ECAI 2025
I. Lynce et al. (Eds.)

doi:10.3233/FAIA250913

AI Fairness Compliance: Operationalizing the Integration of Social and Legal Perspectives into AI Fairness Metrics

Roberta Calegari[a,*]

[a]Department of Computer Science and Engineering, Alma Mater Studiorum—University of Bologna
ORCID (Roberta Calegari): https://orcid.org/0000-0003-3794-2942

Abstract. In recent years, addressing bias, discrimination, and fairness in AI has garnered significant attention. However, integrating these discussions with established legal frameworks, particularly within European Union legislation, remains a critical and underexplored area. This article addresses this gap by proposing legally valid fairness assessment metrics that capture the socio-legal context specific to each case. Our approach seeks to integrate legal, social, and technical perspectives in evaluating AI fairness. While many AI fairness toolkits provide statistical measures of fairness, they often fall short of aligning with the context-sensitive discrimination metrics and evidential requirements outlined by the European Court of Justice. To bridge this gap, we leverage the concept of *contextual equality*. The concept of contextual equality must be reified within the technology through two main steps: (1) providing a formal definition of contextual equality for AI decision-making problems, whether classification or regression, and (2) operationalizing this definition within state-of-the-art metrics for 'measuring' fairness compliance. The paper provides these two critical contributions. An experimental evaluation is carried out on two benchmark datasets in the field of AI and education, validating the approach with AI stakeholders involved in the field and in the decision-making process. This also serves to highlight the negative impacts that an assessment not legally sound could have.

1 Introduction

Concerns over bias in Artificial Intelligence (AI) and Machine Learning (ML) have intensified as these systems are increasingly used in high-stakes domains such as healthcare, finance, recruitment, and education. Ensuring compliance with trust and fairness requirements is now critical, especially with the EU AI Act [7] introducing a binding regulatory framework for high-risk AI applications. Among the seven required elements of trust compliance, fairness is recognized as one of the four 'Ethical Principles in the Context of AI Systems,' as outlined in the Ethics Guidelines for Trustworthy AI [9]. This principle is intrinsically connected to three other principles—respect for human autonomy, prevention of harm, and explicability. It addresses various ways in which bias can manifest undesirable impacts, including discrimination, inequality, exclusion, segregation, marginalization, exploitation, and manipulation. The urgent need for *fairness assessment and testing* of AI systems arises to ensure these systems meet the necessary compliance requirements, as deployment in high-stakes applications is untenable if the systems exhibit discriminatory behaviours based on sensitive attributes like gender, race, or age.

Numerous fairness metrics have been proposed by researchers to test AI algorithms for bias [13]. The volume of these metrics is on the rise, and they represent a concrete method for testing compliance with the fairness requirements of AI systems. However, the extent to which these metrics meet the standards set by EU law remains an unresolved issue [11, 14]. This uncertainty significantly hampers the value of assessments conducted using these metrics[20]. There is a recognized gap between the statistical fairness metrics included in existing fairness toolkits and governance mechanisms (context-sensitive and intuitive metrics) used by the law and the EU Court of Justice [11, 20, 8]. This discrepancy primarily arises from the fact that the assessment of potential discrimination should in principle be conducted on a case-by-case basis, analyzing all contextual information of the specific case. This critical concept is known as *contextual equality*. Therefore, while numerous statistical metrics exist in the technical literature, none can reliably capture the EU conceptualization of discrimination in law, which is inherently contextual [5, 20]. Existing scalable automated methods to detect and combat discriminatory decision-making seemingly necessitate clear-cut rules or quantifiable thresholds, which EU non-discrimination law and jurisprudence deliberately do not provide.

To address the limitations of existing statistical metrics, it is essential to augment them with the notion of context. This enhancement enables the assessment of discrimination by accounting for the socio-legal context as mandated by law, ensuring a responsible evaluation of behaviours as fair or unfair. Contextual equality has been previously discussed in the literature [20]. However, their definition remains a unique instance in the literature and has not been concretely implemented in existing technological metrics that the market offers for fairness assessments [2, 3, 15] [1]. Moreover, their application is confined to classification tasks, limiting its broader applicability. Recognizing the critical need to measure levels of fairness and unfairness in AI systems within social and legal contexts, this study defines *contextual equality in AI systems*. This concept safeguards against unfairness by ensuring that legally and socially relevant case-specific factors – validated through legal norms and stake-

* Corresponding Author. Email: roberta.calegari@unibo.

[1] To the best of our knowledge, while some toolkits mention it for classification, a closer inspection reveals that they reduce it to mere context-less metrics.

Please contact the corresponding author for any appendices or supplementary material mentioned in the paper.

holder input – are included in fairness assessments. For instance, in the context of educational admissions, EU non-discrimination law – as interpreted by the EU Court of Justice – acknowledges the importance of indirect discrimination, where a seemingly neutral criterion disproportionately disadvantages a protected group unless objectively justified and proportionate (e.g., [4]). A predictive model that flags students as at-risk based solely on academic scores may appear neutral but could disproportionately affect students from socio-economically disadvantaged backgrounds. Variables such as school funding level, regional access to preparatory courses, or parental education level may thus be legally relevant contextual factors. Their inclusion allows fairness assessments to detect disparities rooted in structural inequalities rather than in individual merit, aligning the model's logic with principles of substantive equality under EU law. This contextualization is not arbitrary: it is governed by a structured socio-legal framework rooted in EU directives (e.g., [17]) and interpreted through CJEU jurisprudence, ensuring that only legally justifiable differences are considered in fairness assessments.

