Assessing machine learning fairness with multiple contextual norms

Pim Kerkhoven[1 [0009−0006−9193−5830]], Virginia Dignum[1 [0000−0001−7409−5813]], and Monowar Bhuyan[1 [0000−0002−9842−7840]]

Department of Computing Science
Umeå University, Umeå, 90187, Sweden
{pim, virginia, monowar}@cs.umu.se

Abstract. The increasing societal impact of decisions made with machine learning (ML) systems requires these decisions to be fair. To fully address the fairness of an ML system, fairness must be interpreted as context-dependent and shaped by the societal environment in which this ML system operates. However, existing group fairness measures do not consider this context, reducing fairness to equality between groups, neglecting other norms such as equity or need. Therefore, we propose *contextual fairness*, a context-dependent framework for assessing ML fairness with multiple context-specific norms such as equality, equity, and need. Our approach involves four steps: eliciting fairness norms from the relevant stakeholders, operationalizing these norms into quantitative measures, combining these measures into a weighted contextual fairness score, and analyzing this score at global, between-group, and within-group levels. We evaluate our approach on the ACS Income dataset for both binary classification and regression tasks, showing how *contextual fairness* improves the interpretation of fairness assessment and mitigation in context. By explicitly incorporating norms in fairness assessment, our approach enables more nuanced evaluations and better supports practitioners and regulators in considering the societal context.

Keywords: algorithmic fairness · socio-technical systems · fairness assessment · machine learning

1 Introduction

Machine learning (ML) systems increasingly help to make decisions that impact individuals and society unfairly due to biases in data and models [12, 24]. Notable examples of this are the racial biases exhibited by the COMPAS system [3] and facial recognition systems [6]. To address fairness issues during the development of ML systems, commonly group fairness measures, such as demographic parity [19] or equalized odds [18], are used to calculate to which extent an ML system favors "*[a] group based on their inherent or acquired characteristics*" [24]. Subsequently, mitigation techniques are used to enhance the measured fairness [26].

As these group measures typically reduce fairness to equality between groups (e.g., treating males and females the same) [8], other fairness norms such as equity or need are neglected, leading to misleading or incomplete fairness assessments [5]. Consequently, deployed ML systems that are deemed fair, in practice, may treat individuals and groups unfairly, resulting in negative societal impact. For example, consider a company that prefers internal applicants over external applicants for new job positions. When this company uses a fair automated hiring system that treats all individuals equally, internal applicants will be treated unfairly with respect to the company's preference.

Prior work has shown that to improve fairness assessments, fairness should be considered in context and from the perspective of all relevant stakeholders, leading to the consideration of different fairness norms [5, 10, 31]. These norms arise from the inherent situated values and institutional practices that underpin fairness, as emphasized by philosophical and socio-technical perspectives [5, 17, 31]. Although the importance of context and values for fairness is widely discussed in the literature, as even different notions of equality-based fairness exist (e.g., who should be treated equal) [8, 14], which has led to frameworks that assist in selecting suitable equality-based fairness measures for a specific context [4, 14, 28]. These frameworks and the fairness measures they incorporate focus only on equality across groups, leaving other fairness norms unaddressed.

Therefore, we propose *contextual fairness*, a general framework for assessing fairness in context with multiple norms (e.g., equality, equity, or need), and show its implications in a classification and regression scenario. To this end, we use Forsyth's taxonomy of fairness norms [13] to provide general formulations for operationalizing different types of norms. But we acknowledge that to formulate norms for a specific context, these norms should be elicited through participatory or policy-driven processes.

In summary, this work has the following contributions:

- We extend fairness measures beyond equality by defining rank-based norms (e.g., equity, need) for binary classification and regression tasks.
- We propose *contextual fairness* as a fairness measure that combines multiple context-specific norms into a single weighted measure to assess models on multiple fairness norms.
- The proposed *contextual fairness* measure allows for assessing fairness at the global, between-group, and within-group levels, which leads to more detailed analyses than traditional group measures.
- We empirically show, using the ACS Income dataset [11], how *contextual fairness* changes fairness interpretation and reveals limitations in mitigation techniques with respect to broader contextual norms.

2 Background and related work

2.1 Group fairness measures

Binary classification In general, group fairness measures for binary classification (see Table 1 for notation) rely on the calculation of some probability

Table 1: Notations.

Notation	Description
$D = \{X_1, \ldots, X_n\}$	Dataset of n individuals.
$L = \{Y_1, \ldots, Y_n\}$	The set of target labels for each individual X_i.
$A = \{a_1, \ldots, a_h\}$	The set of attributes used to describe each individual X_i.
$S = \{s_1, \ldots, s_z\}$	The set of sensitive attributes used to describe each individual X_i. Note that S is not necessarily a subset of A.
$X_i.\mathrm{a}_i$	Value for attribute a_i of individual X_i.
$M(X_i) = \hat{Y}_i$	Prediction of the model M for individual X_i. Note that $\hat{Y}_i \in \{0, 1\}$ if M is a binary classification model and $\hat{Y}_i \in \mathbb{R}$ if M is a regression model.
$Pr(\hat{Y}_i = v \mid X_i)$	Probability that the outcome with value v is predicted for individual X_i.
Y_{pref}	The preferred class for the outcome of the binary classification model M. When considering binary classification, this is the class individuals want to be assigned to (e.g., being predicted to get a loan). We use $\neg Y_{\text{pref}}$ to denote the other class.
$g :=$ condition	A group condition g that is used to define a group of individuals. Note that we use g to denote any group condition.
G_g	The set of individuals for which the group condition g holds. We use G to denote any group.

