How Do Algorithmic Fairness Metrics Align with Human Judgement? A Mixed-Initiative System for Contextualized Fairness Assessment

Rareș Constantin, Moritz Dück, Anton Alexandrov, Patrik Matošević, Daphna Keidar, Mennatallah El-Assady

ETH Zürich, Switzerland

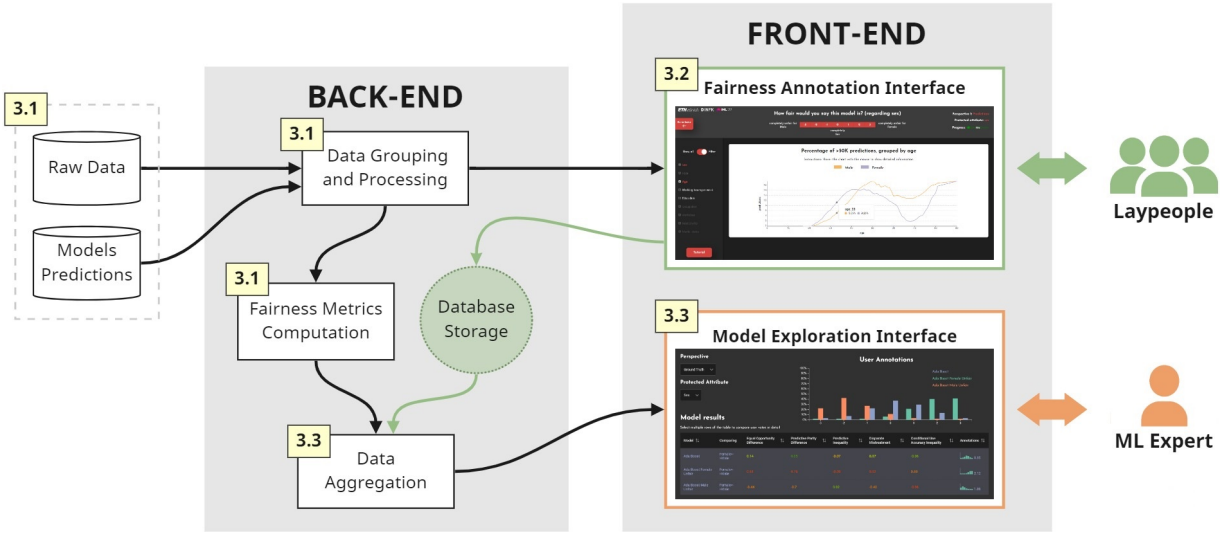


Figure 1: The simplified pipeline of FairAlign, a visual analytics platform for contextualized fairness assessment. The numbers written on the yellow rectangles represent the subsections where the respective components are explained in detail. A workflow example would be as follows: The laypeople sign up, choose an annotation dashboard and start taking decisions regarding algorithmic fairness, based on the provided visualizations of data and model's predictions. After the annotation process is finalized, data scientists and machine learning experts can login and analyze the values of the predefined fairness metrics, along with the aggregated results obtained from the human evaluation.

Abstract

Fairness evaluation presents a challenging problem in machine learning, and is usually restricted to the exploration of various metrics that attempt to quantify algorithmic fairness. However, due to cultural and perceptual biases, such metrics are often not powerful enough to accurately capture what people perceive as fair or unfair. To close the gap between human judgement and automated fairness evaluation, we develop a mixed-initiative system named *FairAlign*, where laypeople assess the fairness of different classification models by analyzing expressive and interactive visualizations of data. Using the aggregated qualitative feedback, data scientists and machine learning experts can examine the similarities and the differences between predefined fairness metrics and human judgement in a contextualized setting. To validate the utility of our system, we conducted a small study on a socially relevant classification task, where six people were asked to assess the fairness of multiple prediction models using the provided visualizations. The results show that our platform is able to give valuable guidance for model evaluation in case of otherwise contradicting and indecisive metrics for algorithmic fairness.

1 Introduction

Fairness represents a major concern in machine learning (ML), especially when classification models are used for automated decision-making in applications where people's lives are directly affected by the models' outcomes [22]. Examples of such applications are automated grading systems in schools [20, 23], loan approval systems [27, 31] or any other situation in which individuals are automatically distributed into groups of different social factors, called protected groups. Hence, using fully automated systems to generate predictions requires a comprehensive understanding of the context in which they are used. Furthermore, it is vital to identify the ethical injustices that could arise with respect to society [6].

Incrementally recognizing its importance for society, much research has started to be done towards the evaluation of algorithmic fairness and ethical decision-making [8, 35]. Throughout the years, over 21 different definitions have been provided for fairness [21]. Following these definitions, various metrics for algorithmic fairness have been proposed [10, 11, 34]. However, these definitions may contradict each other [18], and it is challenging to determine which fairness definition is the appropriate one for a given context.

Schoeffer et al. [26] show that people started to acknowledge the usefulness of automated systems, trusting them even when taking high-stakes decisions. However, the level of trust is reduced for people with low AI literacy. Hence, given that fairness is highly dependent on intricate social nuances, we believe that at least the **the choice of fairness definition should be driven by human feedback.** Furthermore, requiring a rigorous mathematical foundation, the metrics alone do not make use of contextual information and

oversimplify the multifaceted problem of fairness.

Even though there are multiple works on the topic of detecting and mitigating unfairness [1, 3, 30], there is still little work done towards human-centric evaluation of algorithmic fairness. Since manual assessment is expensive and unfeasible, there is an immediate need for finding the fairness metrics that are the most expressive. Srivastava et al. [28] were the first to investigate the correlation between different fairness notions and lay people's perception of fairness. Their findings confirm that the most appropriate metric for algorithmic fairness is highly dependent on the context. Thus, to automate the assessment of algorithmic fairness for a specific task, it is necessary to first investigate people's concept of fairness considering the given setup [36].