Our definition of contextual equality enables a legally grounded assessment of an AI system's compliance with fairness obligations. Building on the definition of [20], we expand its applicability to both classification and regression AI tasks. We also provide a methodology for integrating the notion of contextual equality into existing state-of-the-art metrics to render them legally sound. The methodology is illustrated using one of the most well-known fairness assessment metrics, Demographic Disparity (DD) [14], chosen both for its relevance to the case study and its widespread use. However, our approach is general enough to be applied to any other existing fairness metric in the literature.

The validation and formalization of these concepts have been validated through a real-world case study. The selected case study involves ranking students based on predicted academic performance to identify potential dropouts or to make recommendations. Real-world data[2] along with guidance for determining objectives and assessing quality and fairness metrics, were provided by the Canarian Agency for Quality Assessment and Accreditation (ACCUEE). ACCUEE was involved as a stakeholder, along with socio-economic experts, to formalize the problem and validate the assessment. The case study highlights the significant dangers of using non-legally compliant AI assessments in education, where biases in the algorithms can lead to students being mistakenly identified as underperforming or at-risk. Such misclassifications can result in educational disadvantages, stigmatization, and resource misallocation, negatively impacting students' academic and psychological well-being. Moreover, these flawed assessments can have long-term effects on students' educational trajectories and future opportunities, while also exposing educational institutions to legal and ethical challenges.

2 Related work

This work moves within the context of AI fairness assessment and measurement, which involves determining whether an AI system is perpetuating or amplifying existing societal discriminations in automated decision-making systems. Usually, the level of fairness/unfairness of AI systems is assessed via AI fairness metrics. These metrics are designed to evaluate whether AI systems exhibit bias or discrimination against certain individuals or groups based on protected (aka sensitive) attributes such as race, gender, or age. AI fairness metrics provide a standardized way to quantify disparities. Many current works suggest quantitative statistical metrics to measure the level of unfairness or bias in an AI algorithm based on different notions of fairness or different types of sensitive attributes [13, 14], but few focus on their compliance in terms of law [20]. A recent study highlighting the limitations of existing metrics and how they don't work well in many real-world ML applications is [11]. The study further highlights the limited attention given to legal compliance in current fairness evaluation frameworks. A key work addressing the gap between technical and legal perspectives is [20]. The authors emphasize how the socio-legal context conditions the evaluation of potential discrimination. Accordingly, existing metrics need to be revisited by introducing the notion of context, allowing for conditioning in the fairness assessment. Watcher's theoretical work defines DD in terms of contextual equality, adding the possibility of conditioning the measurement of DD based on the context. More recent works have continued this normative turn. Roy et al. [16] propose *Socio-Economic Parity*, a fairness constraint directly derived from the EU AI Act's obligations, particularly emphasizing socio-economic status as a legally protected dimension in automated decision-making. Their work underscores the need for compliance-ready fairness metrics. Meding [12] complements this with a detailed legal-theoretical analysis of how algorithmic fairness must be framed under the EU AI Act and related anti-discrimination directives. Both works strengthen the argument that legally informed metrics must go beyond statistical parity and must be adaptable to context-specific factors. In addition to these largely conceptual contributions, our work provides a practically implementable framework—empirically validated on real-world datasets—and a generalizable methodology for extending existing statistical fairness metrics to incorporate legal and contextual dimensions. In our study, we adopt the same theoretical definition of [20] but we generalize it. Similar to their approach, we provide fairness metrics for measuring contextual equality, which is essential for AI compliance. However, we expand it for use in both classification and regression tasks, and to measure inter-group and intra-group comparisons. We then transform this definition into quantitative metrics, implementing contextual equality as a statistical fairness metric within an open-source fairness toolkit. Indeed, in some scenarios, having a quantitative metric can assist in making certain automated, albeit supervised, reasoning processes, thereby alleviating the burden of analyzing a large volume of data on the human expert.

3 Move AI fairness metrics to Contextual Equality Notion

In this section, we introduce a methodology to legally validate existing fairness metrics found in the literature by incorporating socio-legal context into their definitions. This approach ensures that the assessments consider this context. To provide specific details of the methodology, it will be applied to a particular metric, Demographic Disparity (DD) [14], without loss of generality.

3.1 DD and CDD for Classification

Let (X, A, Y) represent the input features, the protected attributes, and the output of interest respectively, with $A \in X$ and $f(X) = Y$. The DD metric assesses whether a group has a higher proportion of rejected outcomes compared to accepted outcomes [10]. Formally:

$$DD_a = N_a - P_a \tag{1}$$

where $a \in A$ is the value of the protected attribute (for instance, female or male when the protected attribute is gender) and P and N

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

are defined respectively as:

$$P_a = \frac{|\{i \in I | A = a, y = 1\}|}{|\{i \in I | y = 1\}|}$$

$$N_a = \frac{|\{i \in I | A = a, y = 0\}|}{|\{i \in I | y = 0\}|}$$

where $i \in I$ refers to a specific example (i.e. a row or an individual) within the dataset.