$Pr(\hat{Y} = Y_{\text{pref}} \mid g)$ [8]. Consequently, the choice of which groups g to consider is what sets the different group measures apart. For example, consider a setting with one sensitive attribute s with possible values v_1 and v_2. In this setting, the group conditions for demographic parity [19] are defined as $g_1 := s = v_1$ and $g_2 := s = v_2$. In contrast, for equal opportunity [18] these conditions are defined as $g_1 := Y = 1 \wedge s = v_1$ and $g_2 := Y = 1 \wedge s = v_2$. Additionally, for equalized odds [18] we must also consider the group conditions $g_3 := Y = 0 \wedge s = v_1$ and $g_4 := Y = 0 \wedge s = v_2$.

Regression Most fairness measures are formulated for binary classification [8]. Therefore, these measures consider the model outputs to be binary. In contrast, for regression, we need to consider continuous outputs [8]. To this end, several group fairness measures for regression have been proposed, such as statistical parity for regression and bounded group loss [1]. Statistical parity for regression is achieved when $Pr(\hat{Y} \geq v \mid g) = Pr(\hat{Y} \geq v)$ for all defined groups g and all $v \in \mathbb{R}$. That is, the probability that an individual is predicted a higher value than v should not depend on which group g this individual belongs to. Alternatively, bounded group loss states that given a regression model M with a loss function L, the loss for each defined group G should be below a specified threshold. Consequently, for a sufficiently low threshold, there is no group for which the model M performs much worse than the other groups.

2.2 Fairness norms

In general, group fairness measures aim to measure whether all groups are treated the same [24]. Therefore, we argue that these measures only consider fairness as equality. However, being fair and being equal are not necessarily the same. Depending on the societal context, other fairness norms might be more relevant [10]. Forsyth identifies five such fairness norms, namely equality, equity, need, power, and responsibility [13]. Table 2 provides an overview of these norms and interprets each norm for a binary classification setting. From these descriptions and interpretations, we notice a difference between the equality norm and the other norms. The equality norm aims to treat everyone the same. In contrast, the other norms impose a ranking on the individuals, where individuals that are ranked higher with respect to the norm should get a higher (or lower) probability of being predicted the positive class. Therefore, we will refer to these norms as *rank norms*.

Table 2: Explanation of fairness norms for decisions [13] and their interpretation in a binary classification setting.

Norm	Explanation	Interpretation for binary classification
Equality	Everyone should get the same.	Everyone should be predicted Y_{pref}.
Equity	People who contribute more should receive more.	More equity for an individual X_i should mean a higher probability $Pr(\hat{Y}_i = Y_{\text{pref}}\|X_i)$.
Need	People who need it more should receive more.	More need for an individual X_i should mean a higher probability $Pr(\hat{Y}_i = Y_{\text{pref}}\|X_i)$.
Power	People with more authority, control, or status should receive more.	More power for an individual X_i should mean a higher probability $Pr(\hat{Y}_i = Y_{\text{pref}}\|X_i)$.
Responsibility	People who have more should receive less.	More responsibility for an individual X_i should mean a lower probability $Pr(\hat{Y}_i = Y_{\text{pref}}\|X_i)$.

2.3 Fairness and context

To assess the fairness of an ML system, we should not only consider the input data and the output predictions, but also the societal context this data describes and these predictions affect [31, 32]. This context entails the social, institutional, and normative environment in which decisions are made [15, 34]. Important factors in this are the purpose of the system, the stakeholders involved, and the embedded values [35].

Therefore, to fully understand what makes a decision fair, it is important to consider and understand the societal context. However, as others have argued,

this is currently not the standard. Notably, Binns states that fairness often depends on contextual factors not encoded in the data, and that by ignoring these factors, it is hard to have fair outcomes in a specific context [5]. Additionally, Selbst et al. state that existing fairness measures abstract away the societal context of an ML system, leading to notions of fairness that do not reflect what is fair in context [31]. Similarly, Green and Hu argue that the usage of abstract universal fairness measures ignores contextual social norms, thereby failing to address the social, political, and moral aspects of fairness [17].

To incorporate the societal context of a specific ML system for addressing fairness, contextual values can be translated into context-specific fairness norms by using requirement elicitation techniques [10, 29, 30]. To address all contextual perspectives on fairness techniques that consider all stakeholders, such as value-sensitive design [16] or participatory design [37], should be preferred. One key aspect of such elicitation techniques is that they involve methods to address conflicts and trade-offs between the perspectives of different stakeholders [10].

Although the mentioned works discuss the importance of context and the translation of contextual values into norms, their focus is not on formally operationalizing these norms to measure the fairness of an ML system. Moreover, existing fairness measures and toolkits primarily implement group-based measures focusing on equality [5], which do not consider the societal context and only rely on the input and outputs of an ML model [31]. Therefore, methods for operationalizing fairness assessment in contexts with multiple different norms are lacking.

3 A contextual approach to assessing ML group fairness

Our approach to assessing the fairness of an ML system follows four steps: (1) elicit context-specific norms, (2) operationalize them as quantitative measures, (3) combine these into a contextual fairness score, and (4) analyse this score at three levels, namely global fairness, between-group fairness, and within-group fairness. Figure 1 shows these steps in relation to the ML pipeline [2].

3.1 Formulating norms

To formulate which fairness norms are applicable in a specific context, the elicitation techniques discussed in Section 2.3 are used. Note that applying such techniques correctly is not straightforward and requires careful consideration of the context and who the stakeholders are. However, for our purposes, we are only interested in the results of using such techniques, namely a carefully created list of fairness norms for a specific societal context. For example, if we consider an automated hiring system, these norms could be as follows:

- All applicants must be treated the same regardless of race or gender.
- Internal applicants should be preferred over external applicants.

