Tina Behzad, Mithilesh Kumar Singh, Anthony J. Ripa, and Klaus Mueller. 2025. FairPlay: A Collaborative Approach to Mitigate Bias in Datasets for Improved AI Fairness. Proc. ACM Hum.-Comput. Interact. 9, 2, Article CSCW084 (May 2025), 30 pages. https://doi.org/10.1145/3710982

FairPlay: A Collaborative Approach to Mitigate Bias in Datasets for Improved AI Fairness

TINA BEHZAD*, Department of Computer Science, Stony Brook University, USA
MITHILESH KUMAR SINGH*, Department of Computer Science, Stony Brook University, USA
ANTHONY J. RIPA*, Department of Computer Science, Stony Brook University, USA
KLAUS MUELLER, Department of Computer Science, Stony Brook University, USA

The issue of fairness in decision-making is a critical one, especially given the variety of stakeholder demands for differing and mutually incompatible versions of fairness. Adopting a strategic interaction of perspectives provides an alternative to enforcing a singular standard of fairness. We present a web-based software application, FairPlay, that enables multiple stakeholders to debias datasets collaboratively. With FairPlay, users can negotiate and arrive at a mutually acceptable outcome without a universally agreed-upon theory of fairness. In the absence of such a tool, reaching a consensus would be highly challenging due to the lack of a systematic negotiation process and the inability to modify and observe changes. We have conducted user studies that demonstrate the success of FairPlay, as users could reach a consensus within about five rounds of gameplay, illustrating the application's potential for enhancing fairness in AI systems.

CCS Concepts: • **Human-centered computing → Collaborative and social computing**.

Additional Key Words and Phrases: datasets, causal networks, fairness, bias

ACM Reference Format:
Tina Behzad, Mithilesh Kumar Singh, Anthony J. Ripa, and Klaus Mueller. 2025. FairPlay: A Collaborative Approach to Mitigate Bias in Datasets for Improved AI Fairness. *Proc. ACM Hum.-Comput. Interact.* 9, 2, Article CSCW084 (April 2025), 30 pages. https://doi.org/10.1145/3710982

1 Introduction

Fairness remains an elusive goal in our increasingly data-driven world, hindered by the Impossibility of Fairness [21]. This paradox emerges from the diversity of ideological beliefs surrounding the concept of fairness, creating a scenario where achieving a universally agreed-upon definition becomes unfeasible.

Although the literature has defined a myriad of notions to quantify fairness, each measures and emphasizes different aspects of what can be considered "fair". Many are difficult/impossible to combine [30][14], but ultimately, we must keep in mind (as noted in [13]) there is no universal means to measure fairness, and at present no clear guideline(s) on which measures are "best" [11].

[*] Authors contributed equally to this research.

Authors' Contact Information: Tina Behzad, tbehzad@cs.stonybrook.edu, Department of Computer Science, Stony Brook University, Stony Brook, New York, USA; Mithilesh Kumar Singh, mkssingh@cs.stonybrook.edu, Department of Computer Science, Stony Brook University, Stony Brook, New York, USA; Anthony J. Ripa, aripa@cs.stonybrook.edu, Department of Computer Science, Stony Brook University, Stony Brook, New York, USA; Klaus Mueller, mueller@cs.stonybrook.edu, Department of Computer Science, Stony Brook University, Stony Brook, New York, USA.

ACM 2573-0142/2025/4-ARTCSCW084
https://doi.org/10.1145/3710982

This problem's essence is deeply rooted in context-specific nuances, making it crucial to tailor solutions to the individual characteristics and challenges of each case. Consequently, it becomes vital for human experts to define what constitutes fairness in each distinct scenario. As the range of situations where models are deployed for decision-making expands, so does the necessity for a diverse group of people to scrutinize these models for fairness. To facilitate this, a variety of interfaces have been created, enabling experts across disciplines to assess different fairness metrics and determine the best strategies for mitigating bias in datasets or models [42]. These tools are designed to empower those with in-depth knowledge in their respective fields to define and implement fairness in their models. However, a notable gap in these tools is the lack of a collaborative approach in the bias mitigation activities.

Our approach is rooted in a more practical and collaborative method, inspired by the practice of negotiation for consensus building. We acknowledge and utilize ideological diversity as a strength, channeling it to bring together various stakeholders to collectively define fairness for their specific task. We build our software on the foundation of a previously published web-based software, D-BIAS [23]. D-BIAS is a visual and interactive human-in-the-loop method designed for the pre-processing phase of debiasing algorithmic decision systems (ADS) by ways of a causal model initially derived from the original (potentially biased) training data. Unlike traditional metric-driven methods, D-BIAS provides a detailed view of variable interactions and their impact on the outcome, enabling users to make modifications.

Our system features an enhanced web interface that shifts D-BIAS from a single-user mode to a multi-user framework. Here, professionals from different fields or different stakeholders work together to identify the most fitting and fair causal structure for their specific domain scenario, promoting a consensus-based methodology. Then, once consensus has been reached, the debiased data generated by the causal model can be used to train any ADS. This evolution leads to a more dynamic and inclusive strategy for achieving fairness, where all stakeholders or experts can collaboratively arrive at a mutually satisfactory resolution.

To evaluate our system, we conducted user studies with four different groups, analyzing the effectiveness of our method in reaching consensus on a hiring dataset. The studies reveal that users typically agree to end the game after about five rounds of gameplay, indicating collective agreement on the final causal model.

Our research contributions are:

- The first collaborative methodology and tool for the debiasing of ADS training data, to the best of our knowledge.
- Evolving D-BIAS into a multi-user tool for domain stakeholders to reach consensus.
- Expanding the D-BIAS interactive visual interface by a set of new visual feedback widgets, specifically designed to help stakeholders track their progress more effectively.
- Executing a user study with four groups, each comprising five stakeholders, centered around a hiring ADS scenario.
- Analyzing observed user behavior patterns from the study.
- Evaluating the gathered debiasing outcomes using standard metrics.

In this paper, we present our structured collaborative method in the form of a game, and throughout the text, we use the terms 'users', 'players', and 'stakeholders' interchangeably.

2 Related Work

In this section, we review existing approaches to fairness in machine learning, discuss interactive approaches for addressing bias and fairness, explore consensus-building mechanisms in related domains, and highlight the importance of visualizing these techniques to enable broader understanding and adoption.

2.1 Fairness in Machine Learning

Challenges in Defining Fairness Metrics. Different fairness metrics and definitions have been developed to quantify and measure bias in machine learning models [11][40]. Commonly used fairness metrics include demographic or statistical parity [17][31][7], equal opportunity [24] [1], and equalized odds [17] [24]. These metrics provide quantitative measures to assess the fairness of decisions made by the models across different groups. Different notions and measures can be mutually incompatible and entail unavoidable tradeoffs [32][20]. There is no consensus on a single most appropriate definition of fairness [22]. Determining the right measure to be used must take into account the proper legal, ethical, and social context [40]. For a given application in a given context, algorithms can not be expected to determine the most appropriate definition of fairness and decide a desirable tradeoff between different metrics that is acceptable to all stakeholders. On the other hand, a human trusted by the majority of stakeholders can make an informed decision when presented with the required information [46]. Hence, introducing a human in the loop can improve perceived fairness. As for the aspect of trust, people are more likely to trust a system if they can tinker with it, even if this means making it perform imperfectly [16].

Bias Mitigation Visualization Tools. Understanding and interpreting these fairness approaches might be challenging, especially for non-experts or individuals without a strong technical background such as the stakeholders in a given task. Therefore, in recent years, efforts to visualize and explain these techniques have been developed [42]. Some of these methods include: Silva [55], FairVis [9], FairRankVis [54], DiscriLens [52], FairSight [2], What-If toolkit (WIT) [53], Aequitas[43], AI Fairness 360 (AIF360) [4] and D-BIAS [23]. These tools are crucial to enable a broader audience to understand and engage with fairness in machine learning. Most of these tools focus on bias identification. Some of them, such as FairSight and AIF360, also permit debiasing. D-BIAS, which this paper is built upon, is similar to Silva which also features a graphical causal model in its interface. Silva's empirical study showed that users can interpret causal networks and found them helpful in identifying algorithmic bias [55]. However, like most other visual tools, Silva is limited to bias identification. D-BIAS presents a tool that supports both bias identification and mitigation using a graphical causal model.

Approaches to Achieving Fairness. Various approaches have been proposed to achieve fairness in machine learning.[11][40] Pre-processing methods focus on modifying the training data to remove bias before training the model [10][28]. In-processing methods aim to modify the learning algorithm or objective function to directly optimize for fairness [51]. Post-processing methods modify the model's predictions after training to achieve fairness [29]. Research has shown that teams typically look to their training datasets, not their ML models, as the most important place to intervene to improve fairness in their products [26][42]. D-BIAS and hence our work relates closely with the pre-processing stage where we make changes to the output label based on users' decisions.

2.2 Consensus-Building Mechanisms

Defining Consensus. In the literature, two types of consenses are defined [57]. Several researchers define consensus as the full and unanimous agreement of all the decision-makers regarding all the feasible alternatives [5]. However, unanimity may be difficult to achieve, in particular with large and diversified groups of decision-makers as is the case in real-world settings. In contrast, the concept of consensus has also been considered in a more flexible way with regard to its measurement, which has led to the proposal and use of "soft" consensus degrees [12] with the aim of achieving two important goals: (i) to reflect better partial agreement; and (ii) to guide the Consensus Reaching Process (CRP) until an acceptable high level of agreement is achieved among decision-makers [57].

Challenges in Reaching Consensus. Consensus-building mechanisms have been extensively studied in fields such as multi-agent systems [34], social choice theory [36], and deliberative decision-making [49] to address challenges. The most important challenge is the aim to reach an agreement or consensus among multiple stakeholders with diverse preferences and perspectives. Deliberative decision-making frameworks involve structured dialogue and information sharing among stakeholders to collectively arrive at decisions. Innes argues a number of conditions need to hold for a process to be labeled consensus building [27]. If these do not hold, failure of various kinds is likely. Among these conditions are: including a full range of stakeholders, meaningfulness of the task to all participants, mutual understanding of interests, a dialogue where all are heard and respected, a self-organizing process, and accessible information. To aid in reaching the aforementioned conditions, visualizations have been extensively used to provide a graphical representation of the deliberative process, illustrating arguments, preferences, and their evolution over time [59][19][18].

Incorporating Diverse Perspectives on Fairness. The lack of tools focusing on a collaborative approach in fairness visualization is a significant gap, particularly considering the critical role collaboration plays in this field [25]. Fairness is a concept that varies greatly depending on the context and individual perspectives [45]; a single person or a non-interactive tool might overlook these nuances. Therefore, the use of collaborative visualization tools is essential, as they can amalgamate a range of viewpoints. This integration leads to a more holistic comprehension and implementation of fairness, tailored to the specific requirements of any given task.

Gamification and Cooperative Strategies. Integrating cooperative elements and gamification, as demonstrated in the FairPlay game discussed in this paper, drives active participation and collective decision-making [25]. These strategies do more than just engage individuals; they create an environment where the pursuit of fairness becomes a shared goal and players recognize that others understand their positions. According to Scheff [44], these two elements are essential for reaching consensus. In the absence of tools that focus on collaborative approaches, there is a missed opportunity for richer, more inclusive discussions and solutions around bias mitigation. Such tools could promote a deeper understanding and more effective implementation of fairness in machine learning systems by leveraging the collective intelligence and insights of a wider group of stakeholders [7].

