Long-Term Fairness Strategies in Ranking with Continuous Sensitive Attributes

Luca Giuliani[1,*,†], Eleonora Misino[1,*,†], Roberta Calegari[1] and Michele Lombardi[1]

[1]Alma Mater Studiorum–Università di Bologna, Italy

Abstract

Recent advancements have made significant progress in addressing fair ranking and fairness with continuous sensitive attributes as separate challenges. However, their intersection remains underexplored, although crucial for guaranteeing a wider applicability of fairness requirements. In many real-world contexts, sensitive attributes such as age, weight, income, or degree of disability are measured on a continuous scale rather than in discrete categories. Addressing the continuous nature of these attributes is essential for ensuring effective fairness in such scenarios. This work aims to fill the gap in the existing literature by proposing a novel methodology that integrates state-of-the-art techniques to address long-term fairness in the presence of continuous protected attributes. We demonstrate the effectiveness and flexibility of our approach using real-world data.

Keywords

fair AI, fair ranking, long-term fairness, continuous sensitive attributes

1. Introduction

Ranking in AI is increasingly used across various sectors to enhance decision-making processes, spanning from credit scoring and hiring to education and other high-stakes domains. For instance, in credit scoring, AI models evaluate creditworthiness by analyzing vast amounts of financial data; in hiring, AI ranks candidates by assessing resumes and predicting job fit; and educational programs leverage AI to rank students' performance, providing personalized learning experiences.

The social and ethical implications of these systems have recently gained attention both in research and application domains, particularly concerning their potential to perpetuate or accentuate discrimination. Several approaches and metrics have been proposed to enforce and quantify adherence to fairness requirements, ensuring that trained models do not exhibit discriminations against minorities or individuals [1].

In all these scenarios, it is possible to mitigate discrimination by adjusting the ranking to promote fairness criteria (such as equal opportunity or statistical parity) across sensitive groups or individuals. Various algorithmic mitigation strategies have been proposed in the literature [2]; however, these approaches often focus on a single ranking, as if the AI system only produces one ranking throughout its lifetime, failing to consider that the ranking process is repeated over

Aequitas 2024: Workshop on Fairness and Bias in AI | co-located with ECAI 2024, Santiago de Compostela, Spain

*Corresponding author.

†These authors contributed equally.

luca.giuliani13@unibo.it (L. Giuliani); eleonora.misino2@unibo.it (E. Misino); roberta.calegari@unibo.it (R. Calegari); michele.lombardi2@unibo.it (M. Lombardi)

time. Considering the lifespan of an AI system, it becomes essential to ensure that the system can be deemed fair across all rankings produced, ensuring what is known as long-term fairness [3]. The study and assurance of long-term fairness are necessary to guarantee consistent and unbiased treatment across multiple iterations of the AI system, ensuring that biases do not accumulate or shift over time. If fairness is only considered for individual rankings, it may lead to temporary fairness that can fluctuate, resulting in long-term disparities. Furthermore, current approaches to fair ranking typically only work with categorical sensitive attributes. However, in various real-world scenarios, sensitive attributes like income, or degree of disability are continuous rather than discrete. Consequently, effectively managing their continuous nature is necessary for assessing and ensuring fairness. While there are studies focusing on fairness concerning continuous sensitive attributes, they do not intersect with existing work on fair ranking.

This work aims to fill the gap in the existing literature by proposing a methodology that integrates state-of-the-art techniques to *address long-term fairness in the presence of continuous protected attributes*.

The paper is structured as follows. Section 2 aims at placing our work within the existing literature on fair machine learning, focusing on applications of fairness in ranking and fairness with continuous sensitive attributes. In Section 3, we provide the essential technical background required to understand the details and significance of our approach, as it incorporates different state-of-the-art techniques and frameworks. Following this, we describe the specific aspects of our contribution in Section 4, where we present our methodology grounded on a specific use case. We outline the main results of our empirical evaluation in Section 5. Finally, we summarize our findings in Section 6 and highlight potential directions for future investigation.

2. Related Work

To the best of our knowledge, no previous work has addressed the task of fair ranking with continuous sensitive attributes. Still, there has been a significant growth in publications over the last decade in the two distinct fields, both stemming from the broader domain of fair machine learning. Hereby, we summarize the key developments in these areas as a means to effectively frame our work within the current state of the art.

2.1. Fair Machine Learning

Mehrabi et al. [4] categorize fair machine learning methods into three major groups, namely pre-processing, in-processing, and post-processing. This categorization is based on the timing of debiasing interventions. For example, pre-processing methods can be applicable when there is an opportunity to alter training data [5, 6, 7]. In contrast, in-processing methods are used when the inherent training procedure of the machine learning model is modified, either by loss regularizers or other types of constraint injection [8, 9, 10, 11]. Lastly, post-processing methods are employed when the algorithm must operate on an already trained model, treating it as a black box and reassigning output labels through a specific function in the post-processing stage [12, 13, 14]. Our research aligns with the third category, as we build on the work by [15] regarding the FAiRDAS framework, which aims to ensure sustained fairness

in ranking systems by post-processing the results produced by the learned model in successive batches, independently of the characteristics of the model itself.

2.2. Fairness in Ranking Applications

In their survey [2], Zehlike et al. distinguish between two types of fair ranking algorithms: (1) *score-based* methods, which use a predefined ranking function and allow the bias mitigation step to intervene on either the initial scores of the candidates, the ranking function f, or the final ranked outcome, and (2) supervised *learning-to-rank* methods, which train the ranking function on data and can thus be further categorized as in Section 2.1. Interestingly, the authors note that post-processing methods for learning-to-rank handle fairness constraints similarly to score-based methods. Under this lens, FAiRDAS can be seen as both a learning-to-rank application imposing constraints on model-predicted scores, and a score-based method enforcing fairness by adjusting original scores, whether generated by a model or given as gold standards.