To address this issue, we create a mixed-initiative system that allows data scientists and ML experts to analyze human judgements on algorithmic fairness. This approach enables the contextualized assessment of fairness for different classification models, in hope of capturing the nuances of human judgement to the greatest extent. The toolkit we present has two different types of users: laypeople and data scientists. The laypeople have the role of annotating the fairness of prediction models. They are given an interactive visual analytics system which illustrates the classification model's predictions and the data in a discriminative manner, and are asked to specify whether they perceive the model to be fair. The scientists are the stakeholders of our application, which can analyze the human feedback alongside predefined fairness metrics. Therefore, our proposed solution aims to **find how well various algorithmic fairness metrics align with human judgement in a given context**.

In order to obtain valuable human feedback that provides the moralistic guidance we seek, it is vital to choose the annotators carefully, based on their expertise, cultural background and social status. However, we argue that defining precisely which criterion to use for selecting the laypeople is highly dependent on the context in which the system operates, and is therefore out of the scope of this paper.

The main contribution of our work is two-fold: (1) To find which metrics align best with human perception of fairness, we develop an extensible mixed-initiative system for assessing algorithmic fairness in a contextualized setting; and (2) By conducting a pilot study where our platform is utilized by laypeople to determine the fairness of various classification models, we highlight that automatic evaluation alone is not enough to capture the multifaceted problem of fairness.

2 Related Work

Even though the interest in human-centric assessment of algorithmic fairness spiked only recently, there are several tools for analyzing and mitigating unfairness in prediction models. Moreover, various studies have been conducted on the topic of human perception of fairness. However, there is little research towards finding correspondences between metrics for algorithmic fairness and human judgement.

Toolkits for Analyzing Fairness – Throughout the years, multiple toolkits have been suggested for detecting and mitigating unwanted algorithmic bias and (un)fairness. Aequitas, the approach proposed by Udeshi et al. [30], generates bias reports based on the input data and the selected protected attributes. That can be any sensitive attribute that divides the data into groups of different social significance, such as race, sex, gender, or religion. A downside of this toolkit is that it is not allowed to be utilized for commercial purposes. However, IBM's AI Fairness 360 [3] adds the industrial usability factor that Aequitas lacks. It is an open source toolkit that merges together numerous bias metrics and bias mitigation algorithms. Nevertheless, neither of these applications take human perception of fairness into account. Fairsight [1] is one of the first fairness evaluation programs that utilizes a visual analytics system to help ML experts take fair decisions when designing a new model. The only downside is that its considerations are restricted to the judgement of people with high expertise in the domain of machine learning and data science. For a correct and thorough investigation of fairness, laypeople's opinions should have been examined instead. A similar approach for investigating algorithmic fairness is represented by the benchmarks for detecting possible discrimination within prediction models. Themis [2] tries to measure discrimination by automatically generating test suites. Once again, the downside of Themis is that it is highly relying on the predefined notions of fairness. On the other hand, Hendrycks et al. [14] create a dataset specifically designed to calibrate models to follow basic moral judgements in open-world settings. However, its purpose is only limited to the evaluation of language models.

Human Perception of Fairness – With regard to studies on human perception of algorithmic fairness, Harrison et al. [12] provide a comparative, human-centric analysis of fairness. Moreover, it is one of the first research studies that investigate the underlying human biases that might emerge when manually assessing algorithmic fairness. Their paper shows how people differentiate between bias and fairness, and highlights that people do not trust fully-automated fairness assessment. One shortcoming of this study is that the visualizations provided to the people who assess fairness are not expressive enough. To be more exact, they are restricted to comparing models' predictions and accuracy values with respect to the protected attribute only. None of the other relevant attributes were presented to the participants, potentially misleading them towards the acceptance of the predefined notions of fairness. Watts et al. [33] discuss different human biases in decision-making in the context of ethics. By presenting more information, we argue that we might be able to overcome some of these biases, such as *hastened response time* [16] (i.e., making decisions too quickly before fully processing information) or *ignoring unique information* [29] (i.e., focusing on information commonly known instead of information unique to individual members).

Srivastava et al. [28] are the first to investigate the relationships between fairness definitions and people's understanding of fairness. They approached the problem by obtaining comparative feedback from people. The issue is that their visualizations lack the same additional information about underlying attributes. Nevertheless, the results of their study represent a solid starting point in understanding the contextualized perception of humans regarding algorithmic fairness.

Our work is inspired by the study conducted by Srivastava et al. [28] and aims to materialize their findings into a customizable and extensible toolkit for evaluating algorithmic fairness using human feedback, called *FairAlign*. Moreover, utilizing techniques inspired by all the above-mentioned research, the ultimate goal of our project is finding the most suitable and expressive metrics for the given setting.

3 *FairAlign* Visual Analytics Platform

System Overview – Our application is designed to provide data scientists and ML experts with useful insights into people's opinions on algorithmic fairness for a certain task. To achieve this goal, we propose a simple, yet effective end-to-end architecture, as illustrated in Fig. 1. First, the ML experts are asked to configure the application by providing it with a tabular dataset and the predictions of multiple classification models. The input data and the predictions are split according to the values of the selected *protected attribute*. Then, based on this grouping, multiple statistics are computed and presented to the laypeople as interactive visualizations. After an exhaustive analysis, the laypeople have to take a decision regarding the algorithmic fairness, one model at a time. In the end, all the human feedback is aggregated and presented alongside the predefined fairness metrics in a separate dashboard, dedicated for data scientists and ML experts. Investigating the results, they can choose the models which laypeople perceive as fair and can determine which fairness metrics align best with human perception in the given context.

