Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence (IJCAI-24)
Special Track on AI for Good

Ensuring Fairness Stability for Disentangling Social Inequality in Access to Education: the FAiRDAS General Method

Eleonora Misino , **Roberta Calegari** , **Michele Lombardi** and **Michela Milano**
University of Bologna
eleonora.misino2@unibo.it, roberta.calegari@unibo.it, michele.lombardi2@unibo.it, michela.milano@unibo.it

Abstract

Recent advancements in Artificial Intelligence in Education (AIEd) have revolutionized educational practices using machine learning to extract insights from students' activities and behaviours. Performance prediction, a key domain within AIEd, aims to enhance student achievement levels and address sustainable development goals related to education, health, gender equality, and economic growth. However, the potential of AIEd to contribute to these goals is hindered by the lack of attention to fairness in prediction algorithms, leading to educational inequality. To address this gap, we introduce FAiRDAS, a general framework that models long-term fairness as an abstract dynamic system. Our approach, illustrated through a case study in AIEd with real data, offers a customizable solution to promote long-term fairness while promoting the stability of mitigation actions over time.

1 Introduction

In the past few years, AI in Education (AIEd) has advanced by leveraging machine learning to extract valuable insights from students' activities, educational and social behaviours, and academic backgrounds [Chassignol *et al.*, 2018]. AIEd has made significant strides in various domains, such as path recommendation, teaching strategy optimization, and performance prediction. Performance prediction, in particular, aims to advance student achievement levels by offering more effective and personalized teaching strategies or identifying influential factors (such as dropout probabilities) that can advise supporting actions [Albreiki *et al.*, 2021].

These objectives align seamlessly with Sustainable Development Goals (SDGs), encompassing quality education, good health and well-being, gender equality, decent work and economic growth, reduced inequality, peace, justice, and strong institutions. The social relevance of addressing these issues is underscored by the detrimental effects of misdirecting students towards specific careers or failing to intervene in cases where academic cessation could be prevented and has far-reaching societal and economic implications.

However, while AIEd's potential to significantly contribute to SDGs is considerable, the necessity for careful consideration and proactive measures to consider potential conflicts with equity goals and adhere to the 'Leave No One Behind' principle is a strong requirement. Indeed, it is imperative to address biases in AI algorithms, as they can potentially impact specific groups disproportionately. In the considered context, using student performance predictions in policy decisions becomes problematic when systematic errors emerge, leading to disparate and unfair outcomes.

The fairness of AI predictions in educational contexts has received insufficient attention, and the existing approaches often do not fully analyze the properties of how fairness is achieved [Kizilcec and Lee, 2022]. For example, let us consider an AI system to rank students for performance analysis to allocate scholarships. In this application, fairness can be ensured, for example, by adjusting the ranking to guarantee that all sensitive groups have equal opportunities for access to education over time. Algorithmic solutions for this purpose exist and are often based on choosing mitigation actions based on historical samples. However, these methods often overlook the fact that the ranking process is typically *repeated over time*, and therefore, decisions made in one step can affect long-term fairness [Liu *et al.*, 2018].

In particular, historical data can be subject to significant sampling noise (e.g., resulting from interactions and events affecting individual cohorts) and does not necessarily account for trends in the population distribution. Consequently, decisions made in one step can adversely affect subsequent iterations; for example, a sensitive group could be penalized for being over-represented in one instance; alternatively, decisions can become highly unstable over iterations, making them difficult to motivate from an ethical standpoint. Attempting to predict the population dynamics can help, but it comes with its own issues, such as inaccurate estimates (that may have a detrimental social impact) and increased computational load (due to the need to sample future outcomes).

As an alternative, in this paper, we propose to handle stability by mapping the long-term evolution of fairness metrics on sequential ranking processes as an abstract dynamic system. Based on this idea, we introduce FAiRDAS [1] (Fairness-Aware

[1]This work reports on a substantially improved version of the framework introduced in preliminary form in [Misino *et al.*, 2023], as well as its grounding in the educational field. As a testament to the method's suitability to multiple contexts (such as hiring or lending), the original paper targeted a different application domain.

Ranking as Dynamic Abstract System), a configurable framework that enables controlling the trade-off between multiple fairness/quality metrics and, indirectly, the level of stability of the mitigation actions while ensuring long-term fairness

The paper is organized as follows. The AIEd case study and the problem formulation are presented in Section 2. In Section 3, we provide the related fairness concepts and the relevant state-of-the-art. The general framework is formalized in Section 4. Section 5 presents the experimental setting, describing FAiRDAS groundings to the AIEd case study, and discusses numerical results.

2 Problem Formulation

We focus on a case study in AI and education to contextualize and motivate our approach, and to serve as a guiding example for presenting a general problem formalization.

2.1 Motivating Case Study

The chosen case study involves ranking students based on predicted academic performance for identifying potential dropouts or making recommendations. Real-world data and guidance for the identification of objectives and quality and fairness metrics were provided by the Canarian Agency for Quality Assessment and Accreditation (ACCUEE) [2].

Available Data. The agency collects information to evaluate the performance of the Canary Islands educational system through periodic diagnostic reports. The available data encompasses information spanning four academic years (2015-2019). The diagnostic procedure involves two primary components: (1) assessment of students' academic proficiency across various subjects, including Mathematics, Spanish Language, and English; (2) context questionnaires to students, school principals, families, and teachers, focusing on gathering socio-demographic background information.

Performance Prediction. The agency is interested in the early identification of problematic situations through AIEd solutions so that appropriate support actions can be implemented in a timely manner. For our analysis, we select the mathematics test score as the outcome variable to be predicted due to its minimal number of missing values. Consequently, we assume that students are ranked based on this score and that correctly predicting its value becomes critical for the effectiveness of the support program.