Most fair ranking approaches employ top-k proportional representations as a fairness metric. Namely, they try to ensure an equal representation of protected groups in the first k candidates. For example, among the post-processing fairness methods for learning-to-rank, [16] and [17] adjust the positions of the candidates in the final ranking to meet certain minimal (and optionally maximal) requirements per subgroup. These methods treat top-k rankings as sets, hence disregarding the position of candidates. In contrast, [18] and [19] take the position into account by addressing the visibility bias rather than the score itself; in fact, the exposure of candidates has been shown to decrease geometrically with respect to their ranking position, as defined by their score. Moreover, the latter work proposes a methodology to dynamically change rankings for the same query to achieve equal attention over time, thus inherently incorporating long-term fairness effects within their framework, although at a query level only. For a more comprehensive overview of bias mitigation in ranking at different stages in the pipeline and using different methods, we refer the reader to the original survey.

2.3. Fairness with Continuous Protected Attributes

In the last few years, some works have proposed new metrics and computational methodologies to address continuous sensitive attributes in fairness enforcement tasks. Among them, [20] adopts for the first time the Hirschfeld–Gebelein–Rényi (HGR) correlation coefficient as a way to enforce model debiasing over continuous protected features. This metric, also referred to as the maximal correlation coefficient, is defined as the highest Pearson correlation that can be obtained by transforming random variables into nonlinear spaces through copula transformations. For this reason, its computation poses significant computational difficulties, yet various simplifications and approximations have been developed over recent years. Specifically, [20] introduced a differentiable way to calculate a lower bound of the metric using kernel-density estimation techniques, thus paving the way for its application as a loss regularizer in gradient-based learning algorithms. That work was subsequently improved by [21], whose novel computational technique based on two adversarial neural networks was shown to outperform the former.

A parallel effort was undertaken by [22], who introduced an indicator named Generalized Disparate Impact (GeDI) by slightly modifying the formulation of HGR to better adhere with

the legal concept of "Disparate Impact". Disparate impact arises when a seemingly impartial practice adversely affects a protected group, and a first method to measure it in both regression and classification scenarios was introduced by [23], who proposed a novel fairness metric called *Disparate Impact Discrimination Index* (DIDI). The Generalized Disparate Impact indicator straightforwardly extends this metric to the case of continuous inputs where, as usual, higher GeDI values signify a greater disparity concerning the chosen protected attribute.

3. Background

In this section, we provide a formalization of the ranking problem general enough to model our case study and other similar applications. Next, we introduce FAiRDAS [15], a general framework designed to address long-term fairness in ranking systems. Finally, we describe the Generalized Disparate Impact (GeDI) indicator [22], which we utilize to effectively handle continuous protected attributes.

3.1. Ranking Problem Formulation

We focus on a process wherein a set $\mathscr{R}$ of m resources undergoes repetitive ranking guided by observable information arriving over time. For example, $\mathscr{R}$ may contain students that need to be ranked based on predicted academic performance. The observable information, hereinafter referred to as *batches*, is seen as a stochastic process indexed by time, denoted as $\{X_t\}_{t=1}^{\infty}$. Each batch X_t is a random variable characterized by a domain $\mathscr{X}$ and probability distribution $P(X_t)$. The ranking quality is characterized using a metric function defined in probabilistic terms, typically relying on expectations or event probabilities, namely:

$$y : X, \theta \mapsto y[X; \theta] \tag{1}$$

Here, $\theta \in \Theta$ is an *action vector* whose values can be adjusted to control the ranking procedure behavior. For example, the *action vector* might represent penalty or reward terms linked to sensitive groups. The vector $y[X; \theta] \in \mathbb{R}^n$ denotes the values of n metrics for a given batch X and action vector θ. In real-world scenarios, these metrics will always admit a finite sample formulation, often derived by substituting theoretical expectations with sample averages.

Given that the ranking is performed for every batch, followed by an adjustment of the action vector, the ranking problem can be defined in terms of the tuple:

$$\langle \{X_t\}_{t=1}^{\infty}, \{\theta_t\}_{t=1}^{\infty} \rangle \tag{2}$$

where X_t and θ_t are the batch and action vector at time t respectively. The value of the metrics at time t is determined given X_t and θ_t (Equation (1)).

3.2. FAiRDAS

FAiRDAS [15] is a general framework that models long-term fairness as a dynamic system. It aims at stabilizing fairness and quality metrics below user-defined thresholds and allows users to define a target behavior approximated through a sequence of action; for example,

one may modify an input ranking by adjusting the scores for different protected groups. The approximation of the target behavior involves solving an optimization problem that minimizes the discrepancy between the target values for the metrics $\bar{y}_t$ and the actual metrics y_t determined by the actions θ_t, namely:

$$\theta^*(\bar{y}_t) = \arg\min_{\theta_t \in \Theta} \mathscr{L}(\theta_t, \bar{y}_t) \tag{3}$$

The solution method for Equation (3) relies on the action space characteristics and the chosen distance function. A possible choice for $\mathscr{L}(\theta_t, \bar{y}_t)$ is the Euclidean distance:

$$\mathscr{L}(\theta_t, \bar{y}_t) = \|y[X_t; \theta_t] - \bar{y}_t\|_2^2. \tag{4}$$

The exact evaluation of Equation (4) is often unfeasible, primarily due to the unknown distribution $P(X_t)$; thus, metric values $y[X_t; \theta_t]$ are replaced typically with a Monte Carlo approximation derived from historical data.

