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

Tina Behzad*
Department of Computer Science
Stony Brook University
tbehzad@cs.stonybrook.edu

Mithilesh Kumar Singh*
Department of Computer Science
Stony Brook University
mkssingh@cs.stonybrook.edu

Anthony Ripa*
Department of Computer Science
Stony Brook University
aripa@cs.stonybrook.edu

Klaus Mueller
Department of Computer Science
Stony Brook University
mueller@cs.stonybrook.edu

Abstract

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 perspective 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. We have conducted user studies that demonstrate the success of FairPlay, with users reaching consensus within about five rounds of gameplay, illustrating the application's potential for enhancing fairness in AI systems.

1 INTRODUCTION

Fairness remains an elusive goal in our increasingly data-driven world, hindered by the Impossibility of Fairness [17], emerging 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 [27, 10], but ultimately, we must keep in mind (as noted in [9]) there is no universal means to measure fairness, and at present no clear guideline(s) on which measures are "best" [8].

This problem's essence is deeply rooted in context-specific nuances, making it crucial to tailor solutions to the individual characteristics 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 to assess different fairness metrics and determine the best strategies for mitigating bias in datasets or models [32]. 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 to bias mitigation activities.

*Equal contribution.

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

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 tasks. We build our software on the foundation of a previously published web-based software, D-BIAS [19]. 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 way of a causal model initially derived from the original (potentially biased) ADS training data.

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 task, promoting a consensus-based methodology. Then, once consensus has been reached, the de-biased data generated by the causal model can be used to train any ADS.

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

Different fairness metrics and definitions have been developed to quantify and measure bias in machine learning models [8, 31, 13, 26, 4, 20]. 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 [26, 16]. There is no consensus on a single most appropriate definition of fairness [18]. Determining the right measure to be used must take into account the proper legal, ethical, and social context [31]. 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 [36]. 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 [12].

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 [32]. Some of these methods include: Silva [43], FairVis [6], FairRankVis [42], DiscriLens [40], FairSight [1], What-If toolkit (WIT) [41], Aequitas[33], AI Fairness 360 (AIF360) [3] and D-BIAS [19]. 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 could interpret causal networks and found them helpful in identifying algorithmic bias [43]. 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.

Various approaches have been proposed to achieve fairness in machine learning [8, 31]. Pre-processing [7, 24], in-processing [39] and post-processing methods [25] each tackle the problem at different stages of a machine learning pipeline. 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 [22, 32]. 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.

Consensus-building mechanisms have been extensively studied in fields such as multi-agent systems [28], social choice theory [29], and deliberative decision-making [38] to address challenges. The most important challenge is the aim to reach an agreement or consensus among multiple stakeholders with diverse preferences and perspectives. Innes argues a number of conditions need to hold for a process to be labeled consensus building [23]. If these do not hold, failure of various kinds is likely. 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 are among these conditions. 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 [46, 15, 14].

The lack of collaborative tools in fairness visualization is a significant gap, given the importance of collective approaches in this field [21]. Fairness is a concept that varies greatly depending on the context and individual perspectives [35]; a single person or a non-interactive tool might overlook these nuances. Collaborative visualization tools, like the FairPlay system discussed in this paper, drive active participation and collective decision-making by integrating cooperative elements and gamification [21]. 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 [34], these two elements are essential for reaching a consensus. Without such tools, opportunities for richer, more inclusive discussions and solutions that could promote more effective implementation of fairness in machine learning systems by leveraging the collective intelligence and insights of a wider group of stakeholders, are missed [4].

3 FairPlay Game Design

The transformation of the D-BIAS platform into FairPlay, a collaborative environment, involved an intricate design process aimed at creating an engaging experience. This redesign aligns with the guidelines set forth in the fairness toolkit rubric [32] and incorporates elements identified as crucial for successful consensus-building in related work [23]. The game mechanics and user interface were thoughtfully developed to emphasize interactivity and gamification, aiming to cultivate a cooperative atmosphere where players can actively participate in modifying the causal graph. 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.

Game Configuration. Before entering the game, the configuration page allows players to adjust multiple variables including but not limited to the dataset, the algorithm used for monitoring classification performance metrics, and their roles as illustrated in Figure 1. More information about game configurations is available in Appendix A.1

The Game. Upon entering the game, with the chosen game configuration, the game constructs the initial causal network, along with 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. The game interface is illustrated in Figure 2. Details about each of these panels is available in Appendix A.2.

FairPlay integrates into a machine learning pipeline as follows. Initially, the default causal model is constructed using the raw data intended for training the ML model (first module in Figure 4). The game then begins, with players iteratively tuning the causal model to adjust the default outcomes according to their priorities (center module in Figure 4). The game concludes once the players have met their objectives, resulting in debiased data, which can then be used to train an ML model. It is assumed that the ML model itself does not introduce new biases; otherwise, an additional ML model debiasing step would be necessary. The upper-right module in Figure 4 focuses on analyzing user data generated during the game. Appendix B provides technical details on the visual interface and data storage for post-game analysis as well as a diagram presenting the flow of the game.