Perspectives – For the fairness assessment of the classification mod-

els, the annotators are provided with two different dashboards for each protected attribute. We refer to these types of dashboards as *perspectives*, and they reflect what information should be taken into consideration when taking decisions regarding fairness. The first one is called the *Predictions* perspective, since all the visualizations present in this type of dashboards consider only the models' predictions. The second perspective also takes the true labels into consideration, along with the predictions. Hence, we call it the *Accuracy* perspective.

Design Rationale – Saha et al. [24] show that people without a background in machine learning have difficulties understanding fairness metrics. Thus, we have designed a user-friendly interface which is transparent with respect to the fairness metrics and provides guidance for understanding the task that needs to be solved. First, the laypeople are provided with an onboarding screen, containing a short description of the required key concepts. Secondly, once the users start interacting with the annotation dashboards, they can opt to go through a tutorial. The tutorial explains the steps they need to follow for the current task while presenting each component in the dashboard. Moreover, each visualization analyzed by the users for algorithmic fairness assessment is accompanied by a description and guidance on how to use it in order to analyze the plotted data efficiently. Finally, we try to keep the lay users motivated using gamification techniques, such as displaying intuitive figures for each dashboard and displaying their current progress.

3.1 Design Space: Data, Models, and Metrics

The meaning of fairness is highly dependent on the given contextual information. Hence, a highly generalized toolkit for algorithmic fairness evaluation would likely fail in capturing intricate nuances of human perception. To obtain meaningful results, our application contains visualizations tailored for a subset of problems. In the following paragraphs, we explain which type of data, prediction models, and fairness metrics are compatible with our platform. In addition, we discuss the configuration of our prototype and the system's extensibility.

Dataset – Our toolkit is compatible with any multidimensional dataset that is designed for classification tasks. However, to obtain meaningful results regarding fairness, the input data should contain socially relevant features to be selected as protected attributes.

The prototype we have developed uses the UCI Adult Dataset [9], which contains US census data for 48,598 people. The task is to predict whether a given person earns more than $50,000 a year. The protected attributes we have selected are *sex* and *race*. While the sex attribute has binary values in this dataset (i.e., Male and Female), the race attribute has five distinct values. For simplicity, we clustered the races as follows: White, representing the privileged group, and Non-White, representing the unprivileged group (composed of Black, Amer-Indian-Eskimo, Asian-Pac-Islander and Other). Furthermore, we have considered selecting the native country as an additional protected attribute, but this was shown to be infeasible due to the highly imbalanced distribution of the countries (about 90% of people were born in the US).

Prediction Models – To create a uniform representation for all classifiers and to save resources, we do not save the whole models into the system. Considering the necessary inputs for computing the fairness metrics, it is sufficient to only store the models' predictions in tabular format. It should be noted that our toolkit is not restricted to binary classification tasks, hence also working with multi-class classification. In the latter case, the predicted labels do not require any post-processing for the Accuracy perspective. However, for the Predictions perspective, they need to be separated into favorable and unfavorable labels.

Fairness Metrics – We decided to restrict the scope of our application to the assessment of group fairness, backed by the work of Reuben Binns [5], who investigates the importance and the conflict between individual and group fairness. His research reveals that fully-automated prediction systems are not compatible with individual fairness and that they should be avoided when there is a need of treating people as individuals. Moreover, when using such classifiers, disparities in accuracy might appear between different groups of people, which can result in an unfair disadvantage for the unprivileged group.

Majumder et al. [19] simplify the task of fairness evaluation by clustering numerous metrics by the similarity of the notions they are trying to evaluate. Following the results presented in their paper, we have selected a small set of metrics for group fairness and integrated them into our system. For the Predictions perspective, we have chosen *statistical parity (SP)* [10], while for the Accuracy perspective, we have chosen *equal opportunity (EO)* [7], *predictive parity (PP)* [7], *predictive equality (PE)* [7], *disparate mistreatment (DM)* [11] and *conditional use accuracy equality (CUAI)* [4]. Depending on the nature of the data and the task itself, a few other metrics might become relevant when assessing fairness. However, thanks to the design of the back-end service, integrating new fairness metrics into our system should be a straightforward process.

3.2 Fairness Annotation Interface

To gather valuable human feedback about algorithmic fairness, we have designed multiple dashboards containing interactive visualizations using the input data and the models' predictions, as shown in Fig. 3. Each visualization provided in the dashboards depicts the disparities in model's decisions between the privileged and the unprivileged group uniquely, taking certain data attributes into consideration.

Each dashboard is designed to provide the annotators with a balanced and multifaceted view of all aspects of the data, without feeling overwhelmed. Hence, we offer laypeople the freedom of choosing which attributes they consider relevant when assessing algorithmic fairness. Our design follows the visual information-seeking mantra by displaying only the simplest visualization at first (i.e., the Outcomes Proportions Chart, displayed in Fig. 3e). The granularity of the information can be later increased by gradually showing more complex and more expressive visualizations. To this scope, we have implemented a filtering panel, where the users can choose to either display all the visualizations at once or filter them based on the considered attributes. Namely, the users can select or deselect any of the attributes they want to consider, except for the current protected attribute with respect to which the decision regarding fairness is being made.