FAiRDAS Grounding. To apply FAiRDAS effectively to a specific scenario, it is essential to delineate its core components: 1) the *metrics of interest*, which establish the criteria for evaluating fairness and ranking quality; 2) the corresponding *threshold vectors*; 3) the *target dynamic system* which define the ideal metrics behavior; 4) the *set of actions*, delineating how metrics can be manipulated to enhance ranking fairness and quality; 5) the *distance function*, defining the metric for assessing the effectiveness of the target system's approximation; and finally, 6) the *optimization methods* used to address Equation (3), which heavily depends on the chosen set of actions and distance function.

3.3. Generalized Disparate Impact

The Generalized Disparate Impact (GeDI) was first introduced in [22] as an extension of the Disparate Impact Discrimination Index (DIDI) [23] to expand the availability of fairness metrics for the fully continuous case. It features a mapping function $f(x)$ for the input attribute $x \in \mathscr{X}$, which enables accounting for non-linear correlations between the sensitive input and the target.

This choice is inspired by the copula transformations of the Hirschfeld–Gebelein–Rényi (HGR) maximum correlation coefficient. However, one major difference between GeDI and HGR lies in the absence of a second mapping function on the output feature $y \in \mathscr{Y}$, which prevents it from measuring non-functional dependencies of the type $y \mapsto x$, akin to the DIDI. In addition to that, instead of leveraging the original definition of Pearson's coefficient, the formulation of GeDI is slightly altered to make the indicator sensitive to scale variations. This ensures that reductions in unfairness are proportionally translated to diminished disparate impacts even if the shape of the unfair behavior is not modified, and also guarantees compatibility between GeDI and DIDI since both metrics yield identical results when the input attribute is binary. Finally, the mapping function $f(x)$ is restricted to a linear combination over a polynomial kernel. This allows one to frame the computation as a linear optimization problem, thus keeping a low computational burden although retaining high approximation capabilities thanks to the inherent non-linearities. Additionally, it serves the dual purpose of reducing overfitting while maximizing user-configurability and interpretability of the metric.

Formally, $f(x)$ is defined as the vector product $\mathbf{V}_x^k \cdot \alpha$, where $\mathbf{V}_x^k$ is the polynomial expansion matrix built from the input vector x – i.e., the Vandermonde matrix –, while $\alpha \in \mathbb{R}^k$ is a coefficient vector that weighs the contribution of each polynomial order. GeDI is eventually computed as:

$$\text{GeDI}(x, y; \mathbf{V}^k) = \left| \frac{\text{cov}(\mathbf{V}_x^k \cdot \alpha, y)}{\text{var}(\mathbf{V}_x^k \cdot \alpha)} \right| \qquad \text{s.t. } \|\alpha\|_1 = 1 \tag{5}$$

where the constraint on the L1 norm of the coefficient vector is intended to replace the absence of the scaling factor on the output term. An important detail to note is that the order k of the polynomial expansion is part of the specification of the indicator, as it appears in its notation and aims to offer users a simple way to balance the bias-variance trade-off.

4. FAiRDAS with Continuous Attribute

As a demonstration of our approach, we focus on ranking students by their predicted academic performance to identify those at risk of dropping out. The real-world data is provided by the Canarian Agency for Quality Assessment and Accreditation (ACCUEE)[1], which gathers information to assess the performance of their educational system through regular diagnostic reports. The data spans four academic years (2015-2019) including (1) the evaluation of students' academic proficiency in subjects such as Mathematics, SpaniFirst,ge, and English and (2) context questionnaires completed by students, school principals, families, and teachers to collect socio-demographic background information. In our test case, we rank students based on their Mathematics proficiency measured by a normalized score. The protected attribute considered is the Economic, Social, and Cultural Status (*ESCS*) [24], namely a continuous indicator that serves as a proxy for the socioeconomic status of students. Ensuring long-term stability is crucial in this context: although consistently high accuracy and fairness are desirable, it is essential to maintain stable actions over time to prevent negatively affecting students' academic progress.

In addressing the task at hand, we define two distinct groundings of FAiRDAS framework. The first grounding, inspired by [15], adopts a set of discrete actions that requires a discretization of the sensitive attribute; conversely, the second grounding relies on a set of continuous actions that do not require any discretization. In both groundings, the continuous nature of the attribute is preserved when computing the fairness metric as we rely on GeDI [22]. The groundings we propose represent two potential approaches to addressing long-term fairness with continuous attributes and should not be seen as conflicting: in certain scenarios, depending on the desired level of interpretability and the overall system requirements, discrete actions may be necessary, while in others, continuous actions might be preferred. In the remaining of the section, we describe the two groundings in detail.

4.1. Grounding with Discrete Actions

Set of Actions. Inspired by [15], we design a set of discrete actions that directly modify the scores used by the ranking algorithm. Formally, given the discretization $v \in \mathscr{V} = \{v_1, v_2, ..., v_n\}$ of the continuous protected attribute *ESCS*, the actions are represented by a vector $\theta \in [0, 1]^{|\mathscr{V}|}$

[1]Dataset: https://zenodo.org/records/11171863.

with unit L1 norm. The modified score of a student with $ESCS = v$ is obtained by multiplying their original score by $(1 - \theta_v)$. Thus, the action vector components act as penalizing factors for over-represented sensitive groups in a batch, specifically affecting the scores of students in these protected groups. Higher values in the action vector (closer to 1) correspond to more significant penalization, whereas values closer to zero result in minimal modification to the student's score. In our application, we discretize the continuous protected attribute *ESCS* in four levels; thus, the action vector θ has four components, each applying to the students belonging to the corresponding *ESCS* level.

