February 7, 2026

RE: Revision 1 to Manuscript ID COIN-SA-11-24-10097 for Consideration of Computational Intelligence

Dear Prof. Inkpen and Reviewers,

We would like to thank you for your careful evaluation of our manuscript. It is our pleasure to inform you that we have considered the received comments and carefully addressed them in the manuscript as indicated below. We believe these revisions have strengthened the manuscript and addressed the reviewers' comments, and hope that the revised version meets the expectations of the editor and reviewers. We would like to highlight some important points summarily:

- We clarified the experimental setup, including detailed explanations of metric calculations via illustrative examples. We expanded the experimental evaluation to include the recommended team size of top-k=5 and an additional significance level of $\alpha = 0.05$. To ensure robustness and reproducibility, we perform a careful re-examination of our implementation and reran all the original experiments. We further extended the analysis of the results, providing deeper insights into the observed findings and clearer attribution of outcomes.

- We improved the use of terminology and notation across the manuscript. This includes consistent use of terms, resolving variable name conflicts, harmonizing symbols, and refining definitions to ensure precision and readability throughout the paper. To strengthen the theoretical foundation and improve clarity, we added 5 new equations, 2 propositions, 1 definition, and 7 illustrative examples. These additions are intended to better illustrate the notions of fairness in the team recommendation task, as well as the assumptions and implications underlying both the proposed and baseline reranking methods.

- We expanded the citations by adding more than 15 new works, including pioneering and standard references in the literature, as well as very recent studies, with particular emphasis on fairness in team recommendation systems.

Cordially,
The authors.

REVIEWER 1

Pros
- The paper tackles the significant issue of fairness in team recommendation.
- The proposed method is practical and relatively simple, suggesting broad applicability.
- The paper is well-written, self-contained, and logically structured.
- The experiments are well-executed, and the paper acknowledges the limitations of the proposed approach.

Cons
The main concerns revolve around a lack of clarity in several sections and confusing notation (see required revisions). Specifically:
- The definitions of group fairness for neural recommendation are unclear.
- It is not evident how these definitions are measured in the experiments or how they relate to the evaluation metrics ndkl, skew, and expo).
- The novelty and effectiveness of the proposed method appear marginal.
- The reliability of the results on popularity bias is questionable due to the definition of popularity.
- The paper's contribution would be more impactful if it effectively addressed gender bias, a more prevalent source of bias.

Response: We sincerely thank you for the thorough and constructive feedback. In the following, we provide detailed responses, having carefully considered all the points raised, and the manuscript has been revised accordingly.

R1.1

The paper should justify why popularity is treated as a protected attribute, similar to gender. Discuss the subjective nature of popularity, its potential correlation with expertise, and whether it should be considered a biased feature rather than a protected one.

Response: Thank you for your comment. We have added a new Section **3.3.1 Protected Attribute and Group** with two subsections **Gender as a Canonical Protected Attribute** and **Popularity as an Unconventional Protected Attribute** in **3.3 Fair Team Recommendation**, to formally define (social) groups and their associated protected attributes, and to examine whether popularity can be considered as a protected attribute, discussing its nuances relative to canonical protected attributes like gender. We further moved the subsections **Popularity Labels** and **Gender Labels** from Section **4.1 Dataset Labeling Criteria** to this section to provide a more coherent and integrated discussion, and to avoid sporadic treatment of experts' labeling criteria for their protected attributes across the manuscript.

Revised Text:

3.3.1 Protected Attribute and Group

"We define a (social) group $\mathcal{G}$ as a subset of the experts $\mathcal{E}$ that shares an "identity trait"[1], which may be inherent to an expert and largely immutable [...]

Gender as a Canonical Protected Attribute: *In this paper, given the protected attribute $\mathcal{A}$:gender, the values=$\{a: female, a': male\}$ create disjoint protected groups of experts including female experts $\mathcal{G}_{a:female}$ and male experts $\mathcal{G}_{a':male}$ where [...]*

Popularity as an Unconventional Protected Attribute: *As highlighted by Gallegos et al. and others[1,2], a group can also be formed based on an identity trait that is contextual and dynamic, e.g., disability or religion, or socially constructed based on social network effects and historical contingencies, like outcomes or conditions that arise from past events, as in popularity[3,4,5]. However, [...]"*

R1.2

Provide relevant citations for each group fairness definition on Page 6, Lines 12, 32, and 52.

Response: Thank you for your suggestion. We have revised the manuscript with citations, including seminal works that pioneered this line of research and have since become standard references in the literature, as follows:

[6] Dwork C, Hardt M, Pitassi T, Reingold O, Zemel RS. Fairness through awareness. In: Goldwasser S., ed. Innovations in Theoretical Computer, Science 2012, Cambridge, MA, USA, January 8-10, 2012ACM 2012:214–226841. https://dl.acm.org/doi/10.1145/2090236.2090255.

[7] Hardt M, Price E, Srebro N. Equality of Opportunity in Supervised Learning. In: Advances in Neural Information Processing Systems, Barcelona, Spain 2016:3315–3323. https://dl.acm.org/doi/10.5555/3157382.3157469.

[8] Altenburger KM, De R, Frazier K, Avteniev N, Hamilton J. Are There Gender Differences in Professional Self-Promotion? An Empirical Case Study of LinkedIn Profiles Among Recent MBA Graduates. In: AAAI Press 2017:460–463.839. https://doi.org/10.1609/icwsm.v11i1.14929.

[1] Gallegos IO, Rossi RA, Barrow J, et al. Bias and Fairness in Large Language Models: A Survey. Comput. Linguistics. 2024;50(3):1097–1179. https://doi.org/10.1162/coli_a_00524.

[9] John PG, Vijaykeerthy D, Saha D. Verifying Individual Fairness in Machine Learning Models. In: Proceedings of the Thirty-Sixth Conference on Uncertainty in Artificial Intelligence, UAI 2020, virtual online, August 3-6, 2020. http://proceedings.mlr.press/v124/george-john20a.html.

[10] Zehlike M, Yang K, Stoyanovich J. Fairness in Ranking, Part I: Score-Based Ranking. ACM Comput. Surv.. 2023;55(6):118:1–118:36. https://doi.org/10.1145/3533379.

[11] Rosenblatt L, Witter RT. Counterfactual Fairness Is Basically Demographic Parity. In: Thirty-Seventh AAAI Conference on Artificial Intelligence, AAAI 2023, Washington, DC, USA, February 7-14, 2023:14461–14469. https://doi.org/10.1609/aaai.v37i12.26691.

[12] Ashurst C, Weller A. Fairness Without Demographic Data: A Survey of Approaches. In: ACM 2023:14:1–14:12722. https://doi.org/10.1145/3617694.3623234.

[13] Romano Y, Bates S, Candès EJ. Achieving Equalized Odds by Resampling Sensitive Attributes. In: Larochelle H, Ranzato M, Hadsell R, Balcan M, Lin H., eds. Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual 2020. https://dl.acm.org/doi/abs/10.5555/3495724.3495755.

Revised Text:

3.3 Fair Team Recommendation

"To eschew varied interpretations and to provide actionable criteria to design and evaluate fairness-aware algorithms, fairness has been mathematically formalized, with a level of abstraction from an underlying real-world scenario, based on well-known notions of justice and equity at an individual level[9,10], or at a group level[1,7,8,6] like female vs. male experts [...]"

3.3.2 Demographic Parity

"Demographic (statistical) parity is to provide equal treatment to protected groups, i.e., the proportion of individuals receiving a favorable outcome should be consistent across all protected groups according to their distribution in the population[14,11,12]. [...]"

3.3.3 Equalized Odds

"Equalized odds[7,13] is a stronger notion of fairness. [...]"

3.3.4 Equal Opportunity

"Equal opportunity is a relaxed version of equalized odds that only requires fairness for the desired ground-truth decision, that is, the opportunity among qualified experts[7]. [...]"

R1.3

Correct the typo "equalized odds" on Page 6, Line 55.

Response: We apologize for the oversight. We have proofread the manuscript and fixed this typo as well as others:

> *"[...] based on the equalized dds [...]" → "[...] based on the equalized odds [...]"*
> *"equality of opportunity" → "equal opportunity"*
> *"[...] items with small probability [...]" → "items with small probabilities"*
> *"these values adjusted closer to the ideal." → "these values were adjusted closer to the ideal."*

Also, because groups are formed based on a protected attribute, all resulting groups are, by definition, protected. To differentiate groups, we revised the manuscript to consistently use the terms by *'disadvantaged'* vs. *'advantaged'* groups in place of *'protected'* vs. *'non-protected'* groups. Accordingly,

> *"protected group" → "disadvantaged group"*
> *"nonprotected group" → "advantaged group"*
> *"disadvantaged protected group" → "disadvantaged group"*

Moreover, to avoid the naming conflict between p, which denotes probability throughout the manuscript (e.g., $p(e \in \mathrm{E})$), and the Bernoulli winning probability p used in Section **3.4 Proposed Probabilistic Reranking Method**, we rename the latter to q. Accordingly,

> $\mathbb{F}^{-1}(\alpha; k, p) \rightarrow \mathbb{F}^{-1}(\alpha; k, q)$

Lastly, we observed that the terms *'utility'* and *'accuracy'* were previously used interchangeably, despite referring to distinct concepts. To avoid ambiguity, we adopt a consistent terminology: *'utility'* refers to the model-assigned probabilities reflected in the ranked list produced by the output layer of a neural team recommender, as used in `expo` fairness metric. In contrast, *'accuracy'* measures the extent to which the top-k ranked experts match the ground-truth members of a team in the test set. While related and higher *'accuracy'* typically implies higher *'utility'* for the correctly identified experts, the reverse does not necessarily hold. Accordingly,

> *"utility metric" → "accuracy metric"*
> *"maintaining utility across all domains." → "maintaining accuracy across all domains."*

R1.4

Clarify the notation in Equations 4 through 8. Specifically, define d and $\hat{d}$ to depend on the given expert (i or j). Also, consistently define e_0 and e_1 (whether they refer to group membership or experts).

Response: Thank you for your comment. We agree that our original formalization included unnecessary notation, such as variable enumerators (e.g., i or j) and confusing subscript conventions. We have revised the formal definitions to use simpler, clearer, and more consistent notation, while remaining mathematically sound to the best of our knowledge. For instance, we adopt a consistent convention of using a prime ($'$) to distinguish different protected groups, avoiding subscript enumerators, where distinct values a and a' of the protected attribute $\mathcal{A}$ define the groups $\mathcal{G}_a$ and $\mathcal{G}_{a'}$, respectively. Also, for an expert in a group based on her/his value of protected attribute, we use explicit notation of group membership $e \in \mathcal{G}_a$ or $e \in \mathcal{G}_{a'}$ to prevent the confusion by subscript enumeration. Accordingly, we revised Equations 9 to 20. For more clarity, we also added Examples 1, 2, and 3 for each fairness notions as well as a comprehensive running Example 4 with new Tables 1 and 2 to illustrate how both fairness notions and accuracy are measured. Please also refer to our response to R1.6.

R1.5

Clarify Equation 9. Can a definition of fairness be binary? Address the possibility of achieving absolute zero disparity for equal opportunity and demographic parity. Specify any threshold used to consider Equations 5 and 8 as satisfied and explain how this threshold was chosen.

Response: Thank you for the comment and the question. Prior work, including Dwork et al.[6], defines notions of fairness based on pairwise distribution differences among protected groups, often up to a bias threshold ϵ: *"Definition 3.1 (Statistical parity). We say that a mapping $M : V \to \delta(A)$ satisfies statistical parity between distributions S and T up to bias ϵ if [...]."* In our work, we assume a zero bias threshold, i.e., $\epsilon = 0$, for all notions of fairness. This allows for a direct comparison of prior and posterior distributions using well-known fairness metrics such as `skew` and `ndkl`. To make this clearer, we have added Propositions 2 and 3, which form the basis for calculating fairness metrics. Equation 9 has been revised (now Equation 25) to explicitly reference these propositions. Additionally, we have included illustrative examples for each notion of fairness, demonstrating when zero disparity can be achieved. A comprehensive running Example 4 in Tables 1 and 2 further shows cases of zero disparity and otherwise for the considered notions of fairness.

In our experiments, we report fairness metrics to quantify the divergence of the posterior distributions of protected groups in a recommended team from their prior distributions as the *average* over the *test* set of teams. Zero disparity corresponds to exact equality between posterior and prior distributions for each protected group in *every* recommended team for the entire test set, yielding 0 for both metrics. Otherwise, these metrics report the divergence, with values closer to 0 indicating fairer outcomes. No additional thresholding was applied; the metrics directly reflect the degree of disparity.

R1.6

Address the confusing notation on Line 43. If y (used in the fairness definition with $d = 1$ and $d = 0$ as ground truth) is different from d, clarify their relationship and how the fairness definitions handle both.

Response: Thank you for your comment. We agree that the original formalization contained confusing notation. We have revised the manuscript to clarify that y is the ground-truth label of a team for being successful or unsuccessful, which is required for training neural models, but d is the ground-truth decision for being a member of a team regardless of the team's label of success, which is required for the notions of fairness. We revised the manuscript as follows:

1. We moved the discussion of **Success Labels** from Section **4.1 Dataset Labeling Criteria** to Section **3.1 Preliminaries** by introducing a new Section **3.2 Team Success Labels**, to provide a more coherent and integrated treatment of the success label y used in neural team recommendation models. We clarified that in the absence of explicit labels for unsuccessful teams, neural team recommenders presume all instances of teams in the training dataset are successful (positive samples) for their observable outcomes, e.g., *published* papers in `dblp`, that is, $\mathcal{T} = \mathcal{T}^+ = \{(\mathrm{S}, \mathrm{E})_{y=1}\}$ and proceed with the training procedure. As a result, during the inference, for an input required set of skill S, the models predict (recommend) a subset of experts E for a successful team only, and throughout the paper, the act of recommending a team is understood as recommending a potentially successful team $(\mathrm{S}, \mathrm{E})_y \to (\mathrm{S}, \mathrm{E})_{y=1}$.
2. In Section **3.3.3 Equalized Odds**, we clarified that the ground-truth decision d concerns becoming a member of a team and is defined based on the experts' qualifications with respect to the required skill set S of a recommended team. Specifically, when an expert's skill set has a non-empty intersection with S, i.e., $\mathrm{S}_e \cap \mathrm{S} \neq \varnothing$, the ground-truth decision d is $e \in \mathrm{E}$. Accordingly, we removed the binary values $\{0, 1\}$ previously used for d and revised the decision space to $d \in \mathcal{D} = \{e \in \mathrm{E}, e \notin \mathrm{E}\}$.
3. For additional clarity, in Section **3.3.5 A Fair Team**, we added a comprehensive running Example 4 with new Tables 1 and 2 to illustrate how both fairness notions and accuracy are measured. The example explicitly demonstrates the trade-offs between different fairness criteria and predictive accuracy.

Revised Text:

3.2 Team Success Labels

"From Definitions 1 and 2, neural team recommenders aim at estimating f from samples of teams that are labeled with success or failure, yet most available training data in team recommendation literature only consists of successful teams, missing unsuccessful ones. [...]"

3.3.2 Demographic Parity

*"**Example** 1. Consider a set of $|\mathcal{E}| = 100$ experts, where $\mathcal{A}$ denotes gender, and $|\mathcal{G}_{a:female}| = 30$ and $|\mathcal{G}_{a':male}| = 70$ represent the groups of female and male experts, respectively. Suppose we want to recommend a team of size $k = 10$ for a required set of skills S, i.e., $(S, E)_{y=1}$. To satisfy demographic parity, [...]"*

3.3.3 Equalized Odds

*"**Example** 2. From Example 1, for a set of $|\mathcal{E}| = 100$ experts, where $\mathcal{A}$ denotes gender, $|\mathcal{G}_{a:female}| = 30$ and $|\mathcal{G}_{a':male}| = 70$ represent female and male groups, respectively. To recommend a fair team of size $k = 10$ for a required set of skills S, we first identify the qualified and unqualified experts. Suppose there are 10 female and 10 male experts whose skill sets intersect with S (i.e., qualified experts), and the remaining 20 female and 60 male experts are not qualified. To satisfy equalized odds, [...]"*

3.3.4 Equal Opportunity

*"**Example** 3. Referring to our earlier Example 2 with a set of $|\mathcal{E}| = 100$ experts, where $\mathcal{A}$ denotes gender, and $|\mathcal{G}_{a:female}| = 30$ and $|\mathcal{G}_{a':male}| = 70$ represent female and male groups, respectively, we aim to recommend a fair team of size $k = 10$ for a required set of skills S. To satisfy equal opportunity, [...]"*

3.3.5 A Fair Team

*"**Example** 4. To illustrate the evaluation of fairness and accuracy in recommended teams, we consider a small population of 6 experts, where two are female $\mathcal{G}_{a:female} = \{\boxed{e_1}, \boxed{e_2}\}$ and 4 are male $\mathcal{G}_{a':male} = \{e_3, e_4, e_5, e_6\}$. The required skill set for a successful team is $S = \{s_1, s_2\}$, and only a subset of experts are qualified, i.e., their skill sets intersect with S [...]"*

R1.7

On Line 30, when comparing with pre-processing methods, acknowledge that pre-processing methods are also model-agnostic. Update the statement to emphasize that the proposed method's distinction lies solely in its post-training nature.

Response: Thank you for the clarification. We have revised the manuscript in **1 Introduction** and **2.2 Fairness-aware Recommendation** and acknowledge that pre-processing methods are also model-agnostic. However, we further explain the post-processing advantages over pre-processing, including 1) less demand for computing resources [15,16]; pre-processing mid to large-scale training datasets is costly, 2) more privacy preserving [17,18]; no direct access to the raw data is required, and 3) more practically feasible in the era of pretrained models.

Revised Text:

1 Introduction

"[...] As opposed to pre-processing-based methods [19,20,21,22], which, despite being model-agnostic, require direct access to raw and potentially sensitive training data [18,17] along with substantial computational resources to modify the data or its labels prior to model training [15,16], or in-processing techniques [23,24,25,26,27,28], which are model-dependent and focus on balancing model accuracy with fairness considerations during training, our method belongs to post-processing category of methods [29,30,31,32,33,27,34,35], which are simple and computationally efficient by improving the fairness of model's outputs after training, without adjustments to the data, training procedure, or the model's architecture.[...]"

2.2 Fairness-aware Recommendation

"[...] The latter category is of particular interest due to several practical advantages. Foremost, post-processing methods are computationally efficient, as they do not require access to or manipulation of raw training data, which are typically medium- to large-scale, unlike pre-processing or in-processing methods [15,16]. Secondly, post-processing methods can be readily plugged

into already trained and deployed models without retraining or fine-tuning, a scenario that is increasingly common in realistic settings where pre-processing or in-processing interventions are infeasible. Finally, post-processing methods are more privacy-preserving, as they do not require access to potentially sensitive raw training data, as in pre-processing methods, nor the architecture or parameters of private models, as in in-processing methods[17,18]*. [...]"*

R1.8

Explain why the Bayesian Neural Network (BNN) architecture (Figure 3) is preferable for neural team recommendation compared to a standard Multi-Layer Perceptron (MLP), beyond simply motivating its use.

Response: Thank you for the suggestion. We revised Section **3.1 Preliminaries** and explained that Bayesian neural networks estimate prior uncertainty of model parameters $p(\boldsymbol{\theta})$ via maximum a posteriori, whereas standard multilayer perceptron overlooks it and hence results in overconfident estimates of model parameters. We further include prior studies in which standard multilayer perceptrons (non-variational feedforward neural networks) are employed as baselines to demonstrate the performance gains of variational neural networks.

Revised Text:

3.1 Preliminaries

"During training, given a team (S, E), neural models tune the parameters $\boldsymbol{\theta}$ by maximizing the posterior probability of $\boldsymbol{\theta}$ in $f_{\boldsymbol{\theta}}$ over $\mathcal{T}$. From Bayes theorem:

$$\texttt{argmax}_{\boldsymbol{\theta}}\, p(\boldsymbol{\theta}|\mathcal{T}) \propto p(\mathcal{T}|\boldsymbol{\theta})p(\boldsymbol{\theta}) = p(\boldsymbol{\theta}) \prod_{(S,E)\in\mathcal{T}} p(E \mid S, \boldsymbol{\theta}) \tag{1}$$

$$p(E \mid S, \boldsymbol{\theta}) = \prod_{j\in e} \sigma(\mathbf{z}[j]) \propto \sum_{j\in e} \texttt{log}\, \sigma(\mathbf{z}[j]) \tag{2}$$

where $p(\mathcal{T}|\boldsymbol{\theta})$ is the likelihood and $p(\boldsymbol{\theta})$ is the prior joint probability of weights, which is unknown. While the exact prior probability of weights $p(\boldsymbol{\theta})$ cannot be calculated analytically[36]*, it can be estimated as Gaussian distributions by a variational Bayesian neural architecture via the maximum a posteriori optimization*[37,38,39]*, which contrasts with conventional non-variational multilayer perceptron that assume a uniform probability distribution over all possible real-values of $\boldsymbol{\theta}$ and only estimate the likelihood $p(\mathcal{T}|\boldsymbol{\theta})$ via maximum likelihood optimization, discarding prior uncertainty $p(\boldsymbol{\theta})$, and hence, resulting in overconfident point estimates of $\boldsymbol{\theta}$*[36]*. Existing studies primarily adopt standard multilayer perceptrons (non-variational feedforward neural networks) as baselines to demonstrate the performance advantages of variational neural networks*[40,41,36]*. [...]"*

R1.9

On Line 39 and throughout the experiments, clearly explain how the metrics ndkl, skew, and exposure are used to measure demographic parity and equal opportunity.

Response: Thank you for pointing this out. We have added Propositions 2 in Section **3.3.2 Demographic Parity** and Proposition 3 in Section **3.3.4 Equal Opportunity** to formalize the rationale on measuring fairness as post-hoc manner after reranking based on the prior distribution as reference distribution: demographic parity uses the group distributions in the full expert set, while equal opportunity uses the group distributions among qualified experts. The `ndkl` and `skew` metrics measure deviations between the reranked posterior distributions and the prior distributions of groups. The metric `expo`, however, is agnostic to notions of fairness and solely relies on the model-assigned probabilities and the position of experts with different protected values in the top-k ranked list of recommendations. Though `expo` values used for evaluation before vs. after reranking are different since the reranking algorithm may account for group distributions, as in our proposed methods based on demographic parity or equal opportunity. We have further added Examples 5, 6, and 7 in Section **4.4.3 Fairness and Accuracy Metrics**, for more clarification.

Revised Text:

4.4.3 Fairness and Accuracy Metrics

"Normalized Discounted KL Divergence (`ndkl`) [...]

Example 5. *Considering Example 4 for the top-k=3 recommended experts E : $[\boxed{e_1}, e_3, e_4]$, as shown at row #1 in Table 2, including 1 female expert at position 1. Under demographic parity, let the reference distributions be $q_{a:female} = p(e \in \mathcal{G}_{a:female}) = \frac{30}{100} = 0.3$ and $q_{a':male} = p(e \in \mathcal{G}_{a':male}) = 1 - 0.3 = 0.7$. At top-1, the ranking is entirely female, so the posterior distribution is $p_{a:female}(1) = 1$ and $p_{a':male}(1) = 0$, deviates strongly from the reference. This produces a large divergence, $\mathtt{kl}(p(1) \mid\mid q(1)) = p_{a:female}(1) \times \mathtt{log}\frac{p_{a:female}(1)}{q_{a:female}(1)} + p_{a':male}(1) \times \mathtt{log}\frac{p_{a':male}(1)}{q_{a':male}(1)} = 1 \times \mathtt{log}(\frac{1}{0.3}) + 0 \times \mathtt{log}(\frac{0}{0.7}) = 1.204$, reflecting deviation from the reference distributions. At top-2, the prefix contains 1 female and 1 male experts, giving proportions $p_{a:female}(2) = p_{a':male}(2) = \frac{1}{2} = 0.5$. These values are closer to the reference distributions, and the divergence drops: $\mathtt{kl}(p(2) \mid\mid q(2)) = 0.5 \times \mathtt{log}(\frac{0.5}{0.3}) + 0.5 \times \mathtt{log}(\frac{0.5}{0.7}) = 0.087$. At top-3, the proportions become $p_{a:female}(3) = \frac{1}{3}$ and $p_{a':male}(3) = \frac{2}{3}$ which almost match the reference and the divergence is therefore nearly zero: $\mathtt{kl}(p(3) \mid\mid q(3)) = \frac{1}{3} \times \mathtt{log}(\frac{\frac{1}{3}}{0.3}) + \frac{2}{3} \times \mathtt{log}(\frac{\frac{2}{3}}{0.7}) = 0.002$. [...]*

`Skew@k` *[...]*

Example 6. *Same as Example 5, using the same top-k=3 ranking, the posterior distributions of female experts at different cut-offs are $p_{a:female}(1) = \frac{1}{1} = 1$, $p_{a:female}(2) = \frac{1}{2} = 0.5$ and $p_{a:female}(3) = \frac{1}{3} = 0.33$, respectively. Then, $\mathtt{skew@1}= \mathtt{log}\frac{p_{a:female}(1)}{q_{a:female}(1)} = \mathtt{log}(\frac{1}{0.3}) = 1.204$, $\mathtt{skew@2}= \mathtt{log}\frac{p_{a:female}(2)}{q_{a:female}(2)} = \mathtt{log}(\frac{0.5}{0.3}) = 0.511$ and $\mathtt{skew@5}= \mathtt{log}(\frac{p_{a:female}(3)}{q_{a:female}(3)}) = \mathtt{log}(\frac{0.66}{0.3}) = 0.105$. [...]*

Utility-aware Exposure (`expo`)[...]

Example 7. *As in Examples 5 and 6, considering the top-k=3 recommended experts E : $[\boxed{e_1}, e_3, e_4]$ at row #1 in Table 2, including 1 female expert at position 1, let the model-assigned probabilities be $[\boxed{e_1} : 0.9,\ e_3 : 0.85,\ e_4 : 0.8]$. By Equation 36, the average exposure of the female experts is $\mu_{\mathtt{expo}}(a : female) = \frac{1}{1} \times \frac{1}{\mathtt{log}(1+1)} = 1.443$. Then by Equation 38, the exposure value for the team is then obtained by normalizing average exposure by average utility, i.e., $\mu_{utility}(a : female) = \frac{1}{1} \times 0.9 = 0.9$ and $\mathtt{expo}(a : female) = \frac{\mu_{\mathtt{expo}}(a)}{\mu_{utility}(a)} = \frac{1.443}{0.9} = 1.603$. The average exposure of male experts of the team is computed similarly for the male experts, who occupy the remaining two positions $\{2, 3\}$ as $\mu_{\mathtt{expo}}(a : male) = \frac{1}{2}(\frac{1}{\mathtt{log}(2+1))} + \frac{1}{\mathtt{log}(3+1)}) = 0.816$. Considering the model output probabilities for these two experts, $\mu_{utility}(a : male) = \frac{1}{2}(0.85 + 0.8) = 0.825$, and the exposure value for them is $\mathtt{expo}(a : male) = \frac{\mu_{\mathtt{expo}}(a)}{\mu_{utility}(a)} = \frac{0.816}{0.825} = 0.989$. That is, 1 female expert at top-1 with high probability (utility) is exposed more compared to 2 male experts with lower probability at lower ranks. By Equation 39, the overall exposure value for the gender protected attribute is $\mathtt{expo}(gender) = \frac{\mathtt{expo}(a:female)}{\mathtt{expo}(a:male)} = \frac{1.603}{0.989} = 1.62$. Here, $\mathtt{expo}(\mathcal{A}) > 1$ shows that the female (disadvantaged) experts receive more exposure relative to their utility compared to male (advantaged) experts.*

As a final note and before moving to the experimental result, while `expo` is distribution-agnostic and does not encode a specific fairness notion, its values used for evaluation before vs. after reranking are different since the reranking algorithm may account for group distributions, as in our proposed methods based on demographic parity or equal opportunity."

R1.10

Given that protected attributes are often inferred, clarify in the discussion that the absence of ground truth labels or the use of inferred values might lead to over- or under-estimation of bias, cite references [1, 2].

[1] Chen, J., Kallus, N., Mao, X., Svacha, G., & Udell, M. (2019, January). Fairness under unawareness: Assessing disparity when protected class is unobserved. In *Proceedings of the conference on fairness, accountability, and transparency* (pp. 339-348)

[2] Kenfack, P. J., Kahou, S. E., & Aïvodji, U. (2024). A survey on fairness without demographics. *Transactions on Machine Learning Research*. 2024

Response: Thank you for this observation. We have added a clarification in Section **3.3.1 Protected Attribute and Group** and Section **4.2 Forming Protected Groups** to acknowledge that when protected attributes are inferred rather than directly observed, they may inherently incorporate societal biases, and we have cited the suggested references. We also moved Section **4 Ethical**

Considerations to Section **4.2 Forming Protected Groups** to highlight this concern by examples.

Revised Text:

3.3.1 Protected Attribute and Group

*"**Gender as a Canonical Protected Attribute:** [...] It is worth noting that, when gender values are inferred rather than directly observed, they may inherently incorporate societal biases, or even amplify pre-existing stereotypes in society[42,43], as further illustrated by examples in our experimental setup (Section 4.2). Therefore, fairness assessments and bias mitigation using these inferred values may over- or under-estimate disparities across groups, and results should be interpreted with this limitation in mind. [...]"*

4.2 Forming Protected Groups

"[...] Meanwhile, we recognize that predicting gender using name-to-gender services such as `genderize` *might systematically mislabel certain demographic groups due to cultural, linguistic, and regional variations. For example, names such as 'Andrea' (typically male in Italy but female in English-speaking countries), 'Sasha' (male in Russia but commonly female in North America), and 'Kim' (gender-neutral in East Asia but often inferred as female in Western countries) are frequently misclassified. Hence, our results should be interpreted with this limitation in mind. [...]"*

R1.11

On Line 51, Page 3, clarify the difference between "bnn" and "bnn-emb".

Response: Thank you for your comment. We have revised Section **3.1 Preliminaries** and added the new Definition **2 Neural Team Recommendation** to further clarify on Bayesian neural network architecture (`bnn-`) and explicitly enumerate alternative choices for skill representations at the input layer, including occurrence vectors and dense embeddings (`-emb`), and contrast their properties, explaining why dense embeddings are the preferable choice. Also, in **4.3.1 Neural Team Recommendation**, where we explain the baseline methods, we remind readers of the Section **3.1 Preliminaries** for further details.