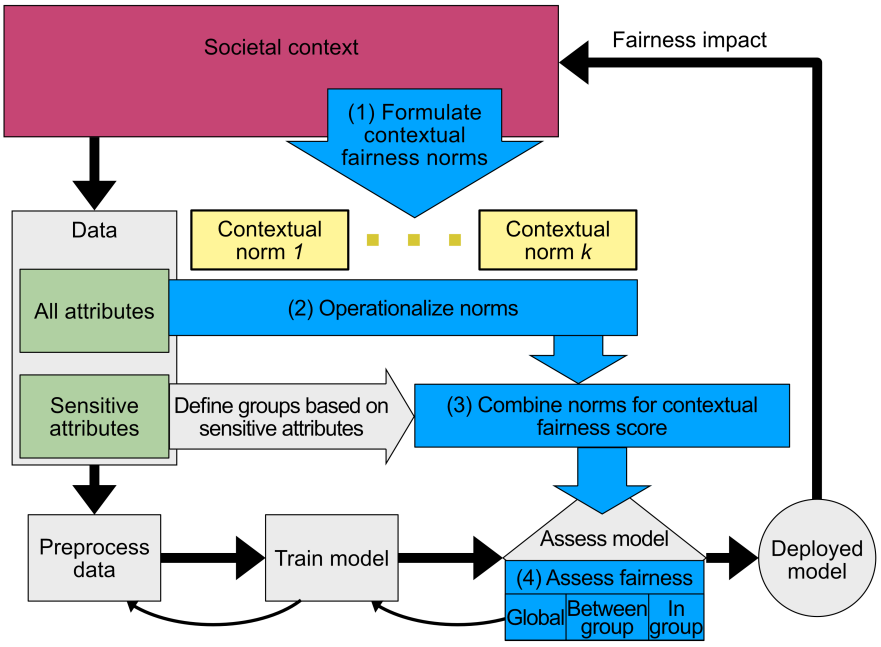


Fig. 1: Our approach (in blue) for assessing fairness in the ML pipeline.

3.2 Contextual fairness

To calculate a fairness score for a specific model, we operationalize the formulated norms by defining a function for each norm. We differentiate between functions for the *equality norm* and the *rank norms* as discussed in Section 2.2. We continue by providing general formulations for these functions for binary classification and regression that can be applied in context. Subsequently, we combine these operationalized norms to calculate a single *contextual fairness* score.

Operationalizing norms for binary classification For binary classification (see Table 1 for notation), we define the *equality norm* as follows:

Definition 1 (Equality norm for binary classification). *Given a binary classification model M, a data set D, and a preferred class Y_{pref}. We calculate a score for the equality norm of a model as:*

$$Nr_{equality}(D) = \frac{\sum_{\hat{X}_i \in D} |M(X_i) - Y_{pref}|}{n}.$$

That is, to calculate the *equality norm*, we count the number of predictions that are not equal to the preferred class and normalize this by dividing by n. Note that whilst counting the number of predictions not equal to the majority

class might seem more intuitive, it leads to misleading interpretations when the preferred and majority classes differ.

To operationalize a *rank norm*, we operationalize the underlying principle of these norms, namely that individuals who are ranked higher with respect to the norm should have a higher (or lower) probability of being predicted the preferred class. To this end, we rank all individuals in two ways. First, based on a norm-specific function $V_{\text{nr}}(X_i) : A \to \mathbb{R}$ that is defined for each context (e.g., hours worked per week for equity). Next, all individuals are ranked based on their predicted probability for the preferred class Y_{pref}. We use the (normalized) Kendall tau ranking distance [21] to calculate a score from these two rankings, i.e., we count the pairs of individuals where the individual with a higher value for V_{nr} does not have a higher probability of being predicted the preferred class. We, more formally, define *rank norms* as follows:

Definition 2 (Rank norm for binary classification). *Given a dataset D, a binary classification model M, and a norm value function V_{nr}, we calculate the rank norm as:*

$$Nr_{rank}(D) = \frac{2}{n(n-1)} \sum_{X_i \in D} \sum_{X_j \in D \setminus \{X_i\}}$$

$$\begin{cases} 1 & V_{nr}(X_i) > V_{nr}(X_j) \ \wedge Pr(M(X_i) = Y_{pref} | X_i) < Pr(M(X_j) = Y_{pref} | X_j) \\ 0 & otherwise \end{cases}.$$

Operationalizing norms for regression We now define the functions for operationalizing *equality* and *rank norms* in a regression setting. To this end, we use the same principles that underpin our definitions for binary classification. Therefore, we define the *equality norm* as follows:

Definition 3 (Equality norm for regression). *Given a regression model M, a dataset D, a maximal prediction $Y_{max} = \max\{M(X_1), \ldots, M(X_n)\}$ and a minimal prediction $Y_{min} = \min\{M(X_1), \ldots, M(X_n)\}$. We calculate a score for the equality norm of a model as:*

$$Nr_{equality}(D) = \frac{\sum_{\hat{X}_i \in D} |M(X_i) - Y_{max}|}{n \times |Y_{max} - Y_{min}|}.$$

That is, assuming that a higher output is more preferable, we calculate the normalized average absolute difference between each individual and the maximum output Y_{max}. In contexts where a lower output would be preferable, we should use Y_{min} instead of Y_{max}.

For operationalizing *rank norms* for regression, we rank individuals using the norm-specific function V_{nr} and the model's output for each individual, giving the following definition:

Definition 4 (Rank norm for regression). *Given a dataset D, a regression model M, and a norm value function V_{nr}, we calculate the rank norm as:*

$$Nr_{rank}(D) = \frac{2}{n(n-1)} \sum_{X_i \in D} \sum_{X_j \in D \setminus \{X_i\}} \begin{cases} 1 & V_{nr}(X_i) > V_{nr}(X_j) \ \wedge M(X_i) < M(X_j) \\ 0 & otherwise \end{cases}.$$

Combining operationalized norms for contextual fairness After operationalizing each norm, we need to combine these operationalized norms to calculate the *contextual fairness* of a model. We define this as follows:

Definition 5 (Contextual fairness). *Given a model M, a dataset D, and k operationalized norms $Nr_1, \ldots, Nr_k$ with corresponding weights $w_1, \ldots, w_k$ that sum to 1 (i.e., $\sum_{i=1}^{k} w_i = 1$). We calculate the contextual fairness as follows:*

$$\textit{contextual fairness}_M(D) = w_1 Nr_1(D) + \ldots + w_k Nr_k(D).$$

That is, we calculate the weighted combination of all operationalized norms. These weights allow us to give more importance to some norms and less to others depending on the context.