CDD extends the DD notion by accounting for *conditioning*, i.e., adding the possibility to specify the contextual variable to be considered when examining demographic disparity [20]. Formally, a classifier satisfies CDD with respect to feature R if:

$$N_{R_a} > P_{R_a} \tag{2}$$

where R is a feature in the input space and P_{R_a} and N_{R_a} are:

$$P_{R_a} = \frac{|\{i \in I | A = a, R = r, y = 1\}|}{|\{i \in I | R = r, y = 1\}|} \tag{3}$$

$$N_{Ra} = \frac{|\{i \in I | A = a, R = r, y = 0\}|}{|\{i \in I | R = r, y = 0\}|} \tag{4}$$

The extension from DD to CDD involves the addition of a contextual variable (condition) R in measuring potential unfairness. Thus, it no longer analyzes unfairness in general but grounds the analysis in a specific context that needs to be considered. CDD must, at least theoretically, be evaluated for all possible combinations of protected features (for instance male/female) and contextual variables (R in the formula), considering a system fair if it ensures CDD is met for all the measurements. However, this "plain" strategy, as pointed out by [20], is prone to finding false positives. In other words, searching over all combinations of attributes for a violation in practice would almost certainly end up labelling a classifier as unfair simply by chance. To account for this, an aggregation statistic should be defined. Various approaches have been considered in the literature [20]. For simplicity, and because this choice does not impact the goal of this work, we adopt a weighted average (based on the number of individuals) of the CDD for each value of R, as proposed by [18]. The smaller the difference $N_{Ra} = P_{Ra}$, the fairer the treatment of the AI system; conversely, the larger the difference, the more the AI system is behaving unfairly.

3.2 DD and CDD as quantitative metrics

The definitions above can benefit from being transformed into a quantitative metric. This can be useful for: *i)* providing a numerical quantification of the magnitude of disparity, offering a concise metric for judicial assessment; *ii)* transitioning to a technical perspective, where the metric needs to be implemented and integrated into a tool; *iii)* enabling AutoML techniques (designed to automatically compare hundreds of predictors to select the optimal one) to function effectively [21]. Accordingly, we define CDD as

$$\overline{CDD}_{R,a} = \overline{N}_{R,a} - \overline{P}_{R,a} \tag{5}$$

with $\overline{N}_{R,a} = \mu_w(N_{r,a}) \quad \forall r \in R$ and $\overline{P}_{R,a} = \mu_w(P_{r,a}) \quad \forall r \in R$ where μ_w represents the weighted average. This difference concisely represents the existing gap between $\overline{N}_{R,a}$ and $\overline{P}_{R,a}$, highlighting a disparity in the case of a difference greater than zero, and is completely aligned with the DD metric (Equation 1). While this disparity already provides an initial framework that allows the human evaluator to identify potential discrimination, we believe that the computational metric could benefit from another analysis: the dynamics between different groups (inter-group or between-group dynamics). For this purpose, our analysis proposes to explicitly introduce an additional difference that needs to be considered, namely, the difference between the CDD of different groups. This difference should be calculated for every possible pair of protected groups. Formally,

$$\widetilde{CDD}_{R,a_i,a_j} = \overline{CDD}_{R,a_i} - \overline{CDD}_{R,a_j} \tag{6}$$

with $a_i, a_j \in A$ and $a_i \neq a_j$. The same for DD, with no conditioning:

$$\widetilde{DD}_{a_i,a_j} = DD_{a_i} - DD_{a_j} \tag{7}$$

with $a_i, a_j \in A$ and $a_i \neq a_j$. The smaller the difference, the more similarly the groups are treated, while larger differences indicate dissimilar treatments.

3.3 DD and CDD for Regression

As for regression, the task can be reduced to a classification task through thresholding, as inspired by [1]. In their work, the authors address the problem of fair regression, which involves predicting a real-valued target while ensuring compliance with fairness constraints related to a protected attribute. Specifically, they adapt regression tasks to be *measurable* using existing fairness metrics that are typically designed for classification tasks. To achieve this goal, the authors first discretize the loss function by discretizing its parameters, including the predicted real-valued score and the ground truth. Then, by using the *approximated* version of the loss, they can define fairness metrics typically used for classification, making them suitable for addressing regression.

Following their core idea, to extend the CDD metric to regression tasks, we discretize the regressor's predicted values using the concept of a *discretization grid*. The discretisation grid is essentially a set of multiples of $\alpha = 1/N$, where N represents the size of the grid. Therefore, the grid can be defined as the set $Z = \{j\alpha : 1, \ldots, N\}$, which contains N integer multiples of α. The basic idea is that the regressor can be *approximated* by N classifiers, each corresponding to a different threshold value from the discretization grid. Specifically, let $h_z(X)$ be a classifier assigning label 1 to a sample with features X if $f(X) \geq z$, 0 otherwise. The extension of the CDD assessment tests to regression tasks involves evaluating $\overline{CDD}$ and $\widetilde{CDD}$ for all values of $z \in Z$, being Z the discretisation grid. This translates to evaluating 5 and 6 for all the classifiers sampled through thresholding with the values in Z. Formally, this would produce the set:

$$\{\overline{CDD}_{R,a,h_z(X)} | R \in \{X \setminus A\}\} \text{ with } a \in A, z \in Z$$

This is the extension for Equation 5. The formulation of this extension for 6 is straightforward.

