FAIRNESS IN RECOMMENDER SYSTEMS: BALANCING BIAS IN ACADEMIC PAPER SELECTION

by

Zachary Bergin

An Honors Thesis

Submitted to the Department of Electrical Engineering and Computer Science

University of Arkansas

In Partial Fulfillment of the Requirements

For Graduating with Honors

May 2025

Abstract

Bias in academic paper selection remains a consistent issue, even within processes designed to promote fairness, such as double-blind peer review, bias stays persistent. In this paper we investigate demographic bias while particularly focusing on racial bias in the process of selecting academic papers and explore the impact of fairness aware recommender systems on the demographic parity. To build an effective system our focus is on the Special Interest Group on Computer Human Interaction (SIGCHI) a pillar in the community, we develop a neural network-based recommender system that uses real demographic data collected by other systems withing the context of SIGCHI and evaluates fairness from using several methods. Using different parameters such as a fairness value of lambda, our system accepts that there will a tradeoff between utility which is actually measured by the h-index and fairness which is a comparison to the overall paper pool. Through a series of experiments across varying methods and number of papers selected, this paper demonstrates that it is possible to close the gap on bias by improving representation of marginalized groups in paper selection while maintaining appropriate quality of papers.

Table of Contents

1.0 Introduction

1.1 Motivation

Understanding human behavior is crucial in acknowledging that individuals hold onto both conscious and subconscious biases. These biases often come from over generalizations and come to the surface in multiple ways, including demographic based prejudices related to gender, and race. Bias is a product to inaccuracy meaning it represents a form of imprecision in data gathering and measurement [1]. It can be classified into two main types, being implicit and explicit. Implicit bias which is the most common happens to be involuntary it shapes perceptions and in effect actions without direct awareness. Explicit bias on the other hand involves consciously held prejudices that can be identified and articulated [2].

Bias is specifically relevant in the selection of academic papers for publication in journals and conferences. Despite strong efforts to promote fairness with processes including double blind reviews, bias persists due to the difficulty of keeping the paper completely anonymous. Minority groups often face roadblocks along the way such as systemic barriers to recognition, which will suppress diversity and leave innovation stagnant. To foster a better environment for strong innovation and improve creativity it’s important to diversify the selected papers for these academic conferences. However, current recommender systems can intrinsically shoulder biases that tip the process of selection.

To gain more insight on bias we can look at a non demographic bias. The bias against resubmitted papers, Stelmakh et al, saw this trend within the ACM context. They noted that the acceptance rates of papers known to be resubmission, although good papers, were denied at a much higher rate. We can take this and understand the gap that these recommender systems can fill in the effort to shoulder bias.[6].

Several solutions are being tried and tested however the development of algorithmic frameworks that are responsible for balancing the bias in these selection processes is the solution we will focus on. These frameworks should continuously adapt and incorporate the latest advancements in machine learning and artificial intelligence. Organizations such as the Special Interest Group on Computer Human Interaction (SIGCHI), a subgroup of the Association for Computing Machinery (ACM), are just one of several impactful groups we can take notice from for their academic paper acceptance and committee membership. By keeping an eye on bias in academic paper selection and following algorithmic solutions, this research aims to contribute to a fairer and more inclusive recommender system.

1.2 Goals

The primary objectives of this study are:

To examine and quantify bias in academic paper selection processes.

To implement and test a fairness aware recommender framework within the context of SIGCHI paper selection.

2.0 Related Work

This section provides an overview of related work found on bias in recommender systems and academic paper selection. As we should be diving into the intersection of the problem and the technology we can use.

2.1 Bias in Academic Paper Selection

Bias in academic peer review and paper selection has been an important topic of discussion and in return is widely documented. Despite the use of double-blind reviews, authorship can often be inferred based on the writing style, citation patterns, and institutional affiliations. Studies have shown that underrepresented minorities including women and others have received disproportionately lower acceptance rates compared to their majority group counterparts. Alsaffar et al. [3] have highlighted that reviewer biases can continue even in the settings that have been under anonymity, and they propose fair recommender algorithms that put diversity in a position of power while maintaining selection utility.

Beyond reviewer bias, demographic features such as where the person is from, and their university also play a prominent role in skewing a fair selection of authors. Studies show that authors from highly prestigious research institutions or from top ranked universities are more likely to be selected and accepted above others, even when paper quality might be lower. This bias of choosing based off affiliation is often left over from the perceived credibility that comes with being a member of a high prestige institution, allowing for an uneven playing field for researchers from not as prestigious universities or developing regions.