4 Experiments and Results

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. To answer this question we conducted four studies with different groups. Details on the users and the study setup are available in Appendix C.

4.1 Results

In our studies, the game continued for 2 to 5 rounds, lasting about 60-90 minutes. By the end of each game, all players achieved improved metrics reflective of their objectives. In the rest of this section, we briefly discuss the results. Details and more information are available in Appendix D.

Evaluating classifiers on the final debiased datasets shows that models trained on these datasets achieve better individual fairness and improved parity, demonstrating the effectiveness in reducing bias. Although there is a slight decrease in accuracy and F1-score, this reflects a deliberate trade-off for increased fairness, consistent with existing research on the accuracy-fairness tradeoff [30][45].

The post-game analysis showed that participants perceived the FairPlay game as contributing to a fairer system, with an average agreement rating of 3.7 out of 5 in response to the statement, "I think that the activities led to a fairer system." Players engaged actively, modifying the causal network according to their priorities, with their decisions shaped by their ethical leanings—Deontologists focused on down-weighting sensitive variables, while Consequentialists adjusted parameters to achieve better outcome metrics. Despite differing approaches, all user studies ended in consensus, demonstrating the game's effectiveness in facilitating agreement through intuitive features and insightful metrics. Additionally, the game had a positive, though moderate, impact on educating participants about fairness and bias, as indicated by an average rating of 3.3 out of 5 on a related survey question.

User satisfaction was measured using the System Usability Scale (SUS)[5], where it received a score of 68.05, indicating above-average usability. Users found the tool's functions well-integrated and user-friendly, with minimal inconsistency, reflecting its effectiveness and cohesive design.

In all four studies, there was a noticeable decrease in edge thickness from the default in the first round, indicating that players quickly acted on their initial assessments of the causal network. After the first round, changes became less drastic, with edge weights stabilizing, suggesting that users reached an early consensus or satisfaction with the network. While the process was consistent across studies, the final outcomes varied, reflecting the subjective nature of the decisions made by participants. The final causal networks showed sparse connections, especially for sensitive attributes like Age, Race, and Gender, resulting in simple network topologies. Players generally avoided unfavorable outcomes for most attributes, though some trade-offs were accepted, particularly concerning individuals with low work experience. Overall, the studies suggest that while satisfying all definitions of fairness is challenging, players were able to reach a consensus on key factors influencing decisions, optimizing benefits for their desired groups without adhering to a specific fairness definition.

5 Discussion

Our platform was designed to help users collaboratively determine the causal structure of their datasets, a task that would be significantly challenging without such a tool. Traditional approaches, like stakeholder discussions, lack the ability to modify and observe changes in the causal network, making consensus unlikely and poorly informed. Our interface addresses these challenges by structuring the negotiation process, ensuring all perspectives are considered, and providing users with the information needed for informed decisions. The Consensus Reaching Process (CRP) aims to achieve two key goals: reflecting better partial agreement and guiding the process until a high level of consensus is reached among decision-makers [44]. In all four user studies, stakeholders expressed satisfaction with the causal network. The game concluded only when every stakeholder confirmed complete satisfaction with the current state, meeting the CRP's conditions. This confirms the success of our platform in achieving consensus among participants.

6 Conclusion

As algorithms increasingly serve as decision-makers, auditing them for ethical concerns is critical, yet research shows that meeting all fairness criteria is often challenging. This highlights the need for context-specific audits, ideally conducted by diverse teams to counteract individual biases. FairPlay supports this collaborative approach, enabling stakeholders or domain experts to systematically identify relevant features in datasets. The varied outcomes and distinct causal structures from the four user studies demonstrate the importance of such tools, as different groups may reach unique agreements. The consensus achieved in all studies confirms FairPlay's effectiveness, highlighting the importance of tools that support informed, collaborative decision-making in algorithmic audits.

7 Acknowledgments

This research was partially supported by NSF grants IIS 1941613 and IIS 1527200. We also thank the SUNY Research Seed Grant (RSG) Program for their generous support. ChatGPT was utilized to generate sections of this work, including text and code.

References

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

[2] Larry Alexander and Michael Moore. Deontological ethics. In *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, 2007.

[3] 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. Ai fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias. *IBM Journal of Research and Development*, PP: 1–1, 09 2019. doi: 10.1147/JRD.2019.2942287.

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

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

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

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

[8] Simon Caton and Christian Haas. Fairness in machine learning: A survey. *ACM Comput. Surv.*, 56(7), apr 2024. ISSN 0360-0300. URL `https://doi.org/10.1145/3616865`.