To sum up, in regression tasks, the CDD evaluation must extend to each classifier used to discretize the regressor. Although increasing N enhances approximation, selecting the highest possible N is impractical. Instead, aim for a sufficiently large N that provides an adequate approximation. Examining the plots containing $\overline{CDD}, \widetilde{CDD} \; \forall z \in \mathrm{Z}$ and analyzing the effects of each individual value on the regressor is a highly effective tool for detailed assessment of the resulting classifer's behaviour. In fact, one could get an idea of what the proportion of individuals receiving an advantageous outcome is for both the protected and non protected groups. If combined with a utility function whose output depends on the model's classification, this approach has the potential to be highly informative as an assessment tool.

However, also having a concise metric to represent the overall magnitude of fairness or unfairness can be very useful. To address this, we propose a metric that aggregates all the DD and $\overline{CDD}$ values obtained across Z by calculating the fraction of fair classifiers. For a fixed z, a classifier is considered "fair" with respect to CDD if, for a given tolerance ϵ, the $\overline{CDD}$ values for all groups defined by the sensitive attributes are below this threshold. By iterating over the discretization grid, we can assess whether each classifier, obtained by thresholding the regressor's predictions, is fair. Finally, the number of fair classifiers is divided by the total number of classifiers (which equals the size of the discretization grid), yielding the value for the aggregation metric. This process is repeated for all contextual variables under consideration. Formally, for each $z \in Z$, the classifier $h_z(X)$ is considered *fair* if:

$$|\overline{CDD}_{R,a,h_z(X)}| < \epsilon \quad \forall a \in A$$

Let $\Psi_{\overline{CDD}}(z)$ be an indicator function such that:

$$\Psi_{\overline{CDD}}(z) = \begin{cases} 1 & \text{if } |\overline{CDD}_{R,a,h_z(X)}| < \epsilon \\ 0 & \text{otherwise.} \end{cases}$$

The aggregation metric $FairRate_{CDD}$ computes the fraction of *fair* classifiers across all $z \in Z$:

$$FairRate_{CDD} = \frac{1}{N} \sum_{j=1}^{N} \Psi_{\overline{CDD}}(z_j),$$

where $z_j = j\alpha$ for $j = 1, 2, \ldots, N$. As per $FairRate_{DD}$, definitions are analogous but are omitted here for the sake of space. The value of ϵ is not important within the context of this work because, as already stated, the main focus of CDD is to be used as an assessment tool providing evidence of possible discrimination. The fairness evaluation is therefore left to the judicial experts that can have a deeper understanding of the use case. Using $FairRate_{CDD}$ and $FairRate_{DD}$ is therefore a more compact way of proving that the fairness of a model largely depends on the conditioning with respect to certain contextual variables.

3.4 Generalization for Other AI Fairness Metrics

The methodology outlined above is generalizable and can be systematically applied to a wide range of fairness metrics. The transformation follows the structured approach previously defined and is summarized below:

- **Start from an existing AI fairness metric:** Select a standard fairness metric – such as demographic parity, equalized odds, or predictive parity – used to measure disparities across protected groups.
- **Identify legally and socially relevant contextual variables:** Choose case-specific features (e.g., department, region, socio-economic status) that are not protected attributes but are recognized by law or domain expertise as relevant to assessing fairness.
- **Condition the metric on context:** Redefine the fairness metric within each subgroup defined by the contextual variable, measuring disparities not in aggregate, but within each relevant social or institutional context.
- **Aggregate across contexts:** Use a weighted average (or alternative statistic) to summarize fairness results across all contextual subgroups, ensuring that population size or significance is taken into account.
- **Compare across protected groups:** Assess fairness consistency between protected groups by comparing their contextualized fairness scores, identifying unjustified disparities in treatment.
- **Extend to regression (if applicable):** In cases where the model outputs continuous predictions, discretize the predicted values using a predefined threshold grid—classifying outcomes below each threshold as negative and those above as positive. For each threshold, a binary classifier is derived and evaluated using the same contextual fairness methodology applied in classification tasks.
- **Summarize with an aggregate fairness score (optional):** Compute a compact summary metric—such as the proportion of fair threshold-based classifiers—to communicate overall compliance in a legally interpretable format.

4 Use Cases and Experimental Setup

Our experimental evaluation uses two benchmarks in AI and education to contextualize and motivate our approach, illustrating the differences between DD and CDD. The case study involves predicting students' academic performance based on data collected from their academic history and socio-economic background. Predicting students' performance can facilitate the early identification of potential issues through AI-based educational solutions, allowing for timely and appropriate interventions. While accurate predictions are crucial in this context, it is also recognized in the literature that avoiding disparities, regardless of students' social backgrounds and other sensitive attributes such as gender, is equally important [19]. Socio-economic status and gender often correlate with access to resources, opportunities, and support systems. Therefore, ensuring that these algorithms are fair is of fundamental importance. However, how fairness is assessed is also crucial to avoid misjudgment and to enable the implementation of appropriate countermeasures. The goal of our experiments is to compare the behaviour of CDD and DD by evaluating the presence (or absence) of bias in a linear regressor fitted to predict each student's score in Mathematics. Our experiments demonstrate how results can vary, sometimes significantly, when the social context is not considered. Therefore, to produce a fairness assessment that meets legal requirements, it is essential to use CDD. We have analyzed two different datasets, both collecting data in real domains, namely the Student Performance Dataset and the Canary Islands Dataset[3].