Another form of bias arises against early career researchers from the positioning of review committees. Papers that challenge on prevailing rules or bring unconventional approaches may face greater denial during the review process, leading to a system that has a preference for established ideas over newer contributions. This issue disproportionately affects early career researchers and those working in newer fields, where the standard for review criteria may not fully appreciate the value of their contribution.

Contributions to mitigate the bias established in academic paper selection includes diverse reviewer pools and algorithms that attempt to balance fairness and quality of selections. However, these measures can fall short because of the complexity of biases and its strong presence in academics. Therefore, the development of a fairness aware recommender system, that will be implemented in this paper, will allow users to target and introduce a more systematic and data driven analysis to provide an acceptable paper selection outcome.

2.2 Bias in Recommender Systems

Previous studies have explored how biases emerge in recommender systems due to algorithmic feedback loops and data that is historically imbalanced. Fairness-aware algorithms attempt to diminish these biases through techniques such as re-weighting training samples, modifying ranking algorithms, or enforcing demographic constraints. A recent study by Alsaffar et al. [3] introduces multidimensional fairness in paper recommendation by incorporating multiple protected variables, such as gender, ethnicity, career stage, university rank, and geolocation, into an author profile. Their approach contrasts with traditional fairness-aware algorithms that typically consider only one protected attribute. This study also focuses on SIGCHI as the selected paper pool. This paper does an excellent job in reducing bias in recommender systems. However, we should investigate technology that is gathering more attention, neural networks.

2.3 Algorithmic Fairness Approaches

Fairness through algorithmic decision making has been implemented and studied across multiple different areas, including hiring, loans, and education. In the context of recommender systems, methods such as demographic parity, equalized odds, and disparate impact minimization have been used to balance bias. Alsaffar et al. [3] propose three fairness-aware recommendation methods that are, Overall Diversity, Round Robin Diversity, and Multifaceted Diversity. Their results demonstrate that these methods increase diversity in accepted papers with minimal if any impact on the loss of paper utility, suggesting that fairness and quality can be balanced effectively.

Demographic parity ensures that selected candidates accurately reflect the distribution of different demographic groups from the original population. This way of addressing fairness helps decrease the intensity of selection bias but may result in minor reductions of utility if fairness constraints override a complete utility selection process.

An important metric for evaluating fairness in recommender systems is the F-measure, which balances fairness and utility by considering both the gain in diversity and preserving of quality. The F-measure is calculated as the harmonic mean of recall and precision ensuring that both components are at their best rather than disproportionately tipping the scales to one. In the context of fairness-aware recommendation, the F-measure provides a way to analyze how well the system provides a specific level of diversity while still selecting papers with acceptable utility. Methods that achieve a high F-measure demonstrate that they can strengthen fairness without significantly degrading the quality of the recommended pool. By integrating the F-measure into the processes of evaluation, fairness-aware algorithms can be fine-tuned to assure they're providing the more favorable outcomes without giving away the overall utility.

These approaches collectively contribute to developing fair recommender systems that balance diversity, fairness, and accuracy. As bias mitigation strategies continue to change, integrating multiple strategies can provide a better solution to reducing bias in academic paper selection and even other decision making processes.

3.0 Approach

3.1 Overview

This study proposes a fairness aware recommender system to address bias in academic paper selection. The approach involves three main components: demographic data, model, and fairness optimization. The system is designed to identify existing biases, quantify their impact, and apply mitigation strategies to enhance diversity while maintaining selection quality.

Initial pilot experiments were conducted to assess the extent of bias in existing paper selection processes. Using a neural network system on paper selection data with zero fairness influence, we evaluated demographic distributions of accepted authors and compared them to applicant pools. Early findings confirmed disparities in race, motivating the need for interventions by adjusting the algorithm.

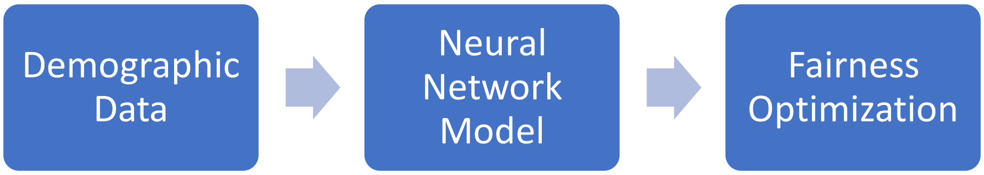
3.2 Pilot Work

In order to achieve our goals, we must begin by setting up a well structured developer environment, as it serves as the foundation for our work. Ensuring that all necessary dependencies are correctly installed and configured will help prevent compatibility issues and streamline development. Our environment requires Python 3.12.2 as the core programming language, along with CUDA 11.7.0 to leverage GPU acceleration for leveraged computational power. Additionally, we will be using PyTorch 2.3.1, TorchVision 0.18.1, and TorchAudio 2.3.1, which together provide essential tools for machine learning and deep learning tasks. Beyond these, several key libraries will be incorporated into our workflow, including pandas for data manipulation, scikit-learn for machine learning utilities, and NumPy for numerical computations.