Revised Text:

3.1 Preliminaries

*"**Definition 2 (Neural Team Recommendation).** Given a subset of skills S and all teams $\mathcal{T}$ as the training set, neural team recommendation estimates $f_{\boldsymbol{\theta}}(S)$ using a multilayer neural network that learns, from $\mathcal{T}$, to map a vector representation of subset of skills S, referred to as v_S, to a vector representation of subset of experts E, referred to as v_E, by maximizing the posterior probability of $\boldsymbol{\theta}$ in $f_{\boldsymbol{\theta}}$ over $\mathcal{T}$, that is, $\mathtt{argmax}_{\boldsymbol{\theta}}\ p(\boldsymbol{\theta}|\mathcal{T})$.*

For the vector representation of subset of skills v_S, neural team recommendation methods adopt either

1. *The occurrence vector representation for S, which is a Boolean vector of size $|\mathcal{S}|$, i.e., $v_S \in \{0,1\}^{|\mathcal{S}|}$ where $v_S[s] = 1$ if $s \in S$, and 0 otherwise, or*
2. *The embedding representation (`emb`) of S, which is a dense low d-dimensional vector, $d << |\mathcal{S}|$, pretrained by e.g., a shallow neural encoder[38] like in distributional representation of words (word2vec)[44] and documents (doc2vec)[45], or a graph neural network[46,39].*

While occurrence vector representations are simple, they lead to high-dimensional sparse input representations, substantially increasing the number of trainable parameters in the input layer of neural networks. In contrast, dense pretrained vectors (embeddings) provide low-dimensional, continuous representations that significantly reduce model complexity and enable more stable and data-efficient learning. Moreover, dense embeddings capture latent semantic relationships among skills, allowing the model to generalize across similar inputs rather than treating each skill as independent, as is the case with the occurrence representation."

4.3.1 Neural Team Recommendation

"For the input layer, as detailed in Section 3.1, we used [...]"

R1.12

On Line 3, Page 3, replace higher is better (should ↑)".

Response: Thank you for raising this error. We have fixed the issue in the caption of all Tables.

R1.13

Explain why other re-ranking methods, such as 'det-cons', are more successful in mitigating gender bias while the proposed method is not. Discuss potential trade-offs between fairness and accuracy and how interventions can manage this trade-off.

Response: Thank you for your valuable comment. We have added the details of our baseline models to Section **4.3.2 Fairness-aware Reranking** to demonstrate why the deterministic methods outperform our method in fairness but at the cost of accuracy. Regarding the trade-offs between fairness and accuracy in our proposed method, we explain the role of the significance level α in details in our response to R1.14.

Revised Text:

4.3.2 Fairness- aware Reranking

"[...] Although the deterministic baselines were originally designed for multi-valued protected attributes (and thus multiple groups), we apply them to Boolean protected attributes in two settings: 1) gender, where female experts constitute the disadvantaged yet minority group and male experts are the advantaged and majority group; and 2) popularity, where nonpopular experts form the disadvantaged yet majority group and popular experts are the advantaged but minority group. [...]

`det-greedy`: *for every top–$\{1, ..., k\}$ prefix of the ranking, this algorithm aims to maintain a proportion of experts from each group as close as possible to the predefined desired distributions, herein, the prior distributions as explained in Propositions 1 and 3. The algorithm generates the new ranking at each top-k rank by first determining the target group for the next selection and then adding the next expert, while keeping the previous steps unchanged. [...]*

`det-cons` (conservative): *A variation of `det-greedy` with modified selection behavior under maximum-threshold condition. When the maximum thresholds for both groups have not yet been reached, instead of selecting the next expert solely based on the highest model-assigned probability, `det-cons` assesses which group is closer to falling below its minimum requirement in subsequent prefixes. [...]*

`det-relaxed`: *A relaxed variant of `det-cons` that introduces additional flexibility when the minimum requirements are satisfied for both groups. In this case, `det-relaxed` computes the smallest future rank at which any group's minimum requirement would be increased, regardless of whether that group's maximum threshold has already been reached. [...]*

The deterministic reranking baselines have notable limitations. Having only two groups in our case, while considering the maximum threshold in the decision process is to prevent disproportionate advantage to one group, it may come at the cost of accuracy, that is, when one group has met the maximum threshold (e.g., males), these algorithms would select an expert from the other group (e.g., females) even with the lower probability, resulting in fairer yet less accurate team recommendation. In contrast, our method enforces only the minimum required representation of the disadvantaged group necessary to satisfy the fairness constraints, while selection is always guided by accuracy. Moreover, deterministic baselines strictly enforce the prior proportions of both groups at every prefix k. In highly skewed datasets, when the proportion of the disadvantaged group is very small, deterministic reranking may replace a predicted disadvantaged expert with an advantaged expert to maintain the prior proportion at a given prefix. In contrast, our algorithm enforces only the minimum required presence of disadvantaged experts at each k, and preserves all high-scoring disadvantaged experts once the requirement is met. [...]"

R1.14

On Line 11, Page 10, justify the choice of $\alpha = 0.1$ and explain or provide experiments on how different values of p and α affect the fairness-accuracy performance.

Response: Thank you for this comment. Foremost, we acknowledge the variable naming conflict between p, which denotes probability throughout the manuscript (e.g., $p(e \in \mathrm{E})$), and the Bernoulli winning probability p used in Section **3.4 Proposed Probabilistic Reranking Method**. In the revised manuscript, we renamed the latter to q, accordingly. Moreover, we added Examples 1 and 3 and revised Table 3 with additional rows to explain how different values of q have been selected based on the notions of fairness. For demographic parity, by Proposition 2, q is set to the prior distribution of disadvantaged experts in the full dataset, and for equal opportunity, by Proposition 3, q is set per test team (S, E) as the proportion of disadvantaged qualified experts among those qualified for the required skill set S. Therefore, while q can technically be varied over the interval [0, 1], it has been set based on notions of fairness, as opposed to a tunable hyperparameter.

Regarding the role of the significance level α, as explained in Section **3.4 Proposed Probabilistic Reranking Method**, a top-k recommended team is considered fair if the number of experts from the disadvantaged group, denoted $|\mathrm{E}|_a$ is consistent with a Bernoulli process with success probability q, where q is set according to a specified notion of fairness. We denote this distribution by $\mathbb{F}(|\mathrm{E}|_a; k, q)$. For example, as shown in Table 3, under demographic parity with $q = 0.3$ and $k = 10$, the expected number of experts from the disadvantaged group in the team is $0.3 \times 10 = 3$. Suppose an observed recommendation contains x such experts. We then conduct a significance test with the null hypothesis

$$\mathrm{H}_0 : x \sim \mathbb{F}(x; k = 10, q = 0.3)$$

and compute the corresponding p-value under this distribution. If the p-value exceeds a predefined significance level α, the observation is considered statistically plausible under H_0, and we *"fail to reject"* the null hypothesis, treating the team as fair. Conversely, if the p-value is less than or equal to α, the observation is deemed unlikely, leading to rejection of H_0 and the team being flagged as *un*fair. As seen, the choice of α directly controls the strictness of the fairness criterion. Larger values of α increase the likelihood of rejecting H_0, requiring stronger alignment with the distribution. For instance, when $\alpha = 0.1$ and no disadvantaged experts are observed ($x = 0$) in the top-k=10, the p-value is $\mathbb{F}(x = 0; k = 10, q = 0.3) = 0.028 < 0.1$, and the null hypothesis is rejected, flagging the team as *un*fair. In contrast, for $\alpha = 0.01$, the same observation yields $0.028 > 0.01$, and the null hypothesis is not rejected, resulting in the team being considered fair. From a mitigation perspective, larger values of α impose stricter fairness requirements, as more disadvantaged experts are needed for a team to pass the significance test. Smaller values of α relax this requirement and may not trigger mitigation even when the observed composition deviates from the expected proportion. This behavior is illustrated in Table 3 for $\alpha \in \{0.5, 0.1, 0.05, 0.01\}$. Hence, larger values of α are preferable when the goal is to actively enforce fairness constraints.

In response to this comment, 1) we have clarified the role and effect of α in the revised manuscript in Section **3.4 Proposed Probabilistic Reranking Method**, accordingly. 2) In addition to the originally reported results with $\alpha = 0.1$, we have added new experiments with $\alpha = 0.05$, which is commonly used in statistical hypothesis testing (e.g., paired t-tests for average performance comparisons). The new results, reported in Tables 8 to 15 exhibit consistent fairness–accuracy trade-offs across $\alpha \in \{0.1, 0.05\}$.

Revised Text:

3.4 Proposed Probabilistic Reranking Method

"[...] Regarding the role of the significance level α, suppose an observed recommendation contains $x = |E_{r,k}|_a$ experts from the disadvantaged group. We perform a significance test with the null hypothesis $H_0 : x \sim \mathbb{F}(x; k, q)$ and compute the corresponding p-value. If the p-value $> \alpha$, the observation is considered statistically plausible under H_0, and we therefore fail to reject the null hypothesis, treating the team as fair. Conversely, if the p-value $\leqslant \alpha$, the observation is deemed unlikely, leading to rejection of H_0 and the team being flagged as unfair. The choice of α directly controls the strictness of the fairness criterion. Larger values of α increase the likelihood of rejecting H_0, requiring stronger alignment with the distribution. For instance, from our earlier Examples 1–3, let the prior distribution of female experts be $p = 0.3$, $\alpha = 0.1$, and no female experts are observed in the top-k=10, i.e., $x = |E_{r,k=10}|_{a:female} = 0$. Based on demographic parity, the posterior distribution should match the prior, i.e., $q = p = 0.3$. Then, the p-value is $\mathbb{F}(x = 0; k = 10, q = 0.3) = 0.028 < 0.1$, and the null hypothesis is rejected, flagging the team as unfair. In contrast, for $\alpha = 0.01$, the same observation yields $0.028 > 0.01$, and the null hypothesis is not

rejected, resulting in the team being considered fair. From a mitigation perspective, larger values of α impose stricter fairness requirements, as more disadvantaged experts are needed for a team to pass the significance test. Smaller values of α relax this requirement and may not trigger mitigation even when the observed composition deviates from the expected proportion, as shown in Table 3 for $\alpha \in \{0.5, 0.1, 0.05, 0.01\}$. Meanwhile, the choice of α also controls the fairness trade-off with accuracy of the recommended experts, allowing more conservative (e.g., $\alpha = 0.1$) against permissive (e.g., $\alpha = 0.5$) additions of experts from the disadvantaged group by avoiding unnecessary enforcement of proportion constraints and preserving accuracy while maintaining statistically grounded fairness. [...]"

R1.15

In Table 6, explain why the exposure before fairness intervention for gender is already close to 1, despite skewness in gender representation suggesting female experts should receive lower exposure and utility.

Response: Thank you for raising this point, which prompted us for a careful re-examination of our implementation. We refactored and debugged the code and reran all experiments; the updated results are now reported in the revised Tables 8 to 15. The same pattern noted in your observation remains, although it is now less pronounced.

Regarding why the exposure metric for gender is already close to 1 before applying fairness intervention, this behavior follows directly from how the exposure metric (`expo`) is defined (Equation 39). When `expo` is close to 1, it indicates that female and male experts' exposures are aligned. Unlike `skew`, `expo` does not compare prior and posterior group distributions. As long as the visibility of a group aligns with its model-assigned probability, the ranking is considered fair under `expo`.

In our datasets, gender is severely imbalanced, with female experts forming a small minority. As a result, top-k recommendations often include only a few female experts, sometimes a single one, whose rank is relatively high compared to their total assigned probability. This is sufficient for the aggregate exposure of female experts to closely match their probability (utility) share, yielding a gender `expo` value near 1 prior to reranking. In contrast, for popularity, nonpopular experts form a large group and appear across more ranking positions, which reduces their exposure relative to utility and results in lower `expo` values.

Revised Text:

4.5 Results

"[...] The extreme imbalance of the datasets also contributes to `expo` values being close to 1 for gender even before reranking. Female experts constitute a small minority, and if only a few female experts, sometimes a single one, appear in the top-k recommendations, they occupy ranks consistent with their predicted probabilities. Consequently, their position-based visibility is not disproportionately low relative to their model-assigned utility, and the ratio between exposure and utility remains similar to that of male experts. Therefore, even limited presence at ranks aligned with predicted scores is sufficient to yield a value near 1. [...]"

R1.16

On Page 8, clarify why the proposed method might not select more qualified female experts to achieve the desired fairness, especially given the assumption of at least k experts from each protected group (Line 34). Explain why the method might avoid selecting experts from disadvantaged groups even when their desired proportion is not met in the current top-i iteration. If the greedy method does prioritize female experts to meet the desired proportion, explain how the proposed method enforces the evaluation fairness metrics (refer back to point 9).

Response: Thank you for this comment. We acknowledge that the original manuscript did not clearly explain the probabilistic nature of our method, especially the role of null hypothesis testing with a significance level α, which may have caused confusion. We have clarified this in our responses to R1.13, R1.14, and R2.5, and revised Sections **3.4 Proposed Probabilistic Reranking Method** and **4.3.2 Fairness-aware Reranking**, accordingly. We hope these revisions clarify the concern raised here. Specifically, in reply to R1.14, we detailed the role of the significance level and explained how fairness is ensured via a statistical significance test as opposed to strict selections of experts based on lower and upper bounds in deterministic reranker baselines. Our method calculates the required number of disadvantaged experts (e.g., females) at each top-k such that the observed number of disadvantaged experts in the top-k follows, statistically and significantly, given a significance level α, the binomial distribution of independent Bernoulli trials with winning probability q, which is set based on the underlying notion of fairness. As long as the observed count of disadvantaged experts passes the significance test (p-value $> \alpha$), the ranking is treated as fair, even if the target proportion is not strictly met at that particular prefix.

Moreover, in reply to R1.13 and Reviewer 2's comment R2.5, we explain limitations of deterministic baselines: 1) they may force the selection of additional experts from disadvantaged group even when the fairness constraint is already satisfied, making such selections unnecessary; and 2) when the disadvantaged group is also a minority group (e.g., female experts), this over-enforcement can lead to a drop in accuracy; and 3) very small prior representation for a disadvantaged group in imbalanced datasets can skew fairness evaluation, as deterministic methods may remove already included disadvantaged members to align the posterior distribution strictly with the prior.

We also revised the manuscript to a more accurate assumption of having at least $k \times p(e \in \mathcal{G}_a)$ experts from the disadvantaged group for demographic parity fairness, ensuring that the prior group distribution matches its posterior distribution among the top-k recommended experts for a successful team (Proposition 2), and $k \times p(e \in \mathcal{G}_a \mid \mathrm{S}_e \cap \mathrm{S} \neq \varnothing)$ *qualified* from the disadvantaged group for equal opportunity fairness, so that the prior distribution of qualified experts matches its posterior distribution among the top-k recommendations for a successful team (Proposition 3).

Revised Text:

3.4 Proposed Probabilistic Reranking Method

"[...] Regarding the role of the significance level α, suppose an observed recommendation contains x experts from the disadvantaged group. We perform a significance test with the null hypothesis $H_0 : x \sim \mathbb{F}(x; k = 10, q = 0.3)$ and [...]

[...] To form a fair ranking of k experts, we assume there exist at least ~~k members from each protected group.~~ $k \times q$ experts from the disadvantaged group. To satisfy demographic parity, we require $k \times (q = p(e \in \mathcal{G}_a))$ experts from the disadvantaged group, ensuring that the prior group distribution matches its posterior distribution among the top-k recommended experts for a successful team (Proposition 2). To satisfy equal opportunity, we require $k \times (q = p(e \in \mathcal{G}_a \mid S_e \cap S \neq \varnothing))$ qualified from the disadvantaged group, so that the prior distribution of qualified experts matches its posterior distribution among the top-k recommendations for a successful team (Proposition 3). [...]"

4.3.2 Fairness-aware Reranking

"[...] The deterministic reranking baselines have notable limitations. Having only two groups in our case, while considering the maximum threshold in the decision process is to prevent disproportionate advantage to one group, it may come at the cost of accuracy, that is, when one group has met the maximum threshold (e.g., males), these algorithms would select an expert from the other group (e.g., females) even with the lower probability, resulting in fairer yet less accurate team recommendation. In contrast, our method enforces only the minimum required representation of the disadvantaged group necessary to satisfy the fairness constraints, while selection is always guided by accuracy. [...]"

REVIEWER 2

This paper addresses the fairness problem in neural team recommendation systems. They suggest a post-processing model-agnostic reranking technique to address this problem. In this regard, they utilized three datasets in three distinct types of team recommendation tasks, illustrating how their method can mitigate existing biases against protected groups. The protected attributes in this work are popularity and gender. While their reranking mechanism works well in case of popularity bias, it falls short in addressing gender bias. To demonstrate the effectiveness of their work, they evaluated its impact against three deterministic reranking methods on various fairness metrics, including skew, ndkl and expo. To assess the effect of probabilistic reranking on utility, they utilized information retrieval metrics, including map and ndcg.

Strength:

- It is well-written, easy to follow, and understand.

- The problem is a critical yet underexplored issue concerning fairness, and this work offers a promising suggestion to address the existing bias in the system.
- The problem is well-defined and well-formalized.
- Fairness measures for evaluation and improvement are clearly defined and explained.
- Utility metrics have been used to show the fairness-utility trade-offs.
- Multiple datasets have been utilized for evaluation to avoid reliance on a single task or dataset.

Response: We are grateful for your careful reading and constructive feedback. We appreciate the positive evaluation of the clarity, problem formulation, and the experimental setup, as well as the recognition of the importance of addressing fairness in team recommendation systems.

R2.1

While mentioning the existence of other probability-based reranking methods, such as the one by Zehlike et al. [93], they did not use it as a baseline that is more similar to their approach, nor did they explain why it has not been used.

Response: We appreciate you pointing this out. We would like to clarify that our probabilistic method builds upon and adapts the framework of Zehlike et al.[31] for the team recommendation task, introducing several necessary modifications. Specifically, our approach generalizes the item-level ranking to team compositions, incorporates team-level success probabilities, which are not present in the original formulation, and alters the reranking procedure. As a result, the original method of Zehlike et al. cannot serve as a separate baseline, since it is effectively a special case of our team-level adaptation. We have revised the manuscript to clarify this point in multiple places, including the **Abstract**, **1 Introduction**, **2.2 Fairness-aware Recommendation**, and **4.3.2 Fairness-aware Reranking** sections, making the lineage and novelty of our method explicit while being transparent about the connection to Zehlike et al.'s work[31].

Revised Text:

Abstract

"[...] Inspired by the promising performance of probabilistic rerankers in user-item recommender systems for fairness guarantees, w~~W~~e further ~~propose~~develop a probabilistic greedy reranking algorithm [...]"

1 Introduction

"[...] Specifically, building upon the promising performance of probabilistic reranking methods for fair user-item recommendation, we ~~propose~~develop a probabilistic fairness-aware reranking method to [...]"

2.2 Fairness-aware Recommendation

"[...] To the best of our knowledge, there is yet to be a neural team recommendation method that specifically takes fairness into account except Loghmani et al.[47], wherein the application of deterministic reranking algorithms[34] to mitigate popularity bias in neural team recommenders[38,37,40] were shown futile due to the substantial compromise to the models' accuracy. Building upon Zehlike et al.[31]'s work, we adapt a probabilistic reranking method for the team recommendation task. However, unlike item-level ranking in their work, team recommendation requires rankings of experts for a team compositions while enforcing fairness over the team as a whole. Our method incorporates team-level success probabilities instead, which are absent in Zehlike et al."

4.3.2 Fairness-aware Reranking

"Our fairness-aware reranking baselines include three deterministic greedy reranking algorithms `det-greedy`, `det-cons`, and `det-relaxed` by Geyik et al.[34] as well as our proposed probabilistic reranking method with the significance level α=0.10 and 0.05. Our probabilistic method builds upon the framework of Zehlike et al.[31] and adapts it for the team recommendation task, introducing necessary modifications to handle team compositions by altering the reranking procedure. Consequently, the original method cannot serve as a separate baseline, as it is effectively a trivial special case of our team-level adaptation."

R2.2

While equalized odds has also been defined as a notion of group metric, it has not been used to evaluate the suggested algorithm, and the justification provided for this is unconvincing. (That we only care about qualified members of the protected group). While non-qualified members of a non-protected group receive positive decisions, their protected counterparts should be treated similarly.

Response: We thank the reviewer for raising this point. Our decision to omit evaluating equalized odds is not to disregard unqualified experts, but rather based on the team recommendation optimization during training. In our setting, positive decisions for unqualified experts (i.e., recommended for a team while unqualified) correspond to false positives, which are explicitly penalized by the training objective (Section **3.1 Preliminaries** and Definition **2 Neural Team Recommendation**), and later evaluated by accuracy metrics, as recommending unqualified experts degrades team accuracy. Consequently, such outcomes are already discouraged by the model. Enforcing parity (fairness) over these undesirable outcomes, as required by equalized odds, is therefore redundant and may conflict with the recommendation objective. For instance, in our running example in Table 2, all candidate recommended teams consist of qualified experts, except row #4, which includes the unqualified expert e_5 and decreases the accuracy metric. Such outcomes are already penalized during training. Ensuring that unqualified experts are also equally distributed across protected groups is an unnecessarily strong and practically misaligned constraint, particularly given that well-performing recommenders avoid recommending unqualified experts. Similar considerations have been studied in other predictive modeling settings[48,49,50]. From Tang and Zhang[48],

> *"It has been shown that if base rates of positives differ among groups, then Equalized Odds and Predictive Rate Parity cannot be achieved simultaneously for non-perfect predictors[49,50]. Any two out of three among Demographic Parity, Equalized Odds, and Predictive Rate Parity are incompatible with each other[51]."*

We have clarified this rationale in the revised manuscript in Section **3.3.5 A Fair Team**:

> *"While equalized odds is a more comprehensive notion of fairness by enforcing parity not only among qualified experts, as in equal opportunity, but also among unqualified experts for the favorable decision of being in a recommended successful team, we do not adopt it in this work. In the context of team recommendation, enforcing parity over unqualified experts is redundant and misaligned with the recommendation objective, wherein models are explicitly trained to identify qualified experts from the required skill set S, and unqualified experts are penalized by the learning objective, as discussed in Section 3.1 and Definition 2. Consequently, outcomes involving unqualified experts are already controlled downstream by the model during training. This observation has also been studied in other predictive models[48,49,50,51], leading to the adoption of equal opportunity as a more focused and actionable fairness notion."*

R2.3

Results are presented for top-k=100 (as per Section 4.2.1), and the study lacks an evaluation of its methodology concerning various values of k.

Response: We apologize for the confusion and acknowledge that our use of the values 100 and 10 for k was unclear. We discuss this together with the comment in R2.4 below.

R2.4

According to the dataset definitions, the actual team sizes in the datasets are small (between 1.88 and 3.06). To achieve a more reliable evaluation of the method, it would have been preferable to use these team sizes (or various team sizes) for evaluation. As per section 4.3.1, they used k=10 (which, however, differs from k=100 as mentioned in 4.2.1). I suggest providing more clarification on these other values of k and the team sizes used.

Response: Thank you for pointing this out. In our experiments, we actually used two different cut-offs, serving distinct purposes. The first cut-off is an internal parameter used purely for computational efficiency in code execution. It limits the expert pool by only the top-K (uppercase K) probabilities from the neural model in the last layer (Equation 5) and reduces memory usage. The neural model outputs a probability for each expert in the expert pool $|\mathcal{E}|$. Given a test set of size $|test|$, storing all probabilities would require a large matrix of $|\mathcal{E}| \times |test|$. For memory and storage efficiency in the underlying reranking methods, we did a first-stage retrieval (best practice in Information Retrieval systems). For each test team, we rank the probabilities once and retain only the top-K=100, which yields a manageable matrix of size $100 \times |test|$. This step is a common practice for efficiency in such systems and was not intended to be mentioned in the manuscript. However, at the reviewer's request, we can add a brief explanation of the first-stage cut-off of K=100 applied to the predicted probabilities. The second cut-off, denoted as k (lowercase k), is the central experimental parameter, representing the final recommended team size and the focus of our evaluation. Unfortunately, the similarity in the roles of these two cut-offs led to confusion, and the value 100 reported in the paper should have been 10 and associated with k. We have revised the manuscript to correct this issue.

Regarding the top-10, as the final recommended team of size k (second cut-off), although the average team size in the datasets is small (between 1.88 and 3.06), applying fairness-aware reranking at such small cardinalities can lead to over- or underestimation of the actual fairness metrics. For example, if the desired ratio between female and male experts is 0.3 to 0.7, then for a team of size 2 this corresponds to 0.6 female and 1.4 male members. The only feasible integer splits are (0 female, 2 male) or (1 female, 1 male), yielding ratios of 0 or 0.5, which deviate substantially from the target 0.3 and can result in large inaccuracies when reporting fairness metrics. Similarly, for a team of size 3, the expected split is 0.9 and 2.1, with feasible splits (0 female, 3 male) or (1 female, 2 male), giving a ratio of approximately 0.33. As shown, using a larger k allows more accurate calculation of posterior distributions and reduces rounding errors and large deviations. On the other hand, as noted, real-world teams are often of size 2–3 members. To balance the need for meaningful posterior calculation with real-world team sizes, we therefore adopt k=10 as the primary evaluation setting, where, for example, $0.3 \times 10 = 3$ female and $0.7 \times 10 = 7$ male members can be exactly realized. We have clarified this rationale in Section **4.3.1 Neural Team Recommendation**.

Finally, in response to the reviewer's suggestion to evaluate additional team sizes, we have extended the revised manuscript to include results for k=5 in Tables 9, 12 and 15, which exhibit consistent trends as in k=10.

Revised Text:

4.3.1 Neural Team Recommendation

"[...] Given a team (S, E) from the test set, we select the top-$k \in \{5, 10\}$~~=100~~ experts with the highest probabilities as the recommended team $E = f_{\theta}(S)$ by the model of each fold. Although the average team size in the datasets is small (between 1.88 and 3.06), evaluating fairness-aware reranking at such small cardinalities can lead to over- or underestimation of the fairness metrics. For example, if the desired ratio between female and male experts is 0.3 to 0.7, then for a team of size k=2, this corresponds to 0.6 female and 1.4 male members and feasible splits are either (0 female, 2 male) or (1 female, 1 male), yielding ratios of 0 or 0.5 and deviates substantially from the target 0.3. Hence, it can result in large inaccuracies when reporting fairness metrics. Similarly, for a team of size 3, the expected split is 0.9 and 2.1, with feasible splits (0 female, 3 male) or (1 female, 2 male), giving a ratio of approximately 0.33. As shown, using a larger k allows for a more accurate calculation of posterior distributions and reduces rounding errors and large deviations. On the other hand, as noted, real-world teams are often of size 2 to 3 members. To balance the need for meaningful posterior calculation with real-world team sizes, we therefore adopt top-$k \in \{5, 10\}$. [...]"

4.4.1 Before Mitigating Bias

"[...] report the average performance of models on all folds in terms of information retrieval metrics including mean average precision (`map`) and normalized discounted cumulative gain (`ndcg`) at top-$k \in \{5, 10\}$~~=10~~, as explained in Section 4.4.3. [...]"

R2.5

According to the results presented, with the fixed setting used to evaluate the effect of their method on mitigating bias against unpopular individuals and women, their method is successful in reducing popularity bias but fails to address gender bias. While gender bias is more important to mitigate (according to the works in the literature, popularity bias is not always a real bias), why should this method be used when it fails to address more critical biases? As justification for this shortcoming, the authors cited the significant gender bias as a reason for the probabilistic reranking technique's failure to mitigate said bias, as it lacks qualified female candidates to select from. This makes it even more crucial to address, because while smaller biases need to be mitigated, addressing bigger ones is even more vital. However, the results indicate that even in circumstances such as demographic parity in dplp, where females are overrepresented, applying this technique can create a negative bias against them, resulting in underrepresentation. To address this issue, the authors proposed employing in-processing techniques as future work. While one key contribution of this work is its model-agnostic approach using a post-processing technique, it is essential to ensure that there are no effective in-processing fairness techniques in place first.

Response: Thank you for the thoughtful comment. We fully agree that addressing gender bias is of high importance. To investigate the limitations of our probabilistic reranking method with respect to gender bias, we conducted an in-depth study to identify the main factors contributing to its reduced effectiveness. To ensure correctness and robustness, we refactored and debugged our implementation, reran all experiments from scratch, and incorporated unit tests to validate the computations. The updated results are now reported in the revised manuscript.

The new findings indicate that our method generally follows the same trend as for popularity bias and is successful in mitigating gender bias across most datasets. Nonetheless, certain limitations remain, particularly in cases with extreme underrepresentation of female experts, where post-processing alone is insufficient. We acknowledge that further improvements may require complementary pre-processing or in-processing techniques to fully address these extreme cases. To our knowledge, no existing pre- or in-processing fairness method for neural team recommendation. There are few works like [52,53] but they are purely algorithmic or rule-based rather than optimization- or machine-learning–based. They do not train predictive models or rely on data-driven parameter learning in the machine learning sense, and therefore, cannot be categorized as pre-processing or in-processing methods within a machine-learning pipeline. We have begun exploring in-processing approaches, with early results reported in our VivaFemme workshop paper on imdb[54]; however, that work relies on manually imposed posterior gender ratios. Preprocessing approaches such as oversampling may improve demographic outcomes, but they introduce artificial distributions and are likely to incur accuracy loss, although this trade-off has not yet been quantified in team recommendation in the literature. We view the demonstrated limitations of reranking under extreme sparsity as one of the contributions of this work. Our findings show that it is not possible to achieve strong mitigation of structural demographic bias using a fully model-agnostic, post-processing framework without modifying the model or the data. We therefore position reranking as a complementary solution and identify pre- and in-processing methods as important future directions.

Regarding the concern that popularity is not always a true bias, we address this point in our response to Reviewer 1 and in the revised manuscript in the newly added Section **3.3.1 Protected Attribute and Group**.

Revised Text:

4.5 Results

"[...] Thus, achieving strong mitigation of structural demographic bias using a fully model-agnostic, post-processing framework is not feasible without modifying the underlying model or the data. To our knowledge, no existing pre- or in-processing fairness method has been developed for neural team recommendation. A few prior works[52,53] exist, but they are purely algorithmic or rule-based rather than optimization- or machine-learning–driven. They do not train predictive models or leverage data-driven parameter learning, and therefore cannot be categorized as pre- or in-processing approaches within a machine-learning pipeline. Developing effective pre- and in-processing fairness methods for neural team recommendation remains an important direction for future work. [...]"

R2.6

While Table 6 indicates that the authors concluded the effectiveness of their technique across all metrics and datasets, it should be taken into account that in certain cases (for imdb, gender bias, equity of opportunity for both baselines, and demographic parity for bnn-emp, as well as uspt for gender and demographic parity) this method even increased bias, and the improvements in other cases are marginal.

Response: Thank you for this comment. As stated in our responses in R2.5, we reran all experiments, and the updated results, reported in Tables 8 to 15 now exhibit a more consistent trend in mitigating both popularity and gender bias after reranking. We refined our analysis and included additional evaluations for top-5 prefixes and experiments under updated α values, as discussed in our response to Reviewer 1 in R1.15. We explicitly acknowledge and analyze these cases in Section **4.5 Results**. Accordingly, we revised the manuscript to avoid claims of uniform improvement across all datasets and fairness notions.