Equal Opportunities. While accurate estimates are important, the need to ensure equal opportunity in this scenario regardless of students' social backgrounds is also recognized as a problem in the literature [Pedro *et al.*, 2019]. For this purpose, the generated ranking can be adjusted to align with long-term fairness objectives.

In our case study, we aim to ensure prediction accuracy while avoiding disparate impacts related to different socioeconomic backgrounds. The chosen protected variable is the Economic, Social, and Cultural Status (*ESCS*), serving as a proxy for students' socioeconomic status. This index is derived from students' access to family resources, which determine the social position of their family/household. Given that *ESCS* is a continuous variable, it has been transformed into a categorical form through a quantile-based function.

Stability. Long-term stability is crucial in this scenario: on the one hand, consistently good values of the accuracy and long-term fairness metrics are obviously desirable, even if difficult to achieve due to their conflicting natures; on the other hand, mitigation actions should also be stable over time. In fact, unstable or inconsistent treatment over time could have lasting effects on student's academic progress and overall educational experience, in addition to being likely unacceptable in terms of public opinion.

2.2 FAiRDAS Approach under an Intuitive Lens

FAiRDAS is applicable across various scenarios aimed at fostering *long-term fairness* and *its stability* over time. This is especially critical in contexts where biases might persist or accumulate over time, resulting in systemic disparities or inequalities in the system's overall outcomes. Examples encompass education, hiring, lending, and similar domains where decisions based on rankings or assessments can cause continuing repercussions, potentially perpetuating existing biases if not appropriately addressed.

Figure 1a illustrates a possible scenario: we have different rankings generated by the system over time (for instance ranking t_0 can represent students enrolled in 2016, t_1 those in 2017, and so forth). Measuring long-term fairness means assessing the fairness of all produced rankings over time (so on $t_1, ..t_n$) instead of measuring it only on t_i, thus enabling considerations of the systemic impact and cumulative effects of the system rankings. It is worth noting that considering long-term fairness enables flexibility in the strictness of single-query fairness requirements. Consequently, the AI system can better reflect the demographic diversity of the population represented in a single query without significantly compromising accuracy. For instance, coming back to the example in Figure 1, it means that it might be acceptable for ranking t_0 to predominantly consist of *ESCS*=1 students, potentially reflecting the demographic composition of that year, with subsequent rankings $t_1, ..t_n$ compensating for this over time.

In the generic scenarios depicted in Figure 1a, ensuring long-term fairness involves taking action on the generated rankings to satisfy the fairness constraints. These actions can take various forms – such as adjusting scoring methods or repositioning candidates at the top/bottom of the ranking – all of which ultimately reduce swapping candidates (as illustrated in Figure 1a (bottom)). During these swap operations, the actions can be either gradual or drastic.

An extreme example of drastic actions is illustrated in Figure 1b, where mitigation measures respond to the presence of many students with *ESCS*=1 (blue circle) at time t_0 by imposing significant penalties on them at t_1 causing students with *ESCS*=1 to disappear from the rankings in t_1 completely. This process repeats at each iteration for different demographic groups; for instance, students with an *ESCS*=4 (green circle) are penalized at t_2, while those with an *ESCS*=3 (purple circle) are moved down in the rankings at t_3 due to their prevalence in previous iterations. While the long-term fairness constraint is respected (all the sensitive groups have equal opportunity over time), such drastic decisions raise ethical concerns

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

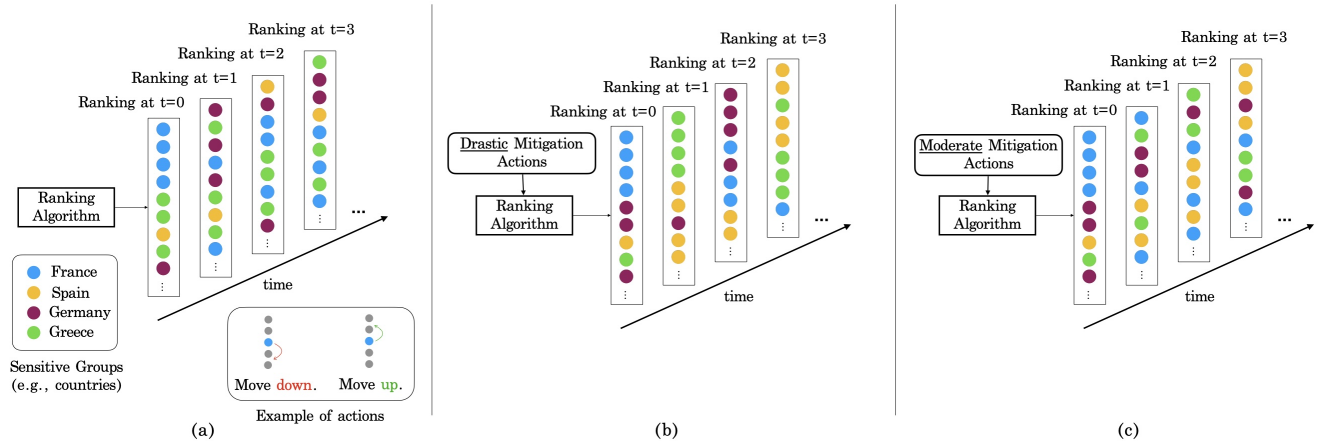


Figure 1: (a) General application scenario for FAiRDAS: generating resources (student) ranking over time; different colours correspond to different sensitive features (ESCS). Possible actions to promote fairness reduce to candidate swapping. (b) Drastic actions impose significant penalties on demographic groups at each iteration. (c) Moderate actions preserve long-term fairness without suppressing demographic groups.

as they unfairly disadvantage certain groups in the single ranking (making some minorities completely disappear for one whole year). Thus, there is a critical need for an approach that promotes not only long-term fairness but also long-term fairness stability, thereby enhancing the overall trustworthiness of the system. For instance, a more stable approach would select moderate actions that gradually respond to the over-representation of demographic groups without abruptly penalizing a group of students. For example, in Figure 1c, the moderate actions preserve long-term fairness without suppressing the initially over-represented group of students with *ESCS*=1 in the following iteration.