3 FairPlay Game Design

The transformation of the D-BIAS platform into FairPlay, a collaborative environment, involved a design process aimed at creating an engaging experience. This redesign aligns with the guidelines set forth in the fairness toolkit rubric [42] and incorporates elements identified as crucial for successful consensus-building in related work [27]. The game mechanics and user interface were thoughtfully developed to emphasize interactivity and gamification, creating a cooperative environment where players can actively engage in modifying the causal graph. A key focus was on developing an intuitive interface that balances complexity with usability, ensuring that even non-technical users can participate. This approach was chosen to make the system accessible to diverse groups, empowering all players to make informed decisions, regardless of their technical expertise. The structure of the game encourages a self-organizing approach, ensuring that each stakeholder is heard and their metrics and evaluations are accessible to all, promoting mutual understanding of interests.

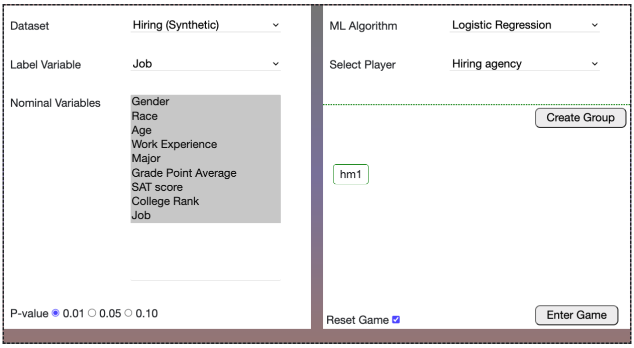
In a real-world scenario, FairPlay can be used by up to five key stakeholders involved in any decision-making process. We chose this number as it provided a good compromise between diversity and manageability; a smaller or larger number of players is also possible. The idea is that each stakeholder brings unique perspectives and concerns, making participation crucial for achieving a fair and unbiased decision. For example in urban planning, five stakeholders could include representatives from city planners, community members, business owners, environmental groups and transportation authorities. In healthcare policy development, stakeholders could be healthcare providers (doctors, nurses, etc.), healthcare administrators, patients, insurance companies, and government regulators.

We demonstrate FairPlay using a hiring task for a programming position, with stakeholders including the hiring agency, the employer, the manager, coworkers, and the union representative. The hiring agency is responsible for matching qualified candidates with job vacancies across various companies. The employer sets the overall hiring policies and goals for the organization. The manager is directly involved in evaluating and selecting candidates, considering factors such as team dynamics and job requirements. Coworkers provide insights into the day-to-day impact of hiring decisions on the work environment and company culture. Finally, the union representative advocates for fair hiring practices and protects the interests of workers. By bringing these five stakeholders together in a collaborative debiasing process, FairPlay ensures that the resulting dataset reflects a balanced and inclusive approach to hiring, taking into account the diverse viewpoints and priorities of each group. It is important to note that while different stakeholders may hold varying levels of influence or power in different scenarios, FairPlay treats all stakeholders as equals, valuing each participant's preferences equally.

The configuration page of the game adheres to the rubric's recommendation of being "Applicable to a diverse range of predictive tasks". The inclusion of visualizations, like causal network diagrams and various charts, aligns with the rubric's focus on providing both a comprehensive and nuanced perspective on fairness and supporting intersectional analysis. This section goes into the specific design features of FairPlay, highlighting its differences from D-BIAS and the rationale behind these changes. A summary of these differences is available in Table 1.

Features	D-Bias	FairPlay
Multi-player	x	✓
Edge modification	✓	✓
Adding/Deleting/Reversing edges	✓	x
Utility and Bias metrics plots	✓	x
Modification History	x	✓
Game metrics, scores and charts	x	✓

Table 1. Comparison of Features between D-Bias and FairPlay



Dataset: Hiring (Synthetic)

Label Variable: Job

Nominal Variables: Gender, Race, Age, Work Experience, Major, Grade Point Average, SAT score, College Rank, Job

P-value ● 0.01 ○ 0.05 ○ 0.10

ML Algorithm: Logistic Regression

Select Player: Hiring agency

Reset Game ☑

Fig. 1. FairPlay: Game Configuration. This is the main configuration panel of the application. For the results presented here, the groups were already pre-configured to make the played games comparable.

3.1 Game Configuration

Before entering the game, the configuration page allows players to select the specific dataset they want to work on, choose the machine learning algorithm to be applied, and express their preferences for certain population groups. as illustrated in Figure 1.

As previously noted, FairPlay primarily concentrates on pre-processing methods, targeting the dataset preparation stage before it is utilized in any training process. The configuration page enables users to select their desired dataset. In this study, a Hiring dataset was used, but the approach is readily applicable to other tabular datasets. The 'Hiring' dataset is a synthetic dataset that was originally introduced in D-BIAS [23], designed to mimic a typical hiring scenario for controlled experimentation and analysis. It consists of 4,000 entries, each representing a fictional job candidate, with key data entry fields including age, gender, race, work experience, Grade Point Average (GPA), SAT score, college rank, major, and a binary hire decision variable. We chose this particular synthetic dataset to seamlessly extend the D-BIAS system from single-player to multiplayer. It was well vetted and many experiments had been done with it. Once the dataset is selected, its features are displayed for the user and a label (outcome) variable is chosen. The initial causal network, represented as a Directed Acyclic Graph (DAG), is inferred using the PC algorithm, a causal discovery method named after its creators Peter Spirtes and Clark Glymour [47]. This algorithm uses a p-value which is a threshold for statistical significance. More details on how the DAG is derived and the p-value's relevance is described later in section 4.1. The default p-value is set to 0.01, ensuring a high level of confidence in the results for most scenarios. Users can proceed with this default setting without needing a deep understanding of p-values. For those with specific requirements, the system allows adjustments to the p-value, assuming users understand the implications of such changes.

On the configuration page, users also have the option to pick from common ML algorithms like Logistic Regression, SVM, Naive Bayes, kNN, Decision Tree, or Neural Network. The chosen algorithm plays a role in continuously monitoring and logging classification performance metrics, providing valuable insights into how data debiasing impacts the model. The purpose of these ML algorithms is to track standard performance metrics, which are computed at each stage of the data debiasing process. This approach allows us to measure the impact of debiasing on classification performance throughout the process [1].

Players can select their role, a feature not available in D-BIAS, a single-user platform. Additionally, players specify a population group based on attributes they consider important. By clicking on the

[1] Technical details of the ML models, their implementation, and data usage can be found in Appendix A.

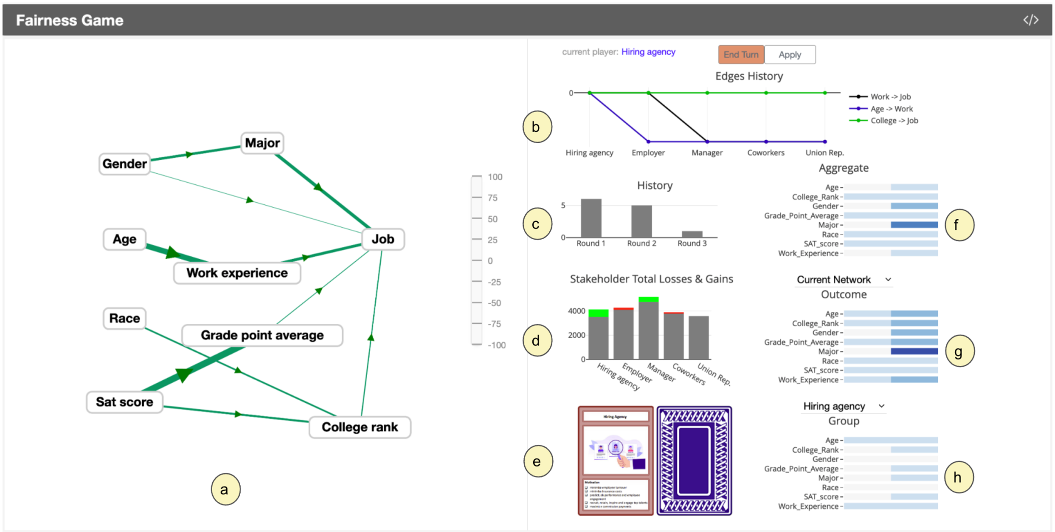


Fig. 2. FairPlay Game Interface. The components are (a) causal network link editor, (b) edge history chart, (c) aggregate edge history chart, (d) stakeholder total loss and gain chart, (e) active stakeholder card stack, (f) aggregate attribute disparity chart, (g) attribute outcome chart, (h) stakeholder attribute priority chart.

'Create Group' button, a pop-up page appears for selecting features and their respective values for user preferences selection or group creation (see figure 7). This involves specifying preferred features and value ranges for those features, aiding in the creation of group-based evaluations that assist in more informed decision-making throughout the game. While currently limited to one group per player, future updates could enable handling multiple groups or sub-groups.

A reset option is also available for restarting the game. By selecting "Enter Game", players move into the main game environment.

3.2 The Game

Upon entering the game with the chosen game configuration, the system reads in the initial causal network for the selected domain scenario and computes the initial game metrics for all players. The game features two main panels: the causal network view panel, with which players interact and manipulate, and the game metrics panel, which tracks and displays the game metrics to guide player decisions.

3.2.1 Causal Network Link Editor. The central mechanism of the game resides in the causal network view, situated in the left panel (see figure 2). All features of the chosen dataset are represented as nodes in the network, and each edge represents a causal relation. The edge's width encodes the magnitude of the corresponding standardized beta regression coefficient which signifies the importance of the source node in predicting the target node, and the arrow indicates the direction of the causal relationship. For instance, in the Hiring dataset, the feature "Age" influences the feature "Work Experience", which directly affects the target node "Job".

A slight deviation from the base system D-BIAS, we have streamlined the causal edge operation in a manner that simplifies the gaming perspective. In this study, we have omitted additional edge operations, including adding, deleting, directing, and reversing causal edges, to focus on weight instead of topology. The default causal network is presumed to encompass all pertinent edge connections, and the weight of each edge can be adjusted within a range. By selecting an edge, players can change the edge weight by sliding the slider up or down, between -100 percent to +100 percent of its original weight [2]. This bounded range of adjustments was chosen deliberately to maintain the coherence and balance of the gameplay, preventing extreme or unrealistic modifications that could disrupt the overall fairness dynamics of the game. By defining a reasonable range, users can focus on the relative impact of the edge weights rather than being overwhelmed by an entire numerical spectrum. Users need to find a relative balance within this scale to adjust the causal network appropriately. This range allows users to experiment with the strength of causal relationships while avoiding extremes that could either oversimplify the network or exacerbate bias.

3.2.2 Game Metrics and Charts. Each player's move depends on the current state of the causal network, group concerns, game metrics and score. This information is located on the right panel of the game interface in figure 2. At the top left, the current player is displayed. In the game, once a player adjusts the causal network and clicks "Apply", the game's metrics are updated, allowing the player to review these metrics and other players' scores before ending their turn with "End Turn". This two-stage process is designed to encourage active reflection on the actions taken. This sequential, turn-based approach ensures that each player has an equal opportunity to modify the causal network without the gameplay descending into chaotic or uncontrollable behavior. In contrast, a simultaneous approach could result in conflicting changes, where multiple players attempt to adjust the network in ways that conflict or overlap, leading to a lack of order in the game flow. The pause for reviewing metrics allows players to carefully consider the impact of their changes and weigh potential trade-offs, thereby fostering thoughtful and deliberate decision-making. Additionally, once players click “Apply”, they are restricted from making further alterations, ensuring that every participant has a fair chance to contribute, maintaining a balanced and orderly progression of the game.