4.1 Student Performance Dataset (SPD)

Dataset description. This dataset[4] [6] examines student achievement in secondary education across two Portuguese schools. For each student, it includes school grades, demographic information, as well as social and school-related data. It was collected through school reports and questionnaires. The dataset is split into two parts, each corresponding to performance in a different subject: Mathematics (mat) and Portuguese language (por). In our experiments, we consider only the partition related to performances in Portuguese. It contains 649 instances with 30 features. The target columns are *G1*, *G2*, and *G3*, which correspond to the Portuguese grade achieved at the end of the first, second, and third periods of the academic year, respectively. When testing CDD, all features except for the sensitive attribute (*sex*) and the target score (*G3*) were used as control variables. Relevant features include: *studytime*, *medu* and *fedu*. The first

[3] All the code is available at https://github.com/aequitas-aod/experiment-cdd-metric

[4] https://archive.ics.uci.edu/dataset/320/student+performance

indicates the number of hours a student spends studying per week with values from 1 (less than two hours) to 4 (more than ten hours). The last two indicate a parent's education level, ranging from 0 (no education) to 4 (highest possible).

Dataset Preprocessing. We decided to keep *G3* as the target and simply drop *G1* and *G2*. We encoded the categorical features with increasing integers after sorting their values. As a final pre-processing step, the values of *G3*, the only continuous feature, were mapped to the $[0, 1]$ range.

4.2 Canary Islands Dataset (CID)

Dataset description. This dataset [5], provided by the Canarian Agency for Quality Assessment and Accreditation (ACCUEE), spans four academic years (2015-2019) and includes data on students in the Canary Islands from the third and sixth grades of primary school and the fourth year of secondary school. The original dataset contains over 80,000 rows and more than 500 features. For this study, the dataset was reduced to 17. The first six are academic performance features: continuous scores achieved in subjects such as Maths, Spanish, and English, and a student's corresponding proficiency levels. We selected the Math score (*score_MAT*) as target variable due to its minimal missing values and used the Economic, Social, and Cultural Status index (*f_ESCS*) as protected variable. This feature indicates a student's socio-economic status. The final 10 features were chosen based on the input from domain experts in order to pick those that, at least in theory, impact a student's marks the most. These features are used as context in the CDD computation when predicting a students' Maths score. The most relevant for this work are *f_mother_education_level* and *f_father_education_level*. They are categorical features with values from 1 to 4, indicating a parent's education level from no/low education to the highest degree possible.

Dataset Preprocessing. When preprocessing the data, we discovered that many students had duplicate records, presumably to track their performance over time. To ensure each student had only one record, duplicate rows were removed based on the student identifier column (i.e *id_student_original*). After that, the first seven columns, each containing a particular kind of identifier, were dropped. Additionally, rows with missing student identifiers were excluded from the dataset used in the experiments as well as all rows containing at least one missing value in the remaining columns. The remaining continuous and categorical features in the resulting dataset were treated exactly as described in the section relative to the preprocessing of the Student Performance Dataset. The continuous features were mapped to the [0, 1] range while each value of the categorical features was given a progressive integer number according to the alphabetical order. For example, considering the *s_gender* attribute, representing a student's gender, its possible values, `FEMALE` and `MALE`, are mapped to 0 and 1 respectively when performing the binarization.

4.3 Experimental Setup

After the cleaning steps, the data was shuffled, split into train and test set (0.7:0.3 ratio) and fed to a linear regressor with four fully connected layers interleaved with the ReLU activation function. The number of hidden units starts at 256 and grows to 512 in the first intermediate layers before going back to 256 and then to 1 in the final two layers. This model was trained for 30 epochs using MSE as the loss function and early stopping with patience set to 5 epochs. The optimizer of choice was Adam with a constant learning rate equal to $1e{-}3$. The model's predictions collected at inference time on the test set are the starting point for the comparisons between CDD and DD. In the experiments, the size of the discretisation grid N was set to 400. This means that the thresholding on the labels and the computation of $\overline{CDD}$, $\widetilde{CDD}$, $\widetilde{DD}$ and DD was repeated 400 times for each control variable. Our code binarizes any continuous protected attribute before computing the DD and CDD metrics. Binarization, or discretization, is required as these metrics count individuals in different groups based on outcomes received. Continuous feature values above average are mapped to 1, others to 0.

[5] https://zenodo.org/records/11171863

5 Results

Due to space constraints, only key results and plots are included here; additional experiments are in the supplementary.

5.1 On the difference between DD and CDD

The computation of CDD is performed in three steps. First, for each value of the condition, counts are gathered for both positive and negative outcomes for all groups of the protected attribute. In our experiments on CID, the protected attribute is *f_ESCS* which, after the binarisation, identifies two groups: the advantaged group ($f\_ESCS = 1$) and the disadvantaged group ($f\_ESCS = 0$). So, for example, if the additional feature R is binary, four separate values are produced for each value of R: the number of positive outcomes in the group with $f\_ESCS = 1$, the number of positive outcomes in the group with $f\_ESCS = 0$, and the analogous quantities for negative outcomes. Using the preprocessed CID dataset with $z = 0.45$, where *f_ESCS* serves as A and *f_mother_education_level (mel)* as R, this first step generates a table like Table 1.