2.3 Ranking Problem Formulation

We can now provide a formalization that is general enough to model our case study and other similar applications.

Resources and Batches. We target a process where a set $\mathcal{R}$ of m resources (e.g., students, professional profiles) needs to be repeatedly ranked over time based on observable information arriving over time (e.g., academic record and personal situation of each cohort, expertise of the professionals with respect to incoming customer requests). We refer to the observable information as *batches* and view them as a time-indexed stochastic process $\{X_t\}_{t=1}^{\infty}$, where each batch X_t is a random variable with support $\mathcal{X}$ and distribution $P(X_t)$.

Actions and Metrics. We assume that the ranking procedure can be controlled by adjusting the values of an *action vector* $\theta \in \Theta$, representing (e.g.) penalty or reward terms associated with sensitive groups, or simply the parameters of an ML model. We abstract from other details of the ranking process by focusing instead on how its outcome affects efficiency, cost, fairness, or any applicable Key Performance Indicator. Formally, the ranking quality is represented by means of a metric function defined in probabilistic terms (e.g., based on expectations or event probabilities):

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

where the vector $y[X;\theta] \in \mathbb{R}^n$ represents the value of n metrics for a given batch X and for a given action vector θ. In practical settings, the metrics will always admit a finite sample formulation, e.g., obtained by replacing theoretical expectations with sample averages.

We assume that ranking is repeated for every batch, followed by an adjustment of the action vector so that the overall problem can be defined in terms of the tuple:

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

where θ_t is the action vector at time t and y_t refers to the value of the metric function for time t, i.e. to $y[X_t;\theta_t]$. Ex-post calibration of the action vector is not an essential assumption of our framework, and it was made here since it offers advantages in terms of response time (the ranking can be computed without needing to optimize θ_t) and transparency (if θ_t is interpretable, it can be made public before ranking).

3 Related Work

3.1 Fairness Metrics and Sensitive Features

In the context of AIEd, notions of fairness stemming from the concept of equal opportunity are typically considered [Kizilcec and Lee, 2022]. Various variants have been proposed to overcome limitations of the specific scenario under analysis (e.g., applicability to continuous features [Holstein and Doroudi, 2019; Giuliani *et al.*, 2023], generalization of the equal opportunity concept [Blandin and Kash, 2023], overestimation of unfairness [Jiang and Pardos, 2021]). In our case study, the fairness metric (or fairness metrics) enforced in the system is entirely at the user's discretion. For this reason, a disparate impact metric [Zafar *et al.*, 2017] has been selected as the fairness metric for the following tests, as it is widely used in literature in these scenarios and is easily understandable. We emphasize that the choice of fairness metric does not impact the results presented in this work.

Regarding sensitive features, many AIEd approaches only utilize gender and race, often managing one feature at a time

[Jiang and Pardos, 2021; Hu and Rangwala, 2020]. The advantage of FAiRDAS is the ability to include multiple sensitive features for consideration and achieve a balance in terms of long-term fairness that considers all features. We consider the Economic, Social, and Cultural Status as the sensitive feature, which is more appropriate for addressing discrimination in this scenario [Pedro *et al.*, 2019].

3.2 Debiasing Algorithms

Debiasing techniques have gained significant attention in AIEd due to their potential to mitigate disparities and enhance fairness in educational settings [Sha *et al.*, 2023].

One prominent set of debiasing techniques involves preprocessing steps to mitigate bias in the training data. For instance, approaches such as reweighting samples or data balancing [Celis *et al.*, 2020] have been explored to mitigate biases related to demographic factors.

Another class of techniques involves post-processing methods that aim to adjust model predictions to achieve fairness after learning. Post-processing techniques typically involve applying corrective measures to model outputs to ensure fairness, such as re-ranking or re-calibrating predictions based on demographic attributes [Xian *et al.*, 2023].

One of the most promising techniques is the in-processing methods involving modifying loss functions, introducing fairness-aware regularization terms, or optimization techniques [Caton and Haas, 2020]. Among these models, sequential minimal optimization works better for predicting students' future performance [Arashpour *et al.*, 2023] (this led to our baseline choice). However, the main problem with current techniques is that few studies focus on the concept of long-term fairness [Yu *et al.*, 2022; Ge *et al.*, 2021]. Even more rarely are the actions taken to ensure long-term fairness analyzed [Yin *et al.*, 2024; Hu and Zhang, 2022]; i.e., solutions often result in unstable actions that adopt overly drastic measures to achieve fairness goals (e.g., the almost complete disappearance of a sensitive group as a response to balance its overexposure in previous rankings). In this work, we will focus on both the concept of long-term fairness and the stability of actions taken to ensure it.

4 FAiRDAS

This section introduces the FAiRDAS framework. The main idea revolves around conceptualizing the evolution of fairness/quality metrics as a dynamic system, allowing the user to define a target behaviour that can then be approximated by operating on the action sequence $\{\theta_t\}_{t=1}^{\infty}$. By configuring the parameters of the target dynamic system, the user can control the trade-off between multiple quality metrics, and the desired level of smoothness and stability.