[9] Flavio Chierichetti, Ravi Kumar, Silvio Lattanzi, and Sergei Vassilvtiskii. Matroids, matchings, and fairness. In Kamalika Chaudhuri and Masashi Sugiyama, editors, *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, pages 2212–2220. PMLR, 16–18 Apr 2019. URL `https://proceedings.mlr.press/v89/chierichetti19a.html`.

[10] Alexandra Chouldechova. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big Data*, 5(2):153–163, 2017. doi: 10.1089/big.2016.0047. URL `https://doi.org/10.1089/big.2016.0047`. PMID: 28632438.

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

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

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

[14] Martin J Eppler. Facilitating knowledge communication through joint interactive visualization. *J. Univers. Comput. Sci.*, 10(6):683–690, 2004.

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

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

[17] Sorelle A. Friedler, Carlos Scheidegger, and Suresh Venkatasubramanian. The (im)possibility of fairness: Different value systems require different mechanisms for fair decision making. *Communications of the ACM*, 64(4):136–143, April 2021. doi: 10.1145/3433949. URL https://dl.acm.org/doi/10.1145/3433949.

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

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

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

[21] Jeffrey Heer and Maneesh Agrawala. Design considerations for collaborative visual analytics. *Information Visualization*, 7(1):49–62, mar 2008. ISSN 1473-8716. doi: 10.1145/1391107.1391112. URL https://doi.org/10.1145/1391107.1391112.

[22] Kenneth Holstein, Jennifer Wortman Vaughan, Hal Daumé, Miro Dudik, and Hanna Wallach. Improving fairness in machine learning systems: What do industry practitioners need? In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI '19, page 1–16, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450359702. doi: 10.1145/3290605.3300830. URL https://doi.org/10.1145/3290605.3300830.

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

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

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

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

[27] Jon Kleinberg, Jens Ludwig, Sendhil Mullainathan, and Ashesh Rambachan. Algorithmic fairness. *AEA Papers and Proceedings*, 108:22–27, May 2018. doi: 10.1257/pandp.20181018. URL https://www.aeaweb.org/articles?id=10.1257/pandp.20181018.

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

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

[30] Aditya Krishna Menon and Robert C. Williamson. The cost of fairness in binary classification. In Sorelle A. Friedler and Christo Wilson, editors, *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, volume 81 of *Proceedings of Machine Learning Research*, pages 107–118. PMLR, 23–24 Feb 2018. URL https://proceedings.mlr.press/v81/menon18a.html.

[31] Dana Pessach and Erez Shmueli. A review on fairness in machine learning. *ACM Comput. Surv.*, 55(3), feb 2022. ISSN 0360-0300. doi: 10.1145/3494672. URL https://doi.org/10.1145/3494672.

[32] Brianna Richardson, Jean Garcia-Gathright, Samuel F. Way, Jennifer Thom, and Henriette Cramer. 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, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450380966. doi: 10.1145/3411764.3445604. URL https://doi.org/10.1145/3411764.3445604.

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

[34] Thomas J. Scheff. Toward a sociological model of consensus. *American Sociological Review*, 32(1):32–46, 1967. ISSN 00031224. URL http://www.jstor.org/stable/2091716.

[35] Andrew D. Selbst, Danah Boyd, Sorelle A. Friedler, Suresh Venkatasubramanian, and Janet Vertesi. Fairness and abstraction in sociotechnical systems. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, page 59–68, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450361255. doi: 10.1145/3287560.3287598. URL https://doi.org/10.1145/3287560.3287598.

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

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

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

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

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

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

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

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

[44] Hengjie Zhang, Yucheng Dong, Francisco Chiclana, and Shui Yu. Consensus efficiency in group decision making: A comprehensive comparative study and its optimal design. *European Journal of Operational Research*, 275(2):580–598, 2019. ISSN 0377-2217. doi: https://doi.org/10.1016/j.ejor.2018.11.052. URL https://www.sciencedirect.com/science/article/pii/S0377221718309937.

[45] Han Zhao and Geoff Gordon. Inherent tradeoffs in learning fair representations. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/file/b4189d9de0fb2b9cce090bd1a15e3420-Paper.pdf.

[46] Roshanak Zilouchian Moghaddam, Brian P Bailey, and Christina Poon. 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*, pages 259–276. Springer, 2011.

A Game Design Details

A.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.

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 [19], 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. Once the dataset is selected, its features are displayed for the user and a label (outcome) variable is chosen. The initial causal network (a directed acyclic graph) is inferred using the PC algorithm, a causal discovery method named after its creators Peter Spirtes and Clark Glymour [37]. 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.

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 'Create Group' button, a pop-up page appears for selecting features and their respective values for user preferences selection or group creation. 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.