Setting up this environment properly is a crucial step, as it will enable us to develop, train, and evaluate the model effectively while maintaining reproducibility and efficiency throughout our system.

3.3 System Architecture

The system architecture consists of three interconnected modules designed to systematically address bias in academic paper recommendations.



3.3.1 Module 1: Demographic Data

A model is only as effective as the data it is built on. To explore fairness in academic paper selection, this study uses a comprehensive set of datasets sourced from the Special Interest Group on Computer Human Interaction (SIGCHI), one of the best venues for research in human computer interaction. These datasets are an essential part of our fairness-aware recommendation system that aims to identify, quantify, and mitigate bias in academic publishing selection. The data captures both paper level and author level metadata, offering a strong building block for analyzing demographic disparities and evaluating the effectiveness.

The primary dataset used for training is final-asian_(continuous/boolean).csv, which contains author-paper structured pairs. Features in each row include items such as h-index (a metric representing the author's research impact), gender, race, country of origin, and a label for the which is indicating the quality of the submission. This dataset serves as the input to a neural network model, enabling it to learn patterns related to academic merit and demographic characteristics. It is especially important that the model learns from this rich feature space while remaining sensitive to fairness constraints that aim to balance representation across different groups.

To assess the model's performance and fairness outcomes, several supplementary datasets are integrated. The full SIGCHI submission pool, including paper titles and the associated race of authors are found in the papers_final-asian-withRace.csv dataset which offers a macro-level view. We can assess how well the system aligns with our goal of demographic parity and representation by comparing the demographic makeup of papers selected by the model to that of the entire pool, .

Real example set of papers that have been accepted for SIGCHI are found in sigchi_selected_papers.csv. This dataset serves as a real-world example of how current methods may fall short and helps our benchmark for evaluating the model's ability to reflect or exceed current fairness standards. To help understand how adjustments through algorithms might impact outcomes in a very real context the model generated recommendations can be compared with selected papers across various fairness metrics .

The data provides meaningful, representative features that allow the model to be trained and gauged against realistic academic standards. The project ensures that its findings are directly relevant to real academic communities by building the system on SIGCHI-specific data, while offering an adaptable framework that can be extended to other fields. The structure and depth of these datasets lay the groundwork for exploring how fairness can be formally implemented in recommendation algorithms without sacrificing research quality or credibility.

3.3.2 Module 2: Model

To balance fairness and utility in academic paper selection we need to provide a strong design for the neural network model used in this recommender system. The model needs to be able to handle Boolean and continuous values for our selected features. We've gone with PyTorch to build it because it's highly flexible and fast. This framework will facilitate the model for maintaining computational efficiency while learning purposeful feature representations.

To optimize paper acceptance probabilities when training the model we'll also need an excellent loss function to be implemented. However, an additional statistical parity loss function is incorporated to address fairness concerns, this is the driving factor when running features that are either Boolean or continuous values. Ensuring fair recommendations the fairness constraint aims to minimize disparities in selection rates between protected and unprotected groups. After data preparation, including relabeling the papers as either a “0” or “1”, we can move on to running the model.

Similarly, with the goal at achieving fairness through interchangeable representations Madras and others proposed a different deep learning framework. They introduced adverse objectives specifically designed to meet various fairness metrics such as equalized odds. Such representations proved fair not only for specific tasks but also maintained fairness across different, unknown tasks. We could further enhance the model's robustness in handling fairness in paper selection scenarios by integrating transferable representation approaches into the current multi-layer perceptron model [5].

While the model optimizes, a weighted fairness term, controlled by the hyper-parameter lambda (lambda_fairness) is added to the total loss function. Once trained, the model is enabled to rank papers based on predicted probabilities of acceptance. Paper selection is performed by sorting predictions in descending order and dynamically adjusting the threshold. This process ensures that fairness constraints are respected while maintaining a high-quality selection of academic work. This model offers a framework for equitable academic paper recommendation by incorporating fairness-aware learning into a traditional MLP-based classifier. The model is enabling a more adaptive approach to fairness in recommender systems through the ability to fine-tune lambda which provides flexibility in balancing bias.