Target Dynamic System. We assume that the goal for the evolution of the system is to stably and smoothly drive the metrics of interest $y_t \in \mathbb{R}^n$ below a user-defined threshold $\mu \in \mathbb{R}^n$. The desire for stable behaviour with limited oscillations suggests characterizing the target behaviour via a linear discrete dynamic system. In particular, we use:

$$\bar{y}_{t+1} = \lambda \odot (\bar{y}_t - \mu) + \bar{y}_t, \tag{3}$$

where $\bar{y}_t$ represent the metric values in the target system, $\lambda \in (0, 2)^n$, and $\odot$ refers to the Hadamard (element-wise) product. Equation (3) corresponds to a particular class of linear discrete system having a positive-definite coefficient matrix with eigenvalues strictly lower than 2. Such systems are known to asymptotically reach a stable equilibrium at μ, i.e. $\lim_{t\to\infty} \bar{y}_t = \mu$. Note that μ represents here an equilibrium point rather than a threshold, meaning that metric values below their threshold will actually be increased: we compensate for this fact in the next step of our approach.

The convergence dynamics can be controlled via the λ vector: values of λ_j close to 0 result in little or no oscillations for $\bar{y}_{t,j}$, but typically also in slower convergence; as λ_j approaches 2 the corresponding metrics evolve much faster, but can also exhibit more frequent and wider oscillations.

Approximating the Target Behavior. The second key idea in FAiRDAS is to approximate the target system's behaviour by operating on the action vector. Formally, this requires solving an optimization problem whose cost function $\mathcal{L}(\theta, \bar{y})$ measures the discrepancy between the actual metrics y_t, as determined by the actions θ_t and the target ones $\bar{y}_t$. In this paper, we use the Euclidean distance:

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

though other functions can be employed. During the approximation step, we also account for the fact that all values of $y_{t,j} \leq \mu_j$ are equivalent in terms of quality since they are all below the desired threshold. This is achieved by avoiding penalties for target metrics that are below their threshold, and in particular by formulating the problem of choosing θ as:

$$\theta^*(\bar{y}_t) = \arg\min_{\theta\in\Theta} \mathcal{L}(\theta_t, u) \tag{5a}$$

$$\text{subject to: } u_j = \bar{y}_{t,j} \qquad \forall j : \bar{y}_{t,j} > \mu_j \tag{5b}$$

$$u_j \in [0, \mu_j] \qquad \forall j : \bar{y}_{t,j} \leq \mu_j \tag{5c}$$

For computing the optimal action vector $\theta^*(\bar{y}_t)$, we replace $\bar{y}_t$ in the distance function with a new variable u. When $\bar{y}_{t,j}$ is above a threshold, the action vector θ should be chosen to obtain a good approximation. When $\bar{y}_{t,j}$ is within the threshold, u_j can be freely adjusted to achieve 0 distance.

The solution method for Equation (5a)-(5c) depends on the characteristics of the action space and of the distance function $\mathcal{L}$. Moreover, in practice, evaluating Equation (4) exactly is impossible in all but simple cases since the distribution $P(X_t)$ will not be precisely known. For this reason, the metric values $y[X_t; \theta_t]$ will need to be replaced (e.g.) by a Monte Carlo approximation computed on historical data. Consequently, the problem solution will also become approximate and subject to sampling noise.

Note that FAiRDAS targets metric stability directly (by relying on the target dynamic system) and action stability *indirectly*. This is intentional and motivated by the fact that the set of available actions may be a poor fit for classical definitions of smoothness (e.g., discrete actions), and it might even change over time (e.g., the addition of new sensitive groups).

Scale Calibration Mechanism. The optimization problem from Equation (5a)-(5c) is sensitive to the metric scales. To address this issue, we propose normalizing the n metrics by:

1) considering a set of k historical batches; 2) optimizing the action vector for each metric individually over each batch; 3) applying the actions to compute all the metrics, thus obtaining a sample of $n \times k$ values for each metric representing its distribution under different action vectors; 4) using the interquartile range of each metric on the sample as a normalization factor. We found this process to be typically effective at rescaling different metrics into similar ranges. This property allows the users to specify priorities by defining threshold rather than using different weights at optimization time.

FAiRDAS General Framework. To ground FAiRDAS on a specific application, we need to define the following list of parameters. The *metrics of interest* define how the fairness and ranking quality should be assessed. The behaviour of the *target dynamical system* from Equation (3) is determined by the vector λ and the threshold vector μ. A careful definition of both these elements is required to satisfy all the objectives.

The *set of actions* defines how the metrics can be altered to adjust the ranking fairness and quality. The available actions can vary widely, from direct manipulation of the resource ordering to the adjustment of penalty factors or of the parameters of an ML model (e.g., neural network weights). The *distance function* defines how we measure the effectiveness of the approximation of the target system; while the Euclidean norm should work in most cases, specific settings may call for a different choice. Finally, the set of actions and the distance function determine to a large extent which *optimization methods* can be used to address Equation (5a)-(5c); among this pool of candidates, one must be chosen considering computational efficiency, accuracy, and optimality guarantees.

5 Empirical Evaluation

In this section, we present the empirical evaluation performed on the AIEd case study described in Section 2.1. We first describe the dataset and FAiRDAS grounding. Then, we describe the evaluation procedure and report the numerical results[3].

5.1 Dataset

Students Dataset. As described in Section 2.1, the motivating case study revolves around assessing students' academic performance predictions for tasks like identifying potential dropouts or providing recommendations. With this objective, we train a multi-layer perceptron (MLP) on the real data to predict the student's test score based on three highly correlated features: the number of books, the mother's education, and the index of economic, social, and cultural status (*ESCS*). Ranking students based on the MLP predictions may lead to disparate and unfair outcomes. Ensuring long-term stability is imperative in this situation: while attaining consistently high values in accuracy and long-term fairness metrics is undeniably desirable, it's crucial to have stable actions over time to avoid a negative impact on students' academic advancement.