Revised Text:

4.5 Results

*"**Finding 5.** In the context of neural team recommendation, our proposed probabilistic reranking method ~~consistently~~ generally outperforms deterministic reranking methods ~~on all~~ across datasets and baselines.*

***Finding 6.** Our probabilistic reranking method's performance in terms of `expo` is ~~consistent~~ largely aligned with its fairness metrics namely `ndkl` and `skew` across ~~all~~ most settings and domains.*

[...] As shown in Tables 8 to 15, all methods, including our baselines, effectively reduce bias (the green color in the fairness metrics columns for `skew` and `ndkl` highlights the substantial improvement after reranking), but unlike the baselines as expected, our method maintains accuracy despite reducing bias (green cell for accuracy metrics, `map` and `ndcg`, as opposed to the red cell for the baselines), demonstrating that it effectively balances the fairness-accuracy trade-off typically encountered in reranking scenarios. [...]"

R2.7

The order of protected and non-protected groups in Table 5 differs from that in Tables 3 and 4; it is better to maintain the same order to make the tables more readable.

Response: Thank you for bringing this error to our attention. We have corrected the ordering of the disadvantaged (~~protected~~) and advantaged (~~nonprotected~~) groups in all result tables to ensure consistency.

Received: Added at production | Revised: Added at production | Accepted: Added at production

DOI: xxx/xxxx

ARTICLE TYPE

A Probabilistic Greedy Attempt to be Fair in Neural Team Recommendation

Hamed Loghmani | **Mahdis Saeedi** | **Gabriel Rueda** | **Edwin Paul** | **Hossein Fani**

[1] School of Computer Science, University of Windsor, Ontario, Canada

Correspondence
Corresponding author Hossein Fani,
Email: hfani@uwindsor.ca

Abstract

Neural team recommendation has brought state-of-the-art efficacy while enhancing efficiency at forming teams of experts whose success in completing complex tasks is almost surely guaranteed.~~; albeit~~However, they overlook fairness, that is, predicted teams are heavily biased toward popular and male experts, falling short of recommending *female* or *non*popular experts. In this work, we introduce and formalize the *fair team recommendation* problem in view of group-based notions of fairness. Inspired by the promising performance of probabilistic rerankers in user-item recommender systems for fairness guarantees, w~~W~~e further ~~propose~~develop a probabilistic greedy reranking algorithm to achieve fairness with respect to popularity or gender biases in neural models ~~with respect to~~based on different notions of fairness, including *demographic parity* and *equal~~ity of~~ opportunity*. Specifically, we aim to ensure a minimum representation of experts from the disadvantaged nonpopular or female groups by reranking the neural model's ranked list of recommended experts. Our experiments on three large-scale benchmark datasets demonstrate: 1) neural team recommenders heavily suffer from biases toward popular and male experts; 2) our reranking method can substantially mitigate such biases while maintaining teams' efficacy; 3) in the presence of extreme biases in specific domains like gender disparities in US patents, post-processing reranking methods alone fall short to demonstrate consistent mitigation performance across *all* fairness evaluation metrics, urging further tandem integration of pre-process and in-process debiasing techniques. The code to reproduce the experiments reported in this paper is available at ~~`https://github.com/fani-lab/Adila/`~~`https://github.com/fani-lab/Adila/tree/coin25`.

KEYWORDS
Fair Team Recommendation, Neural Team Recommendation, Social Information Retrieval.

1 | INTRODUCTION

As modern tasks have surpassed the capacity of individuals, forming teams of experts whose collaboration for a common goal yields success has been a surge of research interest in many disciplines, including psychology[55,56], the science of team science (SciTS)[57], and industrial engineering[58]. Forming teams can be seen as social information retrieval (Social IR) where the right group of experts are searched and hired to solve the task at hand. Traditionally, teams were formed manually by relying on human experience and instinct; a tedious, error-prone, and suboptimal process for an overwhelming number of experts, a multitude of objectives to optimize (e.g., budget, time and team size constraints), and hidden personal and societal biases, among other reasons. As a result, a rich body of various computational methods, from operations research[59,60,61,62,63,64,65,66,67,68,69], social network analysis[70,71,72,73], and recently, machine learning[74,75,38,40,76,77,78,79,39,37,80,81,82,83] have been proposed. Specifically, neural models learn the distributions of experts and their skill sets in the context of successful and unsuccessful teams from training datasets to recommend future teams that are *almost surely* successful. Such models have brought state-of-the-art efficacy while enhancing efficiency, taking the stage and becoming canonical in team recommendation literature.

The primary focus of existing team recommenders is, however, the maximization of the models' accuracy (utility), largely ignoring the fairness in their ranked list of recommended experts, leading to discrimination, reduced visibility for already

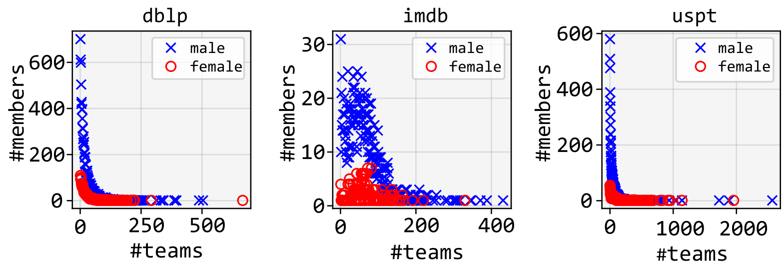

FIGURE 1 Distribution of experts in terms of gender in `dblp`, `imdb`, and `uspt` datasets.

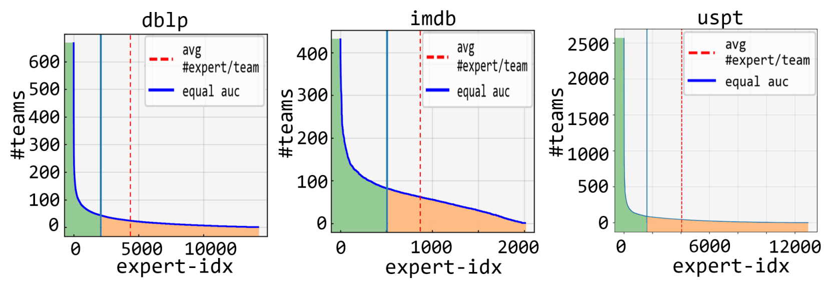

FIGURE 2 Distribution of experts in terms of popularity in `dblp`, `imdb`, and `uspt` datasets. While defining popularity can be controversial, recommender system literature follows *sociometric* popularity[84], where items (herein, experts) of the *head* in the participation distribution are labeled as popular items. As seen, there are two alternatives to split the distribution into *head* and *tail* parts: 1) the average number of teams per expert over the entire dataset, or 2) equal area under the curve (`auc`). In this paper, we opt for the former. For the `dblp` dataset, the average stands at 23.02 teams, 62.45 teams for the `imdb` dataset, and 44.69 teams for the `uspt` dataset. Therefore, in `dblp`, the proportion of popular to nonpopular experts becomes 0.313 to 0.687, in `imdb`, it is 0.426 to 0.574 and for `uspt` it stands at 0.314 to 0.686.

disadvantaged experts, and gender disparities[85,31,86]. These unfair biases, far from being random, originate mainly from training datasets. As seen in Figure 1, such datasets are segmented toward *male* experts, and *female* experts are heavily under-represented, like in the `dblp` dataset of computer research articles with %86 male vs. %14 female researchers. Also, from Figure 2, datasets in team recommendation suffer from popularity bias; that is, the majority of *non*popular experts have scarcely participated in the (successful) teams, whereas a few popular experts dominates many teams. Therefore, popular or male experts would receive more attention and are more frequently recommended by a machine learning model, leading to systematic discrimination against already disadvantaged nonpopular or female experts. To the best of our knowledge, there is no fairness-aware approach in neural team recommendation methods except that of Loghmani et al.[47], who applied deterministic greedy reranking algorithms to mitigate popularity bias and showed such deterministic methods can mitigate bias but at the cost of a substantial drop in the models' accuracy.

In this paper, foremost, we introduce and formalize the *fair team recommendation* problem to foster standards and conventions, which the literature on the team recommendation problem lacks. We set forth a unified set of notations to define the problem in view of group-based notions of fairness, including demographic (statistical) parity[87,14,11,12], equalized odds[7,88,13,48,89], and equal~~ity of~~ opportunity[7,90]. We further incorporate the notions of fairness in tandem with experts' skills in team recommendations to facilitate recommending merit-based teams while ~~equal~~fair opportunity is also maximized. Specifically, building upon the promising performance of probabilistic reranking methods for fair user-item recommendation, we ~~propose~~develop a probabilistic fairness-aware *re*ranking method to adjust the ordering of experts in the final ranked list of recommended experts to address potential biases and promote fairness concerning gender or popularity biases. ~~As opposed to pre-processing-based methods, which modify data or its labels before model training, or in-processing techniques, which focus on balancing model accuracy with fairness considerations during training, our method is model-agnostic and belongs to *post-processing* category of methods, which seek to improve the fairness of model's outputs after training, without adjustments to the data, training procedure, or the model's architecture~~As opposed to pre-processing-based methods[19,20,21,22], which, despite being model-agnostic, require direct access to raw and potentially sensitive training data[18,17] along with substantial computational resources to modify the data or its labels prior to model training[15,16], or in-processing techniques[23,24,25,26,27,28], which are model-dependent and focus on balancing model accuracy with fairness considerations during training, our method belongs to *post-processing* category of methods[29,30,31,32,33,27,34,35], which are simple and computationally efficient by improving the fairness of model's outputs after training, without adjustments to the data, training procedure, or the model's architecture. Moreover, being probabilistic, our approach holds advantages over deterministic methods for managing real-world uncertainties. Instead of providing rigid decisions, our approach offers distributions over possible outcomes, resulting in more adaptive solutions. To illustrate the effectiveness of our proposed approach, we perform experiments on three large-scale benchmark datasets of computer science articles (`dblp`)[91,72], moving pictures (`imdb`)[92,70], and US patents (`uspt`)[93]. Our results show that our proposed approach substantially mitigates popularity and gender biases while maintaining the ~~success rate~~accuracy of the recommended teams. With respect to gender bias in the specific domain of US patents, however, our approach's impact has been marginal due to the highly sparse distribution of female experts in the training datasets ~~(%14, %12, and %14 in `dblp`, `imdb` and `uspt`, respectively)~~, urging further future studies on the integration of pre-process and in-process debiasing techniques.

In summary, our key contributions are as follows:

1. We defined the problem of fair team recommendation in view of group-based notions of fairness including demographic (statistical) parity, equalized odds, and equal~~ity of~~ opportunity.
2. We proposed a model-agnostic post-processing and probabilistic reranking method to mitigate unfair biases in the recommended teams of experts by neural team recommendation models.
3. We demonstrated the performance of our proposed method in the presence of gender or popularity biases with respect to demographic parity and equal~~ity of~~ opportunity on three large-scale datasets from different domains.
4. We developed an open-source reproducible framework hosting canonical neural models as the cutting-edge class of approaches, along with large-scale training datasets from varying domains that integrated our proposed and baseline debiasing reranking algorithms.

Our work addresses the ever-growing need to identify and facilitate successful yet diverse teamwork based on merit while fairness is also maximized, which is one of the pillars of growth in scientific and industrial communities. Employers will be able to identify highly-skilled, diverse workers to fill labour gaps and increase innovation. As AI-based solutions are making notable impacts on how job opportunities are allocated to various groups in society, systematic consideration of fairness in this process is key. The rest of the paper is organized as follows: we first present the related works in Section 2, then we continue with the problem definition, where we elaborate basic foundations and formalize fairness objectives based on which a fair team is defined. We propose our approach in Section 3. The experimental setup and evaluation are described in Section 4, followed by concluding remarks in Section 5.

2 | RELATED WORK

The works related to this paper are largely around 1) neural team recommendation methods and 2) fairness-aware recommendation methods.

2.1 | Neural Team Recommendation

Among the proposed team recommendation methods, we focus on neural models as the cutting-edge computational methods which offer efficiency and effectiveness due to the inherently iterative and online learning procedure. Proposed neural team recommendation models include non-variational feedforward[37,94], variational Bayesian network[76,37,38,94], and graph neural network[80,39,46]. Initially, Rad et al.[37] defined team recommendation as a multilabel classification task and, as a naive baseline for a minimum level of comparison, developed a simple feedforward network with one hidden layer to map the required subset of skills in the input layer onto a subset of experts in the output layer using the standard cross-entropy loss. Rad et al.[37,38] then proposed a variational Bayesian network to mitigate the popularity bias through uncertainty in neural model weights in the form of Gaussian distributions. In this line, Dashti et al[40] further proposed negative sampling heuristics assuming groups of experts who have little or no collaborative experience for the required subset of skills have a low chance for a successful collaboration and can be considered as *virtually un*successful teams. Given that popular experts were dominant in the training datasets, Dashti et al. presume that groups of popular experts are more likely to be selected as negative samples of teams, hence trying to mitigate popularity bias. Successfully as they are, the primary focus of Dashti et al. and Rad et al. was the maximization of the efficacy by tailoring the recommended experts for a team to the required skills only, overlooking to substantiate whether the higher efficacy comes with mitigation of popularity bias.

Sapienza et al.[80] were the first to use a graph neural network in the form of an autoencoder for team recommendation in online multiplayer games. Later, Rad et al.[39] proposed to transfer dense vector representations of skills for the input of variational Bayesian neural network from a heterogeneous graph, whose nodes are teams, experts, skills, and locations and edges connect experts who have collaborated in a team residing in a location, using Dong et al.'s `metapath2vec`[95] and obtained the state-of-the-art performance. More recently, Kaw et al.[46] employed deep graph infomax[96], a graph convolution network with attention layer as an encoder, to learn more effective vector representations of skills in less training epochs owing to the convolutional architecture and contrastive learning procedure.

Nonetheless and despite a few efforts[37,38,40], existing neural team recommendation models still withhold extreme biases. Meanwhile, accounting for fairness in neural models has gained significant importance for their widespread applications in everyday lives, like in healthcare[97,98,99,100], information retrieval[101,102], computer vision[103,104,105,106,107], and recommendation systems[31,108,109,25,34,26,110,111]. To this end, in this paper, we are among the first to formalize the fair team recommendation problem with respect to group-based notions of fairness and undertake an empirical investigation to bridge the fairness gap through a probabilistic post-processing reranking method in favor of recommending more female or nonpopular experts while controlling the accuracy of the recommended teams.

2.2 | Fairness-aware Recommendation

Theoretically, fairness guarantees in machine learning algorithms have been defined at an individual level[9] where an individual should be treated consistently[10] or based on a group of individuals where a disadvantaged group~~, also known as a protected group,~~ should be treated similarly to the advantaged group as a whole[8,6]. Different fairness-aware methods have been proposed to either discover and measure unfair biases[112], or to mitigate them via debiasing algorithms[31,108,33,34,110,35] at individual or group levels.

Debiasing algorithms can further be categorized based on their placement in the machine learning pipeline: 1) pre-processing[19,20,21,22] methods modify data or its labels by re-sampling heuristics before model training, 2) in-processing[23,24,25,26,27] techniques modify models' optimization process to trade-off accuracy with fairness considerations, and 3) post-processing[29,30,31,32,33,27,34,35] methods modify models' outputs during inference, which may involve modifying thresholds, scoring rules, or reranking of the recommended list of items[113,7]. The latter category is of particular interest ~~for it can be model-agnostic and plugged into a model with little to no modification to the model's architecture or negative impact on its predictive power.~~due to several practical advantages. Foremost, post-processing methods are computationally efficient, as they do not require access to or manipulation of raw training data, which are typically medium- to large-scale, unlike pre-processing or in-processing methods[15,16]. Secondly, post-processing methods can be readily plugged into already trained and deployed models without retraining or fine-tuning, a scenario that is increasingly common in realistic settings where pre-processing or in-processing interventions are infeasible. Finally, post-processing methods are more privacy-preserving, as they do not require access to potentially sensitive raw training data, as in pre-processing methods, nor the architecture or parameters of private models, as in in-processing methods[17,18].

Related to this paper, we explain seminal reranking debiasing methods that achieve group-based fairness in recommendation tasks. Geyik et al.[34] propose greedy reranking algorithms to ensure prior desired distributions for disadvantaged protected group within the top-k items. At each iteration i; $1 \leqslant i \leqslant k$, lower and upper bounds are calculated for protected group members to guarantee the desired distribution within top-i. To measure bias in original rankings and rerankings, they use `skew`[34] and normalized discounted cumulative KL-divergence (`ndkl`)[112]. Geyik et al.'s algorithms are, however, deterministic and fall short in the presence of real-world uncertainties. In contrast, Zehlike et al.[31] have proposed a probabilistic method to produce a top-k ranking while maintaining fairness towards multiple protected groups. They rely on statistical tests and aim for a minimum proportion of protected items in each subset of the ranked items. Utilizing cumulative distribution functions, they calculate the minimum number of required protected items at a given position to hold the fairness criteria with a pre-defined confidence level. Instead of providing rigid decisions, they offer distributions over possible outcomes, ensuring that items with small probabilit~~y~~ies are not completely disregarded.

To the best of our knowledge, there is yet to be a neural team recommendation method that specifically takes fairness into account except Loghmani et al.[47], wherein the application of deterministic reranking algorithms[34] to mitigate popularity bias

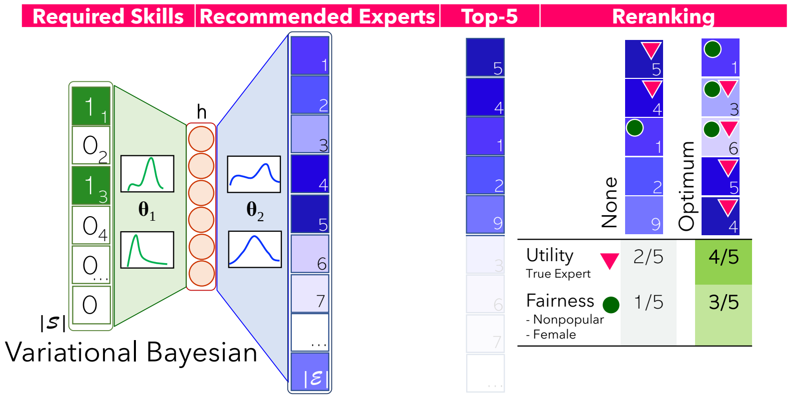

FIGURE 3 Post-processing fairness-aware reranking for fair neural team recommendation. ▽ indicates the correctly predicted expert (accuracy) and ○ shows members of the ~~protected~~disadvantaged group (fairness). The goal is to maximize fairness while maintaining the model's accuracy.

in neural team recommenders[38,37,40] were shown futile due to the substantial compromise to the models' accuracy. Building upon Zehlike et al.[31]'s work, we adapt a probabilistic reranking method for the team recommendation task. However, unlike item-level ranking in their work, team recommendation requires rankings of experts for a team compositions while enforcing fairness over the team as a whole.

3 | FAIR NEURAL TEAM RECOMMENDATION

In this section, we introduce the necessary notations and definitions for neural team recommendation, on the one hand, and group-based fairness, on the other hand. Then, we provide a formal problem statement to recommend a fair team of experts.

3.1 | Preliminaries

Given a set of skills $\mathcal{S} = \{s\}$ and a set of experts $\mathcal{E} = \{e\}$, a team of experts $\mathrm{E} \subseteq \mathcal{E}; \mathrm{E} \neq \varnothing$ that collectively cover a skill set $\mathrm{S} \subseteq \mathcal{S}; \mathrm{S} \neq \varnothing$ is shown by (S, E) along with its success status y where $y \in \{0, 1\}$ which is known *a priori*. Further, $\mathcal{T} = \{(\mathrm{S}, \mathrm{E})_y : y \in \{0, 1\}\}$ indexes all teams, successful and unsuccessful. In team recommendation literature[80,114,64,115,116,117,118,119], an expert's set of skills has been *estimated* through her participation in successful teams denoted by $\mathrm{S}_e = \{s : (s \in \mathrm{S}, e \in \mathrm{E})_{y=1} \in \mathcal{T}\}$, i.e., an expert member of a successful team inherits *all* the required skills of the team, as the expert obtains knowledge about all required skills through collaboration with other expert members of the team.

Definition 1 (**Team Recommendation**)**.** For a given subset of required skills S, the goal of the team recommendation problem is to recommend an optimal subset of experts E whose collaboration as a team leads to success, i.e., $(\mathrm{S}, \mathrm{E})_{y=1}$, while avoiding a potentially unsuccessful subset of experts E′, i.e., $(\mathrm{S}, \mathrm{E}')_{y=0}$. More concretely, the team recommendation problem is to find a mapping function f of parameters $\boldsymbol{\theta}$ from the power set of skills to the powerset of experts such that $f_{\boldsymbol{\theta}} : \mathcal{P}(\mathcal{S}) \rightarrow \mathcal{P}(\mathcal{E})$, $f_{\boldsymbol{\theta}}(\mathrm{S}) = \mathrm{E}$.

~~As shown in Figure the output of a neural team recommendation method for a required set of skills S is a ranked list of *all* experts where each expert $e \in \mathcal{E}$ is assigned a probability of her membership in the final recommended team. The recommended team is a subset of experts $\mathrm{E} \subseteq \mathcal{E}$ with the top-k highest probabilities.~~

Definition 2 (**Neural Team Recommendation**)**.** Given a subset of skills S and all teams $\mathcal{T}$ as the training set, neural team recommendation estimates $f_{\boldsymbol{\theta}}(\mathrm{S})$ using a multilayer neural network that learns, from $\mathcal{T}$, to map a vector representation of subset of skills S, referred to as v_{S}, to a vector representation of subset of experts E, referred to as v_{E}, by maximizing the posterior probability of $\boldsymbol{\theta}$ in $f_{\boldsymbol{\theta}}$ over $\mathcal{T}$, that is, $\mathtt{argmax}_{\boldsymbol{\theta}}\ p(\boldsymbol{\theta}|\mathcal{T})$.

For the vector representation of subset of skills v_S, neural team recommendation methods adopt either

1. The *occurrence* vector representation for S, which is a Boolean vector of size $|\mathcal{S}|$, i.e., $v_S \in \{0,1\}^{|\mathcal{S}|}$ where $v_S[s] = 1$ if $s \in$ S, and 0 otherwise; or
2. The *embedding* representation (emb) of S, which is a dense low d-dimensional vector, $d << |\mathcal{S}|$, pretrained by e.g., a shallow neural encoder[38] like in distributional representation of words (word2vec)[44] and documents (doc2vec)[45], or a graph neural network~~method~~[46,39].

While occurrence vector representations are simple, they lead to high-dimensional sparse input representations, substantially increasing the number of trainable parameters in the input layer of neural networks. In contrast, dense pretrained vectors (embeddings) provide low-dimensional, continuous representations that significantly reduce model complexity and enable more stable and data-efficient learning. Moreover, dense embeddings capture latent semantic relationships among skills, allowing the model to generalize across similar inputs rather than treating each skill as independent, as is the case with the occurrence representation.

In the output layer for vector representation of subset of experts v_E, neural team recommendation methods frame the problem as a multilabel Boolean classification task and used occurrence vector representation for E, that is, $v_E \in [0,1]^{|\mathcal{E}|}$ where $v_E[e] = 1$ if $e \in$ E, and 0 otherwise. Using a neural model of one hidden layer $\mathbf{h}$ of size d, without loss of generality to multiple hidden layers, with the input layer v_S and output layer v_E, a neural team recommendation method can be formalized as[94,40,77,39,37]:

$$\mathbf{h} = \pi(\boldsymbol{\theta}_1 v_S + \mathbf{b}_1) \tag{3}$$

$$logits \rightarrow \mathbf{z} = \boldsymbol{\theta}_2 \mathbf{h} + \mathbf{b}_2 \tag{4}$$

$$v_E = \sigma(\mathbf{z}) \tag{5}$$

where π is a nonlinear activation function, σ is the sigmoid function, and $\boldsymbol{\theta} = \boldsymbol{\theta}_1 \cup \boldsymbol{\theta}_2 \cup \mathbf{b}_1 \cup \mathbf{b}_2$ are learnable parameters for the mapping function f. During training, given a team (S, E), neural models tune the parameters $\boldsymbol{\theta}$ by maximizing the posterior probability of $\boldsymbol{\theta}$ in $f_{\boldsymbol{\theta}}$ over $\mathcal{T}$. From Bayes theorem:

$$\texttt{argmax}_{\boldsymbol{\theta}}\, p(\boldsymbol{\theta}|\mathcal{T}) \propto p(\mathcal{T}|\boldsymbol{\theta})p(\boldsymbol{\theta}) = p(\boldsymbol{\theta}) \prod_{(\mathrm{S},\mathrm{E})_{y=1} \in \mathcal{T}} p(\mathrm{E} \mid \mathrm{S}, \boldsymbol{\theta}) \tag{6}$$

$$p(\mathrm{E} \mid \mathrm{S}, \boldsymbol{\theta}) = \prod_{e \in \mathrm{E}} \sigma(\mathbf{z}[e]) \propto \sum_{e \in \mathrm{E}} \texttt{log}\, \sigma(\mathbf{z}[e]) \tag{7}$$

where $p(\mathcal{T}|\boldsymbol{\theta})$ is the likelihood and $p(\boldsymbol{\theta})$ is the prior joint probability of weights, which is unknown. While the exact prior probability of weights $p(\boldsymbol{\theta})$ cannot be calculated analytically[36], it can be estimated as Gaussian distributions by a variational Bayesian neural architecture (bnn) via the maximum a posteriori optimization[37,38,39], which contrasts with conventional *non*-variational multilayer perceptron that assume a *uniform* probability distribution over all possible real-values of $\boldsymbol{\theta}$ and only estimate the likelihood $p(\mathcal{T}|\boldsymbol{\theta})$ via maximum likelihood optimization, discarding prior uncertainty $p(\boldsymbol{\theta})$, and hence, resulting in overconfident point estimates of $\boldsymbol{\theta}$[36]. Existing studies primarily adopt standard multilayer perceptrons (non-variational feedforward neural networks) as baselines to demonstrate the performance advantages of variational neural networks[40,41,36].

As shown in Figure 3, the output of a neural team recommendation method for a required set of skills S is a ranked list of *all* experts where each expert $e \in \mathcal{E}$ is assigned a probability of her membership in the final recommended team which is a subset of experts $\mathrm{E} \subseteq \mathcal{E}$ with the top-k highest probabilities as a team of size k.

3.2 | Team Success Labels

From Definitions 1 and 2, neural team recommenders aim at estimating f from samples of teams that are labeled with success or failure, yet most available training data in team recommendation literature only consists of successful teams, missing *un*successful ones. For instance, the dblp lacks unsuccessful paper or manuscript submissions. Further, what it means for a team to be successful has remained controversial. For instance, in the movie industry, it is debatable whether a movie's success should be measured based on its immediate reception by the people (box office) or critical reviews (ratings) over a long span of time. In the absence of explicit labels for unsuccessful teams, neural team recommendation methods presume all instances of teams in the training dataset are successful (positive samples) for their observable outcomes, e.g., *published* papers in dblp, *produced* movies in `imdb`, and *issued* patents in `uspt`, that is, $\mathcal{T} = \mathcal{T}^+ = \{(\mathrm{S},\mathrm{E})_{y=1}\}$ and proceed with the training procedure[39,37,38,46]. As a result, during the inference, for an input required set of skill S, the models predict (recommend) a subset of experts E for a successful team only, i.e., $(\mathrm{S},\mathrm{E})_{y=1}$ and the learned probabilities in Equations 5 and 6 should be interpreted as expert's membership likelihood for a successful team. Throughout the remainder of this paper, we therefore assume $y = 1$ whenever a team $(\mathrm{S},\mathrm{E})_y$ is considered, and the act of recommending a team is understood as recommending a potentially successful team, as recommending *un*successful teams $(\mathrm{S},\mathrm{E})_{y=0}$ is not desired.

3.3 | Fair Team Recommendation

To eschew varied interpretations and to provide actionable criteria to design and evaluate fairness-aware algorithms, fairness has been mathematically formalized, with a level of abstraction from an underlying real-world scenario, based on well-known notions of justice and equity at an individual level[9,10], or at a group level[1,7,8,6] like female vs. male experts. In this paper, we focus on group-based notions of fairness.

3.3.1 | Protected Attribute and Group

We define a (social) group $\mathcal{G}$ as a subset of the experts $\mathcal{E}$ that shares an *"identity trait"*[1], which may be inherent to an expert and largely immutable, that is, cannot reasonably be expected to change through choice, effort, or behavior, and is considered sensitive due to its historical association with systemic discrimination. As a result, legal frameworks and policies have been developed to protect the group members from differential treatment based on the group's shared identity trait and to promote fairness, equal opportunity, and anti-discrimination across institutional and algorithmic decision-making processes[120]. Hence, the shared identity trait based on which the groups are formed is referred to as the *protected attribute* $\mathcal{A}$ whose values create $|\mathcal{A}|$ *protected groups* such that experts of a protected group share the same value for the protected attribute.

Gender as a Canonical Protected Attribute: In this paper, given the protected attribute $\mathcal{A}$:gender, the values=$\{a : \textit{female}, a' : \textit{male}\}$ create *disjoint* protected groups of experts including female experts $\mathcal{G}_{a:\textit{female}}$ and male experts $\mathcal{G}_{a':\textit{male}}$ where $\mathcal{E} = \mathcal{G}_{a:\textit{female}} \cup \mathcal{G}_{a':\textit{male}}$ while $\mathcal{G}_{a:\textit{female}} \cap \mathcal{G}_{a':\textit{male}} = \varnothing$. To form the gender groups, we presume an expert's gender value is self-identified and is either available, e.g., in `uspt` dataset of US patent inventors, or can be inferred by the expert role, e.g., actor vs. actress in `imdb` dataset of movies, or the expert's name in `dblp` dataset of scholarly papers. It is worth noting that, when gender values are inferred rather than directly observed, they may inherently incorporate societal biases, or even amplify pre-existing stereotypes in society[42,43], as further illustrated by examples in our experimental setup (Section 4.2). Therefore, fairness assessments and bias mitigation using these inferred values may over- or under-estimate disparities across groups, and results should be interpreted with this limitation in mind. As seen earlier in Figure 1, experts in team recommendation datasets are disproportionately distributed toward male experts. Therefore, we consider the *minority* female group as the disadvantaged group vs. the *majority* male group as the advantaged group.

Popularity as an Unconventional Protected Attribute: As highlighted by Gallegos et al. and others[1,2], a group can also be formed based on an identity trait that is contextual and dynamic, e.g., disability or religion, or socially constructed based on social network effects and historical contingencies, like outcomes or conditions that arise from past events, as in popularity[3,4,5]. However, treating popularity as a protected attribute receives limited or no attention within legal frameworks and policies, as it is weakly aligned with the characteristics of commonly known protected attributes like gender, especially in the context of the team recommendation problem. 1) Foremost, protected attributes are typically defined to protect a *minority* group from being marginalized by a *majority* group, e.g., to ensure that female experts are not under-represented in team recommendations relative to male experts. In contrast, when considering popularity, the aim is reverse, that is, to protect the majority nonpopular experts from the minority popular experts who disproportionately dominate team participation. 2) Secondly, popularity can act as a proxy, a justified signal of merit, or a shortcut for expertise, experience, and success. Highly skilled experts tend to be repeatedly successful, and hence selected for more teams, which increases both their participation count and visibility, and eventually results in their being popular. Respectively, popularity correlates with perceived expertise[121,122], and favoring popular experts appears rational and performance-driven, rather than unfair[123,124]. 3) Finally, popularity can also be interpreted as being widely liked, acclaimed, or recognized by others, often associated with reputation or fame, reducing of which may itself be unfair to popular experts.