The edge history chart (see figure 2 (b)), a line chart, tracks edge changes throughout the game, highlighting the top three edges with the highest percent change in edge weight. The X-axis indicates the player who made the change, while the Y-axis tracks the percent change in edge weight. If a specific edge is selected, the chart updates to show the history of that particular edge instead. This functionality ensures that all edge changes are accounted for, whether they are among the top three or not. The chart showcasing frequently changing edges not only aids in identifying conflicts and areas of disagreement among players but also directs their attention toward these conflicts, thereby expediting the process of reaching a consensus.

The Aggregate edge history chart (see figure 2 (c)) displays the aggregate edge change count per round, indicating if the game is progressing towards a common consensus. In an optimal scenario, the total count of edge changes in the final round should be the lowest of all rounds, ideally reaching zero.

The stakeholder attribute priority chart (see figure 2 (h)) reflects the priorities of the current player, determined by the selections they made on the game's configuration page, specifically the group they formed. For our study, we've simplified all features to a binary scale, assigning each a value of either 0 or 1; we found that this made choices clearer and easier to navigate. Based on the groups users have created, the priority chart will showcase players' level of care. The chart visually represents the extent to which players value each feature, based on the groups they have established. If both levels of a feature's bar are colored blue, it implies the player equally values each subgroup of the target variable. Conversely, if both levels of a feature's bar chart appear gray,

[2] Details on how changing the edge weight affects the data are provided in Appendix F.

it indicates that the player doesn't consider these attributes or features as central to their goals. A more comprehensive explanation of how these colors are assigned will be provided in Section 4.

The attribute outcome chart (see figure 2 (g)), shows the number of individuals from each subgroup being hired based on the current causal network setup. It is a diverging heatmap with 11 color levels, ranging from red (lowest level), to gray (neutral), to dark blue (highest level). This chart indicates how many people from each subgroup were hired.

The aggregate attribute disparity chart (see figure 2 (f)) shows the differences in hiring outcomes relative to the current player's desired outcome. This provides a measure of deviation between the actual outcome versus the player's preferred outcome (more details in Section 4).

Incorporating charts that display the group that each player cares about, the current state, and the difference between their desired and current status enables players to track their advancements, identify the areas that require further modifications, and make informed decisions accordingly. Making this kind of information available to players is crucial to a successful consensus reaching process [27]. Moreover, the visual depiction of the difference between the desired and current status serves as a motivating factor, encouraging players to actively engage in the game and work towards narrowing the gaps.

The stakeholder total loss and gain chart (see Figure 2 (d)) provides players with a simple and effective way to assess their performance in the game and compare it to others. By indicating increases in scores with green and decreases with red on top of the bars, the chart allows players to easily observe their progress and relative standing. This visual representation serves as a tool for players to track their overall performance and gain insights into how they are doing compared to their peers. Further information regarding the calculation of these values will be elaborated upon in Section 4.

Located at the bottom left, a stack of players' cards (see Figure 2 (e)) grants users access to their role-specific general goals and objectives within the game, providing them with insights into the intended outcomes they strive to achieve in their respective roles. It should be mentioned that in actual scenarios, where players are genuine stakeholders with clear intentions, these cards are redundant and therefore not needed. However, in our user studies, volunteers were asked to assume specific roles [3]. To assist these users in remembering their objectives while playing these roles, we incorporated these cards as a helpful reminder [4].

FairPlay design aims to empower players to actively address bias, navigate conflicts, and work collaboratively toward reaching a consensus. Moving on from describing different aspects of the platform, the next section will discuss the methodology behind the calculation of values for the plots described in this section.

4 METHODOLOGY

FairPlay is developed on the foundation of the previously published debiasing application, D-BIAS. In order to understand the system, we discuss its intricate aspects methodically. Figure 3 shows an overview. First, the default (initial) causal model is constructed from the Raw Data that would be used to train the ML model (first module in Figure 3). The default outcomes are also displayed in the game interface. Then the game begins where the players seek to change the default outcomes per their priorities via iterative tuning of the causal model (center module in Figure 3). The game ends when the players have achieved their goals which results in the Debiased Data. The Debiased Data can then be used to train any ML model. Here it is assumed that the ML model will not

[3] For the final user study we recruited individuals with professions that matched these roles to some extent.

[4] The cards are displayed in Appendix E.

introduce biases on its own, else an independent ML model debiasing step would be required. The upper-right-most module of Figure 3 deals with analyzing user data produced during the game.

In the following we describe each of these three modules in detail. We begin by examining the technical aspects of the visual interface, followed by an exploration of the data storage for post-game analysis.

4.1 Before the Game: Game Setup

Before the game begins, two crucial steps must be taken. Firstly, the construction of the causal network is required. Secondly, the players need to select the groups they care about based on their respective roles.

Generating Default Causal Model. The authors of the D-BIAS paper provide a comprehensive explanation of the process used to generate the causal model, employing a widely-used causal discovery algorithm known as the PC algorithm [15]. The causal network is created utilizing the PC algorithm, which infers causal connections between variables based on conditional independence tests and orientation rules using the given p-value. Each node in the network symbolizes a data attribute, and the edges signify causal relations. Since automated causal inference can introduce incorrect or incorrectly directed edges, expert users would usually inspect the generated network and correct any errors. Therefore, our system can also read in a pre-validated DAG created by experts using tools like D-BIAS or other reliable methods, ensuring the accuracy and relevance of the causal model. For our studies, we have used the fully corrected model presented in the D-BIAS paper [23].

The relationships between nodes are quantified using linear Structural Equation Models (SEM), which estimate the value of each node as a linear combination of its parent nodes. The regression coefficients in the model indicate the strength of causal relationships. Within this framework, a distinction is made between endogenous variables, nodes that have at least one edge leading into them, and exogenous variables, independent variables that have no parent nodes.

Creating Groups. In the configuration page, players are required to create a group as explained in section 3.1. A group is a set of prioritized attributes of the features, for example, for the GPA feature a player might prefer the high-GPA attribute. In the current game, it means that the player prefers that jobs are given to candidates with higher GPAs. Note that a player has a fixed budget of priorities. The more attributes the player selects the less priority is given to each. This ensures that players make thoughtful decisions about which attributes are most important to them.

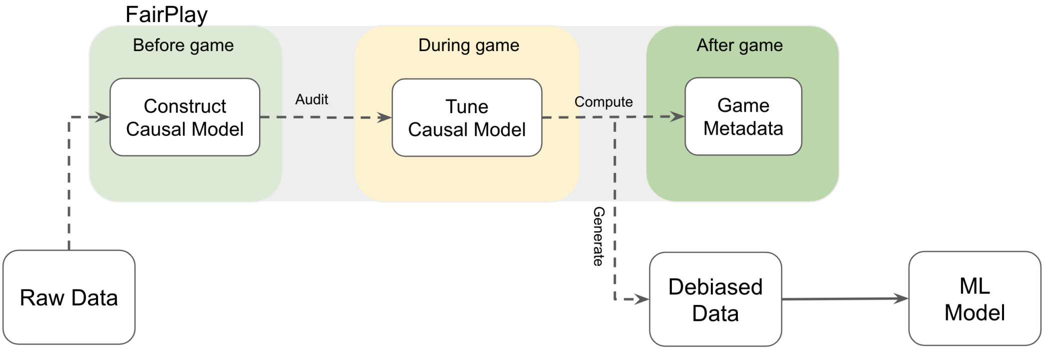


Fig. 3. FairPlay System Overview. Detailed explanations are provided in the main text above.

Although the preferences players set before the game are static and cannot be changed during gameplay, players may still need to adjust their goals as they negotiate and interact with each other. For example, while some of their preferred groups may not perform as well as they initially hoped (as indicated by red shades in the aggregate attribute disparity charts), players can compromise and agree that the final outcome is satisfactory enough to end the game. This process of adapting their goals within a fixed preference framework is key to reaching consensus, even when it means not all preferences are fully met (see Section 6.4 for insights on goal adjustments).

The stakeholder attribute priority chart (see Figure 2 (h)) visually represents the selected and non-selected attributes for each variable, with the blue bars indicating the chosen attributes and the gray bars representing the non-selected ones. The objective of each player is to equally distribute their goal among the total selected attributes. For instance, if a player cares about 10 attributes, their level of concern for each attribute will be 10 percent (refer to Algorithm 1, Line 4-15), leading accordingly to lighter shades of blue for these attributes in the stakeholder attribute priority chart. We track and report various insights on groups for all the candidates using game analysis (see section 4.3).

4.2 During the Game: Players Tune the Causal Model and Debias the Data

During the game, the system aims to monitor modifications to the causal network, calculate metrics, and gather other game metrics. The computed metrics serve as valuable information for players, aiding them in making informed decisions for their next moves.

Tuning the Causal Model. Once a player modifies the edge weight, we proceed to update the causal network with the new edge weight. Subsequently, we create a checkpoint of the updated causal network, labeling it as the “`Current Causal Network`”. Each new version of the updated causal network will be associated with this label, while older versions will be checkpointed for further analysis.

Debiased Dataset. Whenever a modification is made to the causal network, the system generates a new dataset that differs slightly from the original data. The ultimate debiased dataset corresponds to the final causal network obtained after the game concludes. All intermediate datasets generated throughout the process leading to the final game stage are utilized for post-game analysis tracking. The algorithm for generating the debiased dataset is explained in the original D-BIAS paper [23] (Refer to Algorithm 1 in the original paper).

Computing the Game Metrics. The attribute outcome chart, Figure 2(g), offers a graphical display of the hiring outcomes according to the existing causal network at any point in the game. This illustrates the count of individuals from each subgroup whose label variable equals one, indicating in the present scenario that they have gotten the job. Players also have the ability to compare job status changes relative to both the current and original causal networks, allowing for a comprehensive assessment of the causal changes’ impact on hiring outcomes (refer to Algorithm 1, Line 19-32). Two states of the outcome metric are being maintained, one for the current causal network and another for the original causal network. By tracking candidate groups and aggregating over the label (outcome) feature, corresponding to the "Job" feature in here, attribute-wise outcome metrics can be computed and presented in the aggregate chart, revealing deviations between actual and desired outcomes (refer to Algorithm 1, Line 34-42). These visualizations empower players to gain a deeper understanding of the game dynamics and make informed decisions to align their actions with their desired outcomes.

Another essential game metric is the total loss and gain, which represents the scores accumulated by each player throughout the game. This metric is computed by aggregating the sum of pair-wise multiplications between each player's group and the outcome (refer to Algorithm 1, Line 45-52).

4.3 After the Game: Data Collection and Game Analysis

In this section of the system overview, our primary emphasis lies on the post-game analysis and evaluation of the game. This phase involves scrutinizing several aspects, including the causal network changes, game moves, and conducting analysis. Further elaboration on these analyses will be provided in the upcoming section 6.

Every move the players make during the game is saved, and the data gathered opens room for extensive analysis after the game. This systematic tracking of changes in the causal network and each players outcomes allows for a comprehensive understanding of the evolving network and facilitates the evaluation of player interventions.

Another part of our post-game analysis includes ML metrics that play a crucial role in evaluating the performance of the machine learning algorithm. The algorithm used to assess these classification metrics is selected on the game configuration page (refer to Section 3.1 or Figure 1) and we track several standard ML metrics: Predicted Accuracy, representing the model's overall correctness. Predicted F1, a balanced measure of precision and recall. Individual Fairness, defined as the mean percentage of a data point's k-nearest neighbors that have a different output label, measuring equality and consistency in decision-making within a system. Parity, a metric used to assess equality in outcomes across demographic groups.

Analyzing game metrics is crucial for assessing the impact of causal changes on the current debiased data compared to the original data. To prevent information overload, we currently track these metrics internally for analysis purposes, without displaying them to players. By incorporating these diverse metrics, we can effectively analyze the outcome of players' actions, final hiring decisions, and overall game progression.