To evaluate FAiRDAS ability to handle stability over time, we create the *Students Dataset* consisting of 100 batches of 32 students sampled from the distribution of predicted scores to simulate polarized requests in terms of *ESCS*The resulting *score* distribution is highly correlated with the protected attribute *ESCS* (Figure 2); thus, ranking students without taking any mitigation action may affect the support program effectiveness and lead to social inequalities.

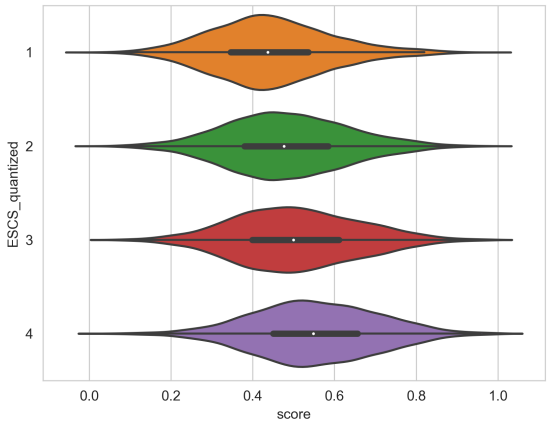


Figure 2: The students' *score* distribution predicted by the MLP shows a correlation with the protected attribute *ESCS*.

5.2 FAiRDAS Grounding

As described in Section 4, we need to define a list of parameters to ground FAiRDAS to a specific application. Here we report the grounding for our case study.

Set of Actions. We choose a set of actions that applies directly to the scores used by the ranking algorithm. Formally, given the protected attribute $v \in \mathcal{V}$, the actions take the form of a vector $\theta \in [0,1]^{|\mathcal{V}|}$ with unit L1 norm, such that each component θ_v applies to the scores of those resources with protected attribute value equal to v. The action vector components apply to the students of the corresponding protected groups as a penalizing factor on their score, thus potentially changing their position in the rank. In particular, values closer to 1 correspond to more drastic penalization, while the student's score is almost unmodified with values close to zero.

In the *Students Dataset*, we have four sensitive groups corresponding to the four levels of *ESCS* indicator. Thus, the action vector has four components, each applying to the students belonging to the corresponding sensitive group. Suppose we want to respond to an overrepresentation of students with *ESCS*=4. An example of drastic action in this scenario is $\theta = \{0, 0, 0, 1\}$ that strongly penalizes all the students with *ESCS*=4 by suppressing their score while leaving the other students' scores unmodified. As a result, all students with *ESCS*=4 will be placed in the last positions of the rank. Conversely, an example of a smooth mitigation action is $\theta = \{0.22, 0.22, 0.22, 0.34\}$, where the score penalization is spread across the four sensitive groups without harsh penalization of a particular category. The resulting ranking will be closer to the original one, with few adjustments in response to an over-representation of *ESCS*=4.

[3]The source code to reproduce the experiments can be found at https://github.com/EleMisi/FAiRDAS_AIforEd under MIT license.

Metrics of Interest. We use the *Disparate Impact Discrimination Index* (DIDI) [Aghaei *et al.*, 2019] as a fairness metric, as it is widely used in literature in similar scenarios and is easily understandable. It is important to emphasize that the choice of fairness metric does not impact the results presented in this work. Given a sample including K values for a protected attribute $v \in \mathcal{V}$ and a continuous target value $f \in \mathbb{R}$, the DIDI is defined as:

$$\mathrm{DIDI}(\theta) = \sum_{v \in \mathcal{V}} \left| \frac{\sum_{k=1}^{K} f_k(\theta) I(v_k = v)}{\sum_{k=1}^{K} I(v_k = v)} - \frac{1}{K} \sum_{k=1}^{K} f_k(\theta) \right|. \tag{6}$$

where $I(\psi)$ is the indicator function for the logical formula ψ, and the target value is a function of the action vector θ applied to the ranking score. In particular, the score s_k of a student is provided by the pre-trained MLP; thus, we have:

$$f_k(\theta) = |s_k|(1 - \theta_{v_k}), \tag{7}$$

where θ_{v_k} is the component of the action vector corresponding to the *ESCS* level of the k-th student. To quantify the ranking accuracy, we measure the absolute difference between the original and modified scores. This translates to:

$$\mathrm{SAE}(\theta) = \frac{1}{K} \sum_{k=1}^{K} |s_k| \theta_{v_k}, \tag{8}$$

where K is the number of students in a request. Note that the two metrics of interest are conflicting: SAE pushes θ_{v_k} close to zero to preserve the original ranking, while DIDI forces $\theta_{v_k} > 0$ for some k to reduce discrimination. Since the action vector has unit L1 norm, the trivial solution with $\theta_{x_k} = 0 \; \forall \; k$, which nullifies both metrics, is excluded.

Target Dynamic System. We are interested in smoothly satisfying the metric thresholds while ensuring long-term stability. To this goal, we adopt the dynamic system described in Equation (3), characterized by a smooth state evolution towards the threshold. Since we have two metrics of interest (DIDI and SAE), λ is size 2 vector, with values selected through a preliminary experiment described in Section 5.4.

Distance Function and Optimization Method. As a distance function, we use the Euclidian distance in Equation (4) that we optimize by relying on the Sequential Least Squares Programming (SLSQP) algorithm.