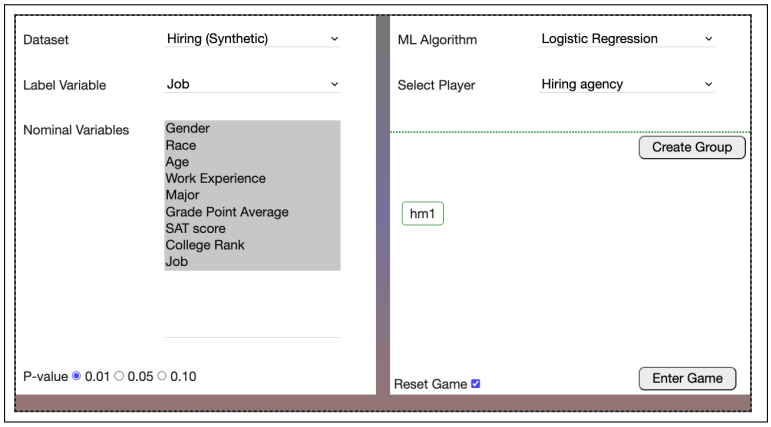

Figure 1: FairPlay: Game Configuration. This is the main configuration panel of the application. For the results discussed in this paper, the groups were already pre-configured to make the played games comparable.

A.2 The Game

A.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.

Limiting the range of edge weight adjustments from -100 percent to +100 percent rather than allowing users to change it to any value is a deliberate design choice. By presenting a bounded range, users can focus on the relative impact of the edge weights rather than being overwhelmed by the entire numerical spectrum. Additionally, the restricted range ensures that users adhere to practical boundaries and prevents extreme or unrealistic adjustments that could disrupt the overall fairness dynamics of the game. By defining a reasonable range, the design maintains the coherence and balance of the gameplay, preventing unintended consequences that may arise from arbitrary edge weight modifications.

A.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. The pause for reviewing metrics ensures players consider the impact of their changes and weigh any potential trade-offs, thereby promoting thoughtful and deliberate decision-making. Moreover, the game's design restricts players from making further alterations to the network once they've clicked "Apply". This rule is implemented to prevent players from continuously adjusting the causal network, ensuring that every participant gets a fair opportunity to play. Within each round, players are allowed a single opportunity to modify as many edges as they wish. After making these changes and clicking "Apply", they can then observe the outcomes of their actions, maintaining a balanced and orderly flow of FairPlay.

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, 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.

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 [23]. 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 [2]. To assist these users in remembering their objectives while playing these roles, we incorporated these cards as a helpful reminder [3].

B 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 4 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 4). 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 4). The game ends when the players have achieved their goals which results in the Debiased Data. Figure 3 shows the overall flow of the game, from when the players enter the game to when the game ends. The Debiased Data can then be used to train any ML model. Here it is assumed that the ML model will not introduce biases on its own, else an independent ML model debiasing step would be required. The upper-right-most module of Figure 4 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.

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

[3]The cards are displayed in Appendix F.

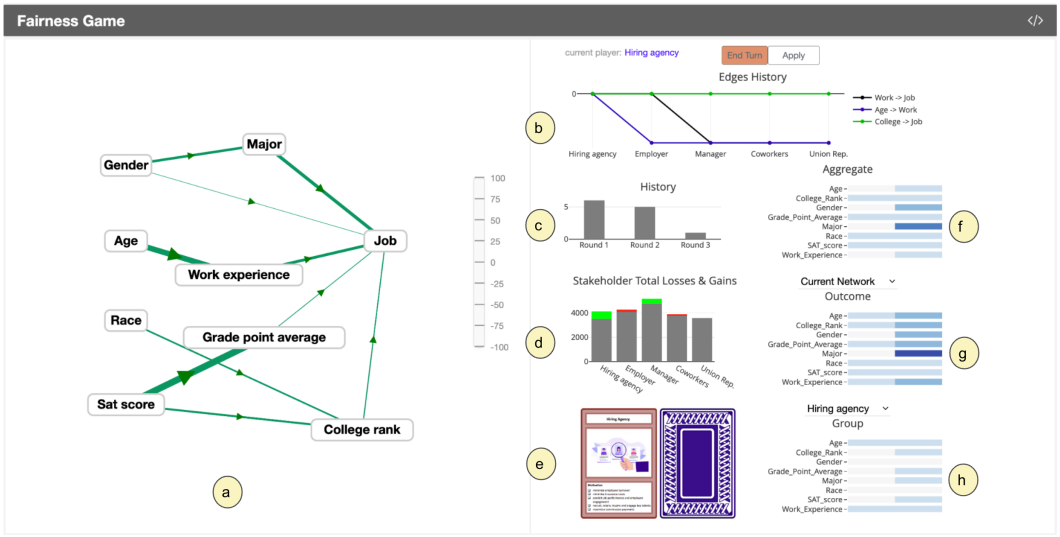

Figure 2: FairPlay Game Interface. The components are (a) causal network link editor, (b) edge history chart, (c) aggregate edge history chart, (d) stakeholder total loss and gain chart, (e) active stakeholder card stack, (f) aggregate attribute disparity chart, (g) attribute outcome chart, (h) stakeholder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