Metrics of Interest. In our case study, we are interested in decreasing socioeconomic discrimination while preserving ranking accuracy; thus, we need 1) a fairness metric able to deal with the continuous protected attribute *ESCS* and 2) an accuracy metric to measure the drop in ranking performance due to the application of the action vector θ. As a fairness metric, we rely on GeDI, whereas to assess the system's drop in performance we measure the sum of absolute differences between the original and modified scores, namely:

$$\text{SAE}(\theta) = \frac{1}{K}\sum_{k=1}^{K} |s_k - (1 - \theta_{v_k}) \cdot s_k| = \frac{1}{K}\sum_{k=1}^{K} |s_k \cdot \theta_{v_k}| = \frac{1}{K}\sum_{k=1}^{K} s_k \cdot \theta_{v_k}, \tag{6}$$

where K is the number of students in a batch, $\theta_{v_k} \in [0, 1]$ is the component of the action vector corresponding to the *ESCS* level of the k-th student, and $s_k \in [0, 1]$ is the score of the k-th student. It is worth noting that the two metrics of interest conflict: SAE drives θ_{v_k} towards zero to maintain the original ranking, whereas GeDI requires $\theta_{v_k} > 0$ for some k to mitigate discrimination. Given that the action vector must have a unit L1 norm, the trivial solution of $\theta_{v_k} = 0$ for all k, which would nullify both metrics, is not allowed.

Target Dynamic System. As we aim to meet the metric thresholds while maintaining long-term stability, we define our desired behavior by means of following dynamic system, which defines a smooth evolution of the target metrics toward the thresholds:

$$\bar{y}_{t+1} = \lambda \odot (\bar{y}_t - \mu) + \bar{y}_t, \tag{7}$$

where $\bar{y}_t$ represent the metric values in the target system, μ is the vector of thresholds, $\lambda \in (0, 2)^n$, and $\odot$ refers to the Hadamard (element-wise) product. Given that we are focusing on two metrics (GeDI and SAE), λ is a 2-dimensional vector, with its values determined through a preliminary experiment detailed in Section 5.

Distance Function and Optimization Method. We use Equation (4) – Euclidean distance – as the distance function, optimizing it with the `scipy` implementation of Sequential Least Squares Programming (SLSQP) optimizer.

4.2. Grounding with Continuous Actions

Set of Actions. To avoid the discretization of the protected attribute *ESCS*, we define the set of possible actions as a family of polynomial functions W_β parameterized by $\beta \in \mathbb{R}^{d+1}$, where d is

the order of the polynomial[2]. The functions map each value of ESCS to a real number, which is then used as a multiplicative discount factor to modify the student's score. First, we rescale *ESCS* into the domain $[0, 1]$, then we impose two constraints on the family of polynomial functions W_β, namely: 1) their integral must be unitary over domain in order to avoid degenerate solutions, and 2) their roots must lie outside the domain in order to guarantee that each discount factor $W_\beta(z_k)$ is strictly positive for all $z_k \in [0, 1]$. These constraints enhance the interpretability of the mitigation strategy by simplifying the comparison between the selected polynomial functions. Additionally, they prevent the trivial solution of a constant function equal to zero, which would nullify the fairness metric.

Metrics of Interest. As in *Discrete Actions*, we use GeDI as fairness metric to deal with the continuous nature of *ESCS*. The ranking performance is measured by the mean squared error between the original and modified scores[3]:

$$\mathrm{MSE}(\beta) = \frac{1}{K}\sum_{k=1}^{K}(s_k - W_\beta(z_k)\cdot s_k)^2. \tag{8}$$

where z_k is the *ESCS* value of the k-th student, and $W_\beta(z_k)$ the weighting polynomial function evaluated on z_k. As in *Discrete Actions*, the two metrics of interest conflict since MSE pushes W_β to be close to the constant function $W_\beta = 1$ while GeDI forces W_β to deviate from it.

Target Dynamic System. We rely on the same dynamic system in Equation (7), as our goal is to stably evolve the two metrics of interest below the predefined thresholds.

Distance Function and Optimization Method. As before, we use Equation (4) – Euclidean distance – as a distance function. However, when optimizing it, we rely on the `scipy` implementation of the Trust Region Method (trust-constr), as it proved to be more reliable in the solution, although at the expense of a slightly higher computational time.

5. Experimental Results

This section outlines the empirical evaluation performed on the case study described in Section 4. We first define the evaluation procedure and then report the numerical results[4].

5.1. Evaluation

We compare each of the two groundings with a baseline method focusing on metrics of interest and *action smoothness* ($\mathrm{m}_{Actions}$) described below. For each approach, we report the mean and standard deviation of the metrics across batches to assess performance and stability over time.

[2]In our application, we choose $d = 4$ as it provides a sufficient trade-off between the expressiveness of the function and the known numerical instability of polynomial kernels, along with their higher computational workload.
[3]We rely on MSE and not on SAE to avoid the computation of an absolute error.
[4]The source code to reproduce the experiments can be found at https://github.com/EleMisi/FairRanking under MIT license.