Since each operationalized norm is defined as a summation over all individuals, *contextual fairness* can be expressed as a weighted sum of norm scores for each individual in the dataset, as follows:

$$\text{contextual fairness}_M(D) = \sum_{X_i \in D} w_1 \text{Nr}_{1_i}(X_i) + \ldots + w_k \text{Nr}_{k_i}(X_i).$$

Additionally, this formulation allows for computing fairness scores for a particular group G:

$$\text{contextual fairness}_M(G) = \sum_{X_i \in G} w_1 \text{Nr}_{1_i}(X_i) + \ldots + w_k \text{Nr}_{k_i}(X_i).$$

Finally, note that if we have a set of t groups $\{G_1, \ldots, G_t\}$ that partition the dataset D, then the sum of their *contextual fairness* scores equals the global *contextual fairness* score.

3.3 Fairness analysis

The definitions given for *contextual fairness* allow for a fairness analysis on three different levels, namely by analyzing the global score, the difference in scores between groups, and the distribution of the score within a group.

The global *contextual fairness* (see Definition 5) gives a general indication of how unfair a model is. However, a higher *contextual fairness* score does not necessarily mean a model is more unfair than a model with a lower *contextual fairness* score. For example, consider two ML models, A and B. Suppose model A has a *contextual fairness* score of 0.5. However, both men and women contribute equally to this score, i.e., both make up 0.25 of this score. Now suppose, model B has a *contextual fairness* score of 0.3, but only men contribute to this score. Although model B appears to be more fair if we only analyze the global *contextual fairness* of the models, when considering how this score is distributed, model A appears to be more fair.

Therefore, we need to consider how the score is distributed between different groups. Preferably, we want the *contextual fairness* to be distributed equally over the groups. In this case, all groups are treated the same. To allow for a fair comparison between groups of different sizes, we scale the score for each group as follows:

Definition 6 (Scaled contextual fairness). *Given a dataset D, and a partition of D into t groups $\{G_1, \ldots, G_t\}$, we calculate the scaled contextual fairness score for each group G_i as follows:*

$$\textit{scaled contextual fairness}_M(G_i) = \frac{\textit{contextual fairness}(G_i)/|G_i|}{\sum_{G_j \in \{G_1,\ldots,G_t\}} \textit{contextual fairness}(G_j)/|G_j|} \textit{contextual fairness}(D).$$

Note, this score is defined such that for a partition of groups, the sum of the *scaled contextual fairness* scores still sums to the global *contextual fairness* score.

Lastly, we also need to consider how the score for a group is distributed within the group. Do all individuals contribute equally to the score, or are there a few individuals who contribute substantially more than the others in the group.

By combining these three analysis steps, we make a detailed analysis of the fairness of an ML model. Both in terms of the magnitude and distribution of the score over and within the different groups.

4 Evaluation

We evaluate *contextual fairness* in a hypothetical loan approval scenario where a bank uses an ML model to predict applicants' income[1]. This prediction is then used in deciding whether an individual's loan request is approved. Our evaluation considers both a binary classification and a regression variant of this scenario. For each experiment, we analyze how the fairness analysis with *contextual fairness* differs from a baseline analysis using a group-based equality measure.

Models are trained on the 2016 California ACS Income dataset [11], which includes ten demographic and socioeconomic attributes, with yearly income as

[1] Code available at https://github.com/pimkerkhoven/contextualfairness

the target attribute. For classification, income is binarized into a $+50k$ and a $-50k$ class, with $+50k$ being the preferred class. We treat sex as the sensitive attribute resulting in two groups, namely male (m) and female (f). We split the data into a training set (80%) and testing set (20%) and use this data to train logistic regression (classification) and linear regression (regression) models from Scikit-learn[2] [25].

4.1 Classification setting

For the classification setting, we use equalized odds [18] as the baseline. For the described scenario, equalized odds is achieved when the following two equalities hold:

$$Pr(\hat{Y} = +50k|\text{sex} = \text{m} \wedge Y = +50k) = Pr(\hat{Y} = +50k|\text{sex} = \text{f} \wedge Y = +50k)$$
$$Pr(\hat{Y} = +50k|\text{sex} = \text{m} \wedge Y = -50k) = Pr(\hat{Y} = +50k|\text{sex} = \text{f} \wedge Y = -50k)$$

We consider three different cases, namely one with only an equality norm for *contextual fairness*, one with a changed context with more contextual norms, and one where we apply a mitigation technique to the trained model.

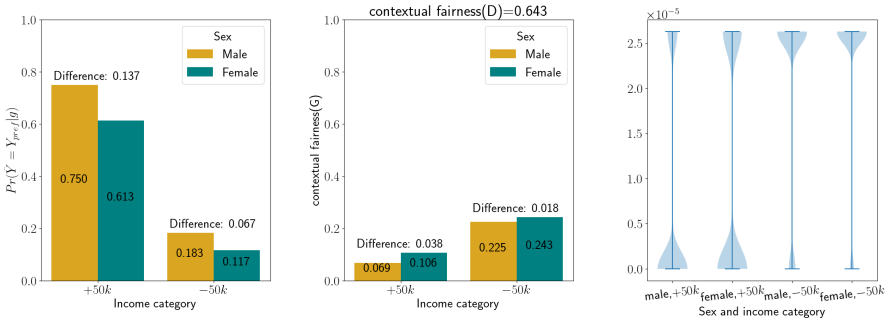
Only equality In this experiment, we assume only the equality norm applies in the system's context. Therefore, to showcase the analysis process described in Section 3.3, we compare the baseline with a *contextual fairness* score that only uses the equality norm.