5.3 Evaluation

We compare FAiRDAS against a baseline approach in terms of metrics of interests (i.e., DIDI and SAE) and action smoothness ($\mathrm{m}_{Actions}$). The latter aims at evaluating the stability of selected actions over time and is computed as the mean absolute difference between subsequent action vectors:

$$\mathrm{m}_{Actions} = \frac{1}{N} \sum_{t=1}^{N-1} \frac{1}{|\mathcal{V}|} \sum_{j=1}^{|\mathcal{V}|} |\theta_{t,j} - \theta_{t+1,j}| \tag{9}$$

where N is the number of incoming batches and $\theta_{t,j}$ is the j-th component of the action vector chosen for the t-th batch. We report each metric's mean and standard deviation over batches to compare the approaches' performance and stability over time.

The baseline approach focuses on searching for the optimal action vector that minimizes the cost function:

$$\mathcal{L}(\theta) = \max\left(\mathrm{DIDI}(\theta), \mu_{\mathrm{DIDI}}\right) + \max\left(\mathrm{SAE}(\theta), \mu_{\mathrm{SAE}}\right) \tag{10}$$

where μ_{DIDI} and μ_{SAE} are the metrics' thresholds. The set of possible actions is the same as for FAiRDAS, and we rely on the same algorithm to tackle the optimization problem.

5.4 Numerical Results

As an initial step, we aim to analyze the impact of the eigenvalues λ of the FAiRDAS dynamical system on action smoothness. We run repeated experiments with different eigenvalues and fixed thresholds and report the action smoothness in Table 1. As expected from the theoretical properties of the target dynamic state, lower eigenvalues lead to more stable actions. This outcome showcases the remarkable adaptability of the FAiRDAS framework, allowing for indirect control of the stability level of mitigation actions. Such flexibility proves advantageous as it enables us to tailor the smoothness of our approach according to the requirements of different use cases. In this work, we select eigenvalues corresponding to the elbow of action smoothness metrics.

λ	$\mathrm{m}_{\mathrm{Actions}}$	$\sigma_{\mathrm{m}_{\mathrm{Actions}}}$
1.0	*0.276 ± 0.029*	*0.272 ± 0.016*
0.5	*0.102 ± 0.016*	*0.119 ± 0.021*
0.2	***0.044 ± 0.008***	***0.074 ± 0.019***
0.1	*0.024 ± 0.004*	*0.062 ± 0.018*
0.01	*0.008 ± 0.002*	*0.056 ± 0.019*

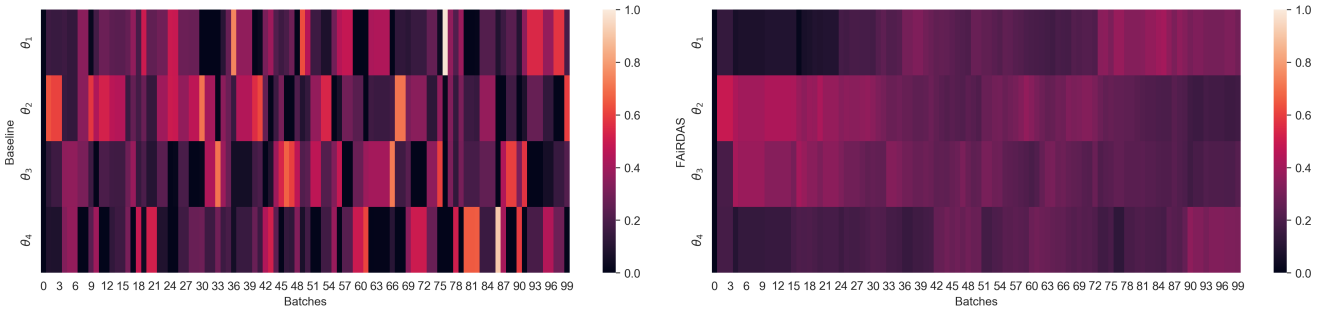
Table 1: Mean and standard deviation of the action smoothness computed over the batches. We analyse 5 eigenvalues (λ) with a fixed threshold and $\{0.5, 0.5\}$. For each eigenvalue, we run eight repeated experiments. We select $\lambda = 0.2$ as the elbow of the curve (in bold).

Next, we are interested in analysing how FAiRDAS and the baseline behave under different pairs of thresholds ($\{\mu_{\mathrm{DIDI}}, \mu_{\mathrm{SAE}}\}$). The first pair of thresholds $\{0, 2\}$ defines an extreme situation where we care only for fairness without regard for the ranking performance. Next, we select a loose pair of thresholds $\{0.7, 0.7\}$ and a strict pair of thresholds $\{0.5, 0.5\}$. Lastly, we investigate the not-reachable pair of thresholds $\{0.2, 0.2\}$. Table 2 reports the mean and standard deviation of the metrics over batches. FAiRDAS and the baseline achieve similar levels of the metrics of interest (DIDI and SAE) across all thresholds. However, the baseline method uses significantly more unstable actions compared to FAiRDAS, especially under strict thresholds. Regarding fairness and acceptable outcomes, the validation of results has involved ACCUEE stakeholders in interpreting the ethical acceptability or discriminatory nature of the mitigation actions. The finding emphasizes FAiRDAS's capability to maintain both effective performance and fairness over time, all while avoiding drastic actions that give rise to ethical concerns.