3.3.3 Module 3: Fairness Optimization

The third module in our system focuses on fairness optimization. Where we analyze the neural network and its outputs then apply evaluation strategies that balance fairness and utility. This module serves as the bridge between actionable insight and raw model outputs, enabling us to compare different fairness strategies and determine how well they mitigate bias relative to a baseline.

Our baseline is made purely based on predicted paper quality without incorporating any fairness constraints, meaning the selection is established using a model trained with lambda_fairness = 0. This provides a useful reference point for evaluating the trade-offs introduced when fairness optimization is applied. We then study the effects on both demographic parity and utility from introducing different fairness constraints into the model.

The fairness optimization module incorporates three primary evaluation strategies:

Macro Data Method: This method analyzes the macro-level representation of protected groups in the selected paper set. Specifically, it evaluates the percentage of accepted papers that include an author from our chosen protected demographic group. When wanting to understand whether our goals are present across the broader set of selected works, this approach should be implemented.

Micro Data Method: The micro method dives deeper by measuring the proportion of individual authors within the selected papers who belong to protected groups. For a finer view of fairness a micro level view is required, relaying whether the overall pool of selected authors reflects demographic parity. Unlike at the macro level, which really only focuses on the paper level, being on the micro level highlights demographic representation across all authors.

F-Measure Method: The F-measure is defined as a metric that balances precision and recall. We calculate it as the harmonic mean between protected group (measured via protected group representation) and utility (quantified by the average weighted h-index of selected authors). We can also adjust the importance of either utility or the protected group with the value of beta. The F-Measure will capture the trade-off between matching the demographic of the pool and preserving academic excellence as a single metric.

$$F_\beta = \frac{(1+\ \beta^2)(protected \times\ utility)}{(\beta^2\ \times\ protected) + utility}$$

For each strategy, results are collected across a range of lambda_fairness values and varying paper acceptance counts (e.g., 10, 25, 50, 150, 250, 350). This modular approach permits a flexible yet systematic evaluation of fairness and performance trade-offs. To visualize trends we introduce line plots generated from these evaluations, showing off which configurations get us to a more desirable balance.

This module not only validates the effectiveness of fairness constraints but also guides future model tuning by identifying fairness levels that provide meaningful diversity gains with minimal losses in research utility.

4.0 Evaluation

4.1 Dataset Implementation

This study and its implementation of datasets is supported by several core Python libraries, most notably pandas and PyTorch. We form the foundation with these tools for the data preprocessing pipeline, allowing for better, scalable, and replicative experimentation with fairness-aware recommendation systems.

All datasets including final-asian_(binary/continuous).csv, papers_final-asian-withRace.csv, features_final-asian.csv, and sigchi_selected_papers.csv utilize the pandas library and are loaded via read_csv() functions. When the sets are loaded, data is manipulated using pandas own DataFrame operations to clean and filter the columns. We can take the "Label" column for example, which using mapping dictionaries is converted into binary target labels, and demographic fields like "Race" are converted to boolean or numerical types for the model to take in.

A big component of the preprocessing pipeline is located in helpers.py, the prepare_data() function. This function accepts the dataset and the protected attribute (race) and performs the following steps:

Splits the data into features (X) and labels (y),

Converts them into floating point,

Separates them into training and validation sets by using scikit-learn's train_test_split, Converts the resulting subsets into PyTorch TensorDataset objects.

To give access to the demographic metadata at the level of the author, the system uses features_final-asian.csv, which is handled through cleaning functions like clean_author_name() and parse_authors(). These functions apply methods to standardize the authors names, allowing a consistent access to values such as race, gender, country, career, and h-index. This information is used both for fairness review and for calculating utility scores based on the weighted averages of author h-index values.

Throughout the pipeline, pandas DataFrame are used to track and export the models results. Paper subsets and experimental results that were selected are saved using the function to_csv() for analysis and visualization in a further module. In conclusion, this system mostly integrates pandas to allow for data processing which is more flexible resulting in a dynamic framework for fairness-aware academic paper recommendation.

4.2 Neural Network Model Implementation

This fairness-aware recommender system implements a custom Multi Layer perceptron utilizing PyTorch for the core of the system. As a feed forward neural network we're provided with flexibility, capacity, and efficiency to handle the complex representations of our dataset. The MLP contains an input layer, two hidden layers, and an output layer. The input layer is custom to the number of features we choose to extract from the selected dataset. The two hidden layers respectively have 64 and 32 neurons, and the output layer a sigmoid activation function in order to produce a probability for the paper's acceptance. Training dynamics can be enhanced by equipping a ReLU activation function and batch normalization to each hidden layer. Batch normalization will speed up and stabilize training while ReLU will help with complex patterns during learning as it brings non-linearity to the model. The final layer has the sigmoid function to make sure the outputs are within the range of (0,1), which pairs nicely for our binary classification.