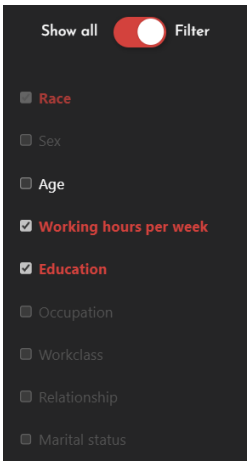
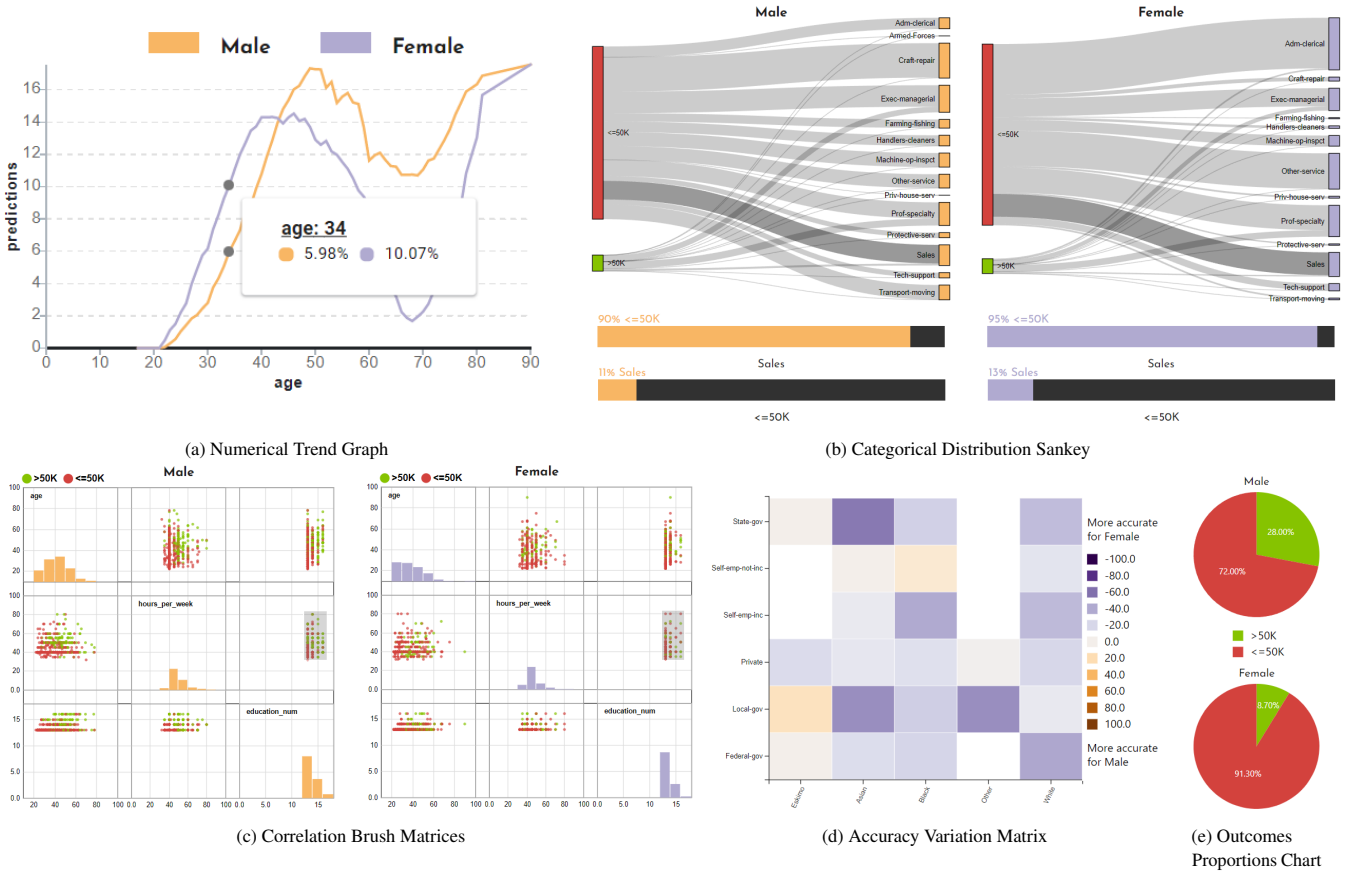


Figure 2: Filtering panel, present in all of our annotation dashboards. Only the visualizations that consider exactly the selected subset of attributes are displayed.

Fig. 3a illustrates one of the most intuitive visualizations we have designed, called Numerical Trend Graph, which is present in all of our annotation dashboards due to its high effectiveness. By selecting an additional numeric attribute from the filtering panel, this visualization displays the percentage of positive predictions or the model accuracy with respect to the values of the attribute. Analyzing the Numerical Trend Graphs, the users can easily discover trends in the model's predictions and can get detailed information by hovering over the plotted lines. To prevent possible sparseness issues, we utilized smoothing techniques such as Savitzky–Golay filtering [25].

After a thorough inspection of the visualizations, laypeople have



(a) Numerical Trend Graph

(b) Categorical Distribution Sankey

(c) Correlation Brush Matrices

(d) Accuracy Variation Matrix

(e) Outcomes Proportions Chart

Figure 3: Examples of visualizations used for algorithmic fairness assessment. Depending on the dashboard's perspective, the visualizations display different statistics about each demographic group in a comparative manner. The Numerical Trend Graph (3a) is a line graph that is used to discover trends in outcomes with respect to numerical features. The Categorical Distribution Sankey (3b) consists of two Sankey diagrams which compare the distribution of outcomes over categorical features. The Correlation Brush Matrices (3c) is used to investigate the correspondences among multiple numerical features. The Accuracy Variation Matrix (3d) displays the variations in model's accuracy, divided in subdomains of two categorical features. The Outcomes Proportions Chart (3e) represents our simplest visualization, which offers an overview of the outcomes.

the task of annotating the fairness of the current model with regard to a certain protected attribute on a 7-point Likert scale. This scale is two-sided and is centered in 0, which represents a completely fair model. Oppositely, -3 and 3 represent totally unfair models for one of the two groups. To complete the fairness assessment task, lay users need to successfully annotate all the models in each existing dashboard. The dashboards can be seen as the Cartesian product of the perspectives and the chosen protected attributes. Since a large number of the visualizations we have created are tailored for a specific perspective, we will further present the utility of the most important visualizations in each perspective.