Baseline: Figure 2a shows that females are disadvantaged compared to males in terms of equalized odds.
Contextual fairness: Figures 2b and 2c capture the same disparity but also quantify the global unfairness and show how it is distributed within groups.
Takeaway: Even with a single equality norm, *contextual fairness* provides for a more detailed analysis than equalized odds alone. By being able to analyze the in-group distribution of fairness, we can potentially address the inequalities at the individual level as well.

Changing contexts and multiple norms In the previous experiment, we considered a context in which only the equality norm was applicable as a fairness norm. For this experiment, we assume that the context changed and multiple fairness norms apply. Although in a real-world scenario these norms should be elicited in an elaborate process (see Section 3.1), for our hypothetical scenario, we assume the following norms apply that we weigh equally for the *contextual fairness* score:

- Everybody should get the same prediction (equality).

[2] https://scikit-learn.org/



(a) Equalized odds. (b) Between-group *contextual fairness*. (c) In-group *contextual fairness*.

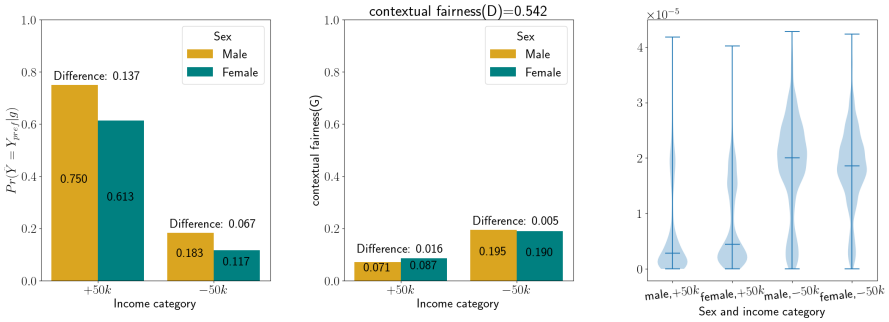
Fig. 2: Results for equalized odds and *contextual fairness* when considering only the equality norm.

- People who work more hours should earn more (equity), i.e., $V_{\text{equity}}(X_i) = X_i.\text{worked\_hours\_per\_week}$.
- People with a lower education level should earn more (need), i.e., $V_{\text{need}}(X_i) = -X_i.\text{education\_level}$.

Baseline: As we use the same model as above, Figure 3a displays the same disparity between males and females as above.
Contextual fairness: Figures 3b and 3c show that by changing the context the *contextual fairness* score is affected compared to the previous experiment.
Takeaway: The outcome for equalized odds remains the same across different contexts. In contrast, the *contextual fairness* score is affected by the changing context. This highlights that in contexts where fairness is not solely defined by the equality norm, equality-based measures cannot fully capture the fairness of a model. Rather, more context-sensitive evaluations are essential for assessing model fairness.

Mitigating unfairness To improve the fairness of an ML system, we apply mitigation techniques to reduce the measured unfairness. To showcase this, we apply Fairlearn's correlation remover [36] to the model used in the previous experiment. We analyze the effect of this mitigation on equalized odds and *contextual fairness* in the same context as the previous experiment.

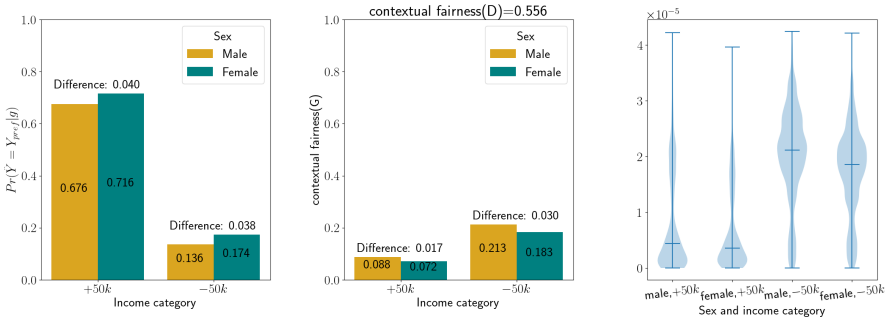
Baseline: Figure 4a shows that the mitigation reduced the disparities between males and females in both income groups compared to Figure 3a.
Contextual fairness: Compared to Figures 3b and 3c, Figures 4b and 4c do not show that the mitigation technique reduced the unfairness. Rather, the unfairness appears to have increased, as both the global score and the between-group differences increased.



(a) Equalized odds. (b) Between-group *contextual fairness*. (c) In-group *contextual fairness*.

Fig. 3: Results for equalized odds and *contextual fairness* when considering multiple contextual fairness norms.

Takeaway: A mitigation technique that reduces unfairness with respect to equalized odds does not necessarily do this for *contextual fairness*. This indicates that mitigation techniques should also take the context into account.



(a) Equalized odds. (b) Between-group *contextual fairness*. (c) In-group *contextual fairness*.

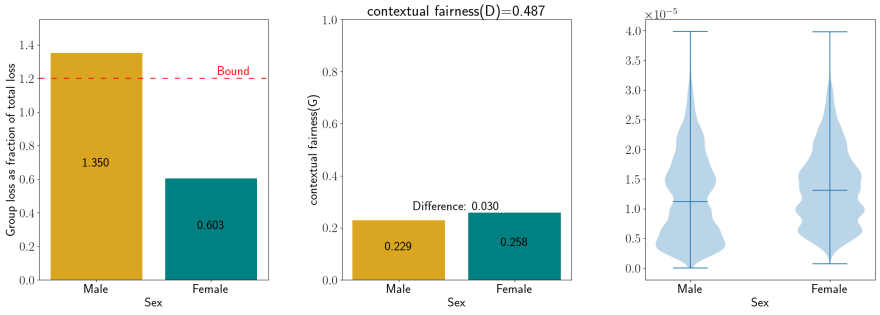
Fig. 4: Results for equalized odds and *contextual fairness* after applying a mitigation technique.