5 Experiments

5.1 Setup

To evaluate the effectiveness of FairPlay, our research question aimed to determine whether consensus can be achieved among players in a multi-player game environment while modifying the causal graph to mitigate bias. The study goals were:

- **G1**: Assess the game's ability to educate and engage players in the complexities of bias mitigation through their interactions and feedback during gameplay.
- **G2**: Gather insights into the consensus-building process within the game, observing how players collaboratively modify the causal graph and reach a consensus on removing bias and how they perceive the process and the outcome of FairPlay.
- **G3**: Analyze the outcome of the process, the debiased datasets, using accuracy and fairness metrics.
- **G4**: Collect feedback on the usability of the game interface and mechanics.

For the user studies, we recruited volunteers online, from individuals affiliated with the Computer Science department at a major university. For the first three studies, this included employees and students. For the fourth study, we looked for participants with experience in industry for the specific roles we needed for the study. A total of 20 volunteers were sought for participation. On average, participants demonstrated familiarity with AI, ML, and fairness, along with a solid understanding of current issues in these domains. The participants were not limited to a specific demographic group with the majority being between 21-30 years old (41.7%) and primarily male (75%). Ethnically, the largest group identified as Asian (75%). A significant portion of participants held a Master's degree (27.3%), followed by those with Bachelor's and Doctoral degrees. More details about the demographic information of participants and their backgrounds is available in Appendix C.

Once we received responses from interested individuals, we randomly divided them into three groups of five participants each. Additionally, for the last user study, we recruited 5 participants separately, each with real-world experiences in a particular role we needed. The games were scheduled to be conducted remotely via Zoom, allowing for remote participation and not requiring players to be co-located. The game was played synchronously, with each player taking their turn to make changes to the causal network. During the game sessions, players took turns and requested remote access to the host's screen to play their respective turns. Once a player ended their turn by clicking "End Turn," the next player could make their move. This synchronous play style ensured that every participant could get a fair opportunity to play, maintaining a balanced and orderly flow of gameplay in which participants could observe other players' actions and understand their goals better.

In real-world applications, users would typically access the game through a web address on their own systems and would only be able to make changes to the game during their turn. However, for the user studies, we conducted sessions over Zoom to monitor user interactions and discussions. To maintain consistency across our first three studies, we predefined profile preferences and the groups each player cared about, rather than allowing users to make these selections themselves. We thoroughly explained these preferences to participants so they could act in line with their assigned roles. Player cards were used to help users keep track of their role-specific general goals. The roles selected for this particular dataset and user studies included Hiring Agency, Employer, Union Representative, Co-workers, and Manager. These roles represent stakeholders involved in debiasing training data for a job-applicant decision system in a real-world context as previously discussed. More information about these roles and each roles' goals and preferences is discussed in Appendix E.

A fourth user study was conducted with participants recruited based on their experience in one of the aforementioned roles to better approximate real-world conditions. In this study, participants selected their own preferences based on their experience in that position before starting the game. This approach allowed us to assess our goals in a more realistic scenario and ensured that the observed results were not solely influenced by how we designed each role's preferences.

5.2 GamePlay

To familiarize participants with the FairPlay game mechanics, we began each user-study session with a ten-minute informational video. This video detailed how the game functions and how players could utilize its features to achieve their objectives. Following the video presentation, we conducted a brief question-and-answer session to address any queries participants had about the game.

Once all questions were addressed, the game commenced with each of the five players selecting a unique role. In a given round, players were tasked with modifying the weights in the causal diagram to align them with their respective goals. Upon satisfaction with their adjustments, players hit the 'Apply' button to review the results of their interventions. Subsequently, they ended their turn, allowing the next player to perform their modifications.

At the conclusion of a full round involving all five players, a popup displaying all players' scores was shown. Players were then asked whether a consensus to accept the current network had been reached or if they wished to continue modifying it. If even a single player opted to continue, the game extended into another round.

We conducted a total of four user studies. In the first three studies, the game continued for 5, 2, and 3 rounds respectively, lasting for 1 hour, 1.5 hours, and 1.5 hours. Our final study with participants with real-world experience continued for 3 rounds and lasted 1 hour. By the end of each game, all players achieved improved metrics reflective of their objectives. In the next section, we will discuss the results of these user studies.

6 Results

6.1 Machine Learning Metrics Analysis

Evaluating how classifiers perform on the final debiased datasets, particularly in terms of accuracy and fairness, is vital (**G3**). Accuracy is a primary concern in real-world applications, and having a fairer dataset is what the process aims to achieve. To evaluate the effectiveness, we consider 4 metrics detailed in section 4.3 and displayed in Table 2. Predicted Accuracy, reflecting the model's overall correctness, is conventionally sought at higher values; however, the debiased models reveal lower scores, signaling a deliberate trade-off for heightened fairness. Similarly, Predicted F1, a metric balancing precision and recall, is typically favored at higher values, yet the debiased models exhibit lower figures. Examining Individual Fairness, where lower scores indicate reduced disparate treatment among individuals, the debiased models consistently achieve significantly better (lower) values, indicative of a noteworthy improvement. Assessing Parity, a metric gauging equality in outcomes across demographic groups, higher values are preferred, and the debiased models generally exhibit enhanced parity values. These findings align with existing research on the tradeoff between accuracy and fairness [38][58].

	User Study 1		User Study 2		User Study 3		User Study 4	
	Original	Debiased	Original	Debiased	Original	Debiased	Original	Debiased
Predicted Accuracy	0.76	0.63 ▼	0.76	0.69 ▼	0.76	0.63 ▼	0.76	0.63 ▼
Predicted F1	0.58	0.44 ▼	0.58	0.53 ▼	0.58	0.44 ▼	0.58	0.44 ▼
Individual Fairness	22.11	0.28 ▼	22.11	3.21 ▼	22.11	1.73 ▼	22.11	0.04 ▼
Parity	35.96	47.45 ▲	35.96	45.9 ▲	35.96	45.86 ▲	35.96	47.69 ▲

Table 2. ML Metrics observed during all four user studies. The green color indicates improvement, and the direction of the triangles shows how the value changed. For example, a green triangle pointing down means the value decreased, and lower values are preferred for this feature, so this decrease represents an improvement.

6.2 Behavioral Observations

Our initial analysis of whether players perceived the final outcome as fair (**G2**) was assessed through a question in our post-game survey. Participants were asked to rate their agreement with the statement: "I think that the activities led to a fairer system." on a scale from 1 (strongly disagree) to 5 (strongly agree). On average, users rated this question 3.7, suggesting that they believed the collaboration had a positive impact.

Also, throughout the user studies, participants were encouraged to vocalize their thoughts while playing, discussing the factors influencing their decisions each round. With their consent, the studies were recorded for more in-depth analysis later on. This was to be able to analyze the consensus-building process in FairPlay in more detail (**G2**).

Qualitative analysis of players' dialogues throughout the game helped us assess whether the dashboard features were assisting or confusing them. The dashboard appeared intuitive to players, even those with no prior experience with causal networks. Players modified edges based on attributes they were supposed to care about, stating things like, "I'm doing this because I care about feature X." They also adjusted edges previously edited by others, saying, "I don't care about this feature, so I don't want this to play a role." Additionally, players predicted how their changes would affect the groups they cared about, with statements such as, "I want more people with feature X to get the job." In most cases, their predictions aligned with the results shown in the right panel plots after clicking the "Apply" button, indicating that they were able to use the causal network correctly to achieve their desired outcomes.

We can analyze user behaviors when it comes to the right panel and how insightful it was in the game through the lens of two different philosophical schools of thought: Consequentialism (which focuses on outcomes) and Deontology (which focuses on the morality of actions) [3]. Prior to the studies, all participants were asked whether their ethical approach aligned more closely with Deontological or Consequentialist principles. 72.7% of participants identified as Deontologists, while 27.3% identified as Consequentialists. This distinction in mindsets was evident in the way users made their decisions. Some players adjusted the network to achieve the best outcome metrics for their groups, while others prioritized setting the network parameters correctly, regardless of the outcomes shown by the plots. The Deontologist players showed a strong preference for down-weighting edges that emerged from the sensitive variables [5], regardless of the consequences of their metric outcomes, hence not paying too much attention to the right panel. If we consider only the Consequentialist players, we notice they were more likely to fiddle with parameters in either direction while searching for the best outcome metrics. They would first look at the panel on the right to see how their groups are doing, modify edges accordingly, and then observe the outcomes on the right panel more thoroughly.

Despite these differences between strategies, a consensus was eventually reached in all user studies, demonstrating the game's ability to facilitate mutual agreement even among diverse objectives by providing intuitive means of modifying the network and insightful metrics to help users make decisions (**G2**).

To determine if the game educated players about the complexities of bias mitigation (**G1**), we asked users to rate the following statement in our post-game survey on a scale from 1 (strongly disagree) to 5 (strongly agree): "The game improved my understanding of fairness and bias in automated decision systems." The average score was 3.3, indicating that the game had an overall positive effect on educating the players.

6.3 System Usability Score Analysis

One of the key indicators of a tool's success, irrespective of its features and objectives, is its perceived effectiveness, efficiency, and user satisfaction (**G4**). To assess this, we utilized the standard System Usability Scale [8] (created by John Brook at Digital Equipment Corporation in 1986), which employs a 5-point Likert scale. After completing the study, users were requested to fill out a feedback form. The user feedback statistics, as shown in Figure 4, reveal that the statement players disagreed with the most was "I thought there was too much inconsistency in FairPlay", while the statement the players agreed with the most was "I found the various functions in FairPlay were well integrated." This reflects that the tool's visual presentations and functionalities were cohesively aligned and user-friendly. The overall SUS score was 68.05, positioning FairPlay as a positive and intuitive system, especially since SUS scores above 68 are considered above average.

6.4 Insights

In this section, we discuss the specifics of the four user studies and examine the outcomes of each game upon conclusion.

[5] Many variables are explicitly defined as "sensitive" by specific legal frameworks [11]. In our dataset, Age, Gender, and Race are considered sensitive based on the framework outlined in [50].