Action Smoothness. To evaluate the stability of the chosen actions over time, we compute the cosine distance between actions performed on consecutive batches. For the *Discrete Actions* grounding, $\mathrm{m}_{Actions}$ is defined as follows:

$$\mathrm{m}_{Actions} = \frac{1}{N}\sum_{t=1}^{N-1}\left(1 - \frac{\theta_t \cdot \theta_{t+1}}{\sqrt{\theta_t^2}\cdot\sqrt{\theta_{t+1}^2}}\right) \tag{9}$$

where N is the number of incoming batches and θ_t is the action vector of the t-th batch. For the *Continuous Actions* grounding, $\mathrm{m}_{Actions}$ is computed by evaluating the weighting polynomial functions on a fine-grained discretization of the interval $[0, 1]$. Formally, it is defined as:

$$\mathrm{m}_{Actions} = \frac{1}{N}\sum_{t=1}^{N-1}\left(1 - \frac{W_{\beta_t} \cdot W_{\beta_{t+1}}}{\sqrt{W_{\beta_t}^2}\cdot\sqrt{W_{\beta_{t+1}}^2}}\right) \tag{10}$$

where N is the number of incoming batches and W_{β_t} is the evaluation of the polynomial function chosen for the t-th batch.

Baseline Approach. We compare FAiRDAS in its *Discrete Actions* grounding against a baseline approach that focuses on finding the optimal action vector that minimizes:

$$\mathscr{L}(\theta) = \max\left(\mathrm{GeDI}(\theta), \mu_{GeDI}\right) + \max\left(\mathrm{SAE}(\theta), \mu_{SAE}\right) \tag{11}$$

where μ_{GeDI} and μ_{SAE} are the metrics' thresholds. The action vector θ is the same described in Section 4.1, and it is optimized via the SLSQP method, as for FAiRDAS. For the *Continuous Actions* grounding, the baseline approach searches for the optimal polynomial function W_β that satisfies the constraints described in Section 4.2 and minimizes:

$$\mathscr{L}(\beta) = \max\left(\mathrm{GeDI}(\beta), \mu_{GeDI}\right) + \max\left(\mathrm{MSE}(\beta), \mu_{MSE}\right). \tag{12}$$

with μ_{GeDI} and μ_{MSE} are the metrics' thresholds. As for FAiRDAS, we rely on the Trust Region Methods to tackle the optimization problem.

5.2. Numerical Results

As a preliminary step, we examine how the eigenvalues λ of the FAiRDAS dynamic system influence action smoothness to determine their optimal values for the experiments. We conduct multiple runs with a fixed threshold while varying the eigenvalues (Table 1). As expected, based on the theoretical characteristics of the dynamic state under consideration, lower eigenvalues result in more stable actions in both groundings. For our experiments, we select the eigenvalues corresponding to the inflection point of the action smoothness metric.

Next, we compare the performance of FAiRDAS and the baseline under different pairs of thresholds for the metrics of interest. For both *Discrete Actions* and *Continuous Actions* settings, the threshold pair $\{0, 2\}$ represents an extreme scenario where fairness is prioritized over ranking performance. Subsequently, we examine a loose threshold pair, $\{0.7, 0.7\}$, and a stringent pair, $\{0.5, 0.5\}$. Finally, we investigate a pair of thresholds, $\{0.2, 0.2\}$, that cannot be reached.

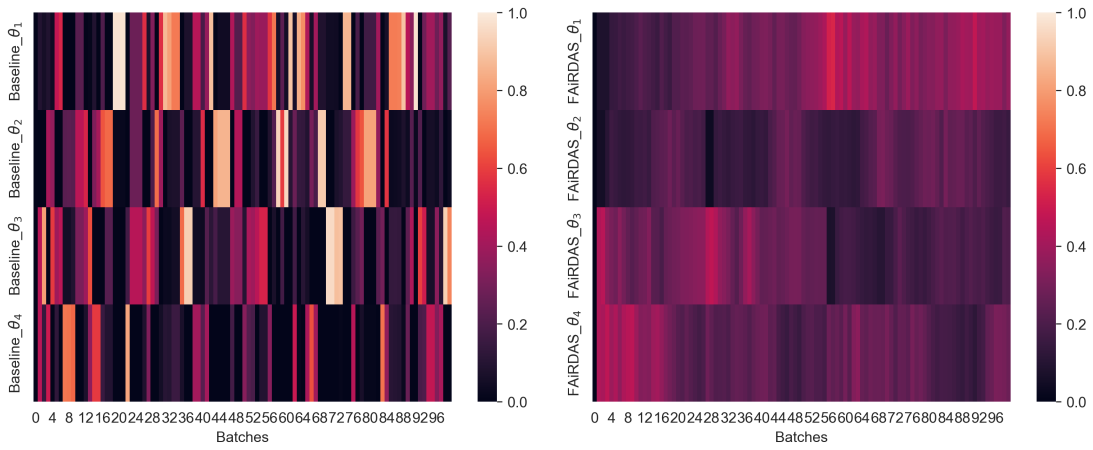
Table 1
Mean and standard deviation of the action smoothness computed over the batches for FAiRDAS. We analyse the results for 5 different eigenvalues (λ) with a fixed threshold pair $\{0.5, 0.5\}$. For each eigenvalue, we run eight experiments. We select $\lambda = 0.2$ as the elbow of the curve for both groundings (in bold).

	Discrete Actions		*Continuous Actions*	
λ	$\mathrm{m}_{\mathrm{Actions}}$	$\sigma_{\mathrm{m}_{\mathrm{Actions}}}$	$\mathrm{m}_{\mathrm{Actions}}$	$\sigma_{\mathrm{m}_{\mathrm{Actions}}}$
1.0	*0.201 ± 0.042*	*0.265 ± 0.050*	*0.044 ± 0.023*	*0.166 ± 0.032*
0.5	*0.054 ± 0.045*	*0.128 ± 0.106*	*0.011 ± 0.003*	*0.053 ± 0.016*
0.2	***0.036 ± 0.048***	***0.115 ± 0.116***	***0.006 ± 0.002***	***0.023 ± 0.010***
0.1	*0.031 ± 0.048*	*0.105 ± 0.123*	*0.005 ± 0.002*	*0.025 ± 0.012*
0.01	*0.029 ± 0.049*	*0.101 ± 0.126*	*0.001 ± 0.001*	*0.002 ± 0.002*

Table 2
Mean and standard deviation of the metrics computed over batches for *Discrete Actions*. We run eight experiments for each pair of thresholds and report the results for the baseline and FAiRDAS approach.