Predictions Perspective – The visualizations in the Predictions perspective use only the predicted labels, without any ground truth information. This perspective can be associated with the concept of equity, where a completely fair model would have an equal outcome for both groups. However, often the data itself is imbalanced, meaning that complete equity may be associated with decreased model performance. Moreover, most visualizations we have designed provide additional information for each group, which might shift the decisions of laypeople. Thus, the underlying attributes may reveal that a perfectly fair scenario does not always correspond to an equal outcome.

Since the Numerical Trend Graph is limited to numerical features, we have designed another visualization that can be used to investigate categorical features, called the Categorical Distribution Sankey. An instance of this type of visualization is illustrated in Fig. 3b. It contains two Sankey diagrams displayed next to each other, one for each protected group (e.g., Male and Female). Based on the chosen nominal attribute, each of the two diagrams show how the predicted labels distribute over the distinct subgroups spanned by this attribute. Moreover, the users can hover over the connections to get more details on demand. When hovering over a connection between a subgroup and label, they can see what percentage of the subgroup spanned by the attribute value is assigned with the respective label and what percentage of all data points assigned with the label are part of that subgroup. To facilitate the comparison between the protected groups, hovering over one diagram activates the corresponding connection in the neighboring diagram as well.

Laypeople can select more than one additional attribute at once, resulting in the display of even more intricate visualizations. Using Weber's Law, Harrison et al. [13] show that the scatterplot is one of the most effective visualizations for showcasing correlation. Thus, we implemented a visualization that utilizes multiple scatterplots to highlight the correspondences among all the selected attributes. Shown in Fig. 3c, the Correlation Brush Matrices is the Cartesian product representation over multiple numerical attributes. On the diagonal of each matrix, the distributions of data with respect to each considered attribute are displayed as histograms. All the other cells contain scatterplots for the respective pairs of attributes. For comparing protected groups, we plot two such matrices side by side. Moreover, to allow in-depth exploration, the plotted data

can be filtered by drawing a rectangular area on any of the shown scatterplots.

Accuracy Perspective – Unlike the Predictions perspective, the dashboards presenting the Accuracy perspective contain visualizations that take into consideration both the predictions and the real labels, showcasing the models' accuracy across protected groups. This perspective can be associated with the concept of equality, where each group is treated the same, regardless of what outcome this might lead to. In our case, this perspective encourages users to be aware of the real labels, not only the ones predicted by the model. Thus, a perfectly fair model seen from this perspective would treat each protected group the same, in the sense that the protected attribute should not influence the performance of the model.

One complex visualization we use in the dashboards presenting the Accuracy perspective is the Accuracy Variation Matrix, which displays a grid over two different categorical attributes. This visualization comes in two different forms. In the first type of Accuracy Variation Matrix, the color of each grid cell encodes the magnitude of the accuracy for one of the two protected groups. To quickly spot variations in the model's accuracy, the users can repeatedly toggle between protected groups. The second type, shown in Fig. 3d, is a more compact version of the toggling matrix. It encapsulates the information for both protected groups in a single grid by computing the difference of accuracy between the two groups. Furthermore, similar to other integrated visualizations, the users can get detailed information by hovering over the cells of the grid.

3.3 Model Exploration Interface

Once the fairness annotation process is finalized, data scientists and ML experts can use the model exploration interface to investigate the correspondences between pre-defined fairness metrics and human judgement. The main modules of this interface are illustrated in Fig. 4. It can be noticed that the results are grouped by perspective (i.e., Predictions or Accuracy) and protected attribute, hence selecting a combination of the two is required in order to display the fairness scores. After deciding upon a certain configuration for these options, both the fairness metrics and the aggregated human annotations are presented together, in a tabular format.

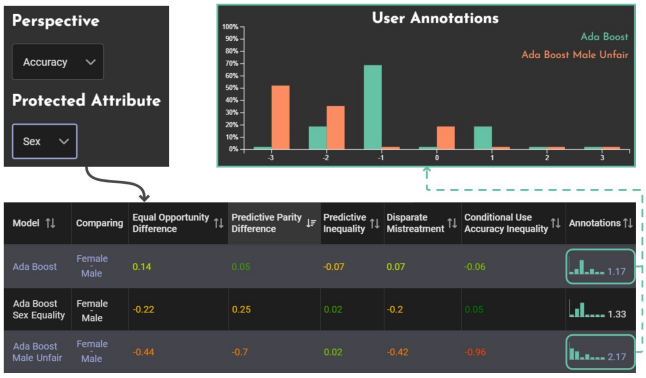


Figure 4: Components of the model exploration interface. Selecting a protected attribute and a perspective provides a table containing information regarding both the automatic and the manual assessment of algorithmic fairness for the classification models. The table's rows, representing the models, can be ordered using numerous criteria and further selected in order to show additional details.

Since the feedback obtained from laypeople is not quantifiable, we have decided to present the results in a comparative manner. To achieve this, the table allows the ascending or descending ordering of the models from the most unfair to the most fair one, based on the values of human annotations or any of the available fairness metrics. By analyzing the extent to which these different rankings agree, the stakeholders of our application can find out which fairness metrics are the most suitable for the given task and social context. They can use this valuable information to improve the practicality of automatic fairness assessment when developing new models.