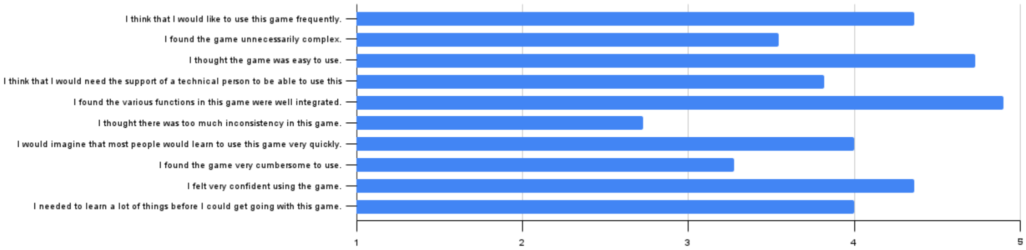


Fig. 4. FairPlay Feedback Results

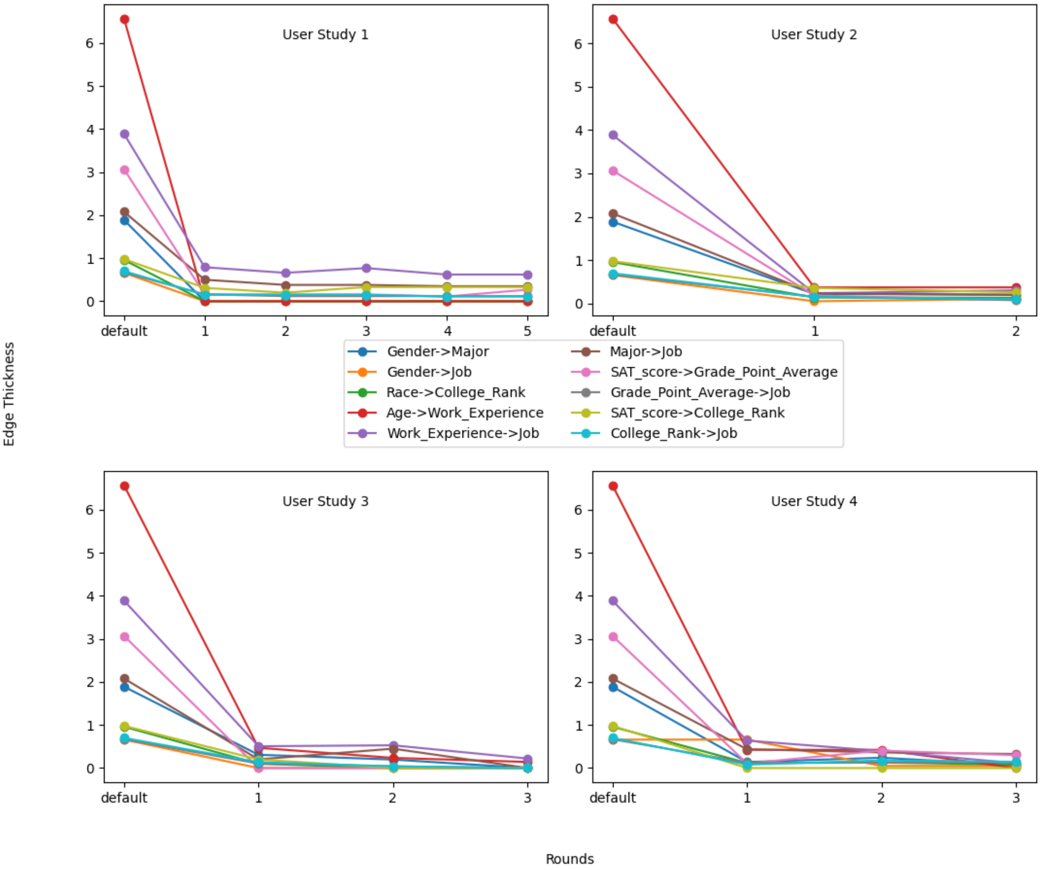


Fig. 5. FairPlay User Studies: Edge Weights vs. Round for all User Studies

Figure 5 shows the progression of edge-weight adjustments made by players in each round.

In all four studies, there's a sharp decline in edge thickness from the default to the first round. This suggests that players were quick to act on their initial assessments of the causal network.

Furthermore, nearly all edge-weights dropped below 1 in the opening round and remained low, hinting at the strength of initial user impressions.

By the second round and onwards, changes become less drastic, and edge weights appear to stabilize. This could indicate that users reached some form of consensus or satisfaction with the state of the causal network early on.

While the general trend across rounds is similar, the final edge weights vary between studies, suggesting that while the process is consistent, the outcomes are subjective and influenced by the unique dynamics and decisions of the participants in each user study.

Figure 6 displays the final causal networks and the aggregate attribute disparity charts at the conclusion of the gameplay.

Upon examining the final causal networks, a consistent pattern becomes evident across all four games. We observe sparse networks with many edges reduced to minimal weights, particularly for sensitive attributes like Age, Race, and Gender. This pruning of dependencies results in simple network topologies.

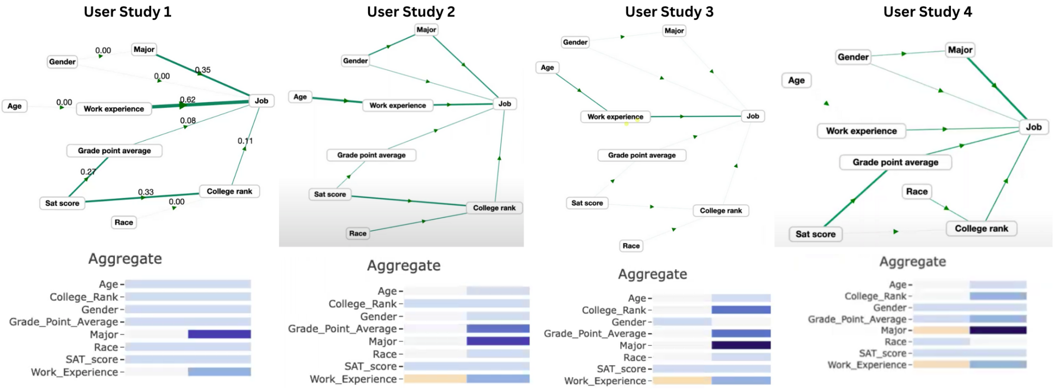


Fig. 6. FairPlay Matrix. This figure compares the causal networks created by each set of players (top row) and the aggregate attribute disparity charts (bottom row)

When we turn to aggregate attribute disparity charts, it's evident that players managed to avoid unfavorable outcomes (marked by red shades) for almost all attributes across all games. However, exceptions are observed in Games 2, 3 and 4, concerning individuals with low work experience. This finding suggests that players accepted that individuals with less work experience may not be selected for the job, despite at least one participant valuing this attribute (otherwise it wouldn't be displayed in a red shade). This shows that users are willing to accept trade-offs in order to reach consensus (The same holds true for Major in the forth study).

Drawing insights from Figure 6 as a whole, players seem to be striving for maximum scores for their groups, as indicated by the aggregate attribute disparity charts. Interestingly, it appears possible to optimize benefits for the desired groups for everyone without explicitly taking any particular definition of fairness into account. This demonstrates that although satisfying all definitions of fairness might be challenging, as suggested by the literature, reaching a consensus on what factors should influence the final decision in a specific context is achievable. Users were able to negotiate and agree on trade-offs, indicating a collective prioritization of certain attributes over others to reach a shared understanding of fairness within the specific context of the game.

7 Discussion

Our platform's objective was to create a tool that facilitates users in collaboratively determining the causal structure of their datasets. It's important to recognize that in the absence of our tool, attaining this goal would be a significant challenge. An alternative might involve engaging stakeholders in a dialogue to debate the causal structure. However, without the ability to modify the causal network and observe its impact on the data, and lacking a structured negotiation process, reaching a consensus is not only improbable but also likely of low quality due to the lack of informed decision-making. Our interface addresses these challenges by 1) Establishing a systematic approach to the negotiation process, 2) Ensuring that every participant's perspective is heard and considered, and 3) Providing users with sufficient information to make well-informed decisions.

As mentioned in section 2.2, soft consensus tries to achieve two important goals: (i) Reflect better partial agreement; and (ii) Guide the Consensus Reaching Process (CRP) until an acceptable high level of agreement is achieved among the decision-makers [57]. Across all four user studies, stakeholders expressed verbal satisfaction with the existing graph. For some participants, this contentment emerged as early as the second round and persisted throughout the game, while for others, satisfaction fluctuated from one round to the next due to modifications made by other players. A crucial point to note is that the game was only concluded when every stakeholder expressed complete satisfaction with the current state. This condition was met in all four user studies, with players verbally confirming their satisfaction with the outcomes. This outcome serves as a clear confirmation of the CRP's success, demonstrating its effectiveness in achieving a high level of consensus among all participants.

8 Limitations

Our approach is subject to certain limitations that are discussed below.
Firstly, the participants in our user studies were not actual stakeholders in a real scenario, which could influence the game dynamics. In real-world settings, stakeholders might approach the game with greater eagerness and potentially be less willing to compromise. Similarly, time-pressure may be less (or more) of an issue in real-world settings. Furthermore, since the sessions were conducted with participants aware that they were being observed, there's a risk that their behavior was influenced by the Hawthorne effect [33] (i.e. studying an agent, changes the agent's behavior). Although this was a conscious choice to facilitate observation of user interaction with the system, it remains a limitation of the study.

Another notable limitation was the learning curve associated with the system. Even with thorough training, participants initially struggled to use the software effectively to pursue their objectives. This challenge was most apparent in the first round, characterized by more experimental than tactical behavior. Despite this, the System Usability Scale (SUS) feedback indicated that users did not perceive the learning curve as steep. Nonetheless, to improve the process, enhancing the introductory phase of the game, potentially with a more interactive approach than a video, could be beneficial.

Lastly, we have deliberately simplified the information available to users. While this decision was made to prevent user overwhelm and aim for simplicity, it may pose limitations when scaling to real-world scenarios. The absence of machine learning metrics in the interface and the full range of edge operations for example, might impede stakeholders in a practical context. Addressing these constraints will be crucial for the interface's applicability in actual stakeholder environments.

9 Future Work

While FairPlay is primarily designed for practitioners and stakeholders across various domains, its utility is not limited to these groups alone. It also holds potential for application in other areas, such as educational settings. Current computer science courses addressing AI fairness and bias typically lean towards statistical analysis or incorporate philosophical perspectives that often lack pragmatic implications for students [37]. However, pedagogical research has long emphasized the benefits of employing tools and visualizations in enhancing student learning [39][41]. In this context, case studies have highlighted the positive effects of using publicly available visualization tools from Human-Computer Interaction (HCI) practice as educational resources to explore algorithmic fairness concepts [37]. Given its collaborative nature and the unique way it engages users, FairPlay emerges as an excellent resource for educational settings. Exploring the potential of FairPlay in aiding students to effectively engage with the identification and mitigation of bias presents a promising research direction.

FairPlay meets many of the design criteria for a fairness toolkit as recommended by practitioners [42], yet there is scope for enhancement. Acknowledging the flexibility that practitioners often have at different stages of their machine learning pipelines [42], it becomes apparent that FairPlay has the potential to expand its impact. By extending its collaborative approach to encompass various phases of the machine learning lifecycle, not just limited to the pre-processing stage, FairPlay could significantly enhance its utility and effectiveness. This could include allowing users to have their own ML models incorporated in the system where right now only a limited number of options are available. Moreover, while FairPlay currently supports only edge weight modification, we plan to introduce additional edge operations, such as deleting or adding edges, to provide players with greater flexibility. This will allow us to explore and analyze the impact of these operations on the consensus-reaching process. Another intentional design choice was to hide ML-related performance metrics from players to avoid overwhelming them with excessive plots and data. However, we aim to tackle the challenge of integrating these metrics into the player's view in a manner that enhances informed decision-making without causing information overload in future work. In the future, we also plan to allow users to specify more detailed preferences, such as age ranges (e.g., 25-35) and additional categorical preferences (e.g., education level) instead of just binary ones.

As highlighted in the related works section, it's clear that no algorithm can be solely relied upon to determine the most fitting definition of fairness or to establish an agreeable balance among various fairness metrics for all stakeholders. Consequently, FairPlay adopts a human-centered approach. Nevertheless, the idea of integrating an automated agent within FairPlay to assist in the consensus-building process among stakeholders is an intriguing concept. There are well-established consensus-reaching algorithms that could be adapted for such an agent [6][48][35]. Investigating how the inclusion of an automated agent might alter the group dynamics, and whether it aids the process, would be a valuable area for future research. This represents an innovative way of exploring the interaction between humans and algorithms in the context of addressing bias, offering potential insights into enhancing collaborative decision-making.

Moreover, we recognize the importance of making FairPlay accessible for replication, collaboration, and improvement by the broader community. We plan to develop FairPlay as an open-source project with detailed contribution guidelines and coding standards. This will facilitate collaboration, allowing researchers and practitioners to enhance functionalities and adapt the tool for various use-cases.

The promising outcomes FairPlay demonstrates in addressing fairness issues through collaboration suggest that this method could be beneficially adopted by other fairness toolkits as well.