Thresholds	*Approach*	GeDI	σ_{GeDI}	SAE	σ_{SAE}	$\mathrm{m}_{\mathrm{Actions}}$	$\sigma_{\mathrm{m}_{\mathrm{Actions}}}$
$\{0, 2\}$	Baseline	*.388 ± .146*	*.606 ± .172*	*.644 ± .117*	*.550 ± .090*	*.139 ± .014*	*.183 ± .010*
	FAiRDAS	***.269 ± .104***	***.469 ± .124***	***.636 ± .087***	***.262 ± .047***	***.015 ± .010***	***.054 ± .058***
$\{0.7, 0.7\}$	Baseline	*.456 ± .108*	*.642 ± .124*	*.608 ± .159*	*.568 ± .127*	*.211 ± .065*	*.288 ± .063*
	FAiRDAS	***.294 ± .103***	***.487 ± .115***	***.612 ± .102***	***.261 ± .062***	***.046 ± .060***	***.144 ± .131***
$\{0.5, 0.5\}$	Baseline	*.510 ± .157*	*.694 ± .151*	*.627 ± .149*	*.660 ± .123*	*.273 ± .046*	*.301 ± .034*
	FAiRDAS	***.281 ± .097***	***.484 ± .111***	***.627 ± .103***	***.263 ± .096***	***.036 ± .048***	***.115 ± .116***
$\{0.2, 0.2\}$	Baseline	*.579 ± .158*	*.743 ± .168*	*.639 ± .157*	*.736 ± .130*	*.358 ± .043*	*.334 ± .013*
	FAiRDAS	***.289 ± .121***	***.511 ± .133***	***.638 ± .119***	***.377 ± .076***	***.027 ± .015***	***.079 ± .057***

Results with Discrete Actions. Table 2 presents the mean and standard deviation of the metrics throughout 100 batches for both baseline and FAiRDAS approach in *Discrete Actions* setting. Across all threshold pairs, the two methods achieve comparable levels of the metrics of interest (GeDI and SAE). However, the baseline exhibits notably higher levels of instability in the chosen actions (higher $\mathrm{m}_{Actions}$) compared to FAiRDAS, especially with stringent thresholds. This finding confirms the ability of FAiRDAS to maintain both performance effectiveness and fairness over time while also avoiding drastic actions that may raise ethical concerns. The increased stability of FAiRDAS approach is demonstrated in Figure 1, which shows the action vectors selected by both approaches in an experiment with stringent thresholds. This figure provides a component-wise comparison of the baseline and FAiRDAS action vectors across all 100 batches. As detailed in Section 4.1, each component of the action vectors affects students from the corresponding ESCS level and acts as penalizing factors on their scores, potentially altering their ranking. Higher values indicate more significant penalization, while values near zero mean the student's score remains untouched. The baseline method tends to favor rapid and drastic interventions, indicated by 1) the sudden color change between batches and 2) action components close to one (lighter color). In contrast, FAiRDAS exhibits a more moderated and balanced behavior, with action vectors evolving smoothly over the experiment (gradual color changes along x-axis) and similar penalization across groups (uniform color along y-axis).



(a) (b)

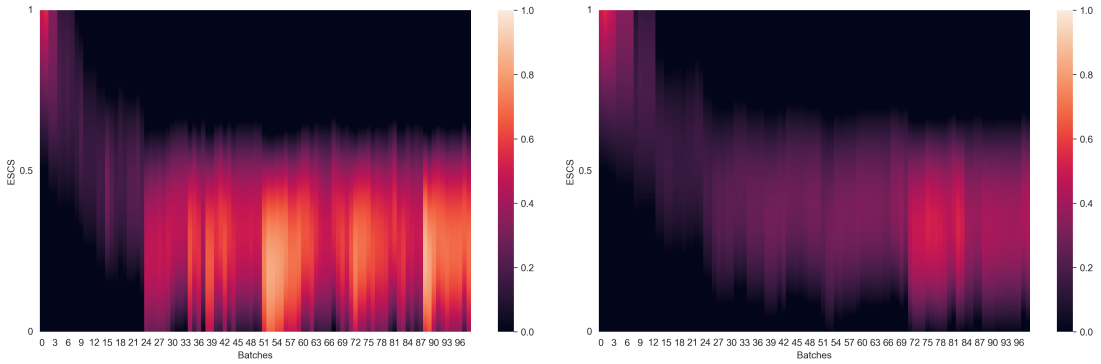
Figure 1: An example of the action vectors selected by the baseline (a) and FAiRDAS (b) in an experiment with thresholds ($\{0.2, 0.2\}$). Each row shows the progression of the corresponding action vector component throughout 100 batches. FAiRDAS exhibits a more moderated and balanced behaviour, with action vectors evolving smoothly over the experiment.

Results with Continuous Actions. In Table 3 we report the mean and standard deviation of the metrics over 100 batches for both baseline and FAiRDAS approach under *Continuous Actions* setting. As with *Discrete Actions*, the numerical results confirm FAiRDAS's capability to maintain both performance effectiveness and fairness over time while avoiding drastic actions. FAiRDAS and the baseline achieve similar levels for the metrics of interest (GeDI and MSE) across all thresholds, but FAiRDAS reaches lower values of action smoothness ($m_{Actions}$). This result is

Table 3
Mean and standard deviation of the metrics computed over batches for *Continuous Actions*. We run eight experiments for each pair of thresholds and report the results for the baseline and FAiRDAS.