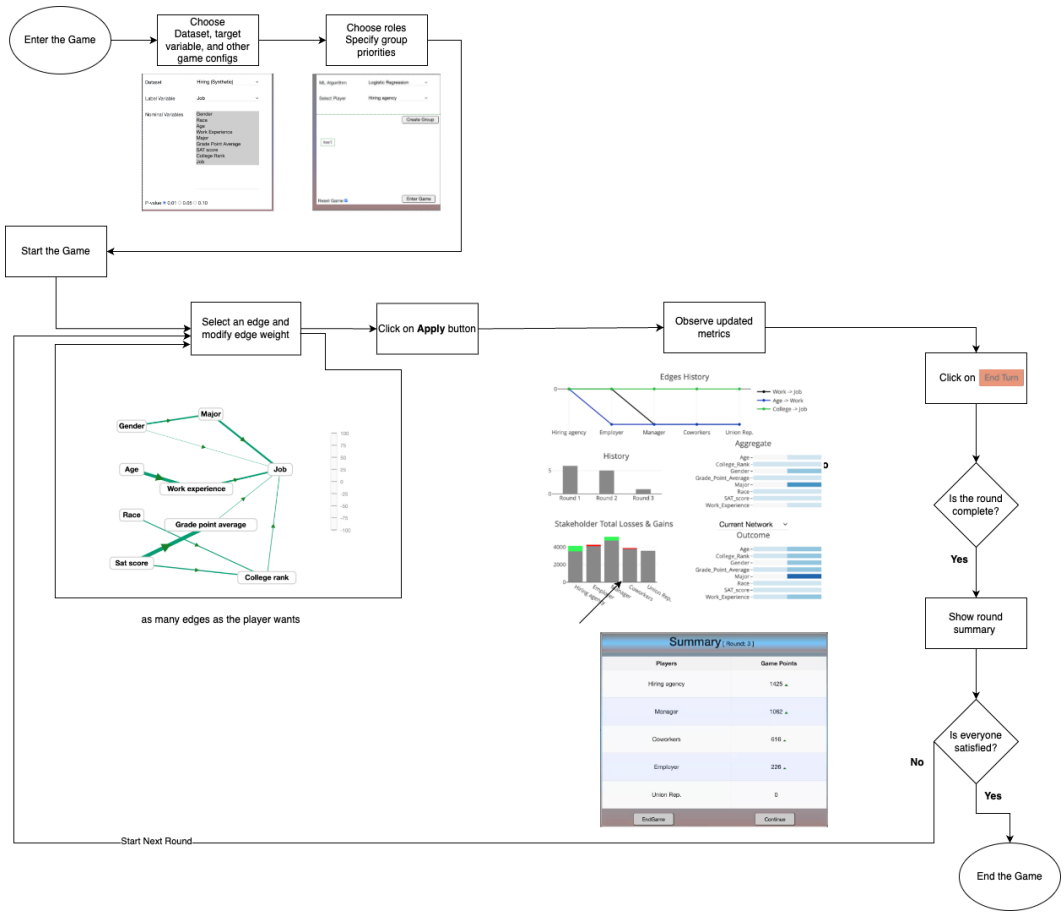

Figure 3: FairPlay game diagram from beginning to end.

B.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 [11]. 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 [19].

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. 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.

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 B.3).

B.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.

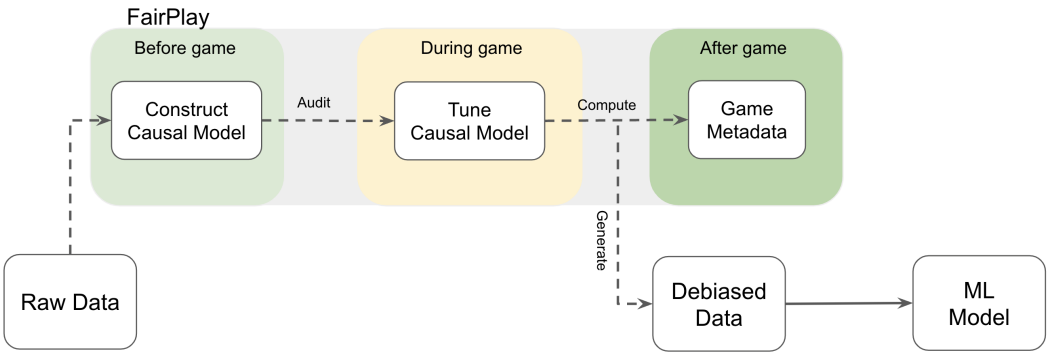

Figure 4: FairPlay System Overview. Detailed explanations are provided in the main text above.