This would provide a diverse group of stakeholders with a structured framework for achieving consensus on complex and often contentious issues like fairness.

10 Conclusion

With the increasing reliance on algorithms as decision-makers across various contexts, the importance of thoroughly auditing these algorithms for ethical concerns has become more pronounced. However, research indicates that fulfilling all fairness criteria can be challenging, if not impossible. This necessitates a context-specific audit, ideally conducted by humans, though it is important to acknowledge that humans too have their own biases and blind spots. Consequently, adopting a team-based approach to this audit process emerges as a promising strategy. FairPlay aims to facilitate such an environment, where stakeholders or domain experts, ideally representing a diverse array of viewpoints, collaborate systematically to discern the relevant features in the underlying dataset. The varied outcomes and distinct final causal structures resulting from the four user studies underscore the necessity of tools like FairPlay that facilitate such processes. This diversity highlights that different groups may converge on varying agreements, leading to unique final structures. The fact that all four user studies achieved consensus serves as a strong validation of FairPlay's effectiveness. Developing tools like FairPlay is crucial for enabling informed auditing processes. Without such platforms, it would be unfeasible to engage in meaningful conversations that lead to prompt action, highlighting the crucial role these tools play in facilitating collaborative decision-making.

Acknowledgments

Work reported in this publication was partially supported by SUNY System Administration under SUNY Research Seed Grant Award 23-01-RSG. This material is based upon work supported by the National Science Foundation under Grant No. NRT-HDR 2125295. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation. ChatGPT was utilized to generate sections of this Work, including text, tables, graphs, code, data, citations, etc.

References

[1] Alekh Agarwal, Alina Beygelzimer, Miroslav Dudík, John Langford, and Hanna Wallach. 2018. A reductions approach to fair classification. In *International conference on machine learning*. PMLR, 60–69.

[2] Yongsu Ahn and Yu-Ru Lin. 2019. Fairsight: Visual analytics for fairness in decision making. *IEEE transactions on visualization and computer graphics* 26, 1 (2019), 1086–1095.

[3] Larry Alexander and Michael Moore. 2007. Deontological ethics. (2007).

[4] Rachel Bellamy, Kuntal Dey, Michael Hind, Samuel Hoffman, Stephanie Houde, Kalapriya Kannan, Pranay Lohia, Jacquelyn Martino, Sameep Mehta, Aleksandra Mojsilovic, Seema Nagar, Karthikeyan Natesan Ramamurthy, John Richards, Diptikalyan Saha, Prasanna Sattigeri, Moninder Singh, Kush Varshney, and Yunfeng Zhang. 2019. AI Fairness 360: An Extensible Toolkit for Detecting and Mitigating Algorithmic Bias. *IBM Journal of Research and Development* PP (09 2019), 1–1. doi:10.1147/JRD.2019.2942287

[5] James C. Bezdek, Bonnie Spillman, and Richard Spillman. 1978. A fuzzy relation space for group decision theory. *Fuzzy Sets and Systems* 1, 4 (1978), 255–268. doi:10.1016/0165-0114(78)90017-9

[6] Rashmi Bhardwaj and Debabrata Datta. 2020. *Consensus Algorithm*. Springer International Publishing, Cham, 91–107. doi:10.1007/978-3-030-38677-1_5

[7] Reuben Binns. 2018. Fairness in machine learning: Lessons from political philosophy. In *Conference on fairness, accountability and transparency*. PMLR, 149–159.

[8] John Brooke. 1996. SUS: a "quick and dirty" usability scale. In *Usability Evaluation in Industry*, P. W. Jordan, B. Thomas, B. A. Weerdmeester, and A. L. McClelland (Eds.). Taylor and Francis, London.

[9] Ángel Alexander Cabrera, Will Epperson, Fred Hohman, Minsuk Kahng, Jamie Morgenstern, and Duen Horng Chau. 2019. FairVis: Visual analytics for discovering intersectional bias in machine learning. In *2019 IEEE Conference on Visual Analytics Science and Technology (VAST)*. IEEE, 46–56.

[10] Flavio Calmon, Dennis Wei, Bhanukiran Vinzamuri, Karthikeyan Natesan Ramamurthy, and Kush R Varshney. 2017. Optimized pre-processing for discrimination prevention. *Advances in neural information processing systems* 30 (2017).

[11] Simon Caton and Christian Haas. 2023. Fairness in Machine Learning: A Survey. *ACM Comput. Surv.* (aug 2023). doi:10.1145/3616865 Just Accepted.

[12] F. Chiclana, J.M. Tapia García, M.J. del Moral, and E. Herrera-Viedma. 2013. A statistical comparative study of different similarity measures of consensus in group decision making. *Information Sciences* 221 (2013), 110–123. doi:10.1016/j.ins.2012.09.014

[13] Flavio Chierichetti, Ravi Kumar, Silvio Lattanzi, and Sergei Vassilvtiskii. 2019. Matroids, Matchings, and Fairness. In *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics (Proceedings of Machine Learning Research, Vol. 89)*, Kamalika Chaudhuri and Masashi Sugiyama (Eds.). PMLR, 2212–2220. https://proceedings.mlr.press/v89/chierichetti19a.html

[14] Alexandra Chouldechova. 2017. Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments. *Big Data* 5, 2 (2017), 153–163. doi:10.1089/big.2016.0047 arXiv:https://doi.org/10.1089/big.2016.0047 PMID: 28632438.

[15] Diego Colombo, Marloes H Maathuis, et al. 2014. Order-independent constraint-based causal structure learning. *J. Mach. Learn. Res.* 15, 1 (2014), 3741–3782.

[16] Berkeley J Dietvorst, Joseph P Simmons, and Cade Massey. 2018. Overcoming algorithm aversion: People will use imperfect algorithms if they can (even slightly) modify them. *Management science* 64, 3 (2018), 1155–1170.

[17] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. 2012. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*. 214–226.

[18] Martin J Eppler. 2004. Facilitating Knowledge Communication through Joint Interactive Visualization. *J. Univers. Comput. Sci.* 10, 6 (2004), 683–690.

[19] Robert D Feick and G Brent Hall. 1999. Consensus-building in a multi-participant spatial decision support system. *URISA journal* 11, 2 (1999), 17–23.

[20] Sorelle A. Friedler, Carlos Scheidegger, and Suresh Venkatasubramanian. 2016. On the (im)possibility of fairness. arXiv:1609.07236

[21] Sorelle A. Friedler, Carlos Scheidegger, and Suresh Venkatasubramanian. 2021. The (Im)possibility of Fairness: Different Value Systems Require Different Mechanisms For Fair Decision Making. *Commun. ACM* 64, 4 (April 2021), 136–143. doi:10.1145/3433949

[22] Pratik Gajane and Mykola Pechenizkiy. 2017. On formalizing fairness in prediction with machine learning. *arXiv preprint arXiv:1710.03184* (2017).

[23] Bhavya Ghai and Klaus Mueller. 2022. D-BIAS: a causality-based human-in-the-loop system for tackling algorithmic bias. *IEEE Transactions on Visualization and Computer Graphics* 29, 1 (2022), 473–482.

[24] Moritz Hardt, Eric Price, and Nati Srebro. 2016. Equality of opportunity in supervised learning. *Advances in neural information processing systems* 29 (2016).

[25] Jeffrey Heer and Maneesh Agrawala. 2008. Design considerations for collaborative visual analytics. *Information Visualization* 7, 1 (mar 2008), 49–62. doi:10.1145/1391107.1391112

[26] Kenneth Holstein, Jennifer Wortman Vaughan, Hal Daumé, Miro Dudik, and Hanna Wallach. 2019. Improving Fairness in Machine Learning Systems: What Do Industry Practitioners Need?. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–16. doi:10.1145/3290605.3300830

[27] Judith E Innes. 2004. Consensus building: Clarifications for the critics. *Planning theory* 3, 1 (2004), 5–20.

[28] Faisal Kamiran and Toon Calders. 2012. Data preprocessing techniques for classification without discrimination. *Knowledge and information systems* 33, 1 (2012), 1–33.

[29] Michael P Kim, Amirata Ghorbani, and James Zou. 2019. Multiaccuracy: Black-box post-processing for fairness in classification. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*. 247–254.

[30] Jon Kleinberg, Jens Ludwig, Sendhil Mullainathan, and Ashesh Rambachan. 2018. Algorithmic Fairness. *AEA Papers and Proceedings* 108 (May 2018), 22–27. doi:10.1257/pandp.20181018

[31] Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. 2016. Inherent trade-offs in the fair determination of risk scores. arXiv:1609.05807

[32] Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. 2016. Inherent trade-offs in the fair determination of risk scores. *arXiv preprint arXiv:1609.05807* (2016).

[33] Henry A. Landsberger. 1958. *Hawthorne Revisited: Management and the Worker, Its Critics, and Developments in Human Relations in Industry.* Cornell University, Ithaca, NY.

[34] Yanjiang Li and Chong Tan. 2019. A survey of the consensus for multi-agent systems. *Systems Science & Control Engineering* 7, 1 (2019), 468–482.

[35] Huchang Liao, Zeshui Xu, Xiao-Jun Zeng, and Dong-Ling Xu. 2016. An enhanced consensus reaching process in group decision making with intuitionistic fuzzy preference relations. *Information Sciences* 329 (2016), 274–286. doi:10.1016/j.ins.2015.09.024 Special issue on Discovery Science.

[36] Tyler Lu and Craig Boutilier. 2011. Budgeted social choice: From consensus to personalized decision making. In *IJCAI*, Vol. 11. 280–286.

[37] Afra Mashhadi, Annuska Zolyomi, and Jay Quedado. 2022. A Case Study of Integrating Fairness Visualization Tools in Machine Learning Education. In *Extended Abstracts of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI EA '22)*. Association for Computing Machinery, New York, NY, USA, Article 37, 7 pages. doi:10.1145/3491101.3503568

[38] Aditya Krishna Menon and Robert C. Williamson. 2018. The cost of fairness in binary classification. In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency (Proceedings of Machine Learning Research, Vol. 81)*, Sorelle A. Friedler and Christo Wilson (Eds.). PMLR, 107–118. https://proceedings.mlr.press/v81/menon18a.html

[39] Željko Obrenović. 2012. Rethinking HCI Education: Teaching Interactive Computing Concepts Based on the Experiential Learning Paradigm. *Interactions* 19, 3 (may 2012), 66–70. doi:10.1145/2168931.2168945

[40] Dana Pessach and Erez Shmueli. 2022. A Review on Fairness in Machine Learning. *ACM Comput. Surv.* 55, 3, Article 51 (feb 2022), 44 pages. doi:10.1145/3494672

[41] Kris Powers, Paul Gross, Steve Cooper, Myles Mcnally, Kenneth Goldman, Viera Proulx, and Martin Carlisle. 2006. Tools for teaching introductory programming: What works? *ACM SIGCSE Bulletin* 38, 560–561. doi:10.1145/1121341.1121514

[42] Brianna Richardson, Jean Garcia-Gathright, Samuel F. Way, Jennifer Thom, and Henriette Cramer. 2021. Towards Fairness in Practice: A Practitioner-Oriented Rubric for Evaluating Fair ML Toolkits. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 236, 13 pages. doi:10.1145/3411764.3445604

[43] Pedro Saleiro, Benedict Kuester, Loren Hinkson, Jesse London, Abby Stevens, Ari Anisfeld, Kit T Rodolfa, and Rayid Ghani. 2018. Aequitas: A bias and fairness audit toolkit. *arXiv preprint arXiv:1811.05577* (2018).