On the other hand, popularity can induce systematic undesirable outcomes, should it otherwise be treated as a protected attribute. Although an expert's participation in many teams, e.g., an author in many research papers in `dblp` or an actor in many movies in `imdb`, may *not* necessarily indicate popularity from the people's *subjective* perspectives, repetition of the expert in many training samples of teams from a machine learning model's perspective does. In particular,

- From societal perspective, when experts' popularity is used naively, either directly as a feature, that is, being popular or not, or indirectly through an imbalance pool of experts with long-tailed distribution of few but dominant experts (popular) in the head against majority experts (nonpopular) in the tail, team recommendation processes leads to, or amplify, persistent under-representation of less- or nonpopular experts, who may possess comparable expertise but lack historical opportunities to accumulate participation records, ultimately receiving reduced access to opportunities such as team membership. Moreover, popularity is influenced by factors orthogonal to expertise. Two experts with comparable skills may diverge in popularity due to institutional affiliation [123], geographic factors [125,126], or historical chances [121,127]. In such cases, consistently preferring the popular expert represents unjustified, unfair, and disproportionate exclusion, not merit-based selection, of nonpopular experts.
- From an algorithmic perspective, as extensively studied in team science [128,129,130], teams composed solely of popular experts are not necessarily optimal in terms of cost or availability of experts, or even successful due to the *"too-much-talent effect"* [131]; rather, complementary skills with heterogeneous levels of expertise provided by nonpopular experts are required. Canonical examples include research teams consisting of senior supervisors and junior students, or industrial teams balancing experts with trainees. Also, many contributions have come from nonpopular experts despite their lower visibility [132]. Moreover, practical constraints such as availability, cost, or workload often make popular experts infeasible choices, further undermining the assumption that popularity-driven recommendations are universally applicable [133,134,135].
- Popularity as a biased feature, as opposed to a protected attribute, is the largely shared understanding and has been well-studied in recommender systems in the context of item recommendation [136,137,138], where popularity is defined over items as a statistical bias arising from long-tailed (skewed) user-item interaction distributions. Yet, interestingly, few item recommender systems explicitly treat item popularity as a protected attribute [3,4] or a sensitive attribute [5]. They either partition items into popular (head) and nonpopular (tail) groups and adjust the learning process to protect the nonpopular items in the tail group, or explicitly model the effect of item popularity as a sensitive feature and remove its influence on rankings.
- Treating popularity as a protected attribute enables mitigation strategies in a unified and technically coherent algorithmic framework in which attributes, whether legally protected or socially constructed, are handled consistently, reducing implementation complexity and improving comparability across models.

Therefore, in social information retrieval systems, including team recommender systems, expert recommender systems, or social talent search engines [139,34,140], where the entities being recommended are *human* experts rather than items, we argue that the popularity should be treated as a protected attribute, with particular emphasis on protecting nonpopular experts as less-exposed disadvantaged group, to prevent unfair societal implications like systematic reduced access to team memberships [124,141]. Accordingly, in this paper, given the protected attribute $\mathcal{A}$:popularity, the values=$\{a{:}nonpopular, a'{:}popular\}$ create *disjoint* protected groups of experts including nonpopular experts $\mathcal{G}_{a:nonpopular}$ and popular experts $\mathcal{G}_{a':popular}$ where $\mathcal{E} = \mathcal{G}_{a:nonpopular} \cup \mathcal{G}_{a':popular}$ while $\mathcal{G}_{a:nonpopular} \cap \mathcal{G}_{a':popular} = \varnothing$. We consider the *majority* nonpopular group as the disadvantaged group vs. the *minority* popular group as the advantaged group.

To obtain an expert's popularity status, we followed social science [84] and recommender system literatures [142,143], where the popularity status of an expert can be *objectively* measured based on the number of teams the expert has participated in, referred to as *sociometric* popularity [84]. We can adopt two alternatives, as shown earlier in Figure 2: 1) an expert is popular if the expert participated in more than the average number of teams per expert over the entire dataset, and nonpopular otherwise (`avg`), or 2) we plot the distribution of experts in teams and split the curve into *short head* and *long tail* based on equal area under the curve (`auc`) should it be long-tail distribution. An expert is popular if she belongs to the short head, and nonpopular otherwise. As seen in Figure 2, experts in team recommendation datasets are disproportionately distributed toward popular experts.

~~Given a protected attribute $a = \{a_1, ..., a_n\}$, e.g., gender={`0`: male, `1`: female} or popularity={`0`: popular, `1`: nonpopular}, we divide experts into groups per their attribute values, each referred to as a protected group $\mathcal{G}_{a_i}$, e.g., female $\mathcal{G}_1$ vs. male $\mathcal{G}_0$ experts, or nonpopular $\mathcal{G}_1$ vs. popular $\mathcal{G}_0$ experts, such that experts of a protected group share the same value for a protected attribute.~~

We now define the notions of group fairness for team recommendation as follows.

3.3.2 | Demographic Parity

Demographic parity, also called statistical parity[6], is to provide equal treatment to protected groups, i.e., the proportion of individuals receiving a favorable outcome should be consistent across all protected groups ~~regardless of protected attributes.~~ according to their distribution in the population[14,11,12]. Given $\mathcal{D}$ the set of possible decisions, demographic parity requires the *predicted* decision $\hat{d} \in \mathcal{D}$ for members of protected groups to be oblivious to the protected attribute[6,1,11,12]. Formally,

$$\cancel{\forall d \in \mathcal{D}, i \neq j \in a; p(\hat{d} \mid e_{a_i}) = p(\hat{d} | e'_{a_j})} \tag{8}$$

$$\forall \hat{d} \in \mathcal{D}\ \forall a \in \mathcal{A}:\ \ p(\hat{d} \mid e \in \mathcal{G}_a) = p(\hat{d}) \tag{9}$$

where $\hat{d}$ is the predicted decision for the ~~correct~~ground-truth decision d, and ~~$e_{a_i} \in \mathcal{G}_{a_i}$ is an expert whose value of protected attribute a is a_i.~~ $e \in \mathcal{G}_a$ is an expert in a disjoint group $\mathcal{G}_a$ whose value of the protected attribute $\mathcal{A}$ is a. It is worth noting that demographic parity is also independent of the ground-truth decision, that is, whether the predicted decision is correct or otherwise.

In fair team recommendation, we assume decisions $\mathcal{D}$ are about the Boolean membership status of experts in ~~the team of experts E, i.e., $e \in$ E or $e \notin$ E~~ a successful team $(\text{S}, \text{E})_{y=1}$, i.e., $\hat{d} \in \mathcal{D}=\{e \in \text{E}, e \notin \text{E}\}$, and protected attributes $\mathcal{A}$ are either gender or popularity. Hence, Equation 9 becomes:

$$\cancel{\forall e_0, e'_1 \in \mathcal{E}; [p(e_0 \in \text{E}) = p(e'_1 \in \text{E})] \wedge [p(e_0 \notin \text{E}) = p(e'_1 \notin \text{E})]} \tag{10}$$

$$\forall a \in \mathcal{A}:\ \ [p(e \in \text{E} \mid e \in \mathcal{G}_a) = p(e \in \text{E})] \wedge [p(e \notin \text{E} \mid e \in \mathcal{G}_a) = p(e \notin \text{E})] \tag{11}$$

Proposition 1. *Demographic parity for fair team recommendation holds if and only if*

$$\forall a \in \mathcal{A}:\ \ p(e \in E \mid e \in \mathcal{G}_a) = p(e \in E) \tag{12}$$

Proof. Since $p(e \in \text{E} \mid e \in \mathcal{G}_a) + p(e \notin \text{E} \mid e \in \mathcal{G}_a) = 1$, it follows that $p(e \in \text{E} \mid e \in \mathcal{G}_a) = 1 - p(e \notin \text{E} \mid e \in \mathcal{G}_a)$. Therefore, Equation 12 directly implies $p(e \notin \text{E} \mid e \in \mathcal{G}_a) = p(e \notin \text{E})$. □

Intuitively, demographic parity requires that the likelihood of being recommended for a successful team is independent of the values of any protected attribute, such as popularity, gender, or ethnicity. In other words, belonging to a particular protected group should neither increase nor decrease an expert's chance of being part of a successful team.

In this paper, we focus on recommended teams within a post-processing reranking framework. To ensure fairness, we need to check, measure, and, if necessary, debias recommended teams after they have been predicted by a neural team recommendation model described in Section 3.1. In the following proposition, we show that demographic parity is satisfied if and only if the *posterior* distribution of experts from each protected group in the recommended team matches their *prior* distribution in the overall pool of experts $\mathcal{E}$. This result is particularly useful for post-hoc assessment and reranking. After recommending a team $(\text{S}, \text{E})_{y=1}$, we can examine the posterior probabilities of experts from each protected group and, if necessary, adjust team membership to satisfy fairness constraints.

Proposition 2. *Demographic parity holds in the fair team recommendation problem if and only if the posterior distributions of experts of protected groups in the successful team are equal to their prior distributions in the entire set of experts $\mathcal{E}$, i.e., $p(e \in \mathcal{G}_a \mid e \in E) = p(e \in \mathcal{G}_a)$.*

Proof. By Bayes theorem,

$$p(e \in \mathcal{G}_a \mid e \in \text{E}) = \frac{p(e \in \text{E} \mid e \in \mathcal{G}_a)\, p(e \in \mathcal{G}_a)}{p(e \in \text{E})} \tag{13}$$

$$= p(e \in \mathcal{G}_a);\ \text{from Equation 12} \tag{14}$$

□

Example 1. Consider a set of $|\mathcal{E}| = 100$ experts, where $\mathcal{A}$ denotes gender, and $|\mathcal{G}_{a:female}| = 30$ and $|\mathcal{G}_{a':male}| = 70$ represent the groups of female and male experts, respectively. Suppose we want to recommend a team of size k=10 for a required set of skills S. To satisfy demographic parity, each expert's probability of being selected can be assumed independent and identically distributed (i.i.d.), with uniform probability across all experts, which gives $\frac{k}{|\mathcal{E}|} = \frac{10}{100} = 0.1$ from Equation 12, regardless of the expert's gender. That is, every female and male expert has an equal chance of being selected. Conversely, from Proposition 2, demographic parity also requires that the composition of the recommended team reflects the overall population. Specifically, the prior distribution of female and male experts, $p(e \in \mathcal{G}_{a:female}) = \frac{|\mathcal{G}_a|}{|\mathcal{E}|} = \frac{30}{100} = 0.3$ and $p(e \in \mathcal{G}_{a':male}) = \frac{|\mathcal{G}_{a'}|}{|\mathcal{E}|} = \frac{70}{100} = 0.7$, should match their posterior distribution in the recommended team E with $p(e \in \mathcal{G}_{a:female} \mid e \in \mathrm{E}) = \frac{3}{10} = 0.3$ and $p(e \in \mathcal{G}_{a':male} \mid e \in \mathrm{E}) = \frac{7}{10} = 0.7$. In this way, the recommended team satisfies demographic parity.

~~Intuitively, demographic parity enforces the membership in a team to be independent of values of a protected attribute for team members, i.e., no regard to their popularity, gender, ethnicity, or any other protected characteristic. However, demographic parity overlooks experts' qualifications; no criteria for experts' memberships has been defined in Equation and Equation~~.

However, demographic parity alone overlooks the qualifications of experts; no criteria for membership have been defined in Equations 9 to 13. In Example 1, giving all experts an equal chance of selection, without considering their expertise with respect to the required set of skills S, would substantially reduce the quality of the recommended team E for a successful team $(\mathrm{S}, \mathrm{E})_{y=1}$. Indeed, Example 1 represents a worst-case scenario based on a random selection model; while it satisfies demographic parity, the team is essentially a uniformly random set of experts, which may not perform effectively. Therefore, ensuring fairness alone is not sufficient and expert qualifications must also be incorporated to maintain team quality while satisfying fairness, as captured by equalized odds and equal opportunity, as defined hereafter.

3.3.3 | Equalized Odds

Equalized odds[7,13] is a stronger notion of fairness. While demographic parity emphasizes equal treatment by ensuring a similar proportion of positive outcomes across protected groups, equalized odds go further by ensuring that the ranking is equitable across groups for both qualified and ~~non-~~unqualified ~~cases~~members. In other words, it applies demographic parity on subsets of protected groups, whose members are qualified ~~(not~~or unqualified~~)~~, for the ground-truth decision d to receive the ~~correct~~predicted decision $\hat{d}$~~in a Boolean decision set and Boolean protected attribute:~~. Formally,

$$\cancel{\forall d \in \mathcal{D} = \{0, 1\}\cancel{; p(\hat{d} \mid e_0, d)} = p(\hat{d} \mid e_1, d)} \tag{15}$$

$$\forall d, \hat{d} \in \mathcal{D} \quad \forall a \in \mathcal{A}: \ p(\hat{d} \mid e \in \mathcal{G}_a, d) = p(\hat{d} \mid d) \tag{16}$$

Unlike demographic parity (Equation 9), equalized odds allows the predicted decision $\hat{d}$ to depend on the ground-truth decision d. As such, for the fair team recommendation with the Boolean decision set, it allows the use of the required subset of skills S that predicts ground-truth decision $d \in \mathcal{D}$=$\{e \in \mathrm{E}, e \notin \mathrm{E}\}$ in $(\mathrm{S}, \mathrm{E})_{y=1}$ for experts who have the skills S versus other experts.

Given an expert e along with her set of skills S_e (Section 3.1), the expert is *qualified* for a team with a required subset of skills S if the expert's skill set has a nonempty intersection with S, that is, $\mathrm{S}_e \cap \mathrm{S} \neq \varnothing$.[†] From Equation 16 and mapping the ground-truth decisions to the experts' qualifications as $d \in \mathcal{D}$=$\{(e \in \mathrm{E} \iff \mathrm{S}_e \cap \mathrm{S} \neq \varnothing), (e \notin \mathrm{E} \iff \mathrm{S}_e \cap \mathrm{S} = \varnothing)\}$, that is, being in the recommended team depends on the overlap between the experts' skills S_e and the required set of skills S, we have:

$$\forall \hat{d} \in \mathcal{D} \quad \forall a \in \mathcal{A}: \ p(\hat{d} \mid e \in \mathcal{G}_a, \mathrm{S}_e \cap \mathrm{S}) = p(\hat{d} \mid \mathrm{S}_e \cap \mathrm{S}) \tag{17}$$

Equivalently,

$$\forall a \in \mathcal{A}: \ [p(e \in \mathrm{E} \mid e \in \mathcal{G}_a, \mathrm{S}_e \cap \mathrm{S}) = p(e \in \mathrm{E} \mid \mathrm{S}_e \cap \mathrm{S})] \wedge [p(e \notin \mathrm{E} \mid e \in \mathcal{G}_a, \mathrm{S}_e \cap \mathrm{S}) = p(e \notin \mathrm{E} \mid \mathrm{S}_e \cap \mathrm{S})] \tag{18}$$

and similar to proposition 1, equalized odds for fair team recommendation holds if and only if,

$$\forall a \in \mathcal{A}: \ p(e \in \mathrm{E} \mid e \in \mathcal{G}_a, \mathrm{S}_e \cap \mathrm{S}) = p(e \in \mathrm{E} \mid \mathrm{S}_e \cap \mathrm{S}) \tag{19}$$

[†] Alternatively, as future work, we can define other qualification measures, e.g., the size of the intersection, to show more (less) qualified experts.

As seen, equalized odds ensure demographic parity given the qualifications of experts for the input required set of skills S for the recommended successful team $(\mathrm{S}, \mathrm{E})_{y=1}$. In other words, the likelihood of being in the recommended team only depends on whether an expert is qualified, regardless of the expert's value of the protected attribute.

Example 2. From Example 1, for a set of $|\mathcal{E}| = 100$ experts, where $\mathcal{A}$ denotes gender, $|\mathcal{G}_{a:female}| = 30$ and $|\mathcal{G}_{a':male}| = 70$ represent female and male groups, respectively. To recommend a fair team of size k=10 for a required set of skills S, we first identify the qualified and unqualified experts. Suppose there are 10 female and 10 male experts whose skill sets intersect with S (i.e., qualified experts), and the remaining 20 female and 60 male experts are *un*qualified. To satisfy equalized odds, an expert's probability of being selected can be assumed to be independent and identically distributed (i.i.d.) and uniform *within each qualification group*. Specifically, 1) the probability that a qualified expert is selected into the team is $\frac{k}{10+10} = \frac{10}{20} = 0.5$, as given by Equation 19, regardless of gender. Thus, every qualified female and male expert has an equal chance of being selected. 2) Equally importantly, the probability that an *un*qualified expert is selected into the team is $\frac{k}{20+60} = \frac{10}{80} = 0.125$, again regardless of gender. Hence, all unqualified female and male experts also have equal selection probability. Now consider an imperfect base recommender model whose top-k=10 recommendations include 6 qualified experts and 4 unqualified experts. To satisfy equalized odds, the qualified subset should be distributed as $0.5 = \frac{x}{6}$, yielding $x = 3$ female and $6 - x = 3$ male experts. The remaining unqualified experts should be distributed by $0.125 = \frac{x}{4}$, resulting in $\lceil 0.5 \rceil = 1$ unqualified female experts and $4 - 1 = 3$ male experts.

~~In fair team recommendation, equalized odds would ensure that among skilled experts, individuals from different protected groups have equal chances of appearing in top positions, while among unqualified members, the likelihood of being incorrectly ranked highly remains consistent across protected groups. In other words it guarantees that the protected groups have equal true positive rates and false positive rates simultaneously. A *qualified* expert e for a team with a required subset of skills S can be simply defined as a Boolean measure based on whether the expert's skill set has a nonempty intersection with S, that is, $\mathrm{S}_e \cap \mathrm{S} \neq \varnothing$. Alternatively, as future work, we can consider a non-Boolean measure, such as the size of the intersection, to identify more (or less) qualified experts.~~

3.3.4 | Equal~~ity of~~ Opportunity

Equal~~ity of~~ opportunity is a relaxed version of equalized odds that only requires fairness for the *desired* ground-truth decision, that is, the opportunity among *qualified* experts[7]. ~~equal true positive rates across protected groups, without constraining the false positive rates.~~ The intuition is to ensure that among ~~individuals~~experts who are qualified for a positive outcome, the probability of receiving a positive prediction should be equal regardless of their protected attributes. This is less restrictive than equalized odds but can be more practical to implement while still preventing discrimination against qualified ~~individuals~~experts in the protected groups.

For a fair team recommendation with a Boolean decision $d \in \mathcal{D}=\{(e \in \mathrm{E} \iff \mathrm{S}_e \cap \mathrm{S} \neq \varnothing), (e \notin \mathrm{E} \iff \mathrm{S}_e \cap \mathrm{S} = \varnothing)\}$, if we prioritize fairness for the qualified experts, i.e., $d = (e \in \mathrm{E} \iff \mathrm{S}_e \cap \mathrm{S} \neq \varnothing)$ to be *'recommended'* in a team as an advantaged outcome, i.e., $\hat{d}$ is $e \in \mathrm{E}$, and ignore the possible discrimination for the *un*qualified experts caused by the same decision, i.e., $\hat{d}$ is $e \in \mathrm{E}$ but $\mathrm{S}_e \cap \mathrm{S} = \varnothing$, we have a less strict variant of equalized odds from Equation 19, as follows:

$$\forall a \in \mathcal{A}: \; p(e \in \mathrm{E} \mid e \in \mathcal{G}_a, \mathrm{S}_e \cap \mathrm{S} \neq \varnothing) = p(e \in \mathrm{E} \mid \mathrm{S}_e \cap \mathrm{S} \neq \varnothing) \tag{20}$$

Proposition 3. *Equal opportunity holds in the fair team recommendation problem if and only if the posterior distributions of experts of protected groups in the successful team are equal to their prior distributions in the entire set of <u>qualified</u> experts $\mathcal{E}$, i.e., $p(e \in \mathcal{G}_a \mid e \in E, S_e \cap S \neq \varnothing) = p(e \in \mathcal{G}_a \mid S_e \cap S \neq \varnothing)$.*

Proof. By Bayes theorem,

$$p(e \in \mathcal{G}_a \mid e \in \mathrm{E}, \mathrm{S}_e \cap \mathrm{S} \neq \varnothing) = \frac{p(e \in \mathrm{E} \mid e \in \mathcal{G}_a, \mathrm{S}_e \cap \mathrm{S} \neq \varnothing)\, p(e \in \mathcal{G}_a, \mathrm{S}_e \cap \mathrm{S} \neq \varnothing)}{p(e \in \mathrm{E}, \mathrm{S}_e \cap \mathrm{S} \neq \varnothing)} \tag{21}$$

$$= p(e \in \mathcal{G}_a, \mathrm{S}_e \cap \mathrm{S} \neq \varnothing);\; \text{from Equation 20} \tag{22}$$

□

Example 3. Referring to our earlier Example 2 with a set of $|\mathcal{E}| = 100$ experts, where $\mathcal{A}$ denotes gender, and $|\mathcal{G}_{a:female}| = 30$ and $|\mathcal{G}_{a':male}| = 70$ represent female and male groups, respectively, we aim to recommend a fair team of size k=10 for a required set of skills S. To satisfy equal opportunity, we consider only qualified experts, i.e., 10 female and 10 male experts whose skill sets intersect with S. Assuming independent and identically distributed (i.i.d.) selection with a uniform distribution among qualified experts, each qualified expert is selected with probability $\frac{k}{10+10} = \frac{10}{20} = 0.5$, as given by Equation 20, regardless of gender. Thus, every qualified female and male expert has an equal chance of being selected, while *unqualified experts are excluded from consideration*. Now, given the imperfect base recommender model whose top-k=10 recommendations include 6 qualified experts and 4 unqualified experts, to satisfy equal opportunity, from Proposition 3, the qualified subset should be distributed as $0.5 = \frac{x}{6}$, yielding $x = 3$ female and $6 - x = 3$ male experts, as in equalized odds. However, unlike equalized odds, the remaining unqualified experts can be distributed arbitrarily without violating equal opportunity.

~~If we prioritize fairness for the decision $d = 1$, that is, recommending an expert to be in a team as an advantaged outcome, and ignore the possible discrimination caused by the *'not recommended'* decision $d = 0$, we have a less strict variant of equalized odds, referred to as equality of opportunity, as follows:~~:

$$\cancel{p(\hat{d} \mid e_0, d = 1) = p(\hat{d} \mid e_1, d = 1)} \tag{23}$$

~~Hence, from Equation and Equation~~

$$\cancel{p(e_0 \in \mathrm{E} \mid e_0, (\mathrm{S}_{e_0} \cap \mathrm{S} \neq \varnothing)) = p(e_1 \in \mathrm{E} \mid e_1, (\mathrm{S}_{e_1} \cap \mathrm{S} \neq \varnothing))} \tag{24}$$

It is worth ~~noting~~reminding that, ~~based on demographic parity, an expert having *no* required skills might be recommended.~~ under demographic parity (Section 3.3.2), recommending an unqualified expert who possesses *none* of the required skills with the same probability as a qualified expert is still considered fair. In contrast, based on the equalized odds and equal opportunity, members must be qualified for the required skills, that is, the intersection of their skills and the required skills must be non-empty (Equation 19 and Equation 20).

3.3.5 | A Fair Team

Once we define the notions of fairness for the team recommendation problem, we can formalize a fair team by the fair *identity* function $\mathbb{I}$ as follows:

$$\mathbb{I}(\mathrm{E}) = \begin{cases} \text{Demographic Parity} & \begin{cases} 1 & \cancel{\textit{iff Equation}} \iff \text{Proposition 2} \\ 0 & \textit{not} \text{ a fair team} \end{cases} \\ \text{Equal}\cancel{\text{ity of}}\text{ Opportunity} & \begin{cases} 1 & \cancel{\textit{iff Equation}} \iff \text{Proposition 3} \\ 0 & \textit{not} \text{ a fair team} \end{cases} \end{cases} \tag{25}$$

where E is a subset of experts in a recommended successful team $(\mathrm{S}, \mathrm{E})_{y=1}$, which is fair with respect to demographic parity *iff* Proposition 2. Alternatively, it is fair with respect to the notion of equal~~ity of~~ opportunity *iff* Proposition 3. While equalized odds is a more comprehensive notion of fairness by enforcing parity not only among qualified experts, as in equal opportunity, but also among unqualified experts for the favorable decision of being in a recommended successful team, we do not adopt it in this work. In the context of team recommendation, enforcing parity over unqualified experts is redundant and misaligned with the recommendation objective, wherein models are explicitly trained to identify qualified experts from the required skill set S, and unqualified experts are penalized by the learning objective, as discussed in Section 3.1 and Definition 2. Consequently, outcomes involving unqualified experts are already controlled downstream by the model during training. This observation has also been studied in other predictive models[48,49,50,51], leading to the adoption of equal opportunity as a more focused and actionable fairness notion.

It is worth noting that merely recommending a fair team while neglecting its success measures is also undesirable ~~, e.g., a team of nonpopular experts who fall short of accomplishing tasks. Hence,~~ and metrics of accuracy (utility) based on the team's ~~true label of success (y)~~ ground-truth set of members should also be measured for a team recommender on top of fairness.

Example 4. To illustrate the evaluation of fairness and accuracy in recommended teams, we consider a small population of 6 experts, where two are female $\mathcal{G}_{a:female} = \{\boxed{e_1}, \boxed{e_2}\}$ and 4 are male $\mathcal{G}_{a':male} = \{e_3, e_4, e_5, e_6\}$. The required skill set for a

TABLE 1 Expert pool of 6 experts showing gender, prior distribution of groups, skills, qualification for the required skill set $S=\{s_1, s_2\}$, and the prior distribution of groups for qualified experts.

$\mathcal{E}$	$\mathcal{A}$=gender	$p(e \in \mathcal{G}_a)$	S_e	$S_e \cap S \neq \varnothing$	$p(e \in \mathcal{G}_a \mid S_e \cap S \neq \varnothing)$
$\boxed{e_1}$	female	$\frac{2}{6} = 0.33$	s_1	✓	$\frac{\lvert\{e_1,e_2\}\rvert}{\lvert\{e_1,e_2,e_3,e_4\}\rvert} = \frac{2}{4} = 0.50$
$\boxed{e_2}$			s_2	✓	
e_3	male	$\frac{4}{6} = 0.67$	s_1	✓	$\frac{\lvert\{e_3,e_4\}\rvert}{\lvert\{e_1,e_2,e_3,e_4\}\rvert} = \frac{2}{4} = 0.50$
e_4			s_2	✓	
e_5			s_3	×	-
e_6			s_3	×	-

TABLE 2 Candidate recommended teams E in top-k=3 for an *unseen* successful team $(S = \{s_1, s_2\}, E)_{y=1}$ showing the accuracy of the recommended team based on ground-truth team $E^* = \{\boxed{e_1}, e_3, e_4\}$ who has been indeed successful, posterior group ratios and whether demographic parity and equal opportunity are satisfied based on prior ratios in Table 1.

#	E	$\lvert E \cap E^* \rvert$	$p(e \in \mathcal{G}_{a:female} \mid e \in E)$ == $p(e \in \mathcal{G}_{a:female}) = 0.33$	demographic parity	$p(e \in \mathcal{G}_{a:female} \mid e \in E, (S_e \cap S \neq \varnothing))$ == $p(e \in \mathcal{G}_{a:female} \mid S_e \cap S \neq \varnothing) = 0.5$	equal opportunity
1	$[\boxed{e_1}, e_3, e_4]$	3	$1/3 \approx 0.33$	✓	$1/3 \approx 0.33$	×
2	$[\boxed{e_1}, \boxed{e_2}, e_3]$	2	$2/3 \approx 0.67$	×	$2/3 \approx 0.67$	×
3	$[\boxed{e_2}, e_3, e_4]$	2	$1/3 \approx 0.33$	✓	$1/3 \approx 0.33$	×
4	$[\boxed{e_1}, e_3, e_5]$	2	$1/3 \approx 0.33$	✓	$1/2 = 0.50$	✓
5	$[\boxed{e_1}, \boxed{e_2}, e_4]$	2	$2/3 \approx 0.67$	×	$2/3 \approx 0.66$	×

successful team is $S = \{s_1, s_2\}$, and only a subset of experts are qualified, i.e., their skill sets intersect with S. The *unseen* ground-truth team $E^* = \{\boxed{e_1}, e_3, e_4\}$ represents the actually successful team from the test set, while the example recommended teams E are generated by neural models for a successful team for the same skill requirements. Table 1 summarizes the expert pool, their gender, skills, qualifications, and prior group probabilities. Table 2 lists several candidate recommended teams at top-k=3, showing the recall (overlap) with the ground-truth team as an accuracy measure, the posterior proportion of each protected group in the team, and whether demographic parity and equal opportunity are satisfied. This example demonstrates the trade-off between accuracy and fairness. As seen, while #1 recommendation achieves perfect accuracy with the ground-truth team, it violates equal opportunity. In contrast, while #4 satisfies demographic parity and equal opportunity, it comes at the cost of reduced accuracy. It also illustrates how posterior distributions can be used post-hoc to assess and, if necessary, adjust team recommendations to improve fairness while considering expert qualifications.