Table 1: Step 1 CDD_{mel}: counting of positive and negative outcomes per protected group conditioned on *f_mother_education_level (mel)*.

mel	$A=1$ $\hat{y}=1$	$A=0$ $\hat{y}=1$	**Total** $\hat{y}=1$	$A=1$ $\hat{y}=0$	$A=0$ $\hat{y}=0$	**Total** $\hat{y}=0$
1	3	434	**437**	0	491	**491**
2	1	843	**844**	1	732	**733**
3	1	1106	**1107**	2	565	**567**
4	1546	2210	**3756**	1134	514	**1648**

Secondly, based on the values computed in step 1, the corresponding statistics are extracted. Each count associated with a given value of R and relative to a certain class is divided by the total number of individuals showcasing the same value of R independent of A. For instance, the number of negative outcomes with a specific value of *mel* and $f\_ESCS = 0$ is divided by the total number of negative outcomes considering individuals with the same value of mel. This process is repeated for each value of attribute R. Essentially, starting with four separate counts from the previous step, this stage yields four corresponding statistics expressed as percentages (Table 2).

Lastly, these statistics, each pertaining to a particular value of *mel*, are weighted by the total number of instances with that value of the conditioning feature, and a weighted average across all its possible values is computed (see Table 3). At this point Equations 5 and 6 can be easily evaluated resulting in:

$$\overline{CDD}_{mel,A=1} = 15,6\% \quad \overline{CDD}_{mel,A=0} = -15,6\%$$

Table 2: Step 2 CDD_{mel} computation: calculation of N_{mel} and P_{mel} for each value of the attribute mel (R).

mel	$S=1$ $\hat{y}=1$	$S=0$ $\hat{y}=1$	$S=1$ $\hat{y}=0$	$S=0$ $\hat{y}=0$
1	0.7	99.3	0.0	100.0
2	0.1	99.9	0.1	99.9
3	0.1	99.9	0.4	99.6
4	41.2	58.8	68.8	31.2

Table 3: Step 3 CDD_R computation: weighted average calculation on the values of R.

$S=1$ $\hat{y}=1$	$S=0$ $\hat{y}=1$	$S=1$ $\hat{y}=0$	$S=0$ $\hat{y}=0$
23.3	76.7	38.9	61.1

where A is used in place of *f_ESCS*. This means that in the case of $f\_ESCS = 1$ the negative outcomes overcome the positive ones by that percentage. In the case of $f\_ESCS = 0$, the situation is instead dual. There are more positive outcomes. By computing the between-group discrepancy we can quickly see that a disparity in treatment between the two socio-economic groups is evident. In fact: $\widetilde{CDD}_{mel,A=1,A=0} = 31,2$. DD is computed following the same steps without the last one comprising the weighted average to aggregate across all values of the additional feature and without any previous conditioning. In this example DD amounts to:

$$DD_{mel,A=1} = 4,2 \quad DD_{mel,A=0} = -4,2$$

With no conditioning the disparity in treatment between the two groups is much lower, $\widetilde{DD}_{mel,A=1,A=0} = 8,4$.

5.2 Main Results: Summary of Findings

In this section, we summarize our experimental findings, emphasizing the need for context-dependent metrics like CDD. Results ignoring context can differ significantly, potentially leading to incorrect fairness conclusions. For each result presented, two types of plots are generated: one[6] compares CDD and DD across Z, potentially highlighting discrepancies. This kind of plot is useful to analyse within group discrepancies in how individuals are treated according to the considered metric (either $\overline{CDD}$ or DD), as the value of z changes. The other type of plot compares the two metrics in terms of $\widetilde{CDD}$ and $\widetilde{DD}$. In this case the plots are useful to compare between group differences in treatment. These plots are an attempt at proving that, through CDD, the presence of *prima facie discrimination* can be determined quite intuitively, which is one of the important requirements for fairness metrics [20].

Table 4: CDD-fair and DD-fair classifiers for SPD

	$FairRate_{DD}$	$FairRate_{CDD}$
studytime	0.150160	0.511182
medu	0.150160	0
fedu	0.150160	0

Table 5: CDD-fair and DD-fair classifiers for CID

	$FairRate_{DD}$	$FairRate_{CDD}$
f_me_level	0.843621	0.905350
f_fe_level	0.843621	0.925926

[6] Proof of symmetry wrt y-axis is in the supplementary material.

SPD results: studytime conditioning. Figure 1 shows that conditioning on *studytime* significantly reduces disparities. The top plot illustrates within-group disparities, where individuals with the same value for the protected attribute (*sex*) but different outcomes are compared. For males, this is evident as the red line is generally closer to 0 compared to the green line, and the same trend is observed for females, as shown by the blue and yellow lines. The bottom plot demonstrates between-group disparities, comparing individuals with different values of the protected attribute. Here, the red line, which accounts for the additional context, is almost always closer to 0 compared to the blue line. This suggests that when the amount of weekly study is considered, gender does not significantly impact academic performance; without this conditioning, bias would appear. Table 4 (first row) supports these points. The number of classifiers deemed "fair" significantly increases when using CDD as the evaluation metric instead of DD. We used $\epsilon = 20$ in our experiments; even with smaller ϵ values, using CDD instead of DD remains essential for more accurate evaluations of a model's behaviour.