The objective function of the model has a standard binary cross-entropy loss but also incorporates a fairness loss confine. This fairness loss comes from a statistical parity, a computation that will instill penalties if there are significant differences in the protected and unprotected groups selection rates. The importance of this loss is calibrated through the hyperparameter of lamda_fairness, resulting in a changeable balance of utility and fairness. Lowering the value of lambda will emphasize utility however can leave existing biases within the system. Raising the lambda value will prioritize the demographic parity of the pool, but this can impact utility in a potentially negative way.

We can proceed with training using the Adam optimizer with a learning rate of 0.001 over 50 epochs. To ensure we don't overly overfit the model, we implement a stopping mechanism that quits the training when validation loss quits improving.

To initiate paper selection the trained model will be applied to the entire dataset and begins ranking the papers based off the predicted probability of being accepted. We dynamically set the threshold of accepted papers, examples include, (10, 25, 50, 150, 250, or 350). This step is important as various real academic scenarios have different fixed number of accepted submissions, while 350 is closest to SIGCHI. We utilize several helper functions that parse names and calculate things such as the h-index scores for utility.

We conducted several experiments among the range of lambda_fairness values, (0, 1, 2, 2.5, 3, 5, and 10), which bring us a deeper insight to what is needed to achieve our goal of demographic parity. We handle the analysis in the next module. The model demonstrates what we can do with PyTorch and methods of implementing a complex recommender system.

4.2.1 Model Implementation For Continuous Values

Our model initially was utilizing a binary only fairness constraint for our function statistical_parity_loss, this for example would take the protected group of race and do the calculations using 0 for non-protected and 1 for protected. This is a good way of using the function to critique the model during training however we wanted to introduce a continuous value to each race. Before, the loss function would take the difference between protected and unprotected and penalize the model if a difference of high proportions was found.

However, to create an individuality to each race a continuous value was needed to replace the Boolean value. This way rather than saying "yes" or "no" in relation to the authors belonging to the protected group, we can represent them with a more specific value relating to the demographic parity. This offered a more distinct way of identifying individuals. Motivated by the limitations of Boolean classifications which over generalize a population. We can find this unique weight relating to race by finding their representation in the total pool of submitted papers and calculate the complement.

Because the original loss function had binary covers to rely on (protected_mask = protectedGroup == 1), it could no longer be meaningfully applied to the new continuous variables. If we were to continue this, we'd get behavior within the model that actually hurts the performance of it, because we don't have a cut and dry protected vs unprotected in this new context.

We clearly needed to implement a "continuous" version of our original function, and we called it statistical_parity_loss_continuous. We incorporate the squared Pearson correlation which essentially measures the proportion of variance from a variable that is being predicted from a second variable [7]. This means we can properly enforce our rule of fairness if the predicted values have a high variance to the weights of the race.

This new approach maintains the goal of demographic parity from our Boolean method but alters it with a higher detailed profiling of authors. The system now operates to a higher degree of inclusive selection.

The weights for each race are as follows:

Race	**Representation (Overall Pool)**	**Weight**
W_NL	53.45%	0.4655
A	32.51%	0.6749
H	8.37%	0.9163
B_NL	5.67%	0.9434

TABLE 1

4.3 Fairness Optimization Implementation

The system isn't complete without installing several evaluation techniques that can test how the model balances our protected groups representation in the selected papers. These strategies are created by integrating custom scripts and analysis channels that produce actionable insights from the outputs being supplied by the neural network. The data we input to these channels are a culmination of several different tweaked variations of the model.

For the optimization we need to supply our pipeline with several runs from the trained model, this included altering the threshold of selected papers. We achieve sets for each desired threshold as well as for each lambda fairness value. This way we have a comprehensive selection of papers from each degree of fairness.

We generate plots and tables from our python scripts using the matplot library with axes such as Protected Macro vs. Paper Count, Protected Micro vs. Paper Count, and F-measure vs. Paper Count which show how fairness degrees can differ in performance across selections. The insights that are given can be used to make informed decisions on which settings the model can best perform for real world academic paper selection, in hopes of achieving a suitable balance.

One of the most important metrics to use is the F-Measure which highlights how well the model is balancing utility and fairness. With so much to absorb in the original 9 column csv the model outputs we combine each csv from each threshold for a total of 6 outputs into one DataFrame. We can view the following charts to see how well the model performs in a Boolean and continuous value format.