Most fairness metrics are two-tailed, meaning that they can differentiate between the discrimination of the privileged group and the discrimination of the unprivileged group. Following a similar approach when gathering human feedback for algorithmic fairness raises an issue when trying to aggregate the annotations by averaging the values. Due to the zero-centered scale for human annotations, any model whose resulted annotations follow a symmetrical distribution would be labeled as a perfectly fair model, no matter how large is the variance. To overcome this problem, we aggregated the human annotations using the average of absolute. For consistency, absolute values are also used when sorting the table after a fairness metric. However, in order to avoid losing information, we utilize glyphs to illustrate the distributions of annotations, as highlighted in Fig. 4. Moreover, for an in-depth exploration of human perception of fairness, multiple models of interest can be selected from the table, displaying the annotation results for the respective models in a separate bar chart.

4 Fairness Annotation Study

To validate the usefulness of *FairAlign*, we conducted a small study, where six people were asked to assess the fairness of eight different classification models. In this section, we describe the setup for this study, as well as how we measured the alignment between human judgement and the algorithmic fairness metrics. Nonetheless, we present the obtained results, followed by a short discussion.

4.1 Study Design

To set up the visual analytics platform for this study, we utilized the configuration of our prototype, explained in subsection 3.1. To capture a wide spectrum of fairness levels, we chose the best performing model from our prototype in terms of accuracy (i.e., Ada Boost) and designed new prediction models based on the baseline model. These models try to tackle the problem of fairness in distinctive ways, such as balancing the training samples across demographic groups or ignoring the features that count as protected attributes during training.

The participants of this study were carefully selected, such that their descriptions fit the one of our conceptual users. Thus, even though some of the participants have a Computer Science background, we made sure that each selected person has very little or no AI expertise. Furthermore, all six participants identified as white for race and male for sex. This information supports the utility of *FairAlign* in terms of contextualized fairness assessment, since all the annotators are part of a specific demographic group.

At the beginning of the study, the participants were required to carefully follow the instructions provided by the platform to guide them through the data exploration and decision-making process regarding algorithmic fairness. Moreover, we offered them the possibility of asking for additional guidance when needed. Since we need to analyze the unbiased decisions of people regarding the fairness of models, our guidance did not manipulate their decisions or interfere with their opinions in any way. At the end of the annotation session, the participants spent about 10 minutes talking about their experience, how they used and interpreted each visualization in order to help them form a decision and what difficulties they encountered along the way.

4.2 Study Results

Based on the received feedback about our platform, we have observed a few generic patterns. In general, the annotators tended to understand rather quickly the basic statistical notions, such as the overall accuracy or the percentage of positive predictions for each

protected group, represented by the Outcomes Proportions Chart or the percentage bars found at the bottom of the Categorical Distribution Sankey visualization. Contrarily, they spent significantly more time understanding the functionality of the interactive actions of the more complex visualizations, such as the Correlation Brush Matrices. Since the scope of this particular visualization is to find correlations between numerical attributes, it naturally takes longer to analyze. As expected, the Numerical Trend Graph was labeled as one of the most intuitive visualizations, since the participants found it easy to observe trends within ordered numerical values. Even though the Accuracy Variation Matrix provides great detail for comparison between the elements of a cross-product between two attributes, the participants mostly looked for a prominent difference in the color intensity between the two groups and rarely looked for a deeper meaning of smaller areas. Nevertheless, one exception from this phenomenon happens for the education attribute, since the participants were particularly interested whether more educated people were "rightfully treated" by the models. In the end, one aspect that all participants agreed on was that the explanations and tutorials are absolutely necessary in order to get a comprehensive understanding of the task.

In order to see which fairness metrics align best with human judgement, we consider the ordering of models from most fair to least fair, induced by laypeople's annotations and fairness metrics, respectively. Using the ranking of a single annotator and the ranking of a fairness metric, we use the Tau-b Kendall rank correlation coefficient [17] to quantify how well these rankings agree. The range of the correlation measure is $\tau_B \in [-1,1]$, where $\tau_B = 1$ in case of perfect agreement between the rankings, $\tau_B = -1$ in case of inverse rankings and $\tau_B = 0$ indicating the absence of any association. Averaging these coefficients over all users tells us how well a metric aligns with the judgements of all annotators. The results for these evaluations are shown in Table 1. The fairness metric that seems to align best with human judgement is statistical parity (SP), which confirms earlier findings of Srivastava et al. [28]. The results also highlight the difficulty of analyzing the fairness metrics without any human reference, as they strongly disagree with each other regarding the ranking of models.

	race		sex	
fairness metric	**avg**	**std**	**avg**	**std**
SP	**0.714**	0.272	**0.432**	**0.187**
EO	-0.025	0.370	0.150	0.355
PP	0.198	0.244	-0.195	0.226
PE	-0.057	0.500	-0.304	0.367
DM	-0.004	**0.193**	-0.109	0.289
CUAI	0.198	0.244	-0.195	0.226

Table 1: Scores representing the level of agreement between the rankings of models' fairness induced by laypeople annotations and automated fairness metrics using the Kendall Tau-B rank correlation coefficient [17]. The table is split into two protected attributes, for both of which annotations and metrics were collected separately.

5 Discussion and Future Work

The correlation scores resulted from our small-scale study show that most fairness metrics do not agree with each other and people's judgement usually correlates with the simplest notions of fairness, such as statistical parity. Moreover, the dissimilarity of results between different protected attributes highlights the fact that human perception of fairness is highly dependent on the provided contextual information. Additionally, the high values for variance might indicate that people have different outlooks regarding algorithmic fairness, even when they are part of the same demographic group. In the next paragraphs, we will underline the lessons learned after this research, as well as potential future work.