3.4 | Proposed Probabilistic Reranking Method

Let S be the subset of required skills and $f_{\boldsymbol{\theta}}(S)$ is the team recommendation method estimated by a neural model that recommends an optimum subset of experts, E who collectively cover the required subset of skills S and are almost surely successful, i.e., $f_{\boldsymbol{\theta}}(S) = E$ such that $(S, E)_{y=1}$. If E is *not* a fair team, i.e., $\mathbb{I}(f_{\boldsymbol{\theta}}(S)) = \mathbb{I}(E) = 0$, our goal is to estimate a function $g : \mathcal{E} \to \mathcal{E}$~~; $g(E) = E^*$~~ such that:

$$\cancel{\mathbb{I}(E^*) =}\ \mathbb{I}(g(E)) = \mathbb{I}(g(f_{\boldsymbol{\theta}}(S))) = 1 \tag{26}$$

We frame our reranking method based on *independent* Bernoulli trials of win/lose to find a fair team in the top-k ranked list of a model's prediction in Equation 5. Given $\cancel{p}q$ as the desired ~~proportion~~distribution of protected experts in a fair team and a significance level α~~, which is selected based on the underlying domain~~, we test, at each position, top-$\{1, \cdots, k\}$, if the ranked list statistically significantly follows Bernoulli distribution with winning probability $\cancel{p}q$. Let $E_{r,k}$ be the first k experts of E ranked by a ranker r based on experts' probabilities in Equation 5 produced by a team recommendation method that estimates $f_{\boldsymbol{\theta}}$. Further, let a protected attribute $\mathcal{A}$ whose values $\{a, a'\}$ split the expert set $\mathcal{E}$ into $\mathcal{G}_a$ as the disadvantaged group and $\mathcal{G}_{a'}$ as the advantaged group, respectively, e.g., $\mathcal{G}_{a:female}$ and $\mathcal{G}_{a':male}$, $|E_{r,k}|_{a\cancel{1}}$ be the current number of ~~protected~~ experts from the disadvantaged group in the $E_{r,k}$, and $\mathbb{F}(|E|_{a\cancel{1}}; |E|, \cancel{p}q)$ be the cumulative distribution function for a binomial probability of having $|E|_{a\cancel{1}}$ experts of ~~a protected~~the disadvantaged group in the predicted team via $|E|$ independent Bernoulli trials with ~~a~~the winning rate $\cancel{p}q$. We calculate ~~the~~ $\mathbb{F}$'s inverse function $\mathbb{F}^{-1}(\alpha; k, \cancel{p}q)$ to determine the minimum number of required protected experts in E from top-$\{1, \cdots, k\}$. Table 3 illustrates ~~the~~sample values of $\mathbb{F}^{-1}(\alpha; k, \cancel{p}q)$ for top-$\{1, \cdots, k{=}10\}$. ~~We denote~~ $\cancel{\mathcal{G}_{a_1} = \{e_1^{(1)}, \cdots, e_1^{(l)}\}}$ ~~as~~

TABLE 3 Minimum number of required *disadvantaged* experts in the top-$k \in \{1, \ldots, 10\}$ under demographic parity and equal opportunity in Examples 1 and 3. The probability of success $\cancel{p}q$ is based on Propositions 2 and 20, leading to different minimum requirements across fairness notions and confidence levels. For equal opportunity, on a *per-team* basis, the prior distribution is obtained over qualified experts whose skill sets intersect with the team's required skill set S, and $\cancel{p}q$ and $\mathbb{F}^{-1}$ are computed accordingly.

fairness notion	q=prior distribution	α	$\mathbb{F}^{-1}(\alpha,k,q)$ k=1	k=2	k=3	k=4	k=5	k=6	k=7	k=8	k=9	k=10
demographic parity	q=$p(e \in \mathcal{G}_{a:female}) = 0.3$ (across dataset)	0.50	0	0	1	1	1	2	2	2	2	3
		0.10	0	0	0	0	0	0	1	1	1	1
		0.05	0	0	0	0	0	0	0	0	1	1
		0.01	0	0	0	0	0	0	0	0	0	0
equal opportunity	q=$p(e \in \mathcal{G}_{a:female} \mid \mathrm{S}_e \cap \mathrm{S} \neq \varnothing) = 0.5$ (per team)	0.50	0	1	1	2	2	3	3	4	4	4
		0.10	0	0	0	1	1	1	2	2	2	3
		0.05	0	0	0	0	1	1	1	2	2	2
		0.01	0	0	0	0	0	0	1	1	1	1

TABLE 4 ~~The minimum number of required disadvantaged experts in the top-$k \in \{1, \cdots, 10\}$; equivalently, the minimum number of successes with statistical confidence of 0.9 in k Bernoulli trials each with winning probability of $p = 0.6$.~~

α	p	k	$\mathbb{F}^{-1}(k)$
0.1	0.6	1	0
		2	0
		3	1
		4	1
		5	2
		6	2
		7	3
		8	3
		9	4
		10	4

~~the set of $|\mathcal{G}_a|$ experts in the disadvantaged protected group (e.g., female or nonpopular experts).~~ Then, we rank experts of the disadvantaged group using the *current* ranker r as $\mathcal{G}_a = \{e_1^{(a)}, e_2^{(a)}, \cdots, e_{|\mathcal{G}_a|}^{(a)}\}$ and, as for reranking function $g(\mathrm{E})$, we propose a new ranking r' based on the following:

$$
\begin{cases}
\cancel{\hat{g}(e_i^{(k)}) = e_i^{(k)}} & \\
g(e_k) = e_k; e_k \in \mathcal{E}; & \mathbb{F}^{-1}(\alpha; k, \cancel{p}q) \leqslant |\mathrm{E}_{r,k}|_{\cancel{a}} \text{ (no change required)} \\
\cancel{\hat{g}(e_i^{(k)}) = e_1^{(1)}; \hat{g}(e_i^{(k+1)}) = e_1^{(2)}; \ldots; \hat{g}(e_i^{(k+m-1)}) = e_1^{(m)}} & \\
g(e_k) = e_1^{(a)} \mid g(e_{k+1}) = e_2^{(a)} \mid \ldots \mid g(e_{k+m-1}^{(a)}) = e_m^{(a)}; & \mathbb{F}^{-1}(\alpha; k, \cancel{p}q) > |\mathrm{E}_{r,k}|_{\cancel{a}} \text{(insertion from } \textit{ranked } \mathcal{G}_a \text{ till no change required)}
\end{cases}
\tag{27}
$$

where $m = \mathbb{F}^{-1}(\alpha; k, \cancel{p}q) - |\mathrm{E}_{r,k}|_{\cancel{a}}$ is the least number of ~~protected~~ experts from the disadvantaged group $\mathcal{G}_a$ to be added to make a team fair. Our method *inserts* the experts by shifting the experts down the list. For example, should a (male) expert in the fourth position ($\cancel{e_1^{(4)}} e_4$) be replaced with the *most* qualified female expert in the ~~seventh~~ position ($\cancel{e_0^{(7)}} e_7^{(a)}$), the female expert would be moved up to the fourth position ($\cancel{g(e_1^{(4)}) = e_0^{(7)}}\, g(e_4) = e_7^{(a)}$) and shift the list down such that the (male) expert in the fourth position moves to the fifth ($\cancel{e_1^{(5)}} e_5$), and so on. Therefore, as we rerank, all experts are still ranked based on their ~~qualifications~~probabilities in Equation 5 *but within each group*. If the number of disadvantaged experts up until the fourth position, i.e., $|\mathrm{E}_{r,k=4}|_{\cancel{a}}$, is greater than or equal to $\mathbb{F}^{-1}(\alpha; k = 4, \cancel{p}q)$, sufficient disadvantaged experts have been inserted up to this position. Otherwise, m disadvantaged experts should be inserted depending on the given notion of fairness. For equal~~ity of~~ opportunity, we select from *qualified* disadvantaged experts. However, for demographic parity, we overlook the qualification criteria and simply select from disadvantaged experts.

Regarding the role of the significance level α, suppose an *observed* recommendation contains $x = |\mathrm{E}_{r,k}|_a$ experts from the disadvantaged group. We perform a significance test with the null hypothesis $\mathrm{H}_0 : x \sim \mathbb{F}(x; k, q)$ and compute the corresponding p-value. If the p-value $> \alpha$, the observation is considered statistically plausible under H_0, and we therefore *fail to reject* the null hypothesis, treating the team as fair. Conversely, if the p-value $\leqslant \alpha$, the observation is deemed unlikely, leading to rejection of H_0 and the team being flagged as unfair. The choice of α directly controls the strictness of the fairness criterion. Larger values of α increase the likelihood of rejecting H_0, requiring stronger alignment with the distribution. For instance, from our earlier Examples 1 to 3, let the prior distribution of female experts be $p = 0.3$, $\alpha = 0.1$, and no female experts are observed in the top-k=10, i.e., $x = |\mathrm{E}_{r,k=10}|_{a:female} = 0$. Based on demographic parity, the posterior distribution should match the prior, i.e., $q = p = 0.3$. Then, the p-value is $\mathbb{F}(x = 0; k = 10, q = 0.3) = 0.028 < 0.1$, and the null hypothesis is rejected, flagging the team as unfair. In contrast, for $\alpha = 0.01$, the same observation yields $0.028 > 0.01$, and the null hypothesis is *not* rejected, resulting in the team being considered fair. From a mitigation perspective, larger values of α impose stricter fairness requirements, as more disadvantaged experts are needed for a team to pass the significance test. Smaller values of α relax this requirement and may not trigger mitigation even when the observed composition deviates from the expected proportion, as shown in Table 3 for $\alpha \in \{0.5, 0.1, 0.05, 0.01\}$. Meanwhile, the choice of α also controls the fairness trade-off with accuracy of the recommended experts, allowing more conservative (e.g., $\alpha = 0.1$) against permissive (e.g., $\alpha = 0.5$) additions of experts from the disadvantaged group by avoiding unnecessary enforcement of proportion constraints and preserving accuracy while maintaining statistically grounded fairness.

To form a fair ranking of k experts, we assume there exist at least ~~k members from each protected group.~~ $k \times q$ experts from the disadvantaged group. To satisfy demographic parity, we require $k \times (q = p(e \in \mathcal{G}_a))$ experts from the disadvantaged group, ensuring that the prior group distribution matches its posterior distribution among the top-k recommended experts for a successful team (Proposition 2). To satisfy equal opportunity, we require $k \times (q = p(e \in \mathcal{G}_a \mid \mathrm{S}_e \cap \mathrm{S} \neq \varnothing))$ *qualified* experts from the disadvantaged group, so that the prior distribution of qualified experts matches its posterior distribution among the top-k recommendations for a successful team (Proposition 3). From datasets in real-world scenarios, as seen in Table 5, the average number of members in a team (team size) is 3.06, 1.88, 2.51 in `dblp`, `imdb`, and `uspt`, respectively, which is almost surely less than the number of disadvantage experts, i.e., female or nonpopular experts, in the entire datasets. We empirically evaluate our reranking method for top-$k \in \{5, 10\}$, which still remains within practical reach given further considerations discussed in Section 4.3.1, on three large-scale datasets using three fairness metrics, namely `ndkl`[112], `skew`[34] and `expo`[110]. Meanwhile, we evaluate the models' accuracy by information retrieval metrics, including `map` and `ndcg`.

4 | EXPERIMENTS

In this section, we lay out the details of our experiments and findings to answer the following research questions:

RQ1: Does our proposed probabilistic reranking method mitigate unfair biases, including popularity bias and gender bias individually, in the recommended team of experts based on demographic parity and equal~~ity of~~ opportunity while maintaining the team's likelihood of success?

RQ2: Does our proposed probabilistic reranking method outperform deterministic reranking methods in mitigating popularity and gender biases in view of demographic parity and equal~~ity of~~ opportunity?

RQ3: Does our proposed probabilistic reranking method effectively reduce bias while enhancing the exposure to success ratio, as measured by utility-aware exposure (`expo`), for disadvantaged groups, e.g., nonpopular and female experts?

RQ4: Is the effect of our proposed reranking method consistent across datasets from different domains?

4.1 | Datasets

We evaluate our proposed method on three well-known large-scale benchmark datasets in team recommendation literature including `dblp`[38,40,37,91], `imdb`[40,92,70], and `uspt`[94,93]. In `dblp`, each team is a publication in computer science consisting of authors as the experts and the fields of study (fos) as the skills. In `imdb`, each instance is a movie. We consider each movie as a team whose members are the cast and crew, and the movies' genres are the teams' skills. The choice of `imdb` in team formation literature is not to be confused with its use cases in review analysis research; herein, the goal is to form a team of casts and crews for a *movie production* as opposed to a movie recommendation[92,70]. In `uspt`, each instance is a patent issued by the United States Patents and Trademarks consisting of inventors (experts) and subcategories (skills). Table 5 reports statistics on the

TABLE 5 Statistics for the benchmark datasets utilized in our experiments as well as mapping data properties to team instances.

	dblp		imdb		uspt	
success $y = 1$	published		produced		issued	
#teams $\|\mathcal{T}\|$	4,877,383	publications	507,034	movies	7,068,508	patents
#experts $\|\mathcal{E}\|$	5,022,955	authors	876,981	cast and crew	3,508,807	inventors
#skills $\|\mathcal{S}\|$	89,504	fields of study	28	genres & subgenres	241,961	classes & subclasses
avg #experts in teams	3.06		1.88		2.51	
avg #teams per expert	23.02%		62.45%		44.69%	
%popular experts (avg)	31.30%		42.60%		31.40%	
%nonpopular experts (avg)	68.70%		57.40%		68.60%	
%female experts	14.20%		12.30%		13.80%	
%male experts	85.80%		87.70%		86.20%	

datasets. From Figure 1, male experts are dominating teams while female experts have participated sparingly in all datasets. Also, from Figure 2, all datasets suffer from the long tail problem in the distributions of teams over experts, i.e., many experts (researchers in dblp, cast and crew in imdb, and inventors in uspt) have participated in very few teams (papers in dblp, movies in imdb, and inventions in uspt).

4.2 | ~~Dataset Labeling Criteria~~ Forming Protected Groups

~~**Success Labels:** Benchmark datasets in team recommendation literature consist of successful teams only, missing unsuccessful ones. For instance, the dblp lacks unsuccessful submissions. Further, what it means for a team to be successful has remained controversial. For instance, in the movie industry, it is debatable whether a movie's success should be measured based on its immediate reception by the people (box office) or critical reviews (ratings) within a long span of time. In the absence of explicit labels for unsuccessful teams, neural team recommendation methods presume *all* instances of teams in the training dataset as successful (positive samples) and proceed with the training procedure.~~

Gender Labels: ~~Contrary to popularity, gender is self-identified.~~ While uspt dataset includes gender labels, other training datasets lack them in part or whole. In imdb, although we inferred the gender of some cast and crew by their role identified as actor or actress, gender labels for other experts were missing. In dblp, no gender label for the experts has been provided. Therefore, we utilized genderize[144], a crowd-based system that predicts a gender based on the given name of the experts, for dblp as well as those that are missing in imdb. Meanwhile, we recognize that predicting gender using name-to-gender services such as genderize might systematically mislabel certain demographic groups due to cultural, linguistic, and regional variations. For example, names such as *'Andrea'* (typically male in Italy but female in English-speaking countries), *'Sasha'* (male in Russia but commonly female in North America), and *'Kim'* (gender-neutral in East Asia but often inferred as female in Western countries) are frequently misclassified. Hence, our results should be interpreted with this limitation in mind. As seen earlier in Figure 1, datasets are heavily biased toward male experts. Specifically, dblp has a male-to-female ratio of 0.858 to 0.142, imdb has a slightly different ratio of 0.877 to 0.123 and uspt has a ratio of 0.862 to 0.138.

Popularity Labels: ~~Defining popularity could be controversial. To avoid varied interpretations, we followed social science and recommender system literatures, where the popularity status of an expert can be *objectively* measured based on the number of teams the expert has participated in, referred to as *sociometric* popularity. Further, although an expert's participation in many teams, e.g., research papers in dblp or movies in imdb, may *not* necessarily indicate popularity from the people's perspectives, repetition of the expert in many training samples of teams from the neural model's perspective does.~~

~~Accordingly, we can adopt two alternatives, as shown in Figure ■ *i)* an expert is popular if the expert participated in more than the average number of teams per expert over the entire dataset, and nonpopular otherwise (avg), or *ii)* an expert is popular if she belongs to the *short head* in the 2-d curve of the distribution of experts in teams, and nonpopular otherwise. We split the curve into *short head* and *long tail* based on equal area under the curve (auc).~~ From Section 3.3.1, among the two options for splitting experts into popular and nonpopular groups based on the long-tail distribution, namely avg and auc, we opt for avg criterion, labeling experts above the average number of teams an expert participates as popular. In contrast to the auc criterion, which identifies popular experts as those located in the first half of the area under the curve (head), we favor the avg since it provides a simpler, more transparent, and easily reproducible threshold, while avoiding sensitivity to the exact shape of the distribution required by the auc. For the dblp dataset, the average stands at 23.02 teams, 62.45 teams for the imdb dataset,

TABLE 6 Expected number of female experts at each top-k by the probabilistic reranking method vs. minimum and maximum female expert allocations by the deterministic baselines under demographic parity, assuming a prior female expert proportion of 0.3.

	expected number of females	k=1	k=2	k=3	k=4	k=5	k=6	k=7	k=8	k=9	k=10
probabilistic	$\mathbb{F}^{-1}(\alpha = 0.1, k, q = 0.3)$	0	0	0	0	0	0	1	1	1	1
deterministic	min=$\lfloor 0.3 \times k \rfloor$	0	0	0	1	1	1	2	2	2	3
	max =$\lceil 0.3 \times k \rceil$	1	1	1	2	2	2	3	3	3	3

and 44.69 teams for the `uspt` dataset. Therefore, in `dblp`, the proportion of popular to nonpopular experts becomes 0.313 to 0.687, in `imdb`, it is 0.426 to 0.574 and for `uspt` it stands at 0.314 to 0.686.

4.3 | Baselines

4.3.1 | Neural Team Recommendation

We compare the impact of our proposed probabilistic reranking method on mitigating neural models' biases using the state-of-the-art variational Bayesian neural network (`bnn`)[38,40,37] with a single hidden layer of size $d = 128$, `leaky relu` as the activation function for the hidden layer, and Kullback-Leibler (KL) divergence as the optimizer. For the input layer, as detailed in Section 3.1, we used sparse occurrence vector representations (one-hot encoded) of skills of size $|\mathcal{S}|$ as well as pretrained dense vector representations (`-emb`)[39]. The output layer is the sparse occurrence vector representations (one-hot encoded) of experts of size ~~$|\mathcal{S}|$ and~~ $|\mathcal{E}|$, respectively. We randomly select 15% of teams for the test set and perform 5-fold cross-validation on the remaining teams for model training over 20 epochs ~~that results~~, resulting in one trained model per ~~each~~ fold. Given a team (S, E) from the test set, we select the top-k∈$\{5, 10\}$~~=100~~ experts with the highest probabilities as the recommended team $\mathrm{E} = f_{\boldsymbol{\theta}}(\mathrm{S})$ by the model of each fold. Although the average team size in the datasets is small (between 1.88 and 3.06), evaluating fairness-aware reranking at such small cardinalities can lead to over- or underestimation of the fairness metrics. For example, if the desired ratio between female and male experts is 0.3 to 0.7, then for a team of size k=2, this corresponds to 0.6 female and 1.4 male members and feasible splits are either (0 female, 2 male) or (1 female, 1 male), yielding ratios of 0 or 0.5 and deviates substantially from the target 0.3. Hence, it can result in large inaccuracies when reporting fairness metrics. Similarly, for a team of size 3, the expected split is 0.9 and 2.1, with feasible splits (0 female, 3 male) or (1 female, 2 male), giving a ratio of approximately 0.33. As shown, using a larger k allows for a more accurate calculation of posterior distributions and reduces rounding errors and large deviations. On the other hand, as noted, real-world teams are often of size 2 to 3 members. To balance the need for meaningful posterior calculation with real-world team sizes, we therefore adopt top-k∈$\{5, 10\}$.

4.3.2 | Fairness-aware Reranking

Our fairness-aware reranking baselines include three deterministic greedy reranking algorithms `det-greedy`, `det-cons`, and `det-relaxed` by Geyik et al.[34], detailed below, as well as our proposed probabilistic reranking method with the significance level $\alpha \in \{0.1, 0.05\}$. Our probabilistic method builds upon the framework of Zehlike et al.[31] and adapts it for the team recommendation task, introducing necessary modifications to handle team compositions by altering the reranking procedure. Consequently, the original method cannot serve as a separate baseline, as it is effectively a trivial special case of our team-level adaptation.

Although the deterministic baselines were originally designed for multi-valued protected attributes (and thus multiple groups), we apply them to Boolean protected attributes in two settings: 1) gender, where female experts constitute the disadvantaged yet minority group and male experts are the advantaged and majority group; and 2) popularity, where nonpopular experts form the disadvantaged yet majority group and popular experts are the advantaged but minority group.

- **`det-greedy`:** for every top-$\{1, ..., k\}$ prefix of the ranking, this algorithm aims to maintain a proportion of experts from each group as close as possible to the predefined desired distributions, herein, the prior distributions as explained in Propositions 1 and 3. The algorithm generates the new ranking at each top-k rank by first determining the target group for the next selection and then adding the next expert, while keeping the previous steps unchanged. The algorithm jointly considers both groups and deterministically computes the minimum and maximum allowable numbers of experts for each group in the top-k ranking by taking the floor and the ceiling, respectively, of k times the desired distribution. Table 6 shows an example under demographic parity for a prior female expert distribution of 0.3. To identify the target group at each rank, the algorithm checks whether the current number of experts from a group in the ranked list falls below its minimum allowable number. Such a group is flagged as the target group, and the next selected expert is the one with the highest output probability assigned by the team recommendation model (Equation 5). If both groups (e.g., female and male) fall below the minimum requirement, the algorithm selects the expert with the highest model-assigned probability, regardless of the attribute value (e.g., gender) without reranking. When the minimum number from both groups are satisfied, the target group is the one whose maximum allowable number has not yet been reached, and if both groups satisfy this condition, the algorithm selects the expert with the highest probability regardless of the attribute value without rerankings. For instance, from Table 6 at rank k=7, the list contains 4 male and 2 female experts, the minimum requirements for both groups are satisfied, and the algorithm selects the next expert solely based on the highest model-assigned probability, irrespective of gender. In contrast, if the list contains 5 male and 1 female experts, the minimum requirement for female experts is not met, and the algorithm restricts selection to the female group, choosing the highest-scoring female expert.

- **`det-cons` (conservative):** A variation of `det-greedy` with modified selection behavior under maximum-threshold condition. When the maximum thresholds for both groups have not yet been reached, instead of selecting the next expert solely based on the highest model-assigned probability, `det-cons` assesses which group is closer to falling below its minimum requirement in subsequent prefixes. To this end, `det-cons` computes, for each group, the smallest future rank $k' > k$ at which the minimum requirement increases, i.e., the smallest k' such that $\lfloor k' \times p \rfloor = \lfloor k \times p \rfloor + 1$, which can be approximated by $k' \sim \frac{\lceil k \times p \rceil}{p}$. As an example, suppose that at rank k=7 the ranking contains 4 male and 2 female experts. The minimum requirement for male experts increases at a smaller rank ($\frac{\lceil 7 \times 0.7 \rceil}{0.7} = \frac{50}{7} = 7.14$) than that for female experts ($\frac{\lceil 7 \times 0.3 \rceil}{0.3} = \frac{30}{3} = 10$), indicating that the male group is closer to falling below its required minimum. Therefore, `det-cons` selects the next expert as the highest model-assigned probability expert among male experts to prevent a future violation of the minimum representation constraint.
- **`det-relaxed`:** A relaxed variant of `det-cons` that introduces additional flexibility when the minimum requirements are satisfied for both groups. In this case, `det-relaxed` computes the smallest future rank at which *any* group's minimum requirement would be increased, regardless of whether that group's maximum threshold has already been reached. In other words, after the minimum thresholds are met for all groups, the method prioritizes the group whose minimum threshold would be violated earliest, while ignoring both groups' maximum thresholds.

The deterministic reranking baselines have notable limitations. Having only two groups in our case, while considering the maximum threshold in the decision process is to prevent disproportionate advantage to one group, it may come at the cost of accuracy, that is, when one group has met the maximum threshold (e.g., males), these algorithms would select an expert from the other group (e.g., females) even with the lower probability, resulting in fairer yet less accurate team recommendation. In contrast, our method enforces only the minimum required representation of the disadvantaged group necessary to satisfy the fairness constraints, while selection is always guided by accuracy. Moreover, deterministic baselines strictly enforce the prior proportions of both groups at every prefix k. In highly skewed datasets, when the proportion of the disadvantaged group is very small, deterministic reranking may replace a predicted disadvantaged expert with an advantaged expert to maintain the prior proportion at a given prefix. In contrast, our algorithm enforces only the minimum required presence of disadvantaged experts at each k, and preserves all high-scoring disadvantaged experts once the requirement is met.

4.4 | Evaluation Strategy

To measure the effectiveness of our proposed method, it is essential to evaluate fairness in the teams to ensure equitable treatment of all ~~individuals~~experts and prevent systematic discrimination against protected groups. Measuring fairness metrics before and after team recommendation helps identify potential biases in the recommendation algorithm, quantifies the effectiveness of fairness interventions, and provides transparency about how well the system promotes fairness. These evaluations also help balance different notions of fairness and demonstrate compliance with ethical guidelines or organizational diversity goals while maintaining accountability in the team recommendation process. Team utility must be considered alongside fairness because teams ultimately need to accomplish tasks effectively. Measuring utility before and after fairness interventions is mandatory to understand potential performance trade-offs, and identify possible performance improvements through diversity, and demonstrate to stakeholders that fairness can be achieved while maintaining team effectiveness. Hence, we measure fairness and utility both before and after applying our methodologies to our baselines.

4.4.1 | Before Mitigating Bias

To answer our research questions, we evaluate the efficacy of the model in recommending *correct* experts for each team of our test set by comparing the ranked list of experts, predicted by the model of each fold, with the observed subset of experts E^* and report the average performance of models on all folds in terms of information retrieval metrics including mean average precision (`map`) and normalized discounted cumulative gain (`ndcg`) at top-$k \in \{5, 10\}$~~=10~~, as explained in Section 4.4.3. Furthermore, to assess the fairness of the predicted teams, we report the average fairness metrics, including normalized discounted Kullback-Leibler (`ndkl`)[112], `skew`[34], and `expo`[110] for popularity and gender attributes with respect to demographic parity and equal~~ity of~~ opportunity, as will be formalized in Section 4.4.3. While `skew` is a symmetric metric to measure the symmetrical distribution of data among protected attributes, `ndkl` is an *a*symmetric measure of differences between the actual and desired distributions of protected attributes. For both of these metrics, the closer the value to 0, the more unbiased the distribution. In contrast to `ndkl` and `skew` ~~that~~, which are not ~~utility~~accuracy-aware, or in other words, they ~~will not~~ consider ~~utility~~neither ground-truth members of the test team nor the model-assigned probabilities in Equation 5 in their evaluations, `expo` calculates the exposure of different protected groups with respect to their ~~utility~~model-assigned probabilities, referred to as utility.

4.4.2 | After Mitigating Bias

Given the predicted ranked list of experts $\cancel{\hat{\mathrm{E}}}\mathrm{E}$ by the model of each fold for the observed ground-truth subset of experts $\cancel{\mathrm{E}}\mathrm{E}^*$ for a team of the test set, we apply fairness-aware reranking methods on $\cancel{\hat{\mathrm{E}}}\mathrm{E}$ to produce an unbiased ranked list of experts $\cancel{\hat{g}(\hat{\mathrm{E}})}g(\mathrm{E})$. We then evaluate the new reranked list in terms of ~~information retrieval~~accuracy metrics and fairness metrics. We presume that fairness-aware reranking baselines improve the fairness metrics without discounting the ~~information retrieval~~accuracy metrics. In total, we compare {`bnn`, `bnn-emb`} baselines for {popularity, gender} protected attributes with respect to {demographic parity, equal~~ity of~~ opportunity} notions of fairness *before* and *after* applying fairness-aware reranking methods {`det-greedy`, `det-cons`, `det-relaxed`, **our method**} in terms of {`map`, `ndcg`} ~~information retrieval~~accuracy metrics and {`ndkl`, `skew`, `expo`} fairness metrics on {`dblp`, `imdb`, `uspt`} datasets.

4.4.3 | Fairness and ~~Utility~~Accuracy Metrics

Let $(\mathrm{S}, \mathrm{E}^*)_{y=1}$ a team of experts E^* for the required set of skills S from the test set, we compare the top-k ranked list of experts, predicted by the model of each fold for the input skills S, i.e., $f_{\theta}(\mathrm{S}) = \mathrm{E}$, with the observed subset of experts E^* and report the average performance of models on all folds in terms of utility metrics (the higher, the better) as formalized below.

1) **Mean Average Precision (`map`)**

$$\mathtt{ap}(k) : \frac{\sum_{i=1}^{k} \mathtt{pr}(i) \times \delta_{\mathrm{E}*}(i)}{|\mathrm{E}^* \cap \cancel{f_{\theta}(S)}\mathrm{E}|} \tag{28}$$

where $\mathtt{pr}(k) = \frac{|\mathrm{E}^* \cap \cancel{f_{\theta}(S)}\mathrm{E}|}{k}$ is the precision, i.e., how many of the k predicted experts are correctly identified from the test instance of the team ~~E~~ and $\delta_{\cancel{f}\mathrm{E}*}(i)$ returns 1 if the i-th predicted expert is in E^*. Finally, we report the mean of average precisions (`map`) on all test instances of teams.