4.2 Regression setting

We now consider the regression setting of our scenario, i.e., instead of predicting the income class ($+50k$ or $-50k$), we predict income directly. As equalized odds does not apply in this setting, we use bounded group loss [1], with a bound of

1.2 as the baseline. We consider the same context as the last two experiments. However, we operationalize the formulated norms for regression.

Baseline: Figure 5a shows that males are treated unfairly as their loss exceeds the bound.
Contextual fairness: In contrast, Figures 5b and 5c show that females are treated unfairly in this context.
Takeaway: Similar to the classification experiments, this experiment shows that context matters for assessing who is treated unfairly. Furthermore, we had to change the baseline measure to a measure that handles regression, changing how we should interpret this baseline. In contrast, by adapting how we operationalize the norms to the regression setting, we could still use *contextual fairness*.



(a) Bounded group loss. (b) Between-group *contextual fairness*. (c) In-group *contextual fairness*.

Fig. 5: Results for bounded group loss and *contextual fairness* for the regression setting.

5 Discussion

Our approach relies on norm elicitation techniques and processes, such as value-sensitive design, leading to a list of contextual norms that can be operationalized to calculate a *contextual fairness* score. We argue that this is a valid approach as these techniques have been used in practice to formulate various norms [33]. Although there is no guarantee that the formulated norms accurately reflect fairness in context, these norms can still prove useful. As incorrect norms elicited from the context through a systematic and transparent process provide an accountable motivation to argue against when contesting these norms [10].

Furthermore, we assume that norms are computable from the data at hand. However, not every formulated norm is necessarily computable, and in such cases,

it should be evaluated if *contextual fairness* can be applied. We argue that, in such cases, this would also be the case for group-based fairness measures.

In this work, we focus on five possible fairness norms in line with Forsyth's framework [13]. However, other takes on fairness do exist, such as Rawls' [27], and may be equally valid. Regardless of the framework used, different perspectives on fairness should be discussed in the elicitation process, such that the formulated fairness norms are well motivated and accurately reflect the real-world values of stakeholders.

Our definition of *contextual fairness* uses a linear combination of different operationalized norms. Although this is a good first step towards using multiple norms for assessing fairness in context, it is inherently limited by this linearity. As this linearity cannot capture conditional or hierarchical relationships between norms, for example, when we want equity except in cases of need. Therefore, future work is needed to explore more expressive aggregations of norms.

ML models are not only developed with tabular data, but also with other data modalities, such as text or image data [23]. For such data types, defining norm value functions becomes non-trivial, e.g., defining equity based on pixels is not as straightforward as equating equity to someone's income. Therefore, currently *contextual fairness* is mostly appropriate for tabular data. Additionally, the calculation of *rank norms* relies on a pairwise comparison of all individuals in the dataset. This may pose scalability issues for large datasets, which could potentially be mitigated with approximation techniques.

The current analysis of *contextual fairness* only considers the weighted score over all norms. Therefore, which norms contribute the most to the unfairness is obscured. Allowing for a more elaborate analysis that breaks down the weighted score into the various norms would lead to a more detailed discussion of how fair a model is.

Finally, for the experiments in Section 4, we only considered one sensitive attribute. However, people at the intersection of multiple sensitive attributes (e.g., black women) are often treated disproportionately unfairly [9, 22]. Therefore, considering multiple sensitive attributes would improve the fairness analyses. Note that, as the number of groups grows exponentially with the number of attributes, this would require some method for selecting which groups should be included in the analyses [7, 20].

6 Concluding remarks

We proposed *contextual fairness*, a fairness measure that assesses the fairness of an ML system with multiple norms elicited from the societal context of this system. We provided definitions for operationalizing the elicited norms for binary classification and regression tasks. Our evaluation showed that by considering the global, between-group, and within-group scores *contextual fairness* allows for a more detailed fairness analysis. In addition, by incorporating multiple contextual norms, fairness assessments with *contextual fairness* better reflect the societal context in which an ML system operates than in its absence.

However, these fairness assessments are not neutral: depending on which norms are elicited, different groups appear disadvantaged and different interventions seem justified. This already shows in the simple scenario we used for evaluating *contextual fairness*. Compared to existing fairness measures, *contextual fairness* makes the assumptions underpinning fairness assessment more explicit. This way *contextual fairness* offers a more structured approach for practitioners and regulators to deliberate on which fairness norms should guide evaluation in a given context. When using *contextual fairness* in practice, three key factors should be considered.

Firstly, when developing ML systems, fairness is not the only factor that should be considered. Other aspects, for example, performance, privacy, or robustness, must also be considered. When considering these aspects, the fairest system is not necessarily the best system with respect to all aspects. Therefore, a deliberate trade-off considering all relevant aspects must be made.

Secondly, the limitations discussed in Section 5 should be taken into account. Before applying *contextual fairness* in scenarios with, for example, non-linear norms, large datasets, or non-tabular data, the norm operationalization definitions should be extended to support such cases.

Lastly, societal contexts are never static, and importantly, the outputs of an ML system will affect the system's context, regardless of how fair these outputs are. Therefore, formulating and operationalizing norms with *contextual fairness* is an ongoing process. Regular checks should be performed to evaluate if the defined *contextual fairness* score still matches the context it is defined for.

Acknowledgments. This project has received funding from the European Union's Horizon Europe research and innovation program under the Marie Skłodowska-Curie grant agreement No 101073307. This work was also supported by the Wallenberg AI, Autonomous Systems and Software Program – Humanities and Society (WASP-HS) funded by the Marianne and Marcus Wallenberg Foundation and the AI Policy Lab at Umeå University, Sweden.