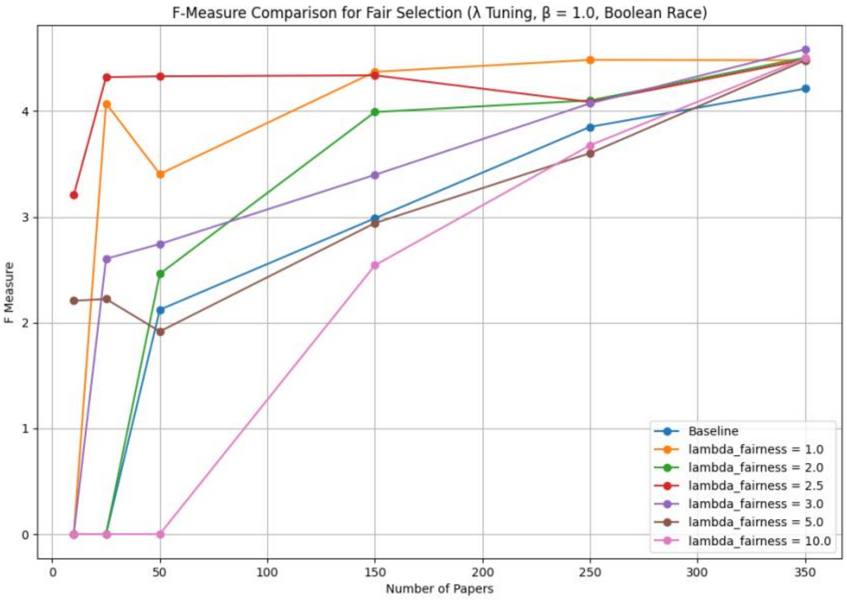


CHART 1

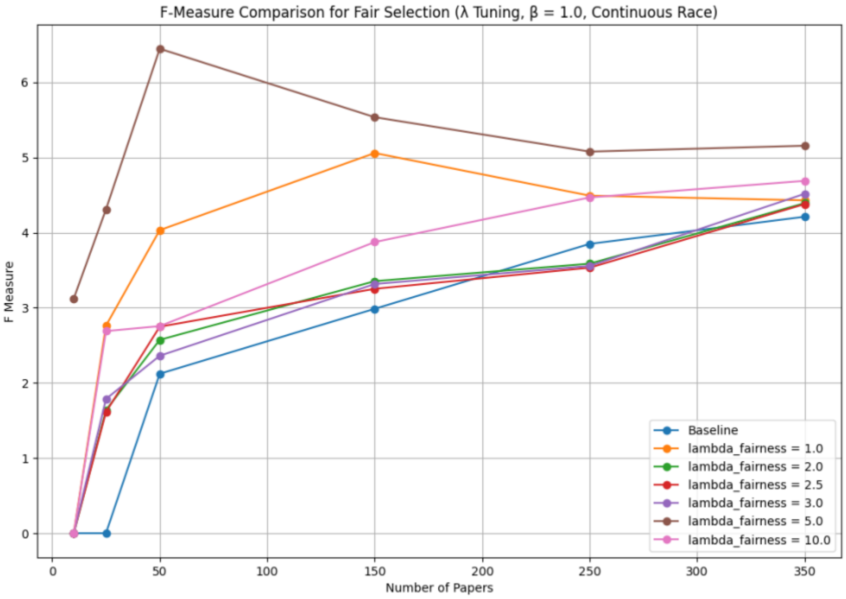


CHART 2

By viewing the different methods performances in a table format we can better see how our different methods performed against one another and the total submission pool. With the parameters of 350 papers, 5.0 lambda fairness value, and a micro level overview we get the following table.

	Average H-Index	**Protected Group (%)**	**H-Index (Method – Pool)**	**Representation (Method – Pool)**
Baseline	2.76	6.29 %	+ 0.0	- 6.52 %
Boolean	2.77	8.49 %	+ 0.01	- 4.32 %
Continuous	3.01	13.6 %	**+ 0.25**	**+ 0.79 %**

Table II

While keeping the paper selection size the same but changing the other parameters around to look at a macro level view with a fairness value of 10 we get the following.

	Average H-Index	**Protected Group (%)**	**H-Index (Method – Pool)**	**Representation (Method – Pool)**
Baseline	2.76	8.89 %	+ 0.0	- 8.16 %
Boolean	2.80	11.46 %	+ 0.04	- 5.59 %
Continuous	2.66	19.8 %	**- 0.10**	**+ 2.75 %**

Table III

5.0 Conclusions and Future Work

5.1 Conclusions

This study demonstrates that our recommender system can effectively balance demographic bias and without significantly compromising research quality in academic paper selection. We show that it is possible to improve the representation of protected groups in academic conferences by applying neural networks with fairness constraints and evaluating results across macro, micro, and F-measure metrics, Our experiments confirm that with increases in the fairness hyperparameter we see modest gains in diversity, with only minimal reductions if any in average h-index an important delegate for academic utility.