The enhanced stability of the FAiRDAS approach is illustrated in Figure 3, wherein we present the action vectors

Thresholds	*Approach*	DIDI	σ_{DIDI}	SAE	σ_{SAE}	$\text{m}_{\text{Actions}}$	$\sigma_{\text{m}_{\text{Actions}}}$
$\{0, 2\}$	Baseline	0.292 ± 0.059	0.396 ± 0.057	$\mathbf{0.663 \pm 0.061}$	0.442 ± 0.042	0.251 ± 0.033	0.199 ± 0.014
	FAiRDAS	$\mathbf{0.188 \pm 0.051}$	$\mathbf{0.327 \pm 0.046}$	0.692 ± 0.077	$\mathbf{0.247 \pm 0.03}$	$\mathbf{0.056 \pm 0.009}$	$\mathbf{0.052 \pm 0.012}$
$\{0.7, 0.7\}$	Baseline	0.390 ± 0.075	0.460 ± 0.067	$\mathbf{0.631 \pm 0.071}$	0.547 ± 0.095	0.233 ± 0.043	0.265 ± 0.027
	FAiRDAS	$\mathbf{0.235 \pm 0.062}$	$\mathbf{0.358 \pm 0.049}$	0.642 ± 0.072	$\mathbf{0.282 \pm 0.076}$	$\mathbf{0.023 \pm 0.005}$	$\mathbf{0.063 \pm 0.015}$
$\{0.5, 0.5\}$	Baseline	0.423 ± 0.056	0.507 ± 0.055	$\mathbf{0.643 \pm 0.081}$	0.594 ± 0.115	0.330 ± 0.024	0.297 ± 0.026
	FAiRDAS	$\mathbf{0.234 \pm 0.049}$	$\mathbf{0.376 \pm 0.078}$	0.654 ± 0.064	$\mathbf{0.309 \pm 0.048}$	$\mathbf{0.044 \pm 0.008}$	$\mathbf{0.074 \pm 0.019}$
$\{0.2, 0.2\}$	Baseline	0.551 ± 0.109	0.579 ± 0.071	$\mathbf{0.673 \pm 0.088}$	0.728 ± 0.123	0.554 ± 0.034	0.337 ± 0.030
	FAiRDAS	$\mathbf{0.290 \pm 0.048}$	$\mathbf{0.409 \pm 0.059}$	0.683 ± 0.055	$\mathbf{0.412 \pm 0.065}$	$\mathbf{0.091 \pm 0.008}$	$\mathbf{0.097 \pm 0.019}$

Table 2: Mean and standard deviation of the metrics computed over batches for *Students Dataset*. We run eight repeated experiments for each pair of thresholds and report the results for baseline and FAiRDAS approach. FAiRDAS and the baseline achieve similar levels of the metrics of interest (DIDI and SAE), but the baseline's actions are more unstable compared to FAiRDAS's ones.



(a) (b)

Figure 3: Action vectors chosen by the baseline (a) and FAiRDAS (b) in an experiment conducted with strict thresholds ($\{0.5, 0.5\}$). In each row, we present the evolution of the corresponding action vector component across 100 batches.

chosen by both approaches in an experiment conducted with strict thresholds. This figure offers a component-wise comparison of baseline and FAiRDAS action vectors across all 100 batches. As described in 5.2, the action vector components apply to the students of the corresponding protected groups as a penalizing factor on their score, thus potentially changing their position in the rank. Higher values correspond to more drastic penalization, while with values close to zero, the student's score is almost unmodified.

As we can notice, the baseline strategy tends to opt for rapid and drastic interventions, demonstrated by the rapid change in colour between batches and by action components close to 1 (lighter colour). In contrast, FAiRDAS demonstrates a more tempered and balanced behaviour, with action vectors evolving smoothly throughout the experiment (smooth change in colour along rows) and with similar penalization across the groups (homogeneous colour along columns).

This difference in behaviour is evident from the very first iteration: since we perform ex-post calibration to enhance action vector interpretability, both approaches start by taking no action in the first batch (i.e., $\theta_i = 0 \; \forall \; i$, corresponding to dark colour); then the baseline increases the action component θ_2 in response to the large presence of students with *ESCS*=2 in the first batch, thus suppressing their score; conversely, FAiRDAS opts for smoother mitigation action, slightly increasing the penalization for all the groups. FAiRDAS achieves higher levels of stability thanks to its underlying mechanism, which leverages a target dynamic system to guide decision-making to minimize abrupt changes. By approximating a smooth dynamic evolution, FAiRDAS effectively prevents the occurrence of drastic interventions, thereby promoting long-term stability and fairness within the system.

6 Conclusion

In this work, we presented a novel framework, FAiRDAS, to model the long-term evolution of fairness metrics as an abstract dynamical system. Our formulation allows control over (i) the trade-off between multiple metrics and (ii) the stability level of mitigation actions. The FAiRDAS approach emerges as advantageous, particularly in contexts where ensuring long-term fairness is essential. These contexts include scenarios that yield different outcomes over time, often with diverse inputs representing possibly a heterogeneous population with different characteristics. The primary advantage of adopting a dynamic system like FAiRDAS lies in its ability to ensure desirable properties, such as mitigation actions stability in achieving fairness constraints, as demonstrated in the AiEd scenario presented. Drastic actions performed by traditional approaches risk compromising the quality of individual rankings, rendering them ethically unacceptable. Future work will be devoted to applying FAiRDAS to other application scenarios and conducting more comprehensive tests to assess its effectiveness.

Ethics Statement

While the focus of this paper is specific to education, our general approach has broader applicability. By prioritizing long-term fairness and stability, we aim to contribute to the development of AI systems that are ethically robust and socially responsible, fostering trust and inclusivity in the broader community. The research conducted adheres to ethical standards throughout its entirety. All procedures and methodologies employed in this study have been designed and executed to comply with established ethical guidelines. In particular, the ACCUEE ensured the responsible and lawful acquisition of data and its elaboration in compliance with GDPR. All data were anonymized before being shared.

Acknowledgements

The work has been partially supported by the AEQUITAS project funded by the European Union's Horizon Europe Programme (Grant Agreement No. 101070363), and by PNRR - M4C2 - Investimento 1.3, Partenariato Esteso PE00000013 - "FAIR - Future Artificial Intelligence Research" - Spoke 8 "Pervasive AI", funded by the European Commission under the NextGeneration EU programme[4].