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 [19] (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).

Algorithm 1 Compute Game Metrics

$group \leftarrow$ Object containing attributes, players care about.
$data \leftarrow$ DataFrame-like object containing tabular data including target variable.
$columns \leftarrow$ 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 $\leftarrow$ 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 $\leftarrow$ 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 $\leftarrow$ 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 $\leftarrow$ COMPUTINGTOTALOSSGAIN ▷ Calling ComputingTotaLossGain and assigning the result
Return $groups, outcome, aggregate, totalLossGain$

B.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.

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 A.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.

C Experiments

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. More details about the demographic information of participants and their backgrounds are available in Appendix E.

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 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 role's goals and preferences is discussed in Appendix F.

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.

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.

D Results

D.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 displayed in Table 1. 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 [30][45].

	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 1: 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.

D.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) [2]. 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 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.

D.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 [5] (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 5, 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.

D.4 Insights

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

Figure 6 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,

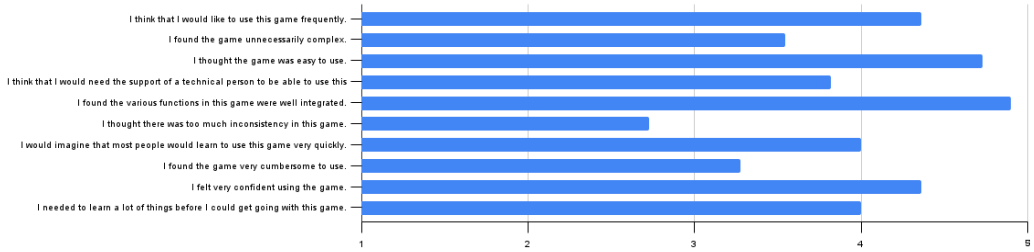

Figure 5: FairPlay Feedback Results
FairPlay Feedback Results

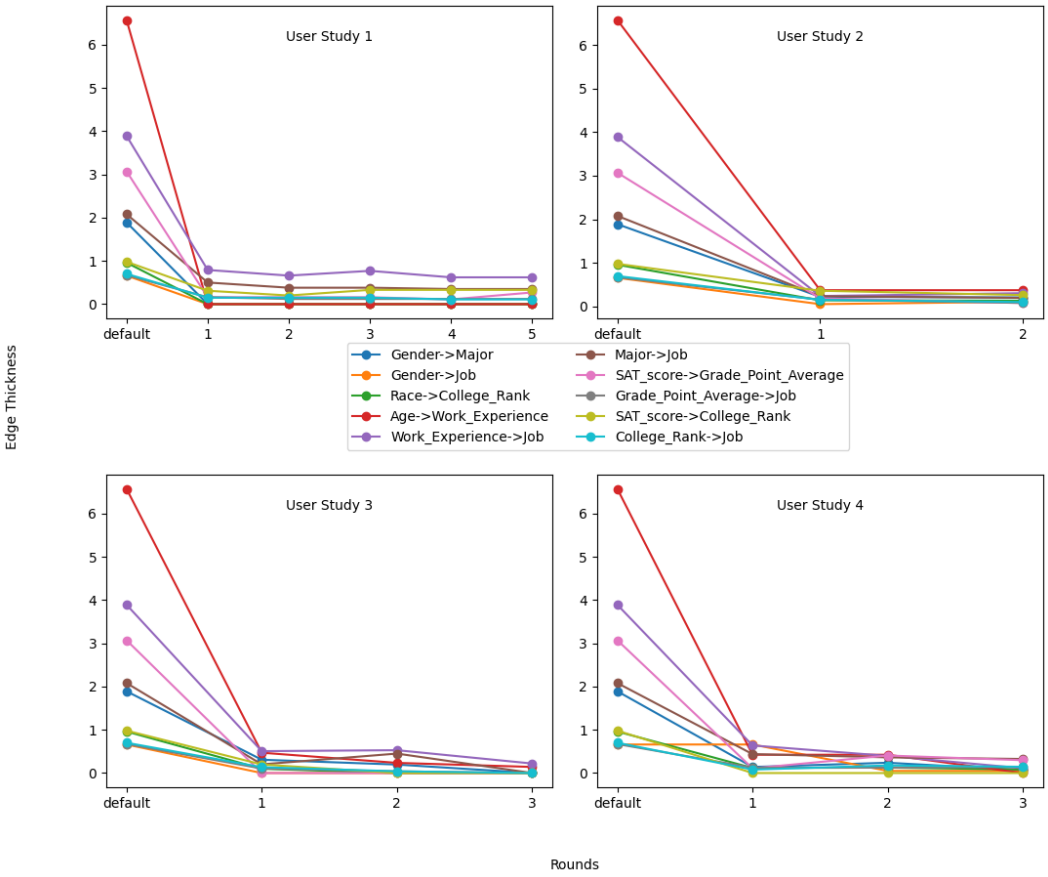

Figure 6: FairPlay User Studies: Edge Weights vs. Round for all User Studies

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 7 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.