Lessons Learned – In a previous research, Kazim et al. [15] stated that "automation of fairness is in itself inherently unfair." Our toolkit illustrates the conflicting nature of different fairness metrics. Since each metric measures algorithmic fairness based on a different aspect, choosing a single metric without any validation of alignment with human perception of fairness is an inconsiderate and unfair decision.

Valentine et al. [32] studied the fairness of human judgement in the context of health professions. However, along studies about human judgement in fairness, their findings can be used as a prism to understand why, when treated as an isolate entity, human judgement cannot be always fair when assessing algorithmic fairness. This happens due to subjective nature, which is influenced by the person's social and cultural background. Thus, a contextualized assessment method, such as *FairAlign*, is required in order to correctly evaluate algorithmic fairness.

Our observations reflect that a conflict exists between the effectiveness and the expressiveness of a visualization system. We have noticed that people with low AI literacy have more trust in simple explanations than in more complicated ones. Even though the former are easier to understand, ignoring the underlying aspects of data can narrow the "field of view" of their judgement considerably. On the other hand, a more complex visualization may take significantly more time to comprehend, but it offers valuable information which might completely shift someone's perception of fairness in a particular scenario. Thus, it is vital that laypeople are instructed to take the necessary time to fully understand the complex visualizations.

Future Work – Since the number of participants that partook in our study is too low to draw any conclusions, a well-structured case study with a statistically relevant number of participants (e.g. over 200) should be conducted. Furthermore, an interesting research direction that our mixed-initiative platform facilitates is whether the fairness perception of people from different demographic groups align with the same metrics or not. The participants of our pilot study consisted a single demographic group, hence conducting a follow-up study should not be challenging.

Regarding further extensibility of our application, we believe that adapting the visualizations to also accept other type of data would greatly broaden the usability of the application. A feasible beginning would be extending the visual system for temporal and relational data. Later, visualizations for more complex data types such as images, videos, and textual data can be added. Furthermore, the platform can be extended to other tasks besides classification, such as regression, detection, or segmentation.

6 Conclusion

In the context of automatic prediction systems, fairness is an elaborated concept that does not have a universal definition. Nevertheless, the recent drive towards ethical AI development forces us to find innovative ways of tackling the difficult problem of fairness. Although a multitude of techniques were proposed for measuring algorithmic fairness, there is currently an unavoidable gap between the mathematically defined metrics and human perception of fairness. With the goal of bringing together automatic and manual fairness assessment, we develop *FairAlign*, a mixed-initiative system that helps data scientists and ML experts analyze the fairness of discriminative models using human evaluation. We acknowledge that the collection of such feedback is costly, but also necessary for the completeness of algorithmic fairness assessment. Thus, once enough human feedback is gathered, our platform can be used to analyze the correlation between human judgement and fairness metrics. The personalized findings can be later used with the scope of shifting to automatic assessment when testing new prediction models in terms of fairness, by discovering which metrics are the most suitable for the respective task and the given contextual information.

REFERENCES

[1] Y. Ahn and Y.-R. Lin. FairSight: Visual analytics for fairness in decision making. *IEEE Transactions on Visualization and Computer Graphics*, pp. 1–1, Aug. 2019. doi: 10.1109/tvcg.2019.2934262

[2] R. Angell, B. Johnson, Y. Brun, and A. Meliou. Themis: Automatically testing software for discrimination. ESEC/FSE 2018, p. 871–875. Association for Computing Machinery, New York, NY, USA, Oct. 2018. doi: 10.1145/3236024.3264590

[3] R. K. E. Bellamy, K. Dey, M. Hind, S. C. Hoffman, S. Houde, K. Kannan, P. Lohia, J. Martino, S. Mehta, A. Mojsilovic, S. Nagar, K. N. Ramamurthy, J. Richards, D. Saha, P. Sattigeri, M. Singh, K. R. Varshney, and Y. Zhang. Ai fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias, Oct. 2018. doi: 10.48550/ARXIV.1810.01943

[4] R. Berk, H. Heidari, S. Jabbari, M. Kearns, and A. Roth. Fairness in criminal justice risk assessments: The state of the art, Mar. 2017. doi: 10.48550/ARXIV.1703.09207

[5] R. Binns. On the apparent conflict between individual and group fairness, Dec. 2019. doi: 10.48550/ARXIV.1912.06883

[6] S. Caton and C. Haas. Fairness in machine learning: A survey, Oct. 2020. doi: 10.48550/ARXIV.2010.04053

[7] A. Chouldechova. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments, Feb. 2017. doi: 10.48550/ARXIV.1703.00056

[8] S. Corbett-Davies, E. Pierson, A. Feller, S. Goel, and A. Huq. Algorithmic decision making and the cost of fairness. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '17, p. 797–806. Association for Computing Machinery, New York, NY, USA, Aug. 2017. doi: 10.1145/3097983.3098095

[9] D. Dheeru and E. Karra Taniskidou. UCI machine learning repository, 2017.

[10] C. Dwork, M. Hardt, T. Pitassi, O. Reingold, and R. Zemel. Fairness through awareness, Apr. 2011. doi: 10.48550/ARXIV.1104.3913

[11] M. Hardt, E. Price, E. Price, and N. Srebro. Equality of opportunity in supervised learning. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, eds., *Advances in Neural Information Processing Systems*, vol. 29. Curran Associates, Inc., Oct. 2016.