SPD, CID results: parents' education conditioning. The remaining experiments compare the effects of a student's parents' education level on their Math performance across the SPD and CID datasets. In CID, the contextual variables are *f_mother_education_level* and *f_father_education_level*, while in SPD, they are *medu* and *fedu*. Since the mother's and father's education levels have a similar impact across both datasets, we focus the discussion solely on the mother's education. For SPD, Figure 3 shows that using CDD reveals a worse situation than DD. Both plots indicate that additional conditioning increases both within-group (top plot) and between-group (bottom plot) disparities. This is also reflected in Table 4, where the second row shows fewer CDD-fair classifiers (none) compared to DD-fair classifiers. Relying solely on DD would underestimate the discrimination displayed by the model across Z. For CID, conditioning on *f_mother_education_level* shows a different picture. This context reduces both between-group and within-group disparities compared to DD. Figure 4 visually demonstrates this, and Table 5 provides additional evidence. The increase in CDD-fair classifiers indicates that, with this context, the classifiers display less discriminatory behaviour across Z. Experiments show that relying solely on demographic disparity can lead to misleading interpretations of fairness, as conditioning on control variables can yield significantly different results.

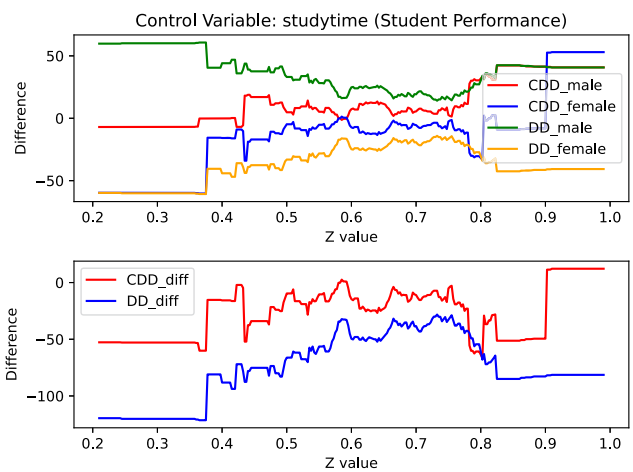

Figure 1: Effect of *studytime* (SPD) as contextual variable

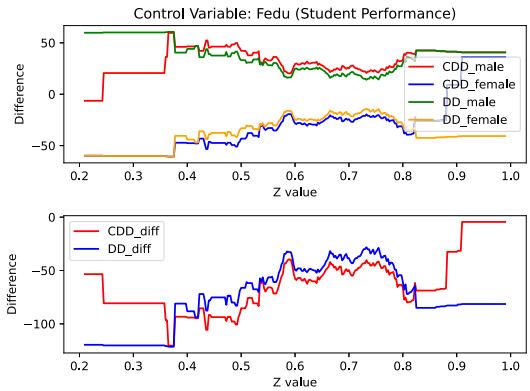

Figure 2: Effect of father education SPD

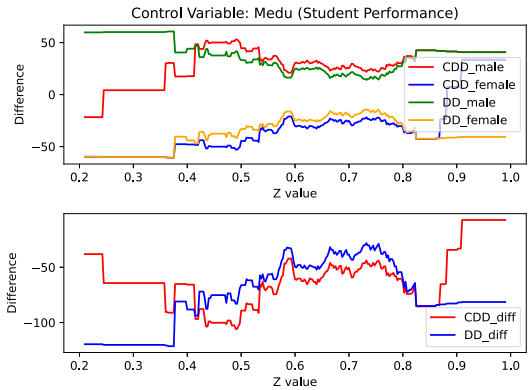

Figure 3: Effect of mother education SPD

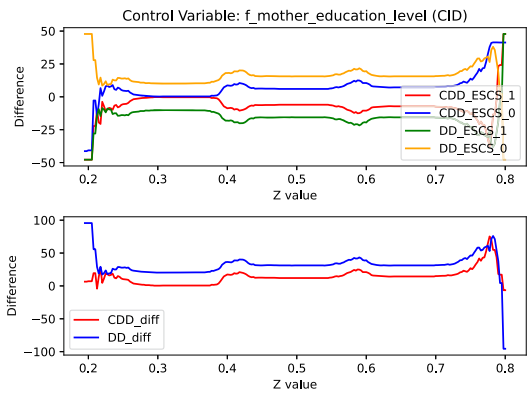

Figure 4: Effect of mother education CID

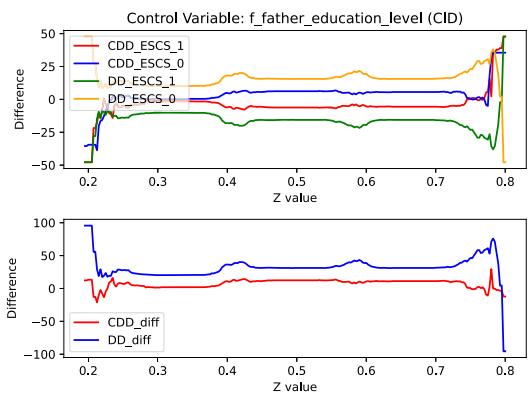

Figure 5: Effect of father education CID

6 Conclusion

In this work, we utilize contextual equality to enhance AI compliance in fairness metrics, providing quantitative measures for inter-group and intra-group discrimination. We then integrate these metrics into an open-source fairness toolkit as statistical measures. Our experimental evaluation on real AI and education benchmark datasets, along with a co-creation process with domain stakeholders, demonstrates our approach's effectiveness and broad applicability. It highlights the social impact of including context in fairness assessments, showing how neglect can yield misleading, non-EU-compliant results detrimental to social equity. The methodology for adding socio-legal context to fairness metrics is applicable to others in the literature. Future work will extend to include multiple conditioning features and protected attributes, embracing intersectionality.