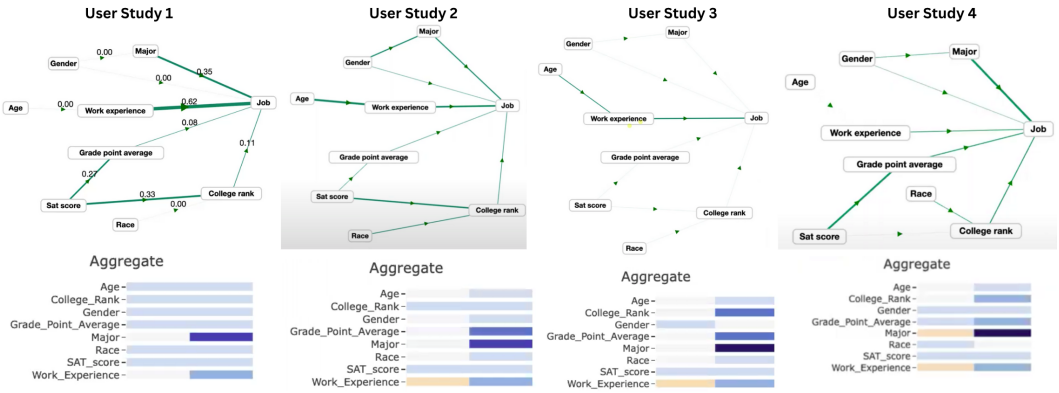

Figure 7: FairPlay Matrix. This figure compares the causal networks created by each set of players (top row) and the aggregate attribute disparity charts (bottom row)

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 a consensus (The same holds true for Major in the fourth study).

Drawing insights from Figure 7 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.

E 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.

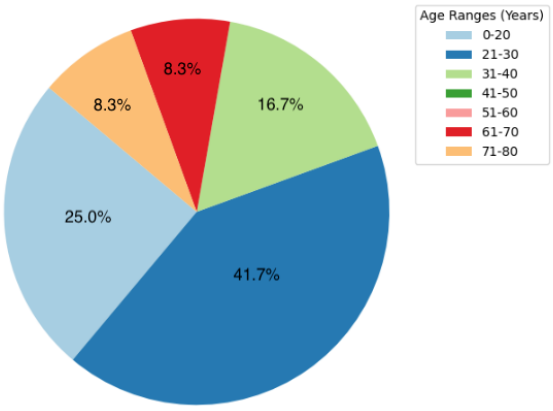

Figure 8: Participants' age range.

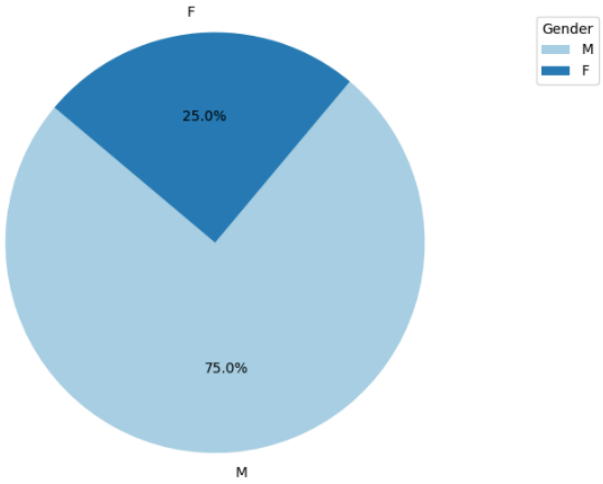

Figure 9: Participants' gender. Provided options included "Non-Binary" and "Prefer Not to Answer" as well.

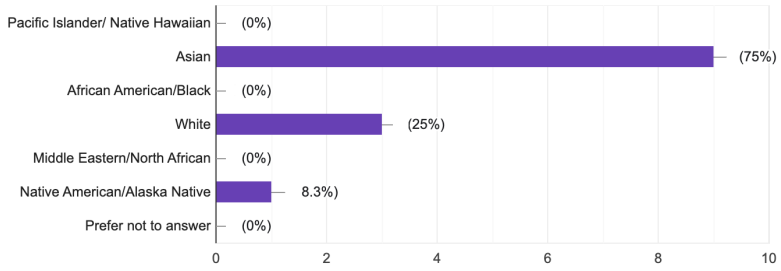

Figure 10: Participants' ethnicity.