References

[Aghaei *et al.*, 2019] Sina Aghaei, Mohammad Javad Azizi, and Phebe Vayanos. Learning optimal and fair decision trees for non-discriminative decision-making. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 1418–1426, 2019.

[Albreiki *et al.*, 2021] Balqis Albreiki, Nazar Zaki, and Hany Alashwal. A systematic literature review of student' performance prediction using machine learning techniques. *Education Sciences*, 11(9), 2021.

[Arashpour *et al.*, 2023] Mehrdad Arashpour, Emad M Golafshani, Rajendran Parthiban, Julia Lamborn, Alireza Kashani, Heng Li, and Parisa Farzanehfar. Predicting individual learning performance using machine-learning hybridized with the teaching-learning-based optimization. *Computer Applications in Engineering Education*, 31(1):83–99, 2023.

[Blandin and Kash, 2023] Jack Blandin and Ian A Kash. Generalizing group fairness in machine learning via utilities. *Journal of Artificial Intelligence Research*, 78:747–780, 2023.

[Caton and Haas, 2020] Simon Caton and Christian Haas. Fairness in machine learning: A survey. *ACM Computing Surveys*, 2020.

[Celis *et al.*, 2020] L Elisa Celis, Vijay Keswani, and Nisheeth Vishnoi. Data preprocessing to mitigate bias: A maximum entropy based approach. In *International conference on machine learning*, pages 1349–1359. PMLR, 2020.

[4]Disclaimer: This paper reflects only the authors' views. The European Commission is not responsible for any use that may be made of the information it contains.

[Chassignol *et al.*, 2018] Maud Chassignol, Aleksandr Khoroshavin, Alexandra Klimova, and Anna Bilyatdinova. Artificial intelligence trends in education: a narrative overview. *Procedia Computer Science*, 136:16–24, 2018. 7th International Young Scientists Conference on Computational Science, YSC2018, 02-06 July2018, Heraklion, Greece.

[Ge *et al.*, 2021] Yingqiang Ge, Shuchang Liu, Ruoyuan Gao, Yikun Xian, Yunqi Li, Xiangyu Zhao, Changhua Pei, Fei Sun, Junfeng Ge, Wenwu Ou, et al. Towards long-term fairness in recommendation. In *Proceedings of the 14th ACM international conference on web search and data mining*, pages 445–453, 2021.

[Giuliani *et al.*, 2023] Luca Giuliani, Eleonora Misino, and Michele Lombardi. Generalized disparate impact for configurable fairness solutions in ml. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org, 2023.

[Holstein and Doroudi, 2019] Kenneth Holstein and Shayan Doroudi. Fairness and equity in learning analytics systems (fairlak). In *Companion proceedings of the ninth international learning analytics & knowledge conference (LAK 2019)*, pages 1–2, 2019.

[Hu and Rangwala, 2020] Qian Hu and Huzefa Rangwala. Towards fair educational data mining: A case study on detecting at-risk students. *International Educational Data Mining Society*, 2020.

[Hu and Zhang, 2022] Yaowei Hu and Lu Zhang. Achieving long-term fairness in sequential decision making. In *Thirty-Sixth AAAI Conference on Artificial Intelligence, AAAI 2022, Thirty-Fourth Conference on Innovative Applications of Artificial Intelligence, IAAI 2022, The Twelveth Symposium on Educational Advances in Artificial Intelligence, EAAI 2022 Virtual Event, February 22 - March 1, 2022*, pages 9549–9557. AAAI Press, 2022.

[Jiang and Pardos, 2021] Weijie Jiang and Zachary A Pardos. Towards equity and algorithmic fairness in student grade prediction. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, pages 608–617, 2021.

[Kizilcec and Lee, 2022] René F Kizilcec and Hansol Lee. Algorithmic fairness in education. In *The ethics of artificial intelligence in education*, pages 174–202. Routledge, 2022.

[Liu *et al.*, 2018] Lydia T. Liu, Sarah Dean, Esther Rolf, Max Simchowitz, and Moritz Hardt. Delayed impact of fair machine learning. In Jennifer G. Dy and Andreas Krause, editors, *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*, pages 3156–3164. PMLR, 2018.

[Misino *et al.*, 2023] Eleonora Misino, Roberta Calegari, Michele Lombardi, and Michela Milano. Fairdas: Fairness-aware ranking as dynamic abstract system. In

Roberta Calegari, Andrea Aler Tubella, Gabriel González-Castañé, Virginia Dignum, and Michela Milano, editors, *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.

[Pedro *et al.*, 2019] Francesc Pedro, Miguel Subosa, Axel Rivas, and Paula Valverde. Artificial intelligence in education: Challenges and opportunities for sustainable development. 2019.

[Sha *et al.*, 2023] Lele Sha, Dragan Gašević, and Guanliang Chen. Lessons from debiasing data for fair and accurate predictive modeling in education. *Expert Systems with Applications*, 228:120323, 2023.

[Xian *et al.*, 2023] Ruicheng Xian, Lang Yin, and Han Zhao. Fair and optimal classification via post-processing. In *International Conference on Machine Learning*, pages 37977–38012. PMLR, 2023.

[Yin *et al.*, 2024] Tongxin Yin, Reilly Raab, Mingyan Liu, and Yang Liu. Long-term fairness with unknown dynamics. *Advances in Neural Information Processing Systems*, 36, 2024.

[Yu *et al.*, 2022] Eric Yang Yu, Zhizhen Qin, Min Kyung Lee, and Sicun Gao. Policy optimization with advantage regularization for long-term fairness in decision systems. *arXiv preprint arXiv:2210.12546*, 2022.

[Zafar *et al.*, 2017] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, and Krishna P Gummadi. Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In *Proceedings of the 26th international conference on world wide web*, pages 1171–1180, 2017.