[12] G. Harrison, J. Hanson, C. Jacinto, J. Ramirez, and B. Ur. An empirical study on the perceived fairness of realistic, imperfect machine learning models. FAT* '20, p. 392–402. Association for Computing Machinery, New York, NY, USA, Jan. 2020. doi: 10.1145/3351095.3372831

[13] L. Harrison, F. Yang, S. Franconeri, and R. Chang. Ranking visualizations of correlation using weber's law. *IEEE Transactions on Visualization and Computer Graphics*, 20(12):1943–1952, Nov. 2014. doi: 10.1109/TVCG.2014.2346979

[14] D. Hendrycks, C. Burns, S. Basart, A. Critch, J. Li, D. Song, and J. Steinhardt. Aligning AI with shared human values. *CoRR*, abs/2008.02275, aug 2020.

[15] E. Kazim, J. Barnett, and A. Koshiyama. Automation and fairness: Assessing the automation of fairness in cases of reasonable pluralism and considering the blackbox of human judgment. *SSRN Electronic Journal*, Jan. 2020. doi: 10.2139/ssrn.3698404

[16] G. Keinan. Decision making under stress: scanning of alternatives under controllable and uncontrollable threats. *Journal of personality and social psychology*, 52(3):639, 1987.

[17] M. G. Kendall. A new measure of rank correlation. *Biometrika*, 30(1/2):81–93, June 1938.

[18] J. Kleinberg, S. Mullainathan, and M. Raghavan. Inherent trade-offs in the fair determination of risk scores, Sept. 2016. doi: 10.48550/ARXIV.1609.05807

[19] S. Majumder, J. Chakraborty, G. R. Bai, K. T. Stolee, and T. Menzies. Fair enough: Searching for sufficient measures of fairness, Oct. 2021. doi: 10.48550/ARXIV.2110.13029

[20] K. Matthews, T. Janicki, L. He, and L. Patterson. Implementation of an automated grading system with an adaptive learning component to affect student feedback and response time. *Journal of Information Systems*, 23:71–84, June 2012.

[21] A. Narayanan. Translation tutorial: 21 fairness definitions and their politics. in conference on fairness, accountability, and transparency, Feb. 2015.

[22] D. Pessach and E. Shmueli. A review on fairness in machine learning. 55(3), Feb. 2022. doi: 10.1145/3494672

[23] V. V. Ramalingam, A. Pandian, P. Chetry, and H. Nigam. Automated essay grading using machine learning algorithm. *Journal of Physics: Conference Series*, 1000:012030, Apr. 2018. doi: 10.1088/1742-6596/1000/1/012030

[24] D. Saha, C. Schumann, D. C. McElfresh, J. P. Dickerson, M. L. Mazurek, and M. C. Tschantz. Human comprehension of fairness in machine learning. *CoRR*, abs/2001.00089, jan 2020. doi: 10.48550/ARXIV.2001.00089

[25] A. Savitzky and M. J. E. Golay. Smoothing and differentiation of data by simplified least squares procedures. *Analytical Chemistry*, 36(8):1627–1639, July 1964. doi: 10.1021/ac60214a047

[26] J. Schoeffer, Y. Machowski, and N. Kuehl. Perceptions of fairness and trustworthiness based on explanations in human vs. automated decision-making, Sept. 2021. doi: 10.48550/ARXIV.2109.05792

[27] M. A. Sheikh, A. K. Goel, and T. Kumar. An approach for prediction of loan approval using machine learning algorithm. In *2020 International Conference on Electronics and Sustainable Communication Systems (ICESC)*, pp. 490–494, July 2020. doi: 10.1109/ICESC48915.2020.9155614

[28] M. Srivastava, H. Heidari, and A. Krause. Mathematical notions vs. human perception of fairness: A descriptive approach to fairness for machine learning. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery Data Mining*, KDD '19, p. 2459–2468. Association for Computing Machinery, New York, NY, USA, July 2019. doi: 10.1145/3292500.3330664

[29] G. Stasser. Information salience and the discovery of hidden profiles by decision-making groups: A "thought experiment". *Organizational behavior and human decision processes*, 52(1):156–181, 1992.

[30] S. Udeshi, P. Arora, and S. Chattopadhyay. *Automated Directed Fairness Testing*, p. 98–108. Association for Computing Machinery, New York, NY, USA, July 2018.

[31] A. Vaidya. Predictive and probabilistic approach using logistic regression: Application to prediction of loan approval. In *2017 8th International Conference on Computing, Communication and Networking Technologies (ICCCNT)*, pp. 1–6, July 2017. doi: 10.1109/ICCCNT.2017.8203946

[32] N. Valentine, S. Durning, E. Shanahan, and L. Schuwirth. Fairness in human judgement in assessment: a hermeneutic literature review and conceptual framework. *Advances in Health Sciences Education*, 26:1–26, May 2021. doi: 10.1007/s10459-020-10002-1

[33] L. L. Watts, K. E. Medeiros, T. J. McIntosh, and T. J. Mulhearn. Decision biases in the context of ethics: Initial scale development and validation. *Personality and Individual Differences*, 153:109609, 2020.

[34] M. B. Zafar, I. Valera, M. Gomez Rodriguez, and K. P. Gummadi. Fairness beyond disparate treatment disparate impact: Learning classification without disparate mistreatment. In *Proceedings of the 26th International Conference on World Wide Web*, WWW '17, p. 1171–1180. International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, CHE, Apr. 2017. doi: 10.1145/3038912.3052660

[35] J. Zhang and E. Bareinboim. Fairness in decision-making — the causal explanation formula. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1), Apr. 2018. doi: 10.1609/aaai.v32i1.11564

[36] J. Zhou, F. Chen, and A. Holzinger. *Towards Explainability for AI Fairness*, pp. 375–386. Jan. 2022. doi: 10.1007/978-3-031-04083-2_18