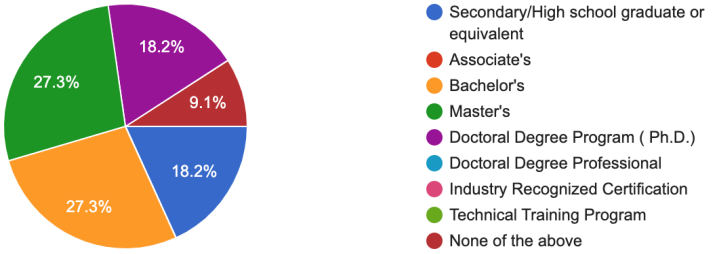

Figure 11: Participants' educational level.

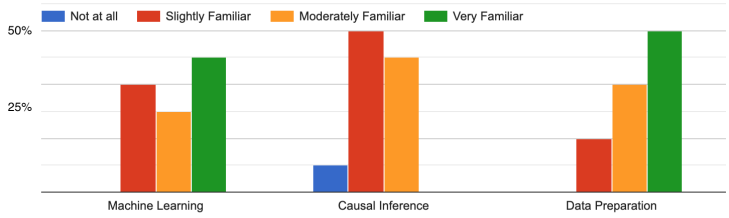

Figure 12: Participants' familiarity with related concepts.

F 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.

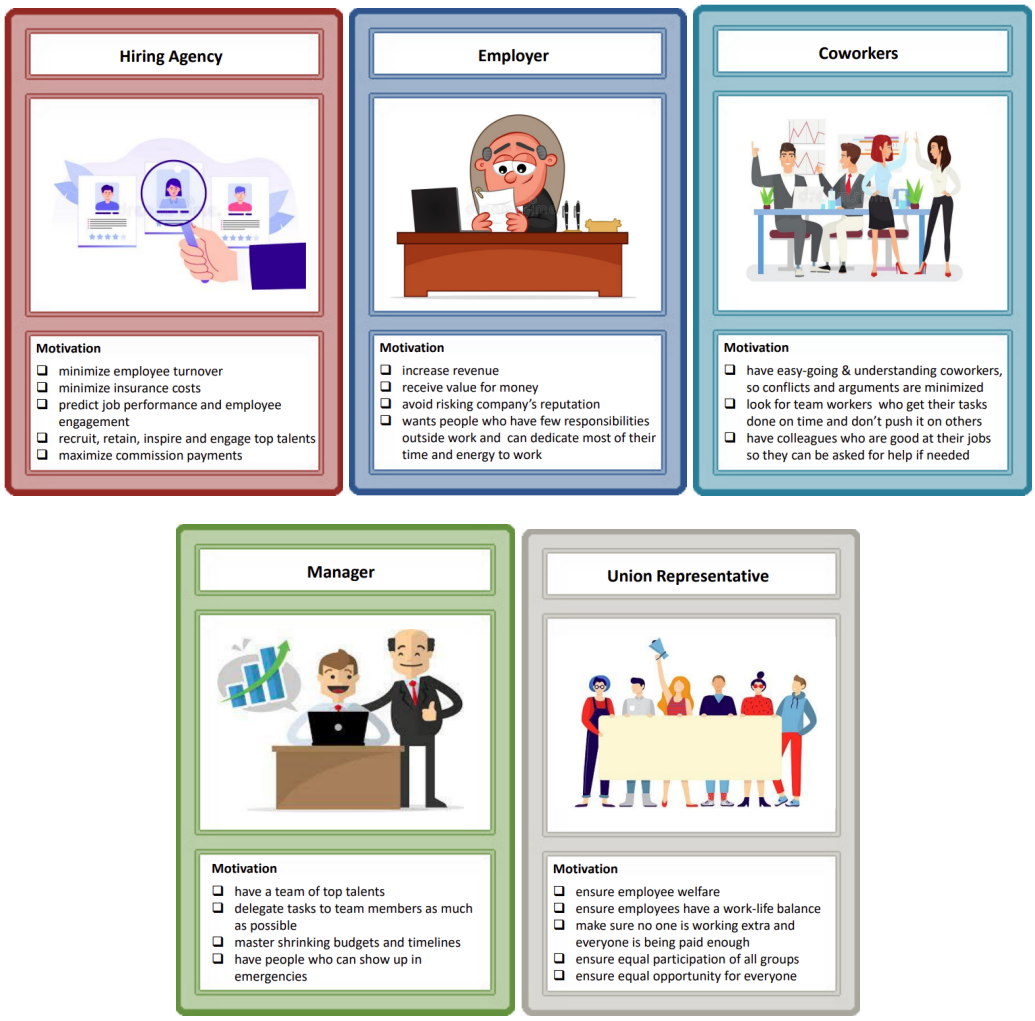

Figure 13: Cards shown on the game interface to remind players of their goals for a particular role.

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 2: 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.