Thresholds	*Approach*	GeDI	σ_{GeDI}	MSE	σ_{MSE}	m_{Actions}	$\sigma_{m_{\text{Actions}}}$
$\{0, 2\}$	Baseline	*.629 ± .267*	*1.202 ± .337*	*.572 ± .197*	*.690 ± .078*	*.015 ± .005*	*.101 ± .056*
	FAiRDAS	***.449 ± .162***	***.717 ± .145***	***.546 ± .202***	***.684 ± .087***	***.005 ± .001***	***.024 ± .006***
$\{0.7, 0.7\}$	Baseline	***.629 ± .235***	*.970 ± .325*	*.550 ± .200*	***.688 ± .082***	*.017 ± .004*	*.108 ± .039*
	FAiRDAS	*.676 ± .194*	***.782 ± .272***	***.532 ± .198***	*.697 ± .088*	***.005 ± .003***	***.020 ± .017***
$\{0.5, 0.5\}$	Baseline	***.644 ± .213***	*1.011 ± .298*	*.555 ± .198*	***.688 ± .08***	*.017 ± .005*	*.107 ± .048*
	FAiRDAS	*.712 ± .275*	***.875 ± .375***	***.536 ± .204***	*.689 ± .087*	***.006 ± .002***	***.023 ± .010***
$\{0.2, 0.2\}$	Baseline	*.675 ± .233*	*1.152 ± .312*	*.567 ± .194*	***.688 ± .079***	*.015 ± .005*	*.100 ± .057*
	FAiRDAS	***.607 ± .201***	***.839 ± .284***	***.534 ± .200***	*.692 ± .088*	***.008 ± .002***	***.034 ± .013***

exemplified in Figure 2, where we present an example of the polynomial functions selected by FAiRDAS and the baseline throughout 100 batches. Each column displays the function chosen for the corresponding batch, evaluated over the ESCS domain $[0, 1]$ (y-axis). As described in Section 4.2, the functions influence the ranking based on the students' ESCS value, serving as a penalizer on their scores: lower values correspond to more substantial penalization, whereas



(a) (b)

Figure 2: An example of the polynomial functions selected by the baseline (a) and FAiRDAS (b) in an experiment with thresholds ($\{0.2, 0.2\}$). Each column represents the polynomial function selected for the corresponding batch evaluated on domain $[0, 1]$.

values close to one indicate that the student's score is unaffected. As for *Discrete Actions*, we observe that the baseline method tends to favor rapid and drastic actions, as indicated by 1) the abrupt color changes between batches and 2) the high penalization values (higher contrast). Conversely, FAiRDAS demonstrates a more moderated and balanced behaviour, with polynomial functions evolving smoothly throughout the batches (gradual color changes along x-axis) and more consistent penalization across different ESCS values (smooth color changes along y-axis).

6. Conclusion

We introduced a novel approach that integrates state-of-the-art techniques to address long-term fairness in the presence of continuous protected attributes. This is achieved by pairing FAiRDAS [15], a framework aimed at ensuring long-term fairness in ranking systems while preserving stable actions, with the Generalized Disparate Impact (GeDI) indicator [22], a fairness metric specifically designed to handle continuous protected attributes. Our contribution includes the definition of two possible sets of actions to handle continuous attributes. The first set prioritizes interpretability but introduces discretization, whereas the second set maintains the continuity of actions at the expense of interpretability. The selection of the set of actions to apply depends on the specific requirements and constraints of the application context. We validated our methodology through a case study in the domain of AI and Education, where we compared the performance and stability of FAiRDAS against a baseline method. Our analysis demonstrates that the integration of FAiRDAS and GeDI with our defined actions presents a robust solution for addressing long-term fairness under continuous protected attributes.

To the best of our knowledge, this is the first work that tackles long-term fairness and stability in ranking with continuous attributes. Thus, we believe that it could lay the groundwork for further research and applications in several domains where handling continuous attributes and stability are of key importance, yet currently understudied.

References

[1] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, A. Galstyan, A survey on bias and fairness in machine learning, volume 54, ACM New York, NY, USA, 2021, pp. 1–35.

[2] M. Zehlike, K. Yang, J. Stoyanovich, Fairness in ranking: A survey, arXiv preprint arXiv:2103.14000 (2021).

[3] L. T. Liu, S. Dean, E. Rolf, M. Simchowitz, M. Hardt, Delayed impact of fair machine learning, in: J. G. Dy, A. Krause (Eds.), Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018, volume 80 of *Proceedings of Machine Learning Research*, PMLR, 2018, pp. 3156–3164. URL: http://proceedings.mlr.press/v80/liu18c.html.

[4] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, A. Galstyan, A survey on bias and fairness in machine learning, ACM Comput. Surv. 54 (2021). URL: https://doi.org/10.1145/3457607. doi:10.1145/3457607.

[5] F. Kamiran, T. Calders, Data preprocessing techniques for classification without discrimination, Knowledge and Information Systems 33 (2011) 1–33. URL: http://dx.doi.org/10.1007/s10115-011-0463-8. doi:10.1007/s10115-011-0463-8.

[6] F. Calmon, D. Wei, B. Vinzamuri, K. Natesan Ramamurthy, K. R. Varshney, Optimized pre-processing for discrimination prevention, in: I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, R. Garnett (Eds.), Advances in Neural Information Processing Systems, volume 30, Curran Associates, Inc., 2017. URL: https://proceedings.neurips.cc/paper/2017/file/9a49a25d845a483fae4be7e341368e36-Paper.pdf.