2) **Normalized Discounted Cumulative Gain (`ndcg`)**

$$\mathtt{dcg}(k) = \sum_{i=1}^{k} \frac{\mathtt{rel}(i)}{\mathtt{log}(i+1)} \tag{29}$$

where $\mathtt{rel}(i)$ captures the degree of relevance for the predicted expert at position i. In our problem setting, however, all members of a test team are considered of the same importance. Therefore, $\mathtt{rel}(i) = 1$ if $i \in \mathrm{E}^*$ and 0 otherwise, and

Equation (29) becomes:

$$\mathtt{dcg}(k) = \sum_{i=1}^{k} \frac{\delta_{\mathrm{E}*}(i)}{\mathtt{log}(i+1)} \tag{30}$$

This metric can be *normalized* relative to the ideal case when the top-k predicted experts include members of the test team E at the lowest possible ranks, i.e.,

$$\mathtt{ndcg}(k) = \frac{\sum_{i=1}^{k} \frac{\delta_{\mathrm{E}*}(i)}{\mathtt{log}(i+1)}}{\sum_{i=1}^{|\mathrm{E}*|} \frac{1}{\mathtt{log}(i+1)}} \tag{31}$$

~~Utility~~Accuracy metrics are, however, oblivious to the protected attributes of experts, and hence, overlook whether the set of top-k predicted experts is a fair team (Section 3.3). To evaluate fairness, we use well-known fairness metrics as follows:

1) **Normalized Discounted KL Divergence (`ndkl`)**[112], which builds upon the foundation of Kullback-Leibler (KL) divergence to measure the expectation of the logarithmic difference between two discrete probability distributions, (the lower, the better) with being 0 in the ideal equal distributions. However, it advances a step further by incorporating a discounting factor, which allows it to assign varying levels of importance to different elements within the distributions being compared. This discounting is particularly valuable in scenarios where the order or priority of elements matters, as a result of which `ndkl` has been extensively employed in recommendation systems and information retrieval[145,146,34,54,147,47,148]. Additionally, `ndkl` includes a normalization component, which scales the results to a more interpretable range and facilitates comparisons across different baselines. Formally, let $p_a = \frac{|\mathrm{E}_{r,i}|_a \cancel{\eta}}{i}$ be the distribution of a protected group in the top-i predicted experts by a ranker r, e.g., the proportions of nonpopular or female experts, and q the ideal fair distribution for a test instance of a team (S, E), the KL divergence of q from p is:

$$\mathtt{kl}(p \,||\, q) = \sum_{i=1}^{k} \sum_{a \in \mathcal{A}} p_a(i)\ \mathtt{log} \frac{p_a(i)}{q_a(i)} \tag{32}$$

where q_a can be manually set~~, like 50%,~~ or calculated based on the ~~overall distribution (proportion) of the protected group in the entire dataset, i.e.,~~fairness notions given by Propositions 2 and 3, for example $q_a = \frac{|\mathcal{E}|_a \cancel{\eta}}{|\mathcal{E}|}$ under demographic parity. This metric has a minimum value of 0 when both distributions are identical up to position i. A higher value indicates a greater divergence between the two distributions, and the metric is always non-negative. We report the *normalized discounted cumulative* KL-divergence (`ndkl`)[34]:

$$\mathtt{ndkl}(p) = \frac{\sum_{i=1}^{|\mathrm{E}|} \frac{1}{\mathtt{log}(i+1)}\ \mathtt{kl}(p \,||\, q)}{\sum_{i=1}^{|\mathrm{E}|} \frac{1}{\mathtt{log}(i+1)}} \tag{33}$$

Example 5. Considering Example 4 for the top-k=3 recommended experts E : $[\boxed{e_1}, e_3, e_4]$, as shown at row #1 in Table 2, including 1 female expert at position 1. Under demographic parity, let the reference distributions be $q_{a:female} = p(e \in \mathcal{G}_{a:female}) = \frac{30}{100} = 0.3$ and $q_{a':male} = p(e \in \mathcal{G}_{a':male}) = 1-0.3 = 0.7$. At top-1, the ranking is entirely female, so the posterior distribution is $p_{a:female}(1) = 1$ and $p_{a':male}(1) = 0$, deviates strongly from the reference. This produces a large divergence, $\mathtt{kl}(p(1) \,||\, q(1)) = p_{a:female}(1) \times \mathtt{log}\frac{p_{a:female}(1)}{q_{a:female}(1)} + p_{a':male}(1) \times \mathtt{log}\frac{p_{a':male}(1)}{q_{a':male}(1)} = 1 \times \mathtt{log}(\frac{1}{0.3}) + 0 \times \mathtt{log}(\frac{0}{0.7}) = 1.204$, reflecting deviation from the reference distributions. At top-2, the prefix contains 1 female and 1 male experts, giving proportions $p_{a:female}(2) = p_{a':male}(2) = \frac{1}{2} = 0.5$. These values are closer to the reference distributions, and the divergence drops: $\mathtt{kl}(p(2) \,||\, q(2)) = 0.5 \times \mathtt{log}(\frac{0.5}{0.3}) + 0.5 \times \mathtt{log}(\frac{0.5}{0.7}) = 0.087$. At top-3, the proportions become $p_{a:female}(3) = \frac{1}{3}$ and $p_{a':male}(3) = \frac{2}{3}$ which almost match the reference and the divergence is therefore nearly zero: $\mathtt{kl}(p(3) \,||\, q(3)) = \frac{1}{3} \times \mathtt{log}(\frac{\frac{1}{3}}{0.3}) + \frac{2}{3} \times \mathtt{log}(\frac{\frac{2}{3}}{0.7}) = 0.002$.

2) **`Skew`**[34] is the logarithmic ratio of ~~1)~~ the proportion of items, herein the experts, from a protected group among the top-k predicted experts to ~~2)~~ the ideal fair proportion for that group. Similar to `ndkl`, given $p = \frac{|\mathrm{E}_{r,i}|_a \cancel{\eta}}{i}$ as the distribution of a protected group in the top-i predicted experts by a ranker r, and $q = \frac{|\mathcal{E}|_a \cancel{\eta}}{|\mathcal{E}|}$ the ideal fair distribution, without loss of generality to any desired distribution ~~for a test instance of a team (S, E)~~:

$$\texttt{skew@}\cancel{i}k(r) = \texttt{log}_{\cancel{e}}(\frac{p}{q}) \tag{34}$$

A negative $\texttt{skew@}\cancel{i}k$ corresponds to a lesser than desired representation of experts with the protected group in the top-k results, while a positive $\texttt{skew@}\cancel{i}k$ corresponds to favoring such experts. The $\texttt{log}$ makes this metric symmetric around zero with respect to ratios for and against a specific protected group and is particularly useful for assessing whether a ranker tends to favor certain protected groups disproportionately [149]. It is important to note that $\texttt{skew}$ is less strict compared to $\texttt{ndkl}$, as it overlooks positional importance within the ranking.

Example 6. Same as Example 5, using the same top-k=3 ranking, the posterior distributions of female experts at different cut-offs are $p_{a:female}(1) = \frac{1}{1} = 1$, $p_{a:female}(2) = \frac{1}{2} = 0.5$ and $p_{a:female}(3) = \frac{1}{3} = 0.33$, respectively. Then, $\texttt{skew@1} = \texttt{log}\frac{p_{a:female}(1)}{q_{a:female}(1)} = \texttt{log}(\frac{1}{0.3}) = 1.204$, $\texttt{skew@2} = \texttt{log}\frac{p_{a:female}(2)}{q_{a:female}(2)} = \texttt{log}(\frac{0.5}{0.3}) = 0.511$ and $\texttt{skew@5} = \texttt{log}(\frac{p_{a:female}(3)}{q_{a:female}(3)}) = \texttt{log}(\frac{0.66}{0.3}) = 0.105$.

3) Utility-aware Exposure ($\texttt{expo}$) [110] measures the ratio of exposure to success across different protected groups in the top-k predicted experts [25,150,151,31,152,153,154,155,156,157] without assuming or estimating any prior or posterior distributions as in $\texttt{ndkl}$ and $\texttt{skew}$, and is therefore agnostic to specific notion of fairness. In recommender systems, exposure for each protected group is defined as the expected probability that an expert of a protected group will be presented at top-k position. This metric quantifies the likelihood of visibility for each member of a protected group and provides a measure of how equitably exposure is distributed across different protected groups. Given the top-k ranked list of predicted experts for a team E, the exposure for an expert is calculated as:

$$\texttt{expo}(e) = \frac{1}{\texttt{log}(i+1)}; e \in \mathrm{E} \tag{35}$$

The simplest fair exposure among groups is that the average exposures of the experts in protected groups are equal. Given a protected attribute $\mathcal{A}$ $\cancel{a = \{a_1, ..., a_n\}}$, the average exposure for the experts of a protected group $\mathcal{G}_{a_{\cancel{i}}}$ based on a protected attribute value $a_{\cancel{i}}$, ~~e.g., female experts $\mathcal{G}_1$ vs. male experts $\mathcal{G}_0$,~~ at the top-k:

$$\mu_{\texttt{expo}}(a_{\cancel{i}}) = \frac{1}{|\mathrm{E} \cap \mathcal{G}_{a_{\cancel{i}}}|} \sum_{e \in \mathrm{E} \cap \mathcal{G}_{a_{\cancel{i}}}} \texttt{expo}(e) \tag{36}$$

The expected utility of an expert in position-based ranking models can be determined by the probability of the expert appearing in a given position. To integrate utility in the exposure metric, an average probability score over experts of a protected group is also calculated as:

$$\mu_{utility}(a_{\cancel{i}}) = \frac{1}{|\mathrm{E} \cap \mathcal{G}_{a_{\cancel{i}}}|} \sum_{e \in \mathrm{E} \cap \mathcal{G}_{a_{\cancel{i}}}} v_{\mathrm{E}}(e) \tag{37}$$

where v_{E} is a vector of probabilities in Equation 5 based on which a ranker select the top-k predicted experts as the recommended team. This is a form of group fairness that considers the utility associated with experts. The exposure value for a protected group is the ratio of average exposure values and average probability scores for a protected group, as follows:

$$\texttt{expo}(a_{\cancel{i}}) = \frac{\mu_{\texttt{expo}}(a_{\cancel{i}})}{\mu_{utility}(a_{\cancel{i}})} \tag{38}$$

Finally, the overall $\texttt{expo}$ value is calculated for a protected attribute based on the ratio of exposure value for a pair of protected groups as:

$$\texttt{expo}\cancel{(a)}(\mathcal{A}) = \frac{\texttt{expo}(a_{\cancel{1}})}{\texttt{expo}(a'_{\cancel{0}})} \tag{39}$$

Example 7. As in Examples 5 and 6, considering the top-k=3 recommended experts $\mathrm{E} : [\boxed{e_1}, e_3, e_4]$ at row #1 in Table 2, including 1 female expert at position 1, let the model-assigned probabilities be $[\boxed{e_1} : 0.9,\ e_3 : 0.85,\ e_4 : 0.8]$. By Equation 36, the average exposure of the female experts is $\mu_{\texttt{expo}}(a : female) = \frac{1}{1} \times \frac{1}{\texttt{log}(1+1)} = 1.443$. Then by Equation 38, the exposure value for the team is then obtained by normalizing average exposure by average utility, i.e., $\mu_{utility}(a : female) = \frac{1}{1} \times 0.9 = 0.9$ and $\texttt{expo}(a : female) = \frac{\mu_{\texttt{expo}}(a)}{\mu_{utility}(a)} = \frac{1.443}{0.9} = 1.603$. The average exposure of male experts of the team is computed similarly for the male experts, who occupy the remaining two positions $\{2,3\}$ as $\mu_{\texttt{expo}}(a:\textit{male}) = \frac{1}{2}(\frac{1}{\log(2+1)} + \frac{1}{\log(3+1)}) = 0.816$. Considering the model output probabilities for these two experts, $\mu_{utility}(a:\textit{male}) = \frac{1}{2}(0.85+0.8) = 0.825$, and the exposure value for them is $\texttt{expo}(a:\textit{male}) = \frac{\mu_{\texttt{expo}}(a)}{\mu_{utility}(a)} = \frac{0.816}{0.825} = 0.989$. That is, 1 female expert at top-1 with high probability (utility) is exposed more compared to 2 male experts with lower probability at lower ranks. By Equation 39, the overall exposure value for the gender protected attribute is $\texttt{expo}(\textit{gender}) = \frac{\texttt{expo}(a:\textit{female})}{\texttt{expo}(a:\textit{male})} = \frac{1.603}{0.989} = 1.62$. Here, $\texttt{expo}(\mathcal{A}) > 1$ shows that the female (disadvantaged) experts receive more exposure relative to their utility compared to male (advantaged) experts.

As a final note and before moving to the experimental result, while `expo` is distribution-agnostic and does not encode a specific fairness notion, its values used for evaluation before vs. after reranking are different since the reranking algorithm may account for group distributions, as in our proposed methods based on demographic parity or equal opportunity.

TABLE 7 ~~Average performance of 5-fold neural models on test set on `dblp` dataset. For the metrics, `ndkl`, the lower the better (↓), `skew`, the closer to `0` the better (→`0`), and `map` and `ndcg`, the higher the better (↓).~~

		`%ndkl` before↓	`%ndkl` after↓	`skew` before→`0` nonprotected	`skew` before→`0` protected	`skew` after→`0` nonprotected	`skew` after→`0` protected	`%map10` Δ ↑	`%ndcg10` Δ ↑
popularity, demographic parity									
bnn	det-cons	109.56	14.64	1.13	-19.97	0.64	-0.54	-0.28	-0.58
	det-greedy		14.64			0.64	-0.54	-0.28	-0.58
	det-relaxed		18.31			0.64	-0.53	-0.28	-0.58
	our method		19.71			0.26	-0.15	0.00	00.00
bnn-emb	det-cons	110.31	14.09	1.14	-20.75	0.62	-0.51	-0.28	-0.58
	det-greedy		14.09			0.62	-0.51	-0.28	-0.58
	det-relaxed		17.65			0.61	-0.50	-0.28	-0.58
	our method		19.61			0.26	-0.15	0.00	0.00
popularity, equality of opportunity									
bnn	det-cons	102.01	13.12	1.05	-19.92	0.57	-0.51	-0.28	-0.58
	det-greedy		13.16			0.57	-0.51	-0.28	-0.58
	det-relaxed		16.15			0.57	-0.50	-0.28	-0.58
	our method		18.96			0.24	-0.16	0.00	0.00
bnn-emb	det-cons	102.85	12.65	1.06	-20.62	0.55	-0.48	-0.28	-0.58
	det-greedy		12.67			0.55	-0.48	-0.28	-0.58
	det-relaxed		15.63			0.55	-0.47	-0.28	-0.58
	our method		18.39			0.25	-0.16	0.00	0.00
gender, demographic parity									
bnn	det-cons	11.80	4.92	-0.08	0.42	-0.07	0.39	-0.28	-0.58
	det-greedy		3.72			-0.07	0.39	-0.28	-0.58
	det-relaxed		6.52			0.00	-0.20	-0.28	-0.58
	our method		8.39			0.00	-0.11	0.00	0.00
bnn-emb	det-cons	7.29	4.97	-0.06	0.30	-0.07	0.37	-0.28	-0.58
	det-greedy		3.59			-0.07	0.37	-0.28	-0.58
	det-relaxed		4.92			-0.04	0.08	-0.28	-0.58
	our method		6.83			-0.03	0.13	0.00	0.00
gender, equality of opportunity									
bnn	det-cons	18.97	9.19	-0.14	0.88	-0.13	0.86	-0.28	-0.58
	det-greedy		7.70			-0.13	0.86	-0.28	-0.58
	det-relaxed		9.53			-0.13	0.85	-0.28	-0.58
	our method		18.97			-0.14	0.88	0.00	0.00
bnn-emb	det-cons	15.93	9.24	-0.11	0.76	-0.13	0.86	-0.28	-0.58
	det-greedy		7.61			-0.13	0.86	-0.28	-0.58
	det-relaxed		10.16			-0.13	0.85	-0.28	-0.58
	our method		15.93			-0.11	0.76	0.00	0.00

TABLE 8 Average performance of 5-fold neural models on test set on `dblp` dataset. For the metrics, `ndkl`, the lower the better (↓), `expo`, the closer to 1 the better (→ 1), `skew`, the closer to 0 the better (→ 0), and `map` and `ndcg`, the higher the better (↑).

`dblp(k=10)`		`%ndkl`↓		`expo` → 1		`skew` before → 0		`skew` after → 0		`%map`	`%ndcg`
		before	after	before	after	disadvantaged	advantaged	disadvantaged	advantaged	$\Delta\uparrow$	$\Delta\uparrow$
popularity, demographic parity											
bnn	det-cons	112.38	18.94	0.08	1.89	-24.70	1.12	-0.66	0.72	-0.28	-0.57
	det-greedy		17.92		2.48			-0.65	0.71	-0.28	-0.57
	det-relaxed		17.97		2.33			-0.62	0.70	-0.28	-0.57
	our method α=0.10		19.71		0.77			-0.15	0.26	0.00	0.00
	our method α=0.05		22.25		0.83			-0.16	0.28	0.00	0.00
bnn-emb	det-cons	112.03	14.09	0.09	1.64	-24.98	1.13	-0.51	0.62	-0.28	-0.58
	det-greedy		14.09		1.64			-0.51	0.62	-0.28	-0.58
	det-relaxed		17.65		3.02			-0.50	0.61	-0.28	-0.58
	our method α=0.10		19.61		0.79			-0.15	0.26	0.00	0.00
	our method α=0.05		22.16		0.85			-0.17	0.29	0.00	0.00
popularity, equal opportunity											
bnn	det-cons	104.80	13.12	0.08	2.10	-24.65	1.04	-0.51	0.57	-0.28	-0.57
	det-greedy		13.16		2.11			-0.51	0.57	-0.28	-0.57
	det-relaxed		16.15		1.85			-0.50	0.57	-0.28	-0.57
	our method α=0.10		18.96		0.78			-0.16	0.24	0.00	0.00
	our method α=0.05		22.61		0.78			-0.18	0.27	0.00	0.00
bnn-emb	det-cons	104.51	12.65	0.09	1.67	-24.94	1.05	-0.48	0.55	-0.28	-0.58
	det-greedy		12.70		1.68			-0.48	0.55	-0.28	-0.58
	det-relaxed		15.59		2.98			-0.47	0.55	-0.28	-0.58
	our method α=0.10		18.87		0.80			-0.16	0.25	0.00	0.00
	our method α=0.05		22.52		0.81			-0.18	0.28	0.00	0.00
gender, demographic parity											
bnn	det-cons	19.57	5.90	0.94	125.73	-5.26	0.03	-0.63	0.07	-0.28	-0.57
	det-greedy		5.21		106.47			-0.62	0.07	-0.28	-0.57
	det-relaxed		5.21		106.83			-0.63	0.07	-0.28	-0.57
	our method α=0.10		5.12		0.94			-0.03	0.00	0.00	0.00
	our method α=0.05		5.27		0.94			-0.08	0.00	0.00	0.00
bnn-emb	det-cons	16.30	5.80	0.92	285.42	-3.41	0.02	-0.64	0.07	-0.28	-0.58
	det-greedy		5.18		274.38			-0.63	0.07	-0.28	-0.57
	det-relaxed		5.20		274.02			-0.63	0.07	-0.28	-0.57
	our method α=0.10		4.24		0.92			-0.04	0.00	0.00	0.00
	our method α=0.05		4.37		0.92			-0.09	0.01	0.00	0.00
gender, equal opportunity											
bnn	det-cons	19.57	5.92	0.94	125.73	-5.26	0.03	-0.64	0.07	-0.28	-0.57
	det-greedy		5.23		106.70			-0.63	0.07	-0.28	-0.57
	det-relaxed		5.24		107.11			-0.63	0.07	-0.28	-0.57
	our method α=0.10		5.12		0.94			-0.03	0.00	0.00	0.00
	our method α=0.05		5.27		0.94			-0.08	0.00	0.00	0.00
bnn-emb	det-cons	16.31	5.82	0.92	285.38	-3.42	0.02	-0.64	0.07	-0.28	-0.58
	det-greedy		5.20		274.54			-0.63	0.07	-0.28	-0.57
	det-relaxed		5.22		274.18			-0.63	0.07	-0.28	-0.57
	our method α=0.10		4.25		0.92			-0.04	0.00	0.00	0.00
	our method α=0.05		4.37		0.92			-0.09	0.01	0.00	0.00

4.5 | Results

We report the comparative experimental results on three benchmark datasets `dblp`, `imdb`, and `uspt` in Tables 8 to 15, respectively. Our analysis reveals several key findings. Foremost, we observe that neural team recommendation baselines exhibit significant biases with respect to two protected attributes: popularity (favoring already well-known experts) and gender (showing systematic under-representation of female experts). Before applying our debiasing reranking method, these baselines demonstrated a clear tendency to amplify the pre-existing biases in the data. Specifically, experts with a high volume of papers in `dblp`, many movies in `imdb`, and patents in `uspt` would be disproportionately recommended, while qualified experts with lower visibility were systematically overlooked. Similarly, the gender distribution in the recommended teams was significantly skewed, indicating algorithmic bias rather than a mere reflection of domain demographics. In terms of `ndkl` and `skew` metrics,

TABLE 9 Average performance of 5-fold neural models on test set on `dblp` dataset. For the metrics, `ndkl`, the lower the better (↓), `expo`, the closer to 1 the better (→ 1), `skew`, the closer to 0 the better (→ 0), and `map` and `ndcg`, the higher the better (↑).

dblp(k=5)		%ndkl↓		expo → 1		skew before → 0		skew after → 0		%map	%ndcg
		before	after	before	after	disadvantaged	advantaged	disadvantaged	advantaged	Δ ↑	Δ ↑
popularity, demographic parity											
bnn	det-cons	113.27	18.94	0.05	0.08	-25.66	1.08	-0.66	0.72	-0.23	-0.41
	det-greedy		17.92		0.10			-0.65	0.71	-0.23	-0.41
	det-relaxed		17.97		1.35			-0.62	0.70	-0.23	-0.41
	our method α=0.10		19.71		0.78			-0.15	0.26	0.00	0.00
	our method α=0.05		22.25		0.78			-0.16	0.28	0.00	0.00
bnn-emb	det-cons	112.45	14.09	0.07	1.92	-25.58	1.10	-0.51	0.62	-0.23	-0.42
	det-greedy		14.09		2.02			-0.51	0.62	-0.23	-0.42
	det-relaxed		17.65		3.65			-0.50	0.61	-0.23	-0.42
	our method α=0.10		19.61		0.82			-0.15	0.26	0.00	0.00
	our method α=0.05		22.16		0.82			-0.17	0.29	0.00	0.00
popularity, equal opportunity											
bnn	det-cons	105.69	13.12	0.05	0.08	-25.62	1.01	-0.51	0.57	-0.23	-0.42
	det-greedy		13.16		0.10			-0.51	0.57	-0.23	-0.42
	det-relaxed		16.15		0.16			-0.50	0.57	-0.23	-0.42
	our method α=0.10		18.96		0.76			-0.16	0.24	0.00	0.00
	our method α=0.05		22.61		0.61			-0.18	0.27	0.00	0.00
bnn-emb	det-cons	104.96	12.65	0.07	1.54	-25.53	1.03	-0.48	0.55	-0.23	-0.42
	det-greedy		12.70		1.46			-0.48	0.55	-0.23	-0.42
	det-relaxed		15.59		3.51			-0.47	0.55	-0.23	-0.42
	our method α=0.10		18.87		0.79			-0.16	0.25	0.00	0.00
	our method α=0.05		22.52		0.64			-0.18	0.28	0.00	0.00
gender, demographic parity											
bnn	det-cons	27.02	5.90	0.64	31.31	-10.93	0.00	-0.63	0.07	-0.23	-0.42
	det-greedy		5.21		0.40			-0.62	0.07	-0.23	-0.42
	det-relaxed		5.21		1.51			-0.63	0.07	-0.23	-0.42
	our method α=0.10		5.12		0.65			-0.03	0.00	0.00	0.00
	our method α=0.05		5.27		0.64			-0.08	0.00	0.00	0.00
bnn-emb	det-cons	22.73	5.80	0.61	36.62	-9.82	0.00	-0.64	0.07	-0.23	-0.42
	det-greedy		5.18		1.65			-0.63	0.07	-0.23	-0.42
	det-relaxed		5.20		11.71			-0.63	0.07	-0.23	-0.42
	our method α=0.10		4.24		0.61			-0.04	0.00	0.00	0.00
	our method α=0.05		4.37		0.61			-0.09	0.01	0.00	0.00
gender, equal opportunity											
bnn	det-cons	27.01	5.92	0.64	31.31	-10.93	0.00	-0.64	0.07	-0.23	-0.42
	det-greedy		5.23		0.40			-0.63	0.07	-0.23	-0.42
	det-relaxed		5.24		1.51			-0.63	0.07	-0.23	-0.42
	our method α=0.10		5.12		0.65			-0.03	0.00	0.00	0.00
	our method α=0.05		5.27		0.64			-0.08	0.00	0.00	0.00
bnn-emb	det-cons	22.74	5.82	0.61	36.62	-9.82	0.00	-0.64	0.07	-0.23	-0.42
	det-greedy		5.20		1.65			-0.63	0.07	-0.23	-0.42
	det-relaxed		5.22		11.71			-0.63	0.07	-0.23	-0.42
	our method α=0.10		4.25		0.61			-0.04	0.00	0.00	0.00
	our method α=0.05		4.37		0.61			-0.09	0.01	0.00	0.00

Tables 8 to 15 reveal popularity bias across different domains and baselines. The `ndkl` metric, where 0 represents the ideal value indicating perfect alignment with the desired distribution, consistently showed substantial positive values across *all* baselines~~. Specifically, we observed ndkl values ranging from 102.01 to 110.31 to for dblp, 61.74 to 74.67 for imdb and 42.11 to 110.13 for uspt datasets,~~ indicating divergence from the desired distribution obtained by demographic parity and equal~~ity of~~ opportunity fairness notions. This deviation demonstrates that the baselines disproportionately favored popular experts, effectively marginalizing less frequently occurring experts in the recommended teams. The consistency of the observed pattern indicates that popularity bias is not confined to any single domain but rather represents a systematic issue in neural team recommendation baselines before our fairness interventions.

The `skew` metric, which is symmetric around 0, representing the desired representation, provides a more detailed view of representation disparities. Positive values indicate over-representation and negative values signal under-representation of protected groups. This metric also shows consistent patterns of popularity bias across our experimental settings. Tables 8 to 15

TABLE 10 ~~Average performance of 5-fold neural models on test set on imdb dataset. For the metrics, ndkl, the lower the better (↓), skew, the closer to 0 the better (→0), and map and ndcg, the higher the better (↓).~~

		%ndkl before↓	%ndkl after↓	skew before→0 nonprotected	skew before→0 protected	skew after→0 nonprotected	skew after→0 protected	%map10 Δ ↑	%ndcg10 Δ ↑
popularity, demographic parity									
bnn	det-cons	67.53	16.59	0.74	-4.01	-0.54	0.26	-0.36	-0.82
	det-greedy		16.60			-0.54	0.26	-0.36	-0.82
	det-relaxed		16.35			-0.53	0.26	-0.36	-0.82
	our method		17.27			0.21	-0.19	0.00	00.00
bnn-emb	det-cons	74.67	15.71	0.7870	-4.30	-0.46	0.23	-0.48	-1.03
	det-greedy		15.72			-0.46	0.23	-0.48	-1.03
	det-relaxed		15.43			-0.45	0.23	-0.48	-1.03
	our method		17.53			0.21	-0.19	0.00	0.00
popularity, equality of opportunity									
bnn	det-cons	61.74	19.85	0.68	-3.96	-0.62	0.32	-0.35	-0.81
	det-greedy		20.11			-0.62	0.32	-0.35	-0.81
	det-relaxed		19.70			-0.62	0.32	-0.35	-0.81
	our method		16.61			0.19	-0.20	0.00	0.00
bnn-emb	det-cons	70.61	18.94	0.72	-4.17	-0.55	0.30	-0.48	-1.03
	det-greedy		19.17			-0.55	0.30	-0.48	-1.03
	det-relaxed		18.68			-0.55	0.30	-0.48	-1.03
	our method		18.15			0.20	-0.20	0.00	0.00
gender, demographic parity									
bnn	det-cons	4.13	4.44	0.00	-0.04	0.03	-0.34	-0.36	-0.81
	det-greedy		3.96			0.04	-0.36	-0.36	-0.81
	det-relaxed		4.00			0.04	-0.35	-0.36	-0.81
	our method		4.09			0.00	-0.04	0.00	0.00
bnn-emb	det-cons	4.92	8.34	0.01	-0.13	0.06	-0.61	-1.17	-1.03
	det-greedy		3.99			0.03	-0.33	-1.17	-1.03
	det-relaxed		4.21			0.03	-0.31	-1.17	-1.03
	our method		4.88			0.01	-0.12	0.00	0.00
gender, equality of opportunity									
bnn	det-cons	3.42	3.97	0.00	-0.06	0.03	-0.27	-0.42	-0.99
	det-greedy		3.67			0.03	-0.25	-0.42	-0.99
	det-relaxed		3.71			0.02	-0.25	-0.42	-0.99
	our method		3.39			0.00	-0.05	0.00	0.00
bnn-emb	det-cons	4.00	4.17	0.01	-0.13	0.02	-0.24	-0.54	-1.20
	det-greedy		3.89			0.02	-0.22	-0.54	-1.20
	det-relaxed		3.93			0.02	-0.21	-0.54	-1.20
	our method		3.96			0.01	-0.13	0.00	0.00

demonstrate that across all domains and baselines before our fairness considerations, there is a systematic over-representation of popular experts (indicated by positive skew values) joint with consistent under-representation of nonpopular experts (negative skew). This bidirectional deviation from the ideal case is particularly informative as it quantifies not just the presence of popularity bias, but its direction and magnitude. For instance, in the dblp dataset, with demographic parity notion of fairness and bnn team recommendation baseline, popular experts showed a positive skew ~~of 1.13~~at top-$k \in \{5, 10\}$, indicating they received more recommendations than would be expected based on their proportion in the entire dataset. Conversely, nonpopular experts exhibited a negative skew ~~of -19.97~~, suggesting they were recommended less frequently than desired. Similar patterns emerged in both the imdb and uspt datasets, with popular experts consistently showing positive values and nonpopular experts showing negative values. A comparable trend is also observed for gender. Across all datasets and for both fairness notions, female experts exhibit negative skew values before reranking, indicating under-representation, while ndkl is consistently positive, reflecting a substantial divergence from the desired prior distribution.

> **Finding 1.** Neural team recommendation baselines, regardless of the underlying architecture, withhold substantial biases in both popularity and gender representation before our debiasing intervention.

746

In response to **RQ1**, whether our proposed probabilistic reranking method can mitigate popularity and gender biases in the recommended team of experts based on demographic parity and equal~~ity of~~ opportunity while maintaining the team's likelihood of success, from Tables 8 to 15, we can observe that our method could substantially reduce the bias of the neural

TABLE 11 Average performance of 5-fold neural models on test set on `imdb` dataset. For the metrics, `ndkl`, the lower the better (↓), `expo`, the closer to 1 the better (→ 1), `skew`, the closer to 0 the better (→ 0), and `map` and `ndcg`, the higher the better (↑).

imdb(k=10)		%ndkl↓		expo → 1		skew before → 0		skew after → 0		%map	%ndcg
		before	after	before	after	disadvantaged	advantaged	disadvantaged	advantaged	Δ ↑	Δ ↑
popularity, demographic parity											
bnn	det-cons	80.46	16.59	0.14	1.89	-22.46	0.79	0.26	-0.54	-0.35	-0.81
	det-greedy		16.60		1.89			0.26	-0.54	-0.35	-0.81
	det-relaxed		16.35		1.60			0.26	-0.53	-0.35	-0.81
	our method α=0.10		17.27		0.71			-0.19	0.21	0.00	0.00
	our method α=0.05		20.16		0.73			-0.21	0.23	0.00	0.00
bnn-emb	det-cons	83.42	15.71	0.05	2.31	-25.29	0.82	0.23	-0.46	-0.47	-1.03
	det-greedy		15.72		2.30			0.23	-0.46	-0.47	-1.03
	det-relaxed		15.43		1.86			0.23	-0.45	-0.47	-1.03
	our method α=0.10		17.53		0.73			-0.19	0.21	0.00	0.00
	our method α=0.05		20.52		0.74			-0.21	0.23	0.00	0.00
popularity, equal opportunity											
bnn	det-cons	74.26	19.85	0.14	1.85	-22.41	0.72	0.32	-0.62	-0.35	-0.81
	det-greedy		20.11		1.94			0.32	-0.62	-0.35	-0.81
	det-relaxed		19.70		1.63			0.32	-0.62	-0.35	-0.81
	our method α=0.10		16.61		0.76			-0.20	0.19	0.00	0.00
	our method α=0.05		19.29		0.72			-0.24	0.22	0.00	0.00
bnn-emb	det-cons	77.06	18.94	0.05	2.26	-25.24	0.75	0.30	-0.55	-0.48	-1.03
	det-greedy		19.17		2.35			0.30	-0.55	-0.48	-1.03
	det-relaxed		18.68		1.91			0.30	-0.55	-0.48	-1.03
	our method α=0.10		18.15		0.77			-0.20	0.20	0.00	0.00
	our method α=0.05		19.66		0.74			-0.24	0.22	0.00	0.00
gender, demographic parity											
bnn	det-cons	14.13	8.48	0.56	0.14	-10.32	0.05	-0.60	0.06	-0.35	-0.81
	det-greedy		8.48		0.08			-0.59	0.06	-0.35	-0.81
	det-relaxed		8.48		0.09			-0.60	0.06	-0.35	-0.81
	our method α=0.10		3.74		0.56			0.11	-0.01	0.00	0.00
	our method α=0.05		3.89		0.56			0.06	-0.01	0.00	0.00
bnn-emb	det-cons	15.96	8.34	0.47	0.03	-13.50	0.05	-0.61	0.06	-0.47	-1.03
	det-greedy		8.45		0.00			-0.59	0.06	-0.47	-1.03
	det-relaxed		8.43		0.00			-0.59	0.06	-0.47	-1.03
	our method α=0.10		4.24		0.47			0.07	-0.01	0.00	0.00
	our method α=0.05		4.53		0.47			0.01	0.00	0.00	0.00
gender, equal opportunity											
bnn	det-cons	13.79	7.78	0.56	0.13	-10.24	0.04	-0.51	0.05	-0.35	-0.81
	det-greedy		7.78		0.07			-0.50	0.05	-0.35	-0.81
	det-relaxed		7.77		0.07			-0.51	0.05	-0.35	-0.81
	our method α=0.10		3.76		0.56			0.20	-0.02	0.00	0.00
	our method α=0.05		3.88		0.56			0.15	-0.02	0.00	0.00
bnn-emb	det-cons	15.71	7.66	0.47	0.03	-13.41	0.04	-0.52	0.05	-0.47	-1.03
	det-greedy		7.74		0.00			-0.50	0.05	-0.47	-1.03
	det-relaxed		7.72		0.00			-0.50	0.05	-0.47	-1.03
	our method α=0.10		4.26		0.47			0.15	-0.02	0.00	0.00
	our method α=0.05		4.51		0.47			0.09	-0.01	0.00	0.00

team recommendation baselines in terms of `ndkl` and `skew` (closer to 0) with no change to the information retrieval metrics on all datasets for top-$k\in\{5, 10\}$.