Ethical Statement

The research conducted herein 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. It is important to note that we exclusively utilised publicly available datasets, already recognised as adhering to legal and ethical standards, ensuring the responsible and lawful acquisition of data.

Acknowledgements

This work has been supported 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 NextGenerationEU programme and by the European Unions Horizon Europe AEQUITAS research and innovation programme under grant number 101070363. The author thanks Liam James for initiating the implementation of this work and for his early contributions.

References

[1] A. Agarwal, M. Dudík, and Z. S. Wu. Fair regression: Quantitative definitions and reduction-based algorithms. In K. Chaudhuri and R. Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pages 120–129. PMLR, 2019. URL http://proceedings.mlr.press/v97/agarwal19d.html.

[2] R. K. Bellamy, K. Dey, M. Hind, S. C. Hoffman, S. Houde, K. Kannan, P. Lohia, J. Martino, S. Mehta, A. Mojsilović, et al. Ai fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias. *IBM Journal of Research and Development*, 63(4/5):4–1, 2019.

[3] S. Bird, M. Dudík, R. Edgar, B. Horn, R. Lutz, V. Milan, M. Sameki, H. Wallach, and K. Walker. Fairlearn: A toolkit for assessing and improving fairness in ai. *Microsoft, Tech. Rep. MSR-TR-2020-32*, 2020.

[4] C. Case. 83/14 chez razpredelenie bulgaria ad v. *Komisia za zashtita ot diskriminatsia*, 2015.

[5] S. Corbett-Davies, J. D. Gaebler, H. Nilforoshan, R. Shroff, and S. Goel. The measure and mismeasure of fairness. *The Journal of Machine Learning Research*, 24(1):14730–14846, 2023.

[6] P. Cortez and A. Silva. Using data mining to predict secondary school student performance. In *Proceedings of 5th Annual Future Business Technology Conference*, 2008.

[7] EC. Regulation - eu - 2024/1689 - en - eur-lex. *OJ*, 2024. URL https://eur-lex.europa.eu/eli/reg/2024/1689/oj/eng#:~:text=Regulation%20%2D%20EU%20%2D%202024%2F1689%20%2D%20EN%20%2D%20EUR%2DLex.

[8] B. Green and L. Hu. The myth in the methodology: Towards a recontextualization of fairness in machine learning. In *Proceedings of the machine learning: the debates workshop*, 2018.

[9] High-Level Expert Group on AI. Ethics guidelines for trustworthy ai. Report, European Commission, Brussels, Apr. 2019. URL https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai.

[10] Z. Jiang, X. Han, C. Fan, F. Yang, A. Mostafavi, and X. Hu. Generalized demographic parity for group fairness. In *International Conference on Learning Representations*, 2022.

[11] J. P. Lalor, A. Abbasi, K. Oketch, Y. Yang, and N. Forsgren. Should fairness be a metric or a model? a model-based framework for assessing bias in machine learning pipelines. *ACM Transactions on Information Systems*, 2024.

[12] K. Meding. It's complicated. the relationship of algorithmic fairness and non-discrimination regulations for high-risk systems in the eu ai act. *arXiv preprint arXiv:2501.12962*, 2025.

[13] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan. A survey on bias and fairness in machine learning. *ACM computing surveys (CSUR)*, 54(6):1–35, 2021.

[14] G. Palumbo, D. Carneiro, and V. Alves. Objective metrics for ethical ai: a systematic literature review. *International Journal of Data Science and Analytics*, pages 1–21, 2024.

[15] B. Richardson and J. E. Gilbert. A framework for fairness: A systematic review of existing fair ai solutions. *arXiv preprint arXiv:2112.05700*, 2021.

[16] A. Roy, S. Rizou, S. Papadopoulos, and E. Ntoutsi. Achieving socioeconomic parity through the lens of eu ai act. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency*, pages 1890–1901, 2025.

[17] D. Schiek. A new framework on equal treatment of persons in ec law? directives 2000/43/ec, 2000/78/ec and 2002/???/ec changing directive 76/207/eec in context. *European Law Journal*, 8(2):290–314, 2002.

[18] R. F. Schoeni, V. A. Freedman, and L. G. Martin. Socioeconomic and demographic disparties in trends in old-age disability. In *Health at older ages: The causes and consequences of declining disability among the elderly*, pages 75–102. University of Chicago Press, 2009.

[19] N. M. Stephens, H. R. Markus, and S. A. Fryberg. Social class disparities in health and education: reducing inequality by applying a sociocultural self model of behavior. *Psychological review*, 119(4):723, 2012.

[20] S. Wachter, B. D. Mittelstadt, and C. Russell. Why fairness cannot be automated: Bridging the gap between EU non-discrimination law and AI. *Comput. Law Secur. Rev.*, 41:105567, 2021. doi: 10.1016/J.CLSR.2021.105567. URL https://doi.org/10.1016/j.clsr.2021.105567.

[21] H. Weerts, F. Pfisterer, M. Feurer, K. Eggensperger, E. Bergman, N. Awad, J. Vanschoren, M. Pechenizkiy, B. Bischl, and F. Hutter. Can fairness be automated? guidelines and opportunities for fairness-aware automl. *Journal of Artificial Intelligence Research*, 79:639–677, 2024.