The continuous fairness model in particular achieved a better selection of papers, in the context of parity, compared to the baseline and binary models, while preserving competitive h-index scores. This supports the idea that fairness does not require a tradeoff with high performing papers, but rather that through algorithmic design choices we can closer align equity with quality.

Overall, this research contributes an interpretable and scalable framework for balancing bias into recommender systems. It confirms the potential for machine learning to not only maintain quality but also promote ethical outcomes in academic evaluation processes.

5.2 Future Work

Future development is something all developers should emphasis, for this fairness-aware recommender system we can focus on several directions that could improve balancing bias. One of the most immediate next steps involves improving the model to incorporate multiple demographic features at the same time, such as race and gender. While this paper focused on a single protected feature (race), future development should seek the impact of an individual belonging to multiple underrepresented groups and the most likely compounding effects that come with it. This would allow the system to detect and learn more distinct patterns across groups that may be overlooked when protected variables are evaluated by themselves.

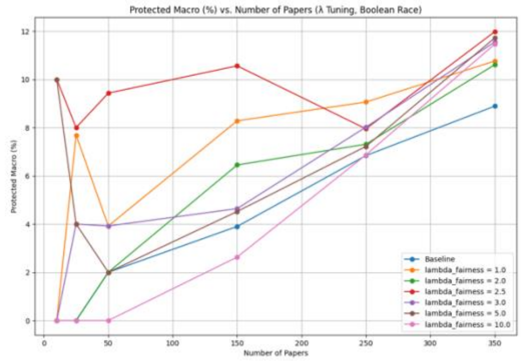
To expand on this we'd need additional support, the system could be adapted to implement multiple objective fairness optimizations, where separate statistical parity constraints, which help the models loss function, are calculated for every attribute deemed protected and then accumulated using a weighted loss function.

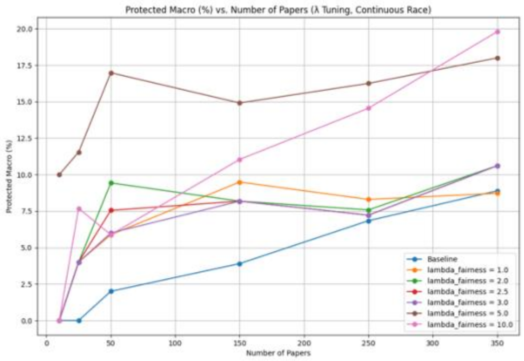
In terms beyond our context, a big step involves deriving the system beyond SIGCHI datasets. The framework could be applied to other areas within the academic community, and bring it to different publication cultures and mitigate their bias.

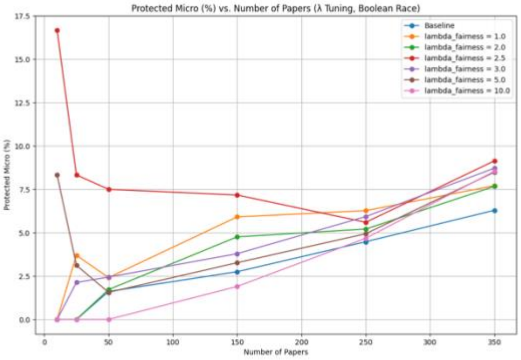
Lastly, our fairness-aware recommender system being deployed into live academic conference processes is really the entire future goal. This would require a better user friendly interactive interface where fairness parameters (like lambda_fairness) can be adjusted by admins and then observe changes in recommended papers. Ultimately, empowering real users with transparent, adjustable fairness parameters can create an academic paper recommender system into a more equitable and inclusive process across academic communities.

