di apprendimento automatico per esaminare modelli e correlazioni, connessioni causali tra le caratteristiche disponibili e i risultati corrispondenti e il grado di dispersione (rumore) nelle decisioni

Mese 6-12

La seconda fase si concentrerà su:

(i)) l'analisi del rapporto tra le caratteristiche estratte e quelle associate per determinare se direttamente o indirettamente qualsiasi caratteristica vietata - e più in generale qualsiasi caratteristica inappropriata o irrilevante - possa contribuire a risultati discriminatori.
(ii) la rimozione di tutte le caratteristiche connesse a casi di ingiustizia e l'aggiunta di caratteristiche necessarie a garantire risultati più equi.
(iii) Sviluppo di un prototipo di sistema di supporto alle decisioni basato su metodi di apprendimento automatico da utilizzare per: (i) supportare le decisioni umane sulle richieste di asilo e (ii) testarne l'equità
(iv) Test e perfezionamento del prototipo sviluppato
(v) Sviluppo di un quadro metodologico per le tecnologie che migliorano l'equità

Research Project—Progetto di ricerca "*Computable models of fairness and their applications in legal procedures*"—"*Modelli computabili della fairness e loro applicazione nelle procedure giuridiche*"

"

Ottobre 2024-Settembre 2025

The EquAI (Equitable Algorithms) project addresses algorithmic evaluations, decisions, and predictions, to promote fairness and counter discrimination affecting individuals and groups. EquAI aims:
- To provide an understanding of the concepts of algorithmic unfairness and discrimination, bridging the notions adopted in social sciences, law, statistics, and artificial intelligence.
- To examine the way in which technologies can promote fairness and support detecting and countering algorithmic unfairness and discrimination, in particular with regard to the assessment of asylum requests.
- To develop a decision support system prototype for supporting the decisions on asylum requests and testing their fairness.

In this context, the proposed research aims to develop algorithmic approaches for assessing the viability of automated decision-making and methods to identify and remedy unfairness and discrimination, in particular with regard to asylum applications. This will involve two perspectives: using algorithmic methods ex post, to assess human decision-making, and using them ex ante, to suggest new decisions, or interim, as a component of an ongoing decision-making process. The ex-post use will be tested through the algorithmic analysis of a set of human decisions, the ex-ante and interim approaches will be analysed by matching possible technologies to the relevant scenarios. A methodology to assess fairness in human and automated decisions will be defined, generalizing the approach adopted in the asylum case study.

Plan of activities — Piano di attività

The research will be organized according to the following steps and activities.

Month 1-6

The first step will focus on:

(i) Development and applications of methods for anonimyzing of data with regard to the selected cases

(ii) a comprehensive review the statistical criteria used in machine learning to assess fairness of automated predictions and a comparison with the philosophical, ethical and social notions of (un)fairness.

(iii) the Automated extraction and identification of the relevant input features (traits, characteristics, and contexts) available to decision-makers in assessing the asylum requests, as emerging from the collected documents

(iv) the creation of a dataset containing the original documents, the extracted features and the associated outcomes. The data set will be balanced in relation to the characteristics of the applicants (e.g.: countries of origin and transit, gender, age, grounds of vulnerability) as well as the outcome of decisions (grant protection/refuse protection/manifestly unfounded claim).

(v) the employment of machine learning statistical and analytics methods to examine patterns and correlations, causal connections between the available features and the corresponding outcomes and the extent of dispersion (noise) in decisions

Month 6-12

The second step will focus on:

(i) the analysis of the relationship between the extracted features and the associated to determine whether directly or indirectly any prohibited —and more generally any inappropriate or irrelevant– feature may contribute to discriminatory outcomes.
(ii) the removal of all the features connected to instances of unfairness and addition of features needed to ensure fairer outcomes.
(iii) Development a of a decision support system prototype based on machine learning methods to be used for: (i) supporting human decision-ma on asylum requests and (ii) testing their fairness
(iv) Test and Refinement of the developed prototype
(v) Development of a methodological framework for fairness enhancing technologies