> **Finding 2.** Our proposed probabilistic reranking ~~can~~ mitigate **popularity** bias while maintaining ~~utility~~accuracy across *all* domains.

With respect to gender bias, our method demonstrates strong effectiveness in mitigating the bias overall, with consistent improvements observed in `ndkl` and `skew` indicating a reduction in divergence. However, our reranking method falls short in `uspt` in terms of `skew`, and we ~~`ndkl` and `skew` values after applying our reranking methods generally remain the same, or become worse (larger value) in several settings like for `bnn` baseline on `uspt` dataset where `ndkl` has increased from 6.50 to 11.46 after applying our method. We~~ attribute this to the extreme gender bias in the dataset~~s (e.g.~~(i.e., 0.862 to 0.138

TABLE 12 Average performance of 5-fold neural models on test set on `imdb` dataset. For the metrics, `ndkl`, the lower the better (↓), `expo`, the closer to 1 the better (→ 1), `skew`, the closer to 0 the better (→ 0), and `map` and `ndcg`, the higher the better (↑).

`imdb(k=5)`		`%ndkl`↓		`expo` → 1		`skew` before → 0		`skew` after → 0		`%map`	`%ndcg`
		before	after	before	after	disadvantaged	advantaged	disadvantaged	advantaged	Δ ↑	Δ ↑
popularity, demographic parity											
bnn	det-cons	82.23	16.59	0.07	4.53	-24.81	0.77	0.26	-0.54	-0.30	-0.63
	det-greedy		16.60		4.55			0.26	-0.54	-0.30	-0.63
	det-relaxed		16.35		2.21			0.26	-0.53	-0.30	-0.63
	our method α=0.10		17.27		0.66			-0.19	0.21	0.00	0.00
	our method α=0.05		20.16		0.07			-0.21	0.23	0.00	0.00
bnn-emb	det-cons	84.10	15.71	0.03	5.21	-26.25	0.81	0.23	-0.46	-0.40	-0.79
	det-greedy		15.72		5.26			0.23	-0.46	-0.40	-0.79
	det-relaxed		15.43		2.40			0.23	-0.45	-0.40	-0.79
	our method α=0.10		17.53		0.67			-0.19	0.21	0.00	0.00
	our method α=0.05		20.52		0.03			-0.21	0.23	0.00	0.00
popularity, equal opportunity											
bnn	det-cons	75.97	19.85	0.07	4.07	-24.75	0.71	0.32	-0.62	-0.30	-0.62
	det-greedy		20.11		4.66			0.32	-0.62	-0.30	-0.62
	det-relaxed		19.70		2.43			0.32	-0.62	-0.30	-0.62
	our method α=0.10		16.61		0.54			-0.20	0.19	0.00	0.00
	our method α=0.05		19.29		0.07			-0.24	0.22	0.00	0.00
bnn-emb	det-cons	77.73	18.94	0.03	4.70	-26.20	0.75	0.30	-0.55	-0.40	-0.79
	det-greedy		19.21		5.39			0.30	-0.55	-0.40	-0.79
	det-relaxed		18.77		2.65			0.30	-0.55	-0.40	-0.79
	our method α=0.10		16.88		0.54			-0.20	0.19	0.00	0.00
	our method α=0.05		19.66		0.03			-0.24	0.22	0.00	0.00
gender, demographic parity											
bnn	det-cons	17.83	8.48	0.25	0.15	-19.52	0.07	-0.60	0.06	-0.30	-0.62
	det-greedy		8.48		0.00			-0.59	0.06	-0.30	-0.62
	det-relaxed		8.48		0.00			-0.60	0.06	0.30	-0.62
	our method α=0.10		3.74		0.25			0.11	-0.01	0.00	0.00
	our method α=0.05		3.89		0.25			0.06	-0.01	0.00	0.00
bnn-emb	det-cons	19.83	8.34	0.31	0.00	-18.20	0.06	-0.61	0.06	-0.40	-0.79
	det-greedy		8.45		0.00			-0.59	0.06	-0.40	-0.79
	det-relaxed		8.43		0.00			-0.59	0.06	-0.40	-0.79
	our method α=0.10		4.24		0.31			0.07	-0.01	0.00	0.00
	our method α=0.05		4.53		0.31			0.01	0.00	0.00	0.00
gender, equal opportunity											
bnn	det-cons	17.46	7.78	0.25	0.15	-19.44	0.06	-0.51	0.05	-0.30	-0.63
	det-greedy		7.78		0.00			-0.50	0.05	-0.30	-0.62
	det-relaxed		7.77		0.00			-0.51	0.05	-0.30	-0.62
	our method α=0.10		3.76		0.25			0.20	-0.02	0.00	0.00
	our method α=0.05		3.88		0.25			0.15	-0.02	0.00	0.00
bnn-emb	det-cons	19.62	7.66	0.31	0.00	-18.12	0.05	-0.52	0.05	-0.40	-0.79
	det-greedy		7.74		0.00			-0.50	0.05	-0.40	-0.79
	det-relaxed		7.72		0.00			-0.50	0.05	-0.40	-0.79
	our method α=0.10		4.26		0.31			0.15	-0.02	0.00	0.00
	our method α=0.05		4.51		0.31			0.09	-0.01	0.00	0.00

male vs. female ratio) ~~in `uspt`) such that the original top-k ranked list lacks enough (qualified) female experts.However, our method exhibits clear limitations when addressing gender bias. The fairness metrics, `ndkl` and `skew` values post-reranking generally remain the same or become worse~~. For instance, under equal opportunity with the `bnn-emb` baseline, the `skew` metric increases, reflecting a deviation from the desired distribution.

> **Finding 3.** Our proposed probabilistic reranking generally mitigates **gender** bias while maintaining ~~utility~~accuracy across domains, except in cases where the domain exhibits extreme bias.

~~applying our method to the `bnn` baseline on the `uspt` dataset resulted in a significant degradation, with `ndkl` increasing from 6.50 to 11.46.~~ This limitation stems from severe pre-existing biases in the underlying data distributions, exemplified by the severe gender imbalance in datasets such as `uspt` ~~with its 0.862 to 0.138 male-to-female ratio. Such extreme gender~~

TABLE 13 ~~Average performance of 5-fold neural models on test set on `uspt` dataset. For the metrics, `ndkl`, the lower the better (↓), `skew`, the closer to `0` the better (→`0`), and `map` and `ndcg`, the higher the better (↓).~~

		`%ndkl` before↓	`%ndkl` after↓	`skew` before →`0` protected	`skew` before →`0` nonprotected	`skew` after →`0` protected	`skew` after →`0` nonprotected	`%map10` Δ ↑	`%ndcg10` Δ ↑
popularity, demographic parity									
bnn	det-cons	90.93	24.36	-10.58	1.05	0.81	0.78	-0.24	-0.53
	det-greedy		24.35			0.81	0.78	-0.24	-0.53
	det-relaxed		28.94			-0.80	0.78	-0.24	-0.53
	our method		18.52			-0.13	-0.59	0.00	0.00
bnn-emb	det-cons	110.13	18.88	-22.79	1.13	-0.61	0.68	-0.67	-1.35
	det-greedy		18.88			-0.61	0.68	-0.67	-1.35
	det-relaxed		24.13			-0.60	0.67	-0.67	-1.35
	our method		19.68			-0.15	0.22	0.00	0.00
popularity, equality of opportunity									
bnn	det-cons	42.11	16.12	-9.95	0.44	-0.32	0.24	-0.24	-0.53
	det-greedy		20.39			-0.32	0.24	-0.24	-0.53
	det-relaxed		19.06			-0.32	0.24	-0.24	-0.53
	our method		17.20			-0.14	-0.62	0.00	0.00
bnn-emb	det-cons	52.44	14.50	-22.16	0.53	-0.18	0.18	-0.67	-1.35
	det-greedy		19.48			-0.18	0.18	-0.67	-1.35
	det-relaxed		18.04			-0.18	0.18	-0.67	-1.35
	our method		14.65			-0.21	0.13	0.00	0.00
gender, demographic parity									
bnn	det-cons	6.50	3.73	0.30	-0.08	0.13	-0.02	-0.24	-0.53
	det-greedy		2.67			0.13	-0.02	-0.24	-0.53
	det-relaxed		3.51			0.13	-0.25	-0.24	-0.53
	our method		11.46			0.36	-0.37	0.00	0.00
bnn-emb	det-cons	8.07	4.52	0.54	-0.13	0.13	-0.02	-0.67	-1.35
	det-greedy		2.68			0.13	-0.02	-0.67	-1.35
	det-relaxed		4.37			0.12	-0.02	-0.67	-1.35
	our method		8.32			0.55	-0.141	0.00	0.00
gender, equality of opportunity									
bnn	det-cons	12.08	10.15	0.20	0.00	0.00	0.07	-0.24	-0.53
	det-greedy		9.03			0.00	0.07	-0.24	-0.53
	det-relaxed		9.71			0.00	0.07	-0.24	-0.53
	our method		16.40			0.33	-0.33	0.00	0.00
bnn-emb	det-cons	12.80	10.81	0.44	-0.04	0.01	0.06	-0.67	-1.35
	det-greedy		9.15			0.01	0.06	-0.67	-1.35
	det-relaxed		10.49			0.01	0.06	-0.67	-1.35
	our method		12.8			0.47	-0.05	0.00	0.00

~~bias~~ creates a fundamental limitation where the top-k ranked lists of experts simply lack a sufficient pool of qualified female experts to draw from. Additionally, our results point to the broader challenge of achieving fairness in domains with extreme under-representation, where the space of possible fair solutions is constrained by the available expert pool. Thus, achieving strong mitigation of structural demographic bias using a fully model-agnostic, post-processing framework is not feasible without modifying the underlying model or the data. To our knowledge, no existing pre- or in-processing fairness method has been developed for neural team recommendation. A few prior works[52,53] exist, but they are purely algorithmic or rule-based rather than optimization- or machine-learning–driven. They do not train predictive models or leverage data-driven parameter learning, and therefore cannot be categorized as pre- or in-processing approaches within a machine-learning pipeline. Developing effective pre- and in-processing fairness methods for neural team recommendation remains an important direction for future work.

> **Finding 4.** Different types of bias may require distinct mitigation strategies due to the *level* of bias across different protected attributes.

To answer **RQ2**, regarding whether our proposed probabilistic reranking method outperforms deterministic reranking methods, from Tables 8 to 15 we observe that our probabilistic method demonstrates superior performance over deterministic rerankers specifically in mitigating popularity bias across various fairness notions on all datasets, as evidenced by the combination of fairness and information retrieval metrics. This superiority is reflected in the consistency of bias reduction and maintaining

TABLE 14 Average performance of 5-fold neural models on test set on `uspt` dataset. For the metrics, `ndkl`, the lower the better (↓), `expo`, the closer to 1 the better (→ 1), `skew`, the closer to 0 the better (→ 0), and `map` and `ndcg`, the higher the better (↑).

uspt (k=10)		`%ndkl`↓		`expo` → 1		`skew` before → 0		`skew` after → 0		`%map`	`%ndcg`
		before	after	before	after	disadvantaged	advantaged	disadvantaged	advantaged	$\Delta\uparrow$	$\Delta\uparrow$
popularity, demographic parity											
bnn	det-cons	95.55	24.36	0.40	9.89	-17.37	1.01	-0.81	0.78	-0.24	-0.53
	det-greedy		24.35		9.91			-0.81	0.78	-0.24	-0.53
	det-relaxed		28.94		12.88			-0.80	0.78	-0.24	-0.53
	our method α=0.10		18.52		0.86			-0.13	-0.59	0.00	0.00
	our method α=0.05		21.73		0.90			-0.14	-0.57	0.00	0.00
bnn-emb	det-cons	112.42	18.88	0.07	8.60	-25.29	1.13	-0.61	0.68	-0.67	-1.35
	det-greedy		18.88		8.51			-0.61	0.68	-0.67	-1.35
	det-relaxed		24.13		10.11			-0.60	0.67	-0.67	-1.35
	our method α=0.10		19.68		0.81			-0.15	0.22	0.00	0.00
	our method α=0.05		23.63		0.82			-0.16	0.24	0.00	0.00
popularity, equal opportunity											
bnn	det-cons	48.3	16.12	0.40	5.04	-16.74	0.41	-0.32	0.24	-0.24	-0.53
	det-greedy		20.39		4.49			-0.32	0.24	-0.24	-0.53
	det-relaxed		19.06		4.60			-0.32	0.24	-0.24	-0.53
	our method α=0.10		17.20		0.81			-0.14	-0.62	0.00	0.00
	our method α=0.05		18.67		0.71			-0.18	-0.58	0.00	0.00
bnn-emb	det-cons	54.37	14.50	0.07	0.19	-24.66	0.53	-0.18	0.18	-0.67	-1.35
	det-greedy		19.48		0.11			-0.18	0.18	-0.67	-1.35
	det-relaxed		18.04		0.12			-0.18	0.18	-0.67	-1.35
	our method α=0.10		14.65		0.65			-0.21	0.13	0.00	0.00
	our method α=0.05		16.87		0.49			-0.26	0.15	0.00	0.00
gender, demographic parity											
bnn	det-cons	17.57	3.73	0.77	79.91	-4.38	-0.03	0.13	-0.02	-0.24	-0.53
	det-greedy		2.67		58.82			0.13	-0.02	-0.24	-0.53
	det-relaxed		3.51		71.47			0.13	-0.02	-0.24	-0.53
	our method α=0.10		11.46		0.85			0.36	-0.37	0.00	0.00
	our method α=0.05		10.51		0.83			0.35	-0.32	0.00	0.00
bnn-emb	det-cons	20.49	4.52	1.03	180.98	-0.16	-0.10	0.13	-0.02	-0.67	-1.35
	det-greedy		2.68		221.92			0.13	-0.02	-0.67	-1.35
	det-relaxed		4.37		219.19			0.12	-0.02	-0.67	-1.35
	our method α=0.10		8.32		1.03			0.55	-0.14	0.00	0.00
	our method α=0.05		8.29		1.03			0.55	-0.13	0.00	0.00
gender, equal opportunity											
bnn	det-cons	23.99	10.15	0.77	76.25	-4.48	0.06	0.00	0.07	-0.24	-0.53
	det-greedy		9.03		62.76			0.00	0.07	-0.24	-0.53
	det-relaxed		9.71		77.72			0.00	0.07	-0.24	-0.53
	our method α=0.10		16.40		0.87			0.33	-0.33	0.00	0.00
	our method α=0.05		15.60		0.83			0.25	-0.23	0.00	0.00
bnn-emb	det-cons	25.40	10.81	1.03	176.74	-0.26	-0.01	0.01	0.06	-0.67	-1.35
	det-greedy		9.15		206.10			0.01	0.06	-0.67	-1.35
	det-relaxed		10.49		211.67			0.01	0.06	-0.67	-1.35
	our method α=0.10		12.80		1.03			0.47	-0.05	0.00	0.00
	our method α=0.05		12.84		1.03			0.46	-0.05	0.00	0.00

information retrieval metrics. While deterministic rerankers show some capability in mitigating popularity bias, this comes at the cost of a drastic drop in information retrieval metrics. Specifically, we observe consistent negative values in Δ`map`~~@10~~ and Δ`ndcg`~~@10~~ at top-$k \in \{5, 10\}$, indicating drops in recommendation quality. This trade-off between fairness and accuracy highlights an important advantage of our probabilistic approach, which maintains recommendation quality while improving fairness metrics.

The effectiveness of gender bias mitigation strategies is lower than that for popularity bias and demonstrates fewer clear-cut results in specific cases. Deterministic rerankers achieve marginally better fairness metrics, showing improvements in `ndkl` and `skew` values, compared to our probabilistic method. However, this modest improvement in fairness comes at a considerable cost to recommendation accuracy, with deterministic approaches showing a significant loss in information retrieval metrics. This suggests that while deterministic approaches might achieve slightly better results in terms of mitigating gender bias, they

TABLE 15 Average performance of 5-fold neural models on test set on `uspt` dataset. For the metrics, `ndkl`, the lower the better (↓), `expo`, the closer to 1 the better (→ 1), `skew`, the closer to 0 the better (→ 0), and `map` and `ndcg`, the higher the better (↑).

uspt (k=5)		%ndkl↓		expo → 1		skew before → 0		skew after → 0		%map	%ndcg
		before	after	before	after	disadvantaged	advantaged	disadvantaged	advantaged	Δ ↑	Δ ↑
popularity, demographic parity											
bnn	det-cons	98.12	24.36	0.31	3.61	-19.45	0.90	-0.81	0.78	-0.19	-0.39
	det-greedy		24.35		3.67			-0.81	0.78	-0.19	-0.39
	det-relaxed		28.94		8.50			-0.80	0.78	-0.19	-0.39
	our method α=0.10		18.52		0.95			-0.13	-0.59	0.00	0.00
	our method α=0.05		21.73		0.88			-0.14	-0.57	0.00	0.00
bnn-emb	det-cons	113.12	18.88	0.05	1.54	-25.83	1.12	-0.61	0.68	-0.55	-1.10
	det-greedy		18.88		1.54			-0.61	0.68	-0.55	-1.10
	det-relaxed		24.13		3.21			-0.60	0.67	-0.55	-1.10
	our method α=0.10		19.68		0.82			-0.15	0.22	0.00	0.00
	our method α=0.05		23.63		0.73			-0.16	0.24	0.00	0.00
popularity, equal opportunity											
bnn	det-cons	51.56	16.12	0.31	1.66	-18.82	0.30	-0.32	0.24	-0.19	-0.39
	det-greedy		20.39		2.04			-0.32	0.24	-0.19	-0.39
	det-relaxed		19.06		1.95			-0.32	0.24	-0.19	-0.39
	our method α=0.10		17.20		0.44			-0.14	-0.62	0.00	0.00
	our method α=0.05		18.67		0.41			-0.18	-0.58	0.00	0.00
bnn-emb	det-cons	55.00	14.50	0.05	0.38	-25.20	0.51	-0.18	0.18	-0.55	-1.10
	det-greedy		19.48		0.38			-0.18	0.18	-0.55	-1.10
	det-relaxed		18.04		0.52			-0.18	0.18	-0.55	-1.10
	our method α=0.10		14.65		0.13			-0.21	0.13	0.00	0.00
	our method α=0.05		16.87		0.08			-0.26	0.15	0.00	0.00
gender, demographic parity											
bnn	det-cons	22.82	3.73	0.51	5.30	-11.43	-0.04	0.13	-0.02	-0.18	-0.38
	det-greedy		2.67		6.85			0.13	-0.02	-0.18	-0.38
	det-relaxed		3.51		8.42			0.13	-0.02	-0.18	-0.38
	our method α=0.10		11.46		0.58			0.36	-0.37	0.00	0.00
	our method α=0.05		10.51		0.57			0.35	-0.32	0.00	0.00
bnn-emb	det-cons	28.53	4.52	0.88	17.62	-3.17	-0.10	0.13	-0.02	-0.55	-1.10
	det-greedy		2.68		7.57			0.13	-0.02	-0.55	-1.10
	det-relaxed		4.37		8.72			0.12	-0.02	-0.55	-1.10
	our method α=0.10		8.32		0.89			0.55	-0.14	0.00	0.00
	our method α=0.05		8.29		0.89			0.55	-0.13	0.00	0.00
gender, equal opportunity											
bnn	det-cons	29.47	10.15	0.51	4.99	-11.54	0.04	0.00	0.07	-0.19	-0.38
	det-greedy		9.03		6.35			0.00	0.07	-0.19	-0.38
	det-relaxed		9.71		6.36			0.00	0.07	-0.19	-0.38
	our method α=0.10		16.40		0.60			0.33	-0.33	0.00	0.01
	our method α=0.05		15.94		0.59			0.31	-0.30	0.00	0.01
bnn-emb	det-cons	33.38	10.81	0.88	12.12	-3.27	-0.01	0.01	0.06	-0.55	-1.10
	det-greedy		9.15		11.82			0.01	0.06	-0.55	-1.10
	det-relaxed		10.49		21.73			0.01	0.06	-0.55	-1.10
	our method α=0.10		12.80		0.89			0.47	-0.05	0.00	0.00
	our method α=0.05		12.84		0.89			0.46	-0.05	0.00	0.00

sacrifice substantial recommendation quality, questioning their practical applicability. Moreover, deterministic baselines enforce group proportions at every ranking prefix; in highly skewed datasets, this may distort fairness metrics and can even result in disadvantaged experts being replaced by advantaged ones in order to maintain the prescribed ratio.

It is particularly noteworthy that all deterministic rerankers exhibit similar performance patterns for `ndkl` and `skew` values across datasets, with no single approach demonstrating clear dominance over the others. This finding aligns with Loghmani et al.'s[47] observations and suggests a fundamental limitation in deterministic approaches. While deterministic approaches might seem appealing due to their simplicity and interpretability, our findings suggest that probabilistic methods offer a better approach to addressing bias, considering both fairness and accuracy. As shown in Tables 8 to 15, all methods, including our baselines, effectively reduce bias (the green color in the fairness metrics columns for `skew` and `ndkl` highlights the substantial improvement after reranking), but unlike the baselines as expected, our method maintains ~~utility~~accuracy despite reducing bias (green cell for ~~utility~~accuracy metrics, `map` and `ndcg`, as opposed to the red cell for the baselines), demonstrating that it effectively balances the fairness-accuracy trade-off typically encountered in reranking scenarios.

> **Finding 5.** In the context of neural team recommendation, our proposed probabilistic reranking method ~~consistently~~ generally outperforms deterministic reranking methods ~~on all~~ across datasets and baselines.

For **RQ3**, regarding the analysis of `expo` metric where the ideal value is 1, indicating that the protected groups' exposure is perfectly proportional to their utility scores, and values less (greater) than 1 suggest bias against (in favor of) the disadvantaged group, from Tables 8 to 15, we observe that our proposed probabilistic reranking method shows improvements for protected groups across datasets, particularly for popularity ~~as the protected attribute~~. For nonpopular ~~items~~experts, we observe consistent and substantial improvements across all datasets. Before reranking, the baseline `bnn` exhibited significant bias against nonpopular groups, with `expo` values substantially less than 1 across all datasets. After applying our probabilistic reranking method, these values were adjusted closer to the ideal. This shift indicates that our method effectively balanced the exposure of nonpopular groups relative to their success, mitigating the initial bias. In contrast, deterministic reranking algorithms often overcompensated, resulting in `expo` values significantly greater than 1. For instance, `det-cons` ~~produced~~ results in large value ~~s of 2.436~~ ~~for~~in `uspt` on `bnn` baseline at top-10 under demographic parity. These elevated values suggest reverse discrimination, favouring the disadvantaged group disproportionately. With respect to gender, the results are, however, relatively poor, similar to **RQ2** and Tables 8 to 15 with no or marginal improvement in `expo` metric. ~~All p~~Post-processing reranking methods fall short of mitigating the gender bias in certain observations, which can be attributed to the presence of extreme gender bias in the dataset. For instance, with male-to-female ratios as skewed as 0.862 to 0.138 in `uspt`, the top-k ranked list of experts lacks sufficient representation of female experts, making it impossible for post-processing methods alone to achieve fairness without compromising quality. The extreme imbalance of the datasets also contributes to `expo` values being close to 1 for gender even before reranking. Female experts constitute a small minority, and if only a few female experts, sometimes a single one, appear in the top-k recommendations, they typically occupy ranks consistent with their predicted probabilities. Consequently, their position-based visibility is not disproportionately low relative to their model-assigned utility, and the ratio between exposure and utility remains similar to that of male experts. Therefore, even limited presence at ranks aligned with predicted scores is sufficient to yield a value near 1. Nonetheless, although our proposed probabilistic method negligibly changes the values of `expo` for better or worse across different settings, its performance is better than deterministic reranking methods. This is evidenced by ~~how deterministic methods tend to push expo values well above 1, indicating~~ the instability of the results for deterministic methods, which is difficult to interpret and may stem from an artificial inflation of female representation that compromises the quality of recommendations.

In summary, our probabilistic reranking method consistently mitigated popularity bias across all datasets and fairness definitions while maintaining utility in terms of `expo`. Unlike deterministic methods that may overcompensate and drop utility drastically, our approach presents a balance by proportionally adjusting exposure based on information retrieval metrics. ~~Also, our method successfully improves the ratio of exposure to utility for disadvantaged groups, moving expo values towards the ideal 1 across multiple datasets and fairness definitions.~~

> **Finding 6.** Our probabilistic reranking method's performance in terms of `expo` is ~~consistent~~ largely aligned with its fairness metrics namely `ndkl` and `skew` across ~~all~~ most settings and domains.

Regarding **RQ4**, that is, whether the effect of our proposed reranking method is consistent across datasets from different domains, Tables 8 to 15 demonstrate that each fairness-aware reranker, whether deterministic or probabilistic, follows a similar trend across the `dblp`, `imdb`, and `uspt` datasets, despite these datasets originating from different domains. Specifically, the performance metrics, including fairness measures and ~~utility~~accuracy metrics, remain consistent in terms of their trends when applying the same reranking algorithms to different datasets. This consistency suggests that the inherent patterns of bias and the distribution of protected attributes are similar across these datasets, as illustrated in Figure 2. Moreover, the similarity in trends indicates that the fairness-aware reranking algorithms are robust to domain variations and can generalize well across different domains. This robustness also implies that the underlying biases in rankings are not unique to a particular domain but are pervasive across various domains. Consequently, fairness interventions that are effective in one domain are likely to be effective in others.

TABLE 16 ~~Average performance of 5-fold neural models in terms of `expo` on test set on `imdb`, `dblp`, `uspt` datasets. For this metric `expo`, the closer to 1 the better (→1)~~

			imdb		dblp		uspt	
			expo→1		expo→1		expo→1	
			before	after	before	after	before	after
popularity, demographic parity								
bnn	det-cons		0.7737	1.7014	0.2612	2.4358	0.6774	1.2303
	det-greedy			1.7014		2.7576		1.2314
	det-relaxed			1.6491		2.2494		1.1268
	our method			**1.0053**		**1.0122**		**0.9858**
bnn-emb	det-cons		0.7705	1.6838	0.2382	2.5549	0.1564	1.1328
	det-greedy			1.6838		2.5552		1.1327
	det-relaxed			1.6320		2.3588		**1.0296**
	our method			**0.9887**		**1.0316**		1.0544
popularity, equal~~ity of~~ opportunity								
bnn	det-cons		0.7737	1.6836	0.2612	2.4456	0.6774	1.3286
	det-greedy			1.6951		2.4488		1.4296
	det-relaxed			1.6434		2.2681		1.3889
	our method			**0.8406**		**1.0139**		**0.9909**
bnn-emb	det-cons		0.7705	1.6677	0.2382	2.5813	0.1564	1.2239
	det-greedy			1.6795		2.5839		1.3298
	det-relaxed			1.6261		2.3951		1.2892
	our method			**0.8377**		**1.0338**		**1.0415**
gender, demographic parity								
bnn	det-cons		**0.9195**	1.1357	0.9741	1.1106	0.9461	0.9925
	det-greedy			1.1640		1.1024		1.0002
	det-relaxed			1.1610		1.1433		0.9481
	our method			0.9066		**0.9750**		**0.9814**
bnn-emb	det-cons		**0.9302**	1.3136	**1.0084**	1.1528	0.9761	1.0175
	det-greedy			1.3003		1.1540		1.0011
	det-relaxed			1.3082		1.1693		1.0128
	our method			0.9233		1.0080		**0.9779**
gender, equal~~ity of~~ opportunity								
bnn	det-cons		**0.9195**	1.1314	0.9741	1.1324	0.9461	1.0065
	det-greedy			1.1563		1.1242		1.0144
	det-relaxed			1.1561		1.1597		1.0189
	our method			0.9066		**0.9743**		**0.9784**
bnn-emb	det-cons		0.9302	1.2875	1.0084	1.1728	0.9761	1.0228
	det-greedy			1.2909		1.1742		**1.0047**
	det-relaxed			1.2971		1.2047		1.0191
	our method			**0.9233**		**1.0014**		0.9752

Finding 7. Our proposed probabilistic reranking method shows consistent effective performance in terms of both fairness and information retrieval metrics across datasets from different domains.

Lastly, our experiments show that while post-processing reranking methods can effectively address biases, their efficacy may become limited to some extent when employed single-handedly when confronting *extreme* biases in a dataset; such methods struggle to rectify biases without a consequential loss in accuracy. A holistic approach that integrates pre-processing, in-processing, and post-processing methods is required to achieve a more balanced and optimal outcome.