[44] Thomas J. Scheff. 1967. Toward a Sociological Model of Consensus. *American Sociological Review* 32, 1 (1967), 32–46. http://www.jstor.org/stable/2091716

[45] Andrew D. Selbst, Danah Boyd, Sorelle A. Friedler, Suresh Venkatasubramanian, and Janet Vertesi. 2019. Fairness and Abstraction in Sociotechnical Systems. In *Proceedings of the Conference on Fairness, Accountability, and Transparency* (Atlanta, GA, USA) *(FAT* '19)*. Association for Computing Machinery, New York, NY, USA, 59–68. doi:10.1145/3287560.3287598

[46] Ben Shneiderman. 2022. *Human-centered AI.* Oxford University Press.

[47] Peter Spirtes, Clark N Glymour, and Richard Scheines. 2000. *Causation, prediction, and search.* MIT press.

[48] Ming Tang, Huchang Liao, Xiaomei Mi, Benjamin Lev, and Witold Pedrycz. 2021. A hierarchical consensus reaching process for group decision making with noncooperative behaviors. *European Journal of Operational Research* 293, 2 (2021), 632–642. doi:10.1016/j.ejor.2020.12.028

[49] Marleen Van de Kerkhof. 2006. Making a difference: on the constraints of consensus building and the relevance of deliberation in stakeholder dialogues. *Policy Sciences* 39, 3 (2006), 279–299.

[50] Michael Veale and Reuben Binns. 2017. Fairer machine learning in the real world: Mitigating discrimination without collecting sensitive data. *Big Data & Society* 4 (12 2017), 205395171774353. doi:10.1177/2053951717743530

[51] Mingyang Wan, Daochen Zha, Ninghao Liu, and Na Zou. 2023. In-processing modeling techniques for machine learning fairness: A survey. *ACM Transactions on Knowledge Discovery from Data* 17, 3 (2023), 1–27.

[52] Qianwen Wang, Zhenhua Xu, Zhutian Chen, Yong Wang, Shixia Liu, and Huamin Qu. 2020. Visual analysis of discrimination in machine learning. *IEEE Transactions on Visualization and Computer Graphics* 27, 2 (2020), 1470–1480.

[53] James Wexler, Mahima Pushkarna, Tolga Bolukbasi, Martin Wattenberg, Fernanda Viégas, and Jimbo Wilson. 2019. The what-if tool: Interactive probing of machine learning models. *IEEE transactions on visualization and computer graphics* 26, 1 (2019), 56–65.

[54] Tiankai Xie, Yuxin Ma, Jian Kang, Hanghang Tong, and Ross Maciejewski. 2021. Fairrankvis: A visual analytics framework for exploring algorithmic fairness in graph mining models. *IEEE Transactions on Visualization and Computer Graphics* 28, 1 (2021), 368–377.

[55] Jing Nathan Yan, Ziwei Gu, Hubert Lin, and Jeffrey M Rzeszotarski. 2020. Silva: Interactively assessing machine learning fairness using causality. In *Proceedings of the 2020 chi conference on human factors in computing systems*. 1–13.

[56] Eliezer Yudkowsky. 2004. Coherent extrapolated volition.

[57] Hengjie Zhang, Yucheng Dong, Francisco Chiclana, and Shui Yu. 2019. Consensus efficiency in group decision making: A comprehensive comparative study and its optimal design. *European Journal of Operational Research* 275, 2 (2019), 580–598. doi:10.1016/j.ejor.2018.11.052

[58] Han Zhao and Geoff Gordon. 2019. Inherent Tradeoffs in Learning Fair Representations. In *Advances in Neural Information Processing Systems*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (Eds.), Vol. 32. Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2019/file/b4189d9de0fb2b9cce090bd1a15e3420-Paper.pdf

[59] Roshanak Zilouchian Moghaddam, Brian P Bailey, and Christina Poon. 2011. Ideatracker: an interactive visualization supporting collaboration and consensus building in online interface design discussions. In *Human-Computer Interaction–INTERACT 2011: 13th IFIP TC 13 International Conference, Lisbon, Portugal, September 5-9, 2011, Proceedings, Part I 13*. Springer, 259–276.

Appendix

This appendix contains additional details and figures referenced in the main text. The following sections provide further insights into the experiments conducted, the configuration settings used, and technical details about machine learning models.

A Machine Learning Models

The ML models used in our study are implemented using the ‘scikitlearn‘ library in Python. Below are the details of each model and the relevant code snippets for their implementation. The original data (‘df‘) and debiased data (‘df_deb‘) are used to compute performance metrics at each stage when changes are made to the causal network by a player.

Logistic Regression. Logistic Regression is a linear model used for binary classification tasks. It models the probability that a given input belongs to a certain class.

```
from sklearn.linear_model import LogisticRegression
model = LogisticRegression()
model.fit(X_train, y_train)
predictions = model.predict(X_test)
```

Support Vector Machine (SVM). SVM is a supervised learning model that can be used for both classification and regression tasks. It works by finding the hyperplane that best separates the classes.

```
from sklearn.svm import SV
model = SVC()
model.fit(X_train, y_train)
predictions = model.predict(X_test)
```

Naive Bayes. Naive Bayes is a probabilistic classifier based on applying Bayes’ theorem with strong (naive) independence assumptions between the features.

```
from sklearn.naive_bayes import GaussianNB
model = GaussianNB()
model.fit(X_train, y_train)
predictions = model.predict(X_test)
```

k-Nearest Neighbors (kNN). kNN is a simple, instance-based learning algorithm used for classification and regression. It predicts the class of a sample based on the majority class among its k nearest neighbors.

```
from sklearn.neighbors import KNeighborsClassifier
model = KNeighborsClassifier(n_neighbors=5)
model.fit(X_train, y_train)
predictions = model.predict(X_test)
```

Decision Tree. Decision Tree is a non-parametric supervised learning method used for classification and regression. It splits the data into subsets based on the value of input features.

```
from sklearn.tree import DecisionTreeClassifier
model = DecisionTreeClassifier()
model.fit(X_train, y_train)
predictions = model.predict(X_test)
```

Neural Network. Neural Network is a set of algorithms, modeled loosely after the human brain, that are designed to recognize patterns.

```
from sklearn.neural_network import MLPClassifier
model = MLPClassifier()
model.fit(X_train, y_train)
predictions = model.predict(X_test)
```

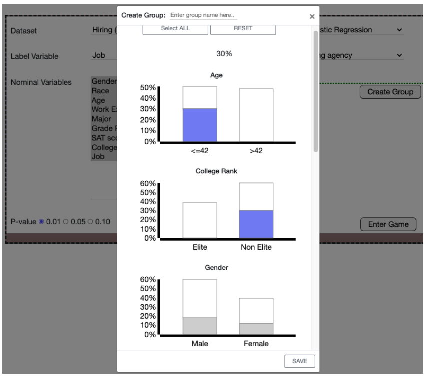
Data Usage. The original data (‘df‘) and the debiased data (‘df_deb‘) are used to evaluate the performance of the ML models. The debiased data is updated at each stage when a player makes changes to the causal network. The performance metrics are computed by comparing predictions on the original data and the debiased data. Here is an example of how the data is used:

```
# Original data
X_train, X_test, y_train, y_test = train_test_split(
    df.drop(columns=['label']), df['label'])
# Debiased data (updated after each change in causal network)
X_train_deb, X_test_deb, y_train_deb, y_test_deb = train_test_split(
    df_deb.drop(columns=['label']), df_deb['label'])
# Model training and evaluation
model = LogisticRegression()
model.fit(X_train_deb, y_train_deb)
predictions = model.predict(X_test_deb)
# Compute performance metrics
accuracy = accuracy_score(y_test_deb, predictions)
```

These code snippets demonstrate the standard implementation of the ML models using ‘scikit-learn‘ to compute the performance metrics at each stage of the data debiasing process.

B Additional Figures

This appendix includes additional figures referenced in the main text.



Dataset Hiring (
Label Variable Job
Nominal Variables Gender Race Age Work E Major Grade F SAT sco College Job
P-value ● 0.01 ○ 0.05 ○ 0.10
stic Regression
g agency
Create Group: Enter group name here..

Fig. 7. Game Configuration: User Preferences Selection or Group Creation. The height of each rectangle indicates the percentage availability of the corresponding attribute in the dataset. Selected attributes’ rectangles will be filled with blue, partially or fully, depending on the percentage of the attribute’s inclusion in the group. Unselected attributes are filled with gray, indicating their availability for selection. The numeric percentage displayed at the top (e.g., 30%) represents the proportion of the dataset included in the selected group. Players need to provide a group name at the top and click ‘Save’ to create a group or their preferences.

C Participants Demographics and Backgrounds

The demographic information of the users is provided in Figures 8,9 and 10. Figures 11 and 12 give some insights on participants backgrounds and familiarity with related concepts.

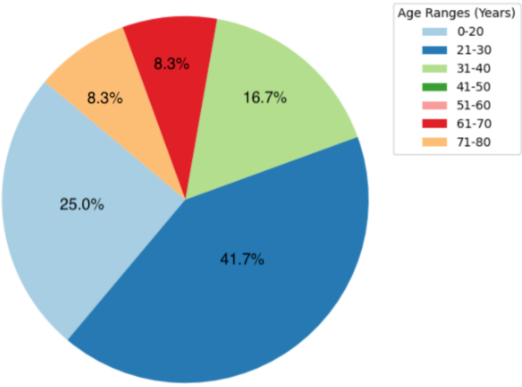


Fig. 8. Participants' age range.

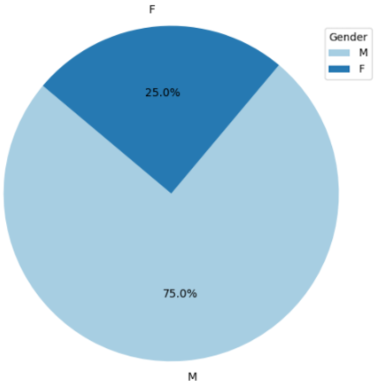


Fig. 9. Participants' gender. Provided options included "Non-Binary" and "Prefer Not to Answer" as well.

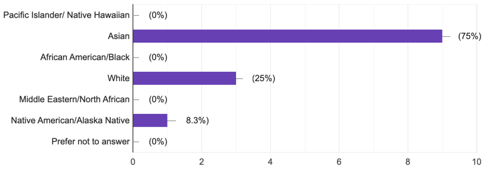


Fig. 10. Participants' ethnicity.

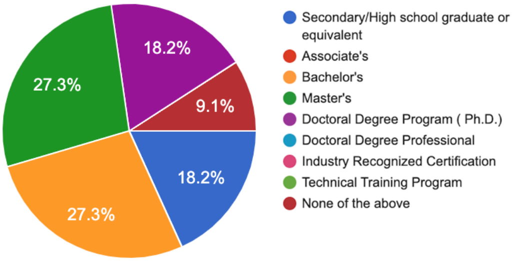


Fig. 11. Participants' educational level.

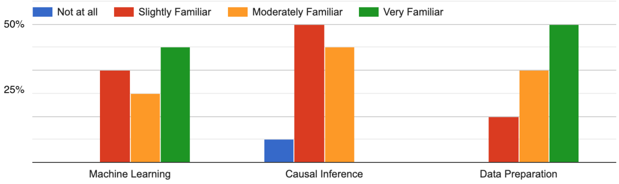


Fig. 12. Participants' familiarity with related concepts.