[7] L. E. Celis, V. Keswani, N. Vishnoi, Data preprocessing to mitigate bias: A maximum entropy based approach, in: International conference on machine learning, PMLR, 2020, pp. 1349–1359.

[8] T. Kamishima, S. Akaho, H. Asoh, J. Sakuma, Fairness-aware classifier with prejudice remover regularizer, in: P. A. Flach, T. De Bie, N. Cristianini (Eds.), Machine Learning and Knowledge Discovery in Databases, Springer Berlin Heidelberg, Berlin, Heidelberg, 2012, pp. 35–50.

[9] J. Komiyama, A. Takeda, J. Honda, H. Shimao, Nonconvex optimization for regression with fairness constraints, in: J. Dy, A. Krause (Eds.), Proceedings of the 35th International Conference on Machine Learning, volume 80 of *Proceedings of Machine Learning Research*, PMLR, 2018, pp. 2737–2746. URL: https://proceedings.mlr.press/v80/komiyama18a.html.

[10] E. Y. Yu, Z. Qin, M. K. Lee, S. Gao, Policy optimization with advantage regularization for long-term fairness in decision systems, arXiv preprint arXiv:2210.12546 (2022).

[11] Y. Ge, S. Liu, R. Gao, Y. Xian, Y. Li, X. Zhao, C. Pei, F. Sun, J. Ge, W. Ou, et al., Towards long-term fairness in recommendation, in: Proceedings of the 14th ACM international conference on web search and data mining, 2021, pp. 445–453.

[12] T. Calders, S. Verwer, Three naive bayes approaches for discrimination-free classification, Data Min. Knowl. Discov. 21 (2010) 277–292. doi:10.1007/s10618-010-0190-x.

[13] M. Hardt, E. Price, E. Price, N. Srebro, Equality of opportunity in supervised learning, in: D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, R. Garnett (Eds.), Advances in Neural Information Processing Systems, volume 29, Curran Associates, Inc., 2016. URL: https://proceedings.neurips.cc/paper/2016/file/9d2682367c3935defcb1f9e247a97c0d-Paper.pdf.

[14] R. Xian, L. Yin, H. Zhao, Fair and optimal classification via post-processing, in: International Conference on Machine Learning, PMLR, 2023, pp. 37977–38012.

[15] E. Misino, R. Calegari, M. Lombardi, M. Milano, Fairdas: Fairness-aware ranking as dynamic abstract system, in: R. Calegari, A. A. Tubella, G. González-Castañé, V. Dignum, M. Milano (Eds.), Proceedings of the 1st Workshop on Fairness and Bias in AI co-located with 26th European Conference on Artificial Intelligence (ECAI 2023), Kraków, Poland, October 1st, 2023, volume 3523 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2023. URL: https://ceur-ws.org/Vol-3523/paper5.pdf.

[16] M. Zehlike, F. Bonchi, C. Castillo, S. Hajian, M. Megahed, R. Baeza-Yates, Fa* ir: A fair top-k ranking algorithm, in: Proceedings of the 2017 ACM on Conference on Information and Knowledge Management, 2017, pp. 1569–1578.

[17] S. C. Geyik, S. Ambler, K. Kenthapadi, Fairness-aware ranking in search & recommendation systems with application to linkedin talent search, in: Proceedings of the 25th acm sigkdd international conference on knowledge discovery & data mining, 2019, pp. 2221–2231.

[18] A. Singh, T. Joachims, Fairness of exposure in rankings, in: Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining, 2018, pp. 2219–2228.

[19] A. J. Biega, K. P. Gummadi, G. Weikum, Equity of attention: Amortizing individual fairness in rankings, in: The 41st international acm sigir conference on research & development in information retrieval, 2018, pp. 405–414.

[20] J. Mary, C. Calauzènes, N. E. Karoui, Fairness-aware learning for continuous attributes and treatments, in: K. Chaudhuri, R. Salakhutdinov (Eds.), Proceedings of the 36th International Conference on Machine Learning, volume 97 of *Proceedings of Machine Learning Research*, PMLR, 2019, pp. 4382–4391. URL: https://proceedings.mlr.press/v97/mary19a.html.

[21] V. Grari, S. Lamprier, M. Detyniecki, Fairness-aware neural rényi minimization for continuous features, in: C. Bessiere (Ed.), Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20, International Joint Conferences on Artificial Intelligence Organization, 2020, pp. 2262–2268. URL: https://doi.org/10.24963/ijcai.2020/313. doi:10.24963/ijcai.2020/313, main track.

[22] L. Giuliani, E. Misino, M. Lombardi, Generalized disparate impact for configurable fairness solutions in ML, in: A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato, J. Scarlett (Eds.), Proceedings of the 40th International Conference on Machine Learning, volume 202 of *Proceedings of Machine Learning Research*, PMLR, 2023, pp. 11443–11458. URL: https://proceedings.mlr.press/v202/giuliani23a.html.

[23] S. Aghaei, M. J. Azizi, P. Vayanos, Learning optimal and fair decision trees for non-discriminative decision-making, in: The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019, AAAI Press, 2019, pp. 1418–1426. URL: https://doi.org/10.1609/aaai.v33i01.33011418. doi:10.1609/aaai.v33i01.33011418.

[24] F. Avvisati, The measure of socio-economic status in pisa: a review and some suggested improvements, Large-scale Assessments in Education 8 (2020). URL: http://dx.doi.org/10.1186/s40536-020-00086-x. doi:10.1186/s40536-020-00086-x.