5 | CONCLUDING REMARKS

In this paper, we formalized the fair team recommendation problem, where we aim to form an unbiased collaborative group of diverse experts to accomplish complex tasks. While state-of-the-art neural team recommenders can efficiently recommend sets of candidate experts to form effective collaborative teams, they are largely biased toward *male* and *popular* experts, potentially overlooking valuable contributors from under-represented groups. We ~~proposed~~developed a model-agnostic post-processing probabilistic reranking method to mitigate unfair biases in the recommended teams of experts by neural team recommendation models, focusing on maintaining team effectiveness while promoting fairness with respect to demographic parity and equal~~ity of~~

opportunity notions of fairness. Our experiments on three large-scale benchmark datasets from different domains, including `dblp`, `imdb` and `uspt`, showed that: 1) neural team recommenders heavily suffer from biases toward popular and male experts, with popular experts; 2) probabilistic greedy reranking algorithms can substantially mitigate popularity biases while maintaining models' efficacy; 3) biases appeared across all neural team recommendation architectures, indicating it is a fundamental challenge of these systems rather than a flaw in specific model designs; 4) ~~I~~even in the presence of extreme biases where initial recommendations show more than 80% skew toward certain groups, ~~post-processing reranking methods fall short of achieving fair representation.~~our method generally mitigates the bias, with the exception of the `uspt` dataset, where the results remain unstable. 5) ~~o~~Our probabilistic method dominantly outperforms deterministic baselines and is robust towards domain changes. Our future research direction includes mitigating multiple biases jointly, i.e., gender bias together with popularity bias, and incorporating in-processing methods to address these challenges at the model training stage rather than solely through post-processing adjustments.

6 | ~~ETHICAL CONSIDERATIONS~~

~~We utilized `genderize` to obtain missing genders for individuals in the `imdb` and all of the experts in the `dblp` dataset. `Genderize` is a crowd-based system that predicts a gender for a given name with a probability based on the number of appearances of a name in their database, name patterns and cultural associations, which may be reasonably accurate, it inherently incorporates societal biases and assumptions that can be problematic.~~

REFERENCES

1. Gallegos IO, Rossi RA, Barrow J, et al. Bias and Fairness in Large Language Models: A Survey. *Comput. Linguistics.* 2024;50(3):1097–1179. doi: 10.1162/COLI_A_00524
2. Hanna A, Denton E, Smart A, Smith-Loud J. Towards a critical race methodology in algorithmic fairness. In: ACM 2020:501–512
3. Zhang J, Wu S, Wang T, Ding F, Zhu J. Relieving popularity bias in recommendation via debiasing representation enhancement. *Complex & Intelligent Systems.* 2025;11(1):34.
4. Ge Y, Liu S, Gao R, et al. Towards Long-term Fairness in Recommendation. In: ACM 2021:445–453
5. Wei T, Feng F, Chen J, Wu Z, Yi J, He X. Model-Agnostic Counterfactual Reasoning for Eliminating Popularity Bias in Recommender System. In: ACM 2021:1791–1800
6. Dwork C, Hardt M, Pitassi T, Reingold O, Zemel RS. Fairness through awareness. In: Goldwasser S., ed. *Innovations in Theoretical Computer Science 2012, Cambridge, MA, USA, January 8-10, 2012*ACM 2012:214–226
7. Hardt M, Price E, Srebro N. Equality of Opportunity in Supervised Learning. In: Lee DD, Sugiyama M, Luxburg vU, Guyon I, Garnett R., eds. *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain* 2016:3315–3323.
8. Altenburger KM, De R, Frazier K, Avteniev N, Hamilton J. Are There Gender Differences in Professional Self-Promotion? An Empirical Case Study of LinkedIn Profiles Among Recent MBA Graduates. In: AAAI Press 2017:460–463.
9. John PG, Vijaykeerthy D, Saha D. Verifying Individual Fairness in Machine Learning Models. In: Adams RP, Gogate V., eds. *Proceedings of the Thirty-Sixth Conference on Uncertainty in Artificial Intelligence, UAI 2020, virtual online, August 3-6, 2020*. 124 of *Proceedings of Machine Learning Research*. AUAI Press 2020:749–758.
10. Zehlike M, Yang K, Stoyanovich J. Fairness in Ranking, Part I: Score-Based Ranking. *ACM Comput. Surv.*. 2023;55(6):118:1–118:36. doi: 10.1145/3533379
11. Rosenblatt L, Witter RT. Counterfactual Fairness Is Basically Demographic Parity. In: Williams B, Chen Y, Neville J., eds. *Thirty-Seventh AAAI Conference on Artificial Intelligence, AAAI 2023, Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence, IAAI 2023, Thirteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2023, Washington, DC, USA, February 7-14, 2023*AAAI Press 2023:14461–14469
12. Ashurst C, Weller A. Fairness Without Demographic Data: A Survey of Approaches. In: ACM 2023:14:1–14:12
13. Romano Y, Bates S, Candès EJ. Achieving Equalized Odds by Resampling Sensitive Attributes. In: Larochelle H, Ranzato M, Hadsell R, Balcan M, Lin H., eds. *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual* 2020.
14. Chan E, Liu Z, Qiu R, Zhang Y, Maciejewski R, Tong H. Group Fairness via Group Consensus. In: ACM 2024:1788–1808
15. Zhao H. Fair and optimal prediction via post-processing. *AI Mag.*. 2024;45(3):411–418. doi: 10.1002/AAAI.12191
16. Cruz AF, Hardt M. Unprocessing Seven Years of Algorithmic Fairness. In: OpenReview.net 2024.
17. Xian R, Li Q, Kamath G, Zhao H. Differentially Private Post-Processing for Fair Regression. In: OpenReview.net 2024.
18. Zhu K, Fioretto F, Hentenryck PV. Post-processing of Differentially Private Data: A Fairness Perspective. In: Raedt LD., ed. *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI 2022, Vienna, Austria, 23-29 July 2022*ijcai.org 2022:4029–4035
19. Oldfield B, Xu Z, Kandanaarachchi S. Revisiting Pre-processing Group Fairness: A Modular Benchmarking Framework. In: Cha M, Park C, Park N, et al., eds. *Proceedings of the 34th ACM International Conference on Information and Knowledge Management, CIKM 2025, Seoul, Republic of Korea, November 10-14, 2025*ACM 2025:6498–6502
20. Leininger C, Rittel S, Bothmann L. Overcoming Fairness Trade-offs via Pre-processing: A Causal Perspective. In: Weerts HJP, Pechenizkiy M, Allhutter D, Corrêa AM, Grote T, Liem CCS., eds. *European Workshop on Algorithmic Fairness, 30-2 July 2025, Eindhoven University of Technology, Eindhoven, The Netherlands*. 294 of *Proceedings of Machine Learning Research*. PMLR 2025:92–115.

21. Ward JJ, Zeng X, Cheng G. FairRR: Pre-Processing for Group Fairness through Randomized Response. In: Dasgupta S, Mandt S, Li Y., eds. *International Conference on Artificial Intelligence and Statistics, 2-4 May 2024, Palau de Congressos, Valencia, Spain*. 238 of *Proceedings of Machine Learning Research*. PMLR 2024:3826–3834.
22. Lahoti P, Gummadi KP, Weikum G. iFair: Learning Individually Fair Data Representations for Algorithmic Decision Making. In: IEEE 2019:1334–1345
23. Fontana M, Naretto F, Monreale A. Optimizing and Tuning Fairness in Machine Learning: An Augmented Lagrangian Method with a Performance Budget. In: Ribeiro RP, Pfahringer B, Japkowicz N, et al., eds. *Machine Learning and Knowledge Discovery in Databases. Research Track - European Conference, ECML PKDD 2025, Porto, Portugal, September 15-19, 2025, Proceedings, Part I*. 16013 of *Lecture Notes in Computer Science*. Springer 2025:213–230
24. Wan M, Zha D, Liu N, Zou N. In-Processing Modeling Techniques for Machine Learning Fairness: A Survey. *ACM Trans. Knowl. Discov. Data.* 2023;17(3):35:1–35:27. doi: 10.1145/3551390
25. Zehlike M, Castillo C. Reducing Disparate Exposure in Ranking: A Learning To Rank Approach. In: Huang Y, King I, Liu T, Steen vM., eds. *WWW '20: The Web Conference 2020, Taipei, Taiwan, April 20-24, 2020*ACM / IW3C2 2020:2849–2855
26. Beutel A, Chen J, Doshi T, et al. Fairness in Recommendation Ranking through Pairwise Comparisons. In: Teredesai A, Kumar V, Li Y, Rosales R, Terzi E, Karypis G., eds. *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, KDD 2019, Anchorage, AK, USA, August 4-8, 2019*ACM 2019:2212–2220
27. Singh A, Joachims T. Policy Learning for Fairness in Ranking. In: Wallach HM, Larochelle H, Beygelzimer A, d'Alché-Buc F, Fox EB, Garnett R., eds. *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada* 2019:5427–5437.
28. Rahman TA, Surma B, Backes M, Zhang Y. Fairwalk: Towards Fair Graph Embedding. In: Kraus S., ed. *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI 2019, Macao, China, August 10-16, 2019*ijcai.org 2019:3289–3295
29. Gunasekara S, Saarela M. Trade-offs between fairness and performance in educational AI: Analyzing post-processing bias mitigation on the OULAD. *Inf. Softw. Technol.*. 2026;189:107933. doi: 10.1016/J.INFSOF.2025.107933
30. Soltan A, Washington P. Challenges in Reducing Bias Using Post-Processing Fairness for Breast Cancer Stage Classification with Deep Learning. *Algorithms.* 2024;17(4):141. doi: 10.3390/A17040141
31. Zehlike M, Sühr T, Baeza-Yates R, Bonchi F, Castillo C, Hajian S. Fair Top-*k* Ranking with multiple protected groups. *Inf. Process. Manag.*. 2022;59(1):102707. doi: 10.1016/J.IPM.2021.102707
32. Feng Y, Shah C. Has CEO Gender Bias Really Been Fixed? Adversarial Attacking and Improving Gender Fairness in Image Search. In: AAAI Press 2022:11882–11890
33. Mehrotra A, Vishnoi NK. Fair Ranking with Noisy Protected Attributes. In: Koyejo S, Mohamed S, Agarwal A, Belgrave D, Cho K, Oh A., eds. *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022* 2022.
34. Geyik SC, Ambler S, Kenthapadi K. Fairness-Aware Ranking in Search & Recommendation Systems with Application to LinkedIn Talent Search. In: ACM 2019:2221–2231.
35. Biega AJ, Gummadi KP, Weikum G. Equity of Attention: Amortizing Individual Fairness in Rankings. In: Collins-Thompson K, Mei Q, Davison BD, Liu Y, Yilmaz E., eds. *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval, SIGIR 2018, Ann Arbor, MI, USA, July 08-12, 2018*ACM 2018:405–414
36. Graves A. Practical Variational Inference for Neural Networks. In: Shawe-Taylor J, Zemel RS, Bartlett PL, Pereira FCN, Weinberger KQ., eds. *Advances in Neural Information Processing Systems 24: 25th Annual Conference on Neural Information Processing Systems 2011. Proceedings of a meeting held 12-14 December 2011, Granada, Spain* 2011:2348–2356.
37. Rad RH, Fani H, Kargar M, Szlichta J, Bagheri E. Learning to Form Skill-based Teams of Experts. In: d'Aquin M, Dietze S, Hauff C, Curry E, Cudré-Mauroux P., eds. *CIKM '20: The 29th ACM International Conference on Information and Knowledge Management, Virtual Event, Ireland, October 19-23, 2020*ACM 2020:2049–2052
38. Rad RH, Fani H, Bagheri E, Kargar M, Srivastava D, Szlichta J. A Variational Neural Architecture for Skill-based Team Formation. *ACM Trans. Inf. Syst.*. 2024;42(1):7:1–7:28. doi: 10.1145/3589762
39. Rad RH, Bagheri E, Kargar M, Srivastava D, Szlichta J. Retrieving Skill-Based Teams from Collaboration Networks. In: ACM 2021:2015–2019
40. Dashti A, Samet S, Fani H. Effective Neural Team Formation via Negative Samples. In: Hasan MA, Xiong L., eds. *Proceedings of the 31st ACM International Conference on Information & Knowledge Management, Atlanta, GA, USA, October 17-21, 2022*ACM 2022:3908–3912
41. Blundell C, Cornebise J, Kavukcuoglu K, Wierstra D. Weight Uncertainty in Neural Network. In: Bach FR, Blei DM., eds. *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015*. 37 of *JMLR Workshop and Conference Proceedings*. JMLR.org 2015:1613–1622.
42. Chen J, Kallus N, Mao X, Svacha G, Udell M. Fairness Under Unawareness: Assessing Disparity When Protected Class Is Unobserved. In: boyd d, Morgenstern JH., eds. *Proceedings of the Conference on Fairness, Accountability, and Transparency, FAT* 2019, Atlanta, GA, USA, January 29-31, 2019*ACM 2019:339–348
43. Kenfack PJ, Kahou SE, Aïvodji U. A Survey on Fairness Without Demographics. *Trans. Mach. Learn. Res.*. 2024;2024.
44. Mikolov T, Sutskever I, Chen K, Corrado GS, Dean J. Distributed Representations of Words and Phrases and their Compositionality. In: Burges CJC, Bottou L, Ghahramani Z, Weinberger KQ., eds. *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States* 2013:3111–3119.
45. Le QV, Mikolov T. Distributed Representations of Sentences and Documents. In: . 32 of *JMLR Workshop and Conference Proceedings*. JMLR.org 2014:1188–1196.
46. Kaw S, Kobti Z, Selvarajah K. Transfer Learning with Graph Attention Networks for Team Recommendation. In: IEEE 2023:1–8
47. Loghmani H, Fani H. Bootless Application of Greedy Re-ranking Algorithms in Fair Neural Team Formation. In: Boratto L, Faralli S, Marras M, Stilo G., eds. *Advances in Bias and Fairness in Information Retrieval - 4th International Workshop, BIAS 2023, Dublin, Ireland, April 2, 2023, Revised Selected Papers*. 1840 of *Communications in Computer and Information Science*. Springer 2023:108–118
48. Tang Z, Zhang K. Attainability and Optimality: The Equalized Odds Fairness Revisited. In: . 177 of *Proceedings of Machine Learning Research*. PMLR 2022:754–786.

49. Kleinberg JM, Mullainathan S, Raghavan M. Inherent Trade-Offs in the Fair Determination of Risk Scores. In: . 67 of *LIPIcs*. Schloss Dagstuhl - Leibniz-Zentrum für Informatik 2017:43:1–43:23
50. Chouldechova A. Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments. *Big Data.* 2017;5(2):153–163. doi: 10.1089/BIG.2016.0047
51. Barocas S, Hardt M, Narayanan A. *Fairness and machine learning: Limitations and opportunities*. MIT press, 2023.
52. Machado L, Stefanidis K. Fair Team Recommendations for Multidisciplinary Projects. In: Barnaghi PM, Gottlob G, Manolopoulos Y, Tzouramanis T, Vakali A., eds. *2019 IEEE/WIC/ACM International Conference on Web Intelligence, WI 2019, Thessaloniki, Greece, October 14-17, 2019*ACM 2019:293–297
53. Barnabò G, Fazzone A, Leonardi S, Schwiegelshohn C. Algorithms for Fair Team Formation in Online Labour Marketplaces10033. In: ACM 2019:484–490
54. Moasses R, Rajaei D, Loghmani H, Saeedi M, Fani H. vivaFemme: Mitigating Gender Bias in Neural Team Recommendation via Female-Advocate Loss Regularization. In: Bellogín A, Boratto L, Kleanthous S, Lex E, Malloci FM, Marras M., eds. *Advances in Bias and Fairness in Information Retrieval - 5th International Workshop, BIAS 2024, Washington, DC, USA, July 18, 2024, Revised Selected Papers*. 2227 of *Communications in Computer and Information Science*. Springer 2024:78–90
55. Gómez-Zará D, Das A, Pawlow B, Contractor N. In search of diverse and connected teams: A computational approach to assemble diverse teams based on members' social networks. *PLOS ONE.* 2022;17(11):1-29. doi: 10.1371/journal.pone.0276061
56. Burke CS, Georganta E, Marlow S. A bottom up perspective to understanding the dynamics of team roles in mission critical teams. *Frontiers in Psychology.* 2019;10:441832.
57. Stokols D, Hall KL, Taylor BK, Moser RP. The science of team science: overview of the field and introduction to the supplement. *American journal of preventive medicine.* 2008;35(2):S77–S89.
58. Chen SG. An Integrated Methodological Framework for Project Task Coordination and Team Organization in Concurrent Engineering. *Concurr. Eng. Res. Appl.*. 2005;13(3):185–197. doi: 10.1177/1063293X05056462
59. Wang L, Zeng Y, Chen B, Pan Y, Cao L. Team Recommendation Using Order-Based Fuzzy Integral and NSGA-II in StarCraft. *IEEE Access.* 2020;8:59559–59570. doi: 10.1109/ACCESS.2020.2982647
60. Campêlo MB, Figueiredo TF, Silva A. The sociotechnical teams formation problem: a mathematical optimization approach. *Ann. Oper. Res.*. 2020;286(1):201–216. doi: 10.1007/s10479-018-2759-5
61. Esgario JGM, Silva dIE, Krohling RA. Application of Genetic Algorithms to the Multiple Team Formation Problem. *CoRR.* 2019;abs/1903.03523.
62. Kalayathankal SJ, Abraham JT, Kureethara JV. A Fuzzy Approach To Project Team Selection. *International Journal of Scientific & Technology Research.* 2019;8.
63. Rahman H, Roy SB, Thirumuruganathan S, Amer-Yahia S, Das G. Optimized group formation for solving collaborative tasks. *VLDB J.*. 2019;28(1):1–23. doi: 10.1007/s00778-018-0516-7
64. Durfee EH, Jr. JCB, Sleight J. Using hybrid scheduling for the semi-autonomous formation of expert teams. *Future Gener. Comput. Syst.*. 2014;31:200–212. doi: 10.1016/J.FUTURE.2013.04.008
65. Strnad D, Guid N. A fuzzy-genetic decision support system for project team formation. *Appl. Soft Comput.*. 2010;10(4):1178–1187. doi: 10.1016/j.asoc.2009.08.032
66. Wi H, Oh S, Mun J, Jung M. A team formation model based on knowledge and collaboration. *Expert Syst. Appl.*. 2009;36(5):9121–9134. doi: 10.1016/j.eswa.2008.12.031
67. Baykasoglu A, Dereli T, Das S. Project Team Selection Using Fuzzy Optimization Approach. *Cybern. Syst.*. 2007;38(2):155–185. doi: 10.1080/01969720601139041
68. Chen SG, Lin L. Modeling team member characteristics for the formation of a multifunctional team in concurrent engineering. *IEEE Trans. Engineering Management.* 2004;51(2):111–124. doi: 10.1109/TEM.2004.826011
69. Zakarian A, Kusiak A. Forming teams: An analytical approach. *IIE Transactions.* 1999;31:85-97. doi: 10.1023/A:1007580823003
70. Kargar M, An A. Discovering top-k teams of experts with/without a leader in social networks. In: Macdonald C, Ounis I, Ruthven I., eds. *Proceedings of the 20th ACM Conference on Information and Knowledge Management, CIKM 2011, Glasgow, United Kingdom, October 24-28, 2011*ACM 2011:985–994.
71. Sozio M, Gionis A. The community-search problem and how to plan a successful cocktail party. In: ACM 2010:939–948
72. Lappas T, Liu K, Terzi E. Finding a team of experts in social networks. In: IV JFE, Fogelman-Soulié F, Flach PA, Zaki MJ., eds. *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Paris, France, June 28 - July 1, 2009*ACM 2009:467–476
73. Gaston ME, Simmons J, desJardins M. Adapting Network Structure for Efficient Team Formation. In: . FS-04-02. AAAI Press 2004:1–8.
74. Barzegar R, Kurepa MN, Fani H. Adaptive Loss-based Curricula for Neural Team Recommendation. In: Nejdl W, Auer S, Cha M, Moens M, Najork M., eds. *Proceedings of the Eighteenth ACM International Conference on Web Search and Data Mining, WSDM 2025, Hannover, Germany, March 10-14, 2025*ACM 2025:914–923
75. Thang K, Hosseini H, Fani H. Translative Neural Team Recommendation: From Multilabel Classification to Sequence Prediction. In: Ferro N, Maistro M, Pasi G, Alonso O, Trotman A, Verberne S., eds. *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2025, Padua, Italy, July 13-18, 2025*ACM 2025:2800–2806
76. Dashti A, Saxena K, Patel D, Fani H. OpeNTF: A Benchmark Library for Neural Team Formation. In: ACM 2022:3913–3917
77. Rad RH, Seyedsalehi S, Kargar M, Zihayat M, Bagheri E. A Neural Approach to Forming Coherent Teams in Collaboration Networks. In: OpenProceedings.org 2022:2:440–2:444
78. Rad RH, Bagheri E, Kargar M, Srivastava D, Szlichta J. Subgraph Representation Learning for Team Mining. In: ACM 2022:148–153
79. Rad RH, Mitha A, Fani H, Kargar M, Szlichta J, Bagheri E. PyTFL: A Python-based Neural Team Formation Toolkit. In: ACM 2021:4716–4720
80. Sapienza A, Goyal P, Ferrara E. Deep Neural Networks for Optimal Team Composition. *Frontiers Big Data.* 2019;2:14. doi: 10.3389/fdata.2019.00014
81. Dara M, Rad RH, Zarrinkalam F, Bagheri E. Retrieval-Augmented Neural Team Formation. In: . 15574 of *Lecture Notes in Computer Science*. Springer 2025:362–371
82. Etemadi R, Zihayat M, Feng K, Adelman J, Zarrinkalam F, Bagheri E. It Takes a Team to Triumph: Collaborative Expert Finding in Community QA Networks. In: ACM 2024:164–174

83. Ahmed MJ, Saeedi M, Fani H. Vector Representation Learning of Skills for Collaborative Team Recommendation: A Comparative Study. In: Barhamgi M, Wang H, Wang X., eds. *Web Information Systems Engineering - WISE 2024 - 25th International Conference, Doha, Qatar, December 2-5, 2024, Proceedings, Part II*. 15437 of *Lecture Notes in Computer Science*. Springer 2024:193–207
84. Zuckerman EW, Jost JT. What Makes You Think You're so Popular? Self-Evaluation Maintenance and the Subjective Side of the "Friendship Paradox". *Social Psychology Quarterly.* 2001;64(3):207–223.
85. Stoyanovich J, Zehlike M, Yang K. Fairness in Ranking: From Values to Technical Choices and Back. In: Das S, Pandis I, Candan KS, Amer-Yahia S., eds. *Companion of the 2023 International Conference on Management of Data, SIGMOD/PODS 2023, Seattle, WA, USA, June 18-23, 2023*ACM 2023:7–12
86. Mehrabi N, Morstatter F, Saxena N, Lerman K, Galstyan A. A Survey on Bias and Fairness in Machine Learning. *ACM Comput. Surv.*. 2022;54(6):115:1–115:35. doi: 10.1145/3457607
87. Calders T, Kamiran F, Pechenizkiy M. Building Classifiers with Independency Constraints. In: Saygin Y, Yu JX, Kargupta H, et al., eds. *ICDM Workshops 2009, IEEE International Conference on Data Mining Workshops, Miami, Florida, USA, 6 December 2009*IEEE Computer Society 2009:13–18
88. Long CX, Hsu H, Alghamdi W, Calmon FP. Individual Arbitrariness and Group Fairness. In: Oh A, Naumann T, Globerson A, Saenko K, Hardt M, Levine S., eds. *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023* 2023.
89. Grant DG. Equalized odds is a requirement of algorithmic fairness. *Synthese.* 2023;201(3):101.
90. Trautmann ST. Procedural fairness and equality of opportunity. *Journal of Economic Surveys.* 2023;37(5):1697–1714.
91. Kou Y, Shen D, Snell Q, et al. Efficient Team Formation in Social Networks based on Constrained Pattern Graph. In: IEEE 2020:889–900
92. Kargar M, Golab L, Srivastava D, Szlichta J, Zihayat M. Effective Keyword Search Over Weighted Graphs. *IEEE Trans. Knowl. Data Eng.*. 2022;34(2):601–616. doi: 10.1109/TKDE.2020.2985376
93. Keane P, Ghaffar F, Malone D. Using machine learning to predict links and improve Steiner tree solutions to team formation problems - a cross company study. *Appl. Netw. Sci.*. 2020;5(1):57. doi: 10.1007/s41109-020-00306-x
94. Fani H, Barzegar R, Dashti A, Saeedi M. A Streaming Approach to Neural Team Formation Training. In: Goharian N, Tonellotto N, He Y, et al., eds. *Advances in Information Retrieval - 46th European Conference on Information Retrieval, ECIR 2024, Glasgow, UK, March 24-28, 2024, Proceedings, Part I*. 14608 of *Lecture Notes in Computer Science*. Springer 2024:325–340
95. Dong Y, Chawla NV, Swami A. metapath2vec: Scalable Representation Learning for Heterogeneous Networks. In: ACM 2017:135–144
96. Velickovic P, Fedus W, Hamilton WL, Liò P, Bengio Y, Hjelm RD. Deep Graph Infomax. In: OpenReview.net 2019.
97. Chen RJ, Wang JJ, Williamson DF, et al. Algorithmic fairness in artificial intelligence for medicine and healthcare. *Nature biomedical engineering.* 2023;7(6):719–742.
98. Grote T, Keeling G. Enabling Fairness in Healthcare Through Machine Learning. *Ethics Inf. Technol.*. 2022;24(3):39. doi: 10.1007/S10676-022-09658-7
99. Lee MK, Rich K. Who Is Included in Human Perceptions of AI?: Trust and Perceived Fairness around Healthcare AI and Cultural Mistrust. In: Kitamura Y, Quigley A, Isbister K, Igarashi T, Bjørn P, Drucker SM., eds. *CHI '21: CHI Conference on Human Factors in Computing Systems, Virtual Event / Yokohama, Japan, May 8-13, 2021*ACM 2021:138:1–138:14
100. Ahmad MA, Patel A, Eckert C, Kumar V, Teredesai A. Fairness in Machine Learning for Healthcare. In: Gupta R, Liu Y, Tang J, Prakash BA., eds. *KDD '20: The 26th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Virtual Event, CA, USA, August 23-27, 2020*ACM 2020:3529–3530
101. Bigdeli A, Arabzadeh N, Seyedsalehi S, Zihayat M, Bagheri E. Gender Fairness in Information Retrieval Systems. In: Amigó E, Castells P, Gonzalo J, Carterette B, Culpepper JS, Kazai G., eds. *SIGIR '22: The 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, Madrid, Spain, July 11 - 15, 2022*ACM 2022:3436–3439
102. Ekstrand MD, Burke R, Diaz F. Fairness and Discrimination in Retrieval and Recommendation. In: Piwowarski B, Chevalier M, Gaussier É, Maarek Y, Nie J, Scholer F., eds. *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2019, Paris, France, July 21-25, 2019*ACM 2019:1403–1404
103. Lv B, Liu F, Li Y, Nie J, Gou F, Wu J. Artificial Intelligence-Aided Diagnosis Solution by Enhancing the Edge Features of Medical Images. *Diagnostics.* 2023;13(6). doi: 10.3390/diagnostics13061063
104. Jalal A, Karmalkar S, Hoffmann J, Dimakis A, Price E. Fairness for Image Generation with Uncertain Sensitive Attributes. In: Meila M, Zhang T., eds. *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*. 139 of *Proceedings of Machine Learning Research*. PMLR 2021:4721–4732.
105. Yee K, Tantipongpipat U, Mishra S. Image Cropping on Twitter: Fairness Metrics, their Limitations, and the Importance of Representation, Design, and Agency. *Proc. ACM Hum. Comput. Interact.*. 2021;5(CSCW2):450:1–450:24. doi: 10.1145/3479594
106. Kyriakou K, Barlas P, Kleanthous S, Otterbacher J. Fairness in Proprietary Image Tagging Algorithms: A Cross-Platform Audit on People Images. In: Pfeffer J, Budak C, Lin Y, Morstatter F., eds. *Proceedings of the Thirteenth International Conference on Web and Social Media, ICWSM 2019, Munich, Germany, June 11-14, 2019*AAAI Press 2019:313–322.
107. Karako C, Manggala P. Using Image Fairness Representations in Diversity-Based Re-ranking for Recommendations. In: Mitrovic T, Zhang J, Chen L, Chin D., eds. *Adjunct Publication of the 26th Conference on User Modeling, Adaptation and Personalization, UMAP 2018, Singapore, July 08-11, 2018*ACM 2018:23–28
108. Nandy P, DiCiccio C, Venugopalan D, Logan H, Basu K, Karoui NE. Achieving Fairness via Post-Processing in Web-Scale Recommender Systems. In: ACM 2022:715–725
109. Saito Y, Joachims T. Fair Ranking as Fair Division: Impact-Based Individual Fairness in Ranking. In: Zhang A, Rangwala H., eds. *KDD '22: The 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Washington, DC, USA, August 14 - 18, 2022*ACM 2022:1514–1524
110. Singh A, Joachims T. Fairness of Exposure in Rankings. In: Guo Y, Farooq F., eds. *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, KDD 2018, London, UK, August 19-23, 2018*ACM 2018:2219–2228
111. Zehlike M, Bonchi F, Castillo C, Hajian S, Megahed M, Baeza-Yates R. FA*IR: A Fair Top-k Ranking Algorithm. In: CIKM '17. Association for Computing Machinery 2017; New York, NY, USA:1569–1578
112. Yang K, Stoyanovich J. Measuring Fairness in Ranked Outputs. In: ACM 2017:22:1–22:6
113. Barocas S, Hardt M, Narayanan A. *Fairness and Machine Learning: Limitations and Opportunities*. MIT Press, 2023.

114. Juárez J, Brizuela CA. A multi-objective formulation of the team formation problem in social networks: preliminary results. In: Aguirre HE, Takadama K., eds. *GECCO*ACM 2018:261–268
115. Hayano M, Hamada D, Sugawara T. Role and member selection in team formation using resource estimation for large-scale multi-agent systems. *Neurocomputing.* 2014;146:164–172. doi: 10.1016/j.neucom.2014.04.059
116. Majumder A, Datta S, Naidu KVM. Capacitated team formation problem on social networks. In: Yang Q, Agarwal D, Pei J., eds. *The 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '12, Beijing, China, August 12-16, 2012*ACM 2012:1005–1013
117. Li C, Shan M. Team Formation for Generalized Tasks in Expertise Social Networks. In: Elmagarmid AK, Agrawal D., eds. *Proceedings of the 2010 IEEE Second International Conference on Social Computing, SocialCom / IEEE International Conference on Privacy, Security, Risk and Trust, PASSAT 2010, Minneapolis, Minnesota, USA, August 20-22, 2010*IEEE Computer Society 2010:9–16
118. Ani ZC, Yasin A, Husin MZ, Hamid ZA. A method for group formation using genetic algorithm. *International Journal on Computer Science and Engineering.* 2010;2(9):3060–3064.
119. Fitzpatrick E, Askin RG. Forming effective worker teams with multi-functional skill requirements. *Comput. Ind. Eng.*. 2005;48(3):593–608. doi: 10.1016/j.cie.2004.12.014
120. Keswani V, Celis LE. Algorithmic Fairness From the Perspective of Legal Anti-discrimination Principles. In: AAAI Press 2024:724–737
121. Bask M. Skill, status and the Matthew