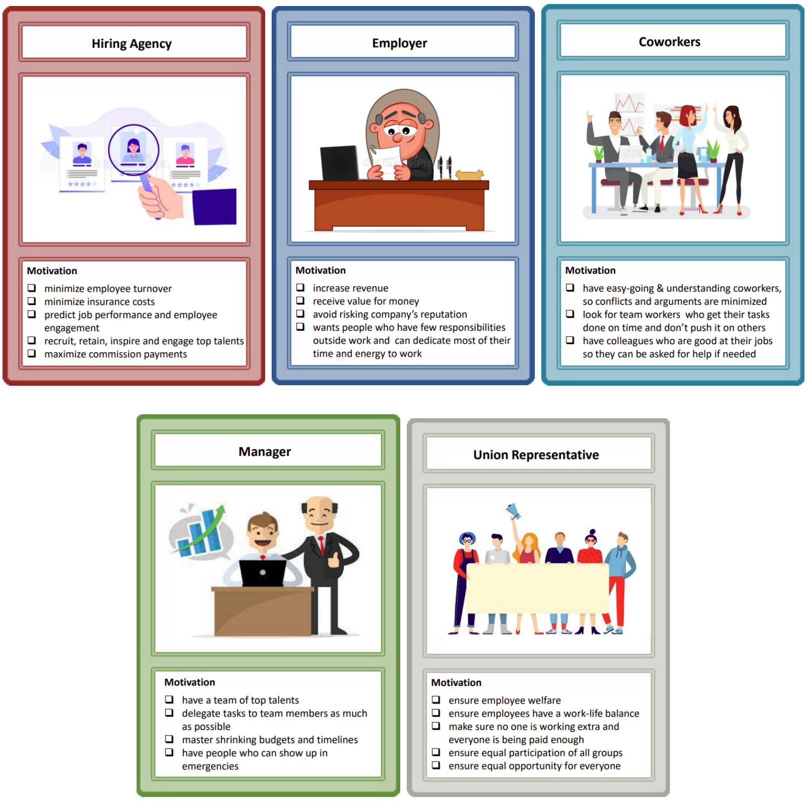
D Algorithm Computing the Game Metrics

Algorithm 1 Compute Game Metrics

```
$group$ ← Object containing attributes, players care about.
$data$ ← DataFrame-like object containing tabular data including target variable.
$columns$ ← data.columns
procedure ComputingGroups
    Input: $group$
    Output: Attribute wise percent care of each players
    Initialize attCare
    for $player$ from 1 to $group.length$ do                                    ▷ 1:n based index
        for col, val in group[player].items() do
            if data[col] == val then
                $attCare[key][val] \leftarrow attCare[key][val] + 1$
            end if
        end for
    end for
    Return $attCare$
end procedure
Initialize groups
groups ← ComputingGroups                  ▷ Calling ComputingGroups and assigning the result
procedure ComputingOutcome
    Input: $data$
    Output: Attribute wise job distribution
    Initialize $attJob$
    for col in columns do
        if data[col] == 0 then
            $attJob[col][0] \leftarrow data[data[col] == 0 \,\&\, data[data.target] == 1].shape[0]$
        else
            $attJob[col][1] \leftarrow data[data[col] == 1 \,\&\, data[data.target] == 1].shape[0]$
        end if
    end for
    Return $attJob$
end procedure
Initialize outcome
outcome ← ComputingOutcome                ▷ Calling ComputingOutcome and assigning the result
procedure ComputingAggregate
    Input: $groups, outcome$
    Output: Attribute wise hiring diferences.
    Initialize $attAggregate$
    for $player$ from 1 to $group.length$ do
        $attAggregate[player] \leftarrow groups[player] - outcome$
    end for
    Return $attAggregate$
end procedure
Initialize aggregate
aggregate ← ComputingAggregate            ▷ Calling ComputingAggregate and assigning the result
procedure ComputingTotaLossGain
    Input: $groups, outcome$
    Output: Attribute wise hiring diferences.
    Initialize $attAggregate$
    for $player$ from 1 to $group.length$ do
        $attAggregate[player] \leftarrow groups[player] - outcome$
    end for
    Return $attAggregate$
end procedure
Initialize totalLossGain
totalLossGain ← ComputingTotaLossGain     ▷ Calling ComputingTotaLossGain and assigning the result
Return $groups, outcome, aggregate, totalLossGain$
```

E Player's Roles

The goals and objectives for each role were developed by analyzing job descriptions and requirements on recruitment websites such as LinkedIn and Indeed. The preferences selected for the players in our initial three user studies were aligned with the goals illustrated in Figure 13. These preferences were not engineered to simplify reaching consensus by aligning the goals for all players. As indicated by the goals on the cards, some roles prioritize experience and talent, while others emphasize equal opportunities for all groups. There are goals and preferences that are aligned, as well as those that are in opposition, to ensure that the studies closely resemble real-world scenarios.



Hiring Agency

Motivation
- ❑ minimize employee turnover
- ❑ minimize insurance costs
- ❑ predict job performance and employee engagement
- ❑ recruit, retain, inspire and engage top talents
- ❑ maximize commission payments

Employer

Motivation
- ❑ increase revenue
- ❑ receive value for money
- ❑ avoid risking company's reputation
- ❑ wants people who have few responsibilities outside work and can dedicate most of their time and energy to work

Coworkers

Motivation
- ❑ have easy-going & understanding coworkers, so conflicts and arguments are minimized
- ❑ look for team workers who get their tasks done on time and don't push it on others
- ❑ have colleagues who are good at their jobs so they can be asked for help if needed

Manager

Motivation
- ❑ have a team of top talents
- ❑ delegate tasks to team members as much as possible
- ❑ master shrinking budgets and timelines
- ❑ have people who can show up in emergencies

Union Representative

Motivation
- ❑ ensure employee welfare
- ❑ ensure employees have a work-life balance
- ❑ make sure no one is working extra and everyone is being paid enough
- ❑ ensure equal participation of all groups
- ❑ ensure equal opportunity for everyone

Fig. 13. Cards shown on the game interface to remind players of their goals for a particular role.

Roles	Age	Gender	College Rank	Grade Point Average	Major	Work Experience	Race
Hiring Agency	Both groups	–	Elite	Above 3	Computer Science	Both groups	–
Employer	Above 42	Both groups	–	Above 3	–	Above 24	Both Groups
Manager	Above 42	Male	Elite	Above 3	Computer Science	Above 24	White
Coworkers	–	Both groups	Both groups	Above 3	Both groups	Above 24	–
Union Rep.	Both groups	Both groups	–	–	–	Both groups	Both Groups

Table 3. Preferences set for players in the first three user studies based on their roles. A line indicates that the role did not have any preference for that particular feature.

F Analysis of Edge Strength Adjustments and Player Preferences

The provided figure 14 illustrates the impact of adjusting the edge strength of the causal relationship between Gender and Job within the FairPlay framework. The x-axis represents the percentage change in edge strength, ranging from -100% to +100%. The y-axis indicates the number of job changes resulting from these adjustments. The blue and orange lines represent job changes for the two categories of gender, while the grey line shows the net difference in job changes, which remains around zero. This demonstrates that while jobs are redistributed between categories, the total number of jobs remains constant, preserving the system's integrity. D-BIAS generates datasets by simulating data distributions via the causal network. As explained in [23], when the weight of an edge in the causal diagram is decreased, it introduces more randomness into the corresponding antecedent node variable. For example, in figure 14, decreasing the edge weight of Gender to Job (where Gender=1 corresponds to males and Job = 1 corresponds to getting the job) will result in a more balanced distribution of females getting the job, given all other qualities being equal. Furthermore, by adding randomness instead of simply removing the sensitive Gender variable it also lowers proxy biases (if present) in variables downstream from them in the causal graph.

In the context of player preferences, adjustments favoring a particular category (e.g., Category 2) may lead to more jobs and game points for that category, negatively impacting players who prioritize the opposite category (e.g., Category 1). Thus, players must negotiate and find a balance that considers the goals of all participants. Changes in other variables within the causal network will also impact job distribution, requiring players to consider the broader network context. The point where the job change lines cross the x-axis represents the default job distribution for this variable, and this origin will shift based on the default distributions. This analysis assumes that other parts of the network remain static and are not altered during this specific analysis. This detailed analysis highlights how FairPlay enables users to explore the impact of edge strength adjustments on job distribution, maintaining coherence and balance in gameplay and aiding in informed decision-making for fairer job outcomes.

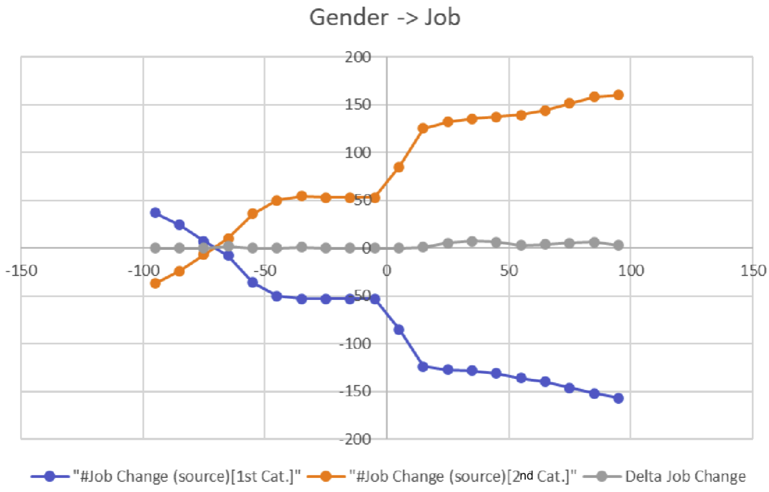


Fig. 14. Impact of Edge Strength Adjustments on Job Distribution for the Gender -> Job Relationship. The x-axis represents the percentage change in edge strength, while the y-axis indicates the number of job changes. The blue line shows job changes for one gender category, the orange line for the other gender category, and the grey line represents the net difference in job changes.

G The Concept of Preference Optimization in FairPlay

It is challenging to definitively determine whether fairness is objective or subjective. However, it is clear that fairness is often contested. This section explores potential approaches to address this issue.

G.1 Concept Overview. The principle underlying the FairPlay system is rooted in the notion of preference optimization, as introduced by Eliezer Yudkowsky [56] in the context of AI safety. Unlike traditional approaches to reducing bias or enhancing fairness, FairPlay aligns the system's biases with the preferences of its users. This approach recognizes that what constitutes a "sensitive variable" or an appropriate decision criterion is ultimately a matter of user preference, rather than an objective standard.

G.2 Analogy to Legislative Bodies. The functioning of FairPlay can be likened to the dynamics within legislative bodies such as senates and parliaments. Representatives in these bodies do not follow a fixed, universally agreed-upon law-making algorithm. Instead, they evolve their strategies over time based on the preferences and interests of their constituents. This adaptive process mirrors how FairPlay aligns its outputs to user biases, allowing for shifts in priorities and strategies.

Just as legislators may oscillate between different principles and levels of engagement, FairPlay's system is designed to adapt to the changing preferences of its users. The ultimate goal, similar to the preference for deliberation over dictatorship in legislative processes, is to ensure that the AI system reflects the collective preferences of its users rather than imposing a singular, potentially arbitrary standard of fairness.

G.3 Implications for AI Alignment. The concept of aligning AI systems to user preferences highlights a fundamental shift in how we think about fairness and bias in automated systems. It suggests that achieving fairness may be less about finding an objective measure of bias and more about ensuring that the system's outputs are consistent with the values and preferences of those it serves. This approach acknowledges the complexity and variability of values and aims to create a more flexible and responsive AI system.

By situating the discussion of FairPlay within this broader context of preference optimization and legislative analogy, we underscore the importance of user-aligned AI and the potential limitations of traditional fairness metrics. This perspective not only informs the design and implementation of FairPlay but also contributes to the ongoing discourse on AI ethics and alignment.

Received January 2024; revised July 2024; accepted October 2024