References

1. Agarwal, A., Dudik, M., Wu, Z.S.: Fair Regression: Quantitative Definitions and Reduction-Based Algorithms. In: Proceedings of the 36th International Conference on Machine Learning. pp. 120–129. PMLR (May 2019), https://proceedings.mlr.press/v97/agarwal19d.html, iSSN: 2640-3498
2. Amershi, S., Begel, A., Bird, C., DeLine, R., Gall, H., Kamar, E., Nagappan, N., Nushi, B., Zimmermann, T.: Software Engineering for Machine Learning: A Case Study. In: 2019 IEEE/ACM 41st International Conference on Software Engineering: Software Engineering in Practice (ICSE-SEIP). pp. 291–300. IEEE, Montreal, QC, Canada (May 2019). https://doi.org/10.1109/ICSE-SEIP.2019.00042, https://ieeexplore.ieee.org/document/8804457/
3. Angwin, J., Larson, J., Mattu, S., Kirchner, L.: Machine Bias: There's software used across the country to predict future criminals. And it's biased against blacks. (2016), https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

4. Barr, C.J.S., Erdelyi, O., Docherty, P.D., Grace, R.C.: A Review of Fairness and A Practical Guide to Selecting Context-Appropriate Fairness Metrics in Machine Learning (Feb 2025). https://doi.org/10.48550/arXiv.2411.06624, http://arxiv.org/abs/2411.06624, arXiv:2411.06624 [cs]
5. Binns, R.: Fairness in Machine Learning: Lessons from Political Philosophy. In: Proceedings of the 1st Conference on Fairness, Accountability and Transparency. pp. 149–159. PMLR (Jan 2018), https://proceedings.mlr.press/v81/binns18a.html, iSSN: 2640-3498
6. Buolamwini, J., Gebru, T.: Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. In: Proceedings of the 1st Conference on Fairness, Accountability and Transparency. pp. 77–91. PMLR (Jan 2018), https://proceedings.mlr.press/v81/buolamwini18a.html, iSSN: 2640-3498
7. Cabrera, A.A., Epperson, W., Hohman, F., Kahng, M., Morgenstern, J., Chau, D.H.: FAIRVIS: Visual Analytics for Discovering Intersectional Bias in Machine Learning. In: 2019 IEEE Conference on Visual Analytics Science and Technology (VAST). pp. 46–56 (Oct 2019). https://doi.org/10.1109/VAST47406.2019.8986948
8. Caton, S., Haas, C.: Fairness in Machine Learning: A Survey. ACM Comput. Surv. **56**(7), 166:1–166:38 (Apr 2024). https://doi.org/10.1145/3616865, https://dl.acm.org/doi/10.1145/3616865
9. Crenshaw, K.: Demarginalizing the intersection of race and sex: a black feminist critique of antidiscrimination doctrine, feminist theory and antiracist policies. University of Chicago Legal Forum **1989**(1), 139–167 (1989), http://chicagounbound.uchicago.edu/uclf/vol1989/iss1/8/
10. Dignum, V.: Responsible Artificial Intelligence: How to Develop and Use AI in a Responsible Way. Artificial Intelligence: Foundations, Theory, and Algorithms, Springer International Publishing, Cham (2019). https://doi.org/10.1007/978-3-030-30371-6, http://link.springer.com/10.1007/978-3-030-30371-6
11. Ding, F., Hardt, M., Miller, J., Schmidt, L.: Retiring Adult: New Datasets for Fair Machine Learning. In: Advances in Neural Information Processing Systems. vol. 34, pp. 6478–6490. Curran Associates, Inc. (2021)
12. Ferrara, E.: Fairness and Bias in Artificial Intelligence: A Brief Survey of Sources, Impacts, and Mitigation Strategies. Sci **6**(1), 3 (Mar 2024). https://doi.org/10.3390/sci6010003, https://www.mdpi.com/2413-4155/6/1/3, number: 1 Publisher: Multidisciplinary Digital Publishing Institute
13. Forsyth, D.R.: Group dynamics. Wadsworth, Cengage Learning, Australia ; Belmont, CA, 5th ed edn. (2010)
14. Franklin, J.S., Powers, H., Erickson, J.S., McCusker, J., McGuinness, D.L., Bennett, K.P.: An Ontology for Reasoning About Fairness in Regression and Machine Learning. In: Ortiz-Rodriguez, F., Villazón-Terrazas, B., Tiwari, S., Bobed, C. (eds.) Knowledge Graphs and Semantic Web. pp. 243–261. Springer Nature Switzerland, Cham (2023). https://doi.org/10.1007/978-3-031-47745-4_18
15. Friedman, B., Kahn, P.H., Borning, A.: Value Sensitive Design: Theory and Methods. Tech. rep., University of Washington (2003)
16. Friedman, B., Kahn, P.H., Borning, A., Huldtgren, A.: Value sensitive design and information systems. In: Doorn, N., Schuurbiers, D., van de Poel, I., Gorman, M.E. (eds.) Early engagement and new technologies: Opening up the laboratory, pp. 55–95. Springer Netherlands, Dordrecht (2013). https://doi.org/10.1007/978-94-007-7844-3_4, https://doi.org/10.1007/978-94-007-7844-3_4
17. Green, B., Hu, L.: The Myth in the Methodology: Towards a Recontextualization of Fairness in Machine Learning. Proceedings of the machine learning: the debates workshop **895** (2018)