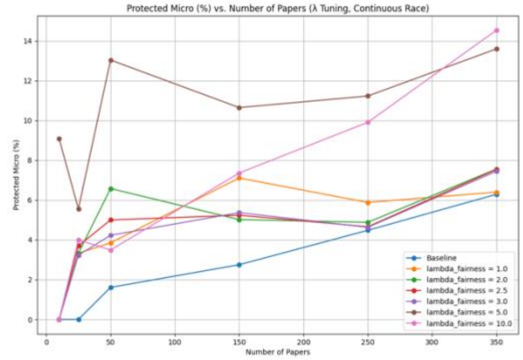
Appendices

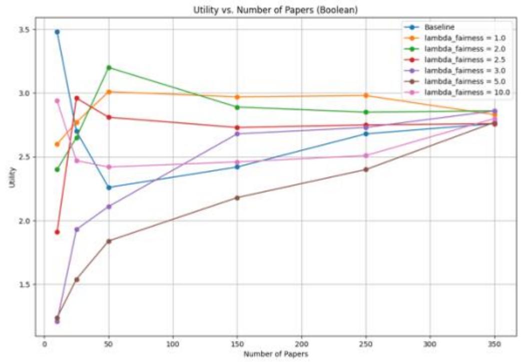
Charts

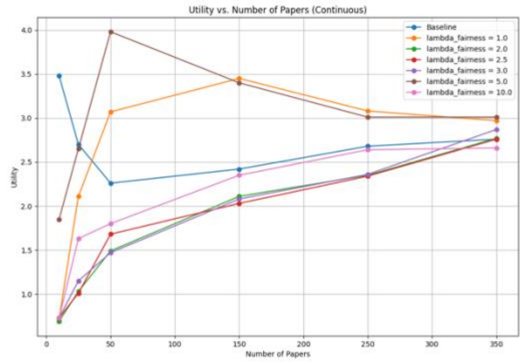

Protected Macro (%) vs. Number of Papers (λ Tuning, Boolean Race)
Protected Macro (%)
Number of Papers
0
2
4
6
8
10
12
0
50
100
150
200
250
300
350
Baseline
lambda_fairness = 1.0
lambda_fairness = 2.0
lambda_fairness = 2.5
lambda_fairness = 3.0
lambda_fairness = 5.0
lambda_fairness = 10.0


Protected Macro (%) vs. Number of Papers (λ Tuning, Continuous Race)
Protected Macro (%)
Number of Papers
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
20.0
0
50
100
150
200
250
300
350
Baseline
lambda_fairness = 1.0
lambda_fairness = 2.0
lambda_fairness = 2.5
lambda_fairness = 3.0
lambda_fairness = 5.0
lambda_fairness = 10.0


Protected Micro (%) vs. Number of Papers (λ Tuning, Boolean Race)
Protected Micro (%)
Number of Papers
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
0
50
100
150
200
250
300
350
Baseline
lambda_fairness = 1.0
lambda_fairness = 2.0
lambda_fairness = 2.5
lambda_fairness = 3.0
lambda_fairness = 5.0
lambda_fairness = 10.0


Protected Micro (%) vs. Number of Papers (λ Tuning, Continuous Race)
Protected Micro (%)
Number of Papers
0
2
4
6
8
10
12
14
0
50
100
150
200
250
300
350
Baseline
lambda_fairness = 1.0
lambda_fairness = 2.0
lambda_fairness = 2.5
lambda_fairness = 3.0
lambda_fairness = 5.0
lambda_fairness = 10.0


Utility vs. Number of Papers (Boolean)
Utility
Number of Papers
1.5
2.0
2.5
3.0
3.5
0
50
100
150
200
250
300
350
Baseline
lambda_fairness = 1.0
lambda_fairness = 2.0
lambda_fairness = 2.5
lambda_fairness = 3.0
lambda_fairness = 5.0
lambda_fairness = 10.0


Utility vs. Number of Papers (Continuous)
Utility
Number of Papers
1.0
1.5
2.0
2.5
3.0
3.5
4.0
0
50
100
150
200
250
300
350
Baseline
lambda_fairness = 1.0
lambda_fairness = 2.0
lambda_fairness = 2.5
lambda_fairness = 3.0
lambda_fairness = 5.0
lambda_fairness = 10.0

References

[1] APA Style. "Bias-Free Language: General Principles." Available at: https://apastyle.apa.org/style-grammar-guidelines/bias-free-language/general-principles

[2] National Education Association. "Recognizing Your Biases." Available at: https://www.nea.org/recognizing-your-biases

[3] Alsaffar, R., Gauch, S., & Al-Kawaz, H. "Multidimensional Fairness in Paper Recommendation." arXiv preprint arXiv:2305.01141, 2023.

[4] Zhang, B. H., Lemoine, B., & Mitchell, M. (2018). Mitigating unwanted biases with adversarial learning. Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society (AIES).

[5] Madras, D., Creager, E., Pitassi, T., & Zemel, R. (2018). Learning adversarially fair and transferable representations. Proceedings of the 35th International Conference on Machine Learning (ICML).

[6] Stelmakh, I., Shah, N. B., Singh, A., & Daumé III, H. (2021). Prior and prejudice: The novice reviewers' bias against resubmissions in conference peer review. Proceedings of the ACM on Human-Computer Interaction (CSCW).

[7] Cohen, J. (1988). Statistical power analysis for the behavioral sciences (2nd ed.). New York: Psychology Press.