18. Hardt, M., Price, E., Srebro, N.: Equality of opportunity in supervised learning. In: Proceedings of the 30th International Conference on Neural Information Processing Systems. pp. 3323–3331. NIPS'16, Curran Associates Inc., Red Hook, NY, USA (Dec 2016)
19. Kamiran, F., Calders, T.: Data preprocessing techniques for classification without discrimination. Knowledge and Information Systems **33**(1), 1–33 (Oct 2012). https://doi.org/10.1007/s10115-011-0463-8, http://link.springer.com/10.1007/s10115-011-0463-8
20. Kearns, M., Neel, S., Roth, A., Wu, Z.S.: Preventing Fairness Gerrymandering: Auditing and Learning for Subgroup Fairness. In: Proceedings of the 35th International Conference on Machine Learning. pp. 2564–2572. PMLR (Jul 2018), https://proceedings.mlr.press/v80/kearns18a.html, iSSN: 2640-3498
21. Kendall, M.G.: A New Measure of Rank Correlation. Biometrika **30**(1-2), 81–93 (Jun 1938). https://doi.org/10.1093/biomet/30.1-2.81, https://doi.org/10.1093/biomet/30.1-2.81
22. Klein, L., D'Ignazio, C.: Data Feminism for AI. In: The 2024 ACM Conference on Fairness, Accountability, and Transparency. pp. 100–112. ACM, Rio de Janeiro Brazil (Jun 2024). https://doi.org/10.1145/3630106.3658543, https://dl.acm.org/doi/10.1145/3630106.3658543
23. Manzoor, M.A., Albarri, S., Xian, Z., Meng, Z., Nakov, P., Liang, S.: Multimodality Representation Learning: A Survey on Evolution, Pretraining and Its Applications. ACM Trans. Multimedia Comput. Commun. Appl. **20**(3), 74:1–74:34 (Oct 2023). https://doi.org/10.1145/3617833, https://doi.org/10.1145/3617833
24. Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., Galstyan, A.: A Survey on Bias and Fairness in Machine Learning. ACM Computing Surveys **54**(6), 115:1–115:35 (Jul 2021). https://doi.org/10.1145/3457607, https://doi.org/10.1145/3457607
25. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E.: Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research **12**(85), 2825–2830 (2011), http://jmlr.org/papers/v12/pedregosa11a.html
26. Pessach, D., Shmueli, E.: Algorithmic Fairness. In: Rokach, L., Maimon, O., Shmueli, E. (eds.) Machine Learning for Data Science Handbook: Data Mining and Knowledge Discovery Handbook, pp. 867–886. Springer International Publishing, Cham (2023). https://doi.org/10.1007/978-3-031-24628-9_37, https://doi.org/10.1007/978-3-031-24628-9_37
27. Rawls, J.: Justice as Fairness: Political not Metaphysical. Philosophy & Public Affairs **14**(3), 223–251 (1985), https://www.jstor.org/stable/2265349, publisher: Wiley
28. Saleiro, P., Kuester, B., Hinkson, L., London, J., Stevens, A., Anisfeld, A., Rodolfa, K.T., Ghani, R.: Aequitas: A Bias and Fairness Audit Toolkit (Apr 2019). https://doi.org/10.48550/arXiv.1811.05577, http://arxiv.org/abs/1811.05577, arXiv:1811.05577 [cs]
29. Schäfer, M., Haun, D.B.M., Tomasello, M.: Fair Is Not Fair Everywhere. Psychological Science **26**(8), 1252–1260 (Aug 2015). https://doi.org/10.1177/0956797615586188, https://journals.sagepub.com/doi/10.1177/0956797615586188
30. Scott, J.T., Matland, R.E., Michelbach, P.A., Bornstein, B.H.: Just Deserts: An Experimental Study of Distributive Justice Norms. American Journal

of Political Science **45**(4), 749–767 (2001). https://doi.org/10.2307/2669322, https://www.jstor.org/stable/2669322, publisher: [Midwest Political Science Association, Wiley]
31. Selbst, A.D., Boyd, D., Friedler, S.A., Venkatasubramanian, S., Vertesi, J.: Fairness and Abstraction in Sociotechnical Systems. In: Proceedings of the Conference on Fairness, Accountability, and Transparency. pp. 59–68. ACM, Atlanta GA USA (Jan 2019). https://doi.org/10.1145/3287560.3287598, https://dl.acm.org/doi/10.1145/3287560.3287598
32. Shelby, R., Rismani, S., Henne, K., Moon, A., Rostamzadeh, N., Nicholas, P., Yilla-Akbari, N., Gallegos, J., Smart, A., Garcia, E., Virk, G.: Sociotechnical Harms of Algorithmic Systems: Scoping a Taxonomy for Harm Reduction. In: Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society. pp. 723–741. AIES '23, Association for Computing Machinery, New York, NY, USA (Aug 2023). https://doi.org/10.1145/3600211.3604673, https://dl.acm.org/doi/10.1145/3600211.3604673
33. Umbrello, S., Capasso, M., Balistreri, M., Pirni, A., Merenda, F.: Value Sensitive Design to Achieve the UN SDGs with AI: A Case of Elderly Care Robots. Minds and Machines **31**(3), 395–419 (Sep 2021). https://doi.org/10.1007/s11023-021-09561-y, https://doi.org/10.1007/s11023-021-09561-y
34. Umbrello, S., Van De Poel, I.: Mapping value sensitive design onto AI for social good principles. AI and Ethics **1**(3), 283–296 (Aug 2021). https://doi.org/10.1007/s43681-021-00038-3, https://link.springer.com/10.1007/s43681-021-00038-3
35. Van De Poel, I.: Embedding Values in Artificial Intelligence (AI) Systems. Minds and Machines **30**(3), 385–409 (Sep 2020). https://doi.org/10.1007/s11023-020-09537-4, https://link.springer.com/10.1007/s11023-020-09537-4
36. Weerts, H., Dudík, M., Edgar, R., Jalali, A., Lutz, R., Madaio, M.: Fairlearn: Assessing and Improving Fairness of AI Systems. Journal of Machine Learning Research **24**(257), 1–8 (2023), http://jmlr.org/papers/v24/23-0389.html
37. Zytko, D., J. Wisniewski, P., Guha, S., P. S. Baumer, E., Lee, M.K.: Participatory Design of AI Systems: Opportunities and Challenges Across Diverse Users, Relationships, and Application Domains. In: Extended Abstracts of the 2022 CHI Conference on Human Factors in Computing Systems. pp. 1–4. CHI EA '22, Association for Computing Machinery, New York, NY, USA (Apr 2022). https://doi.org/10.1145/3491101.3516506, https://dl.acm.org/doi/10.1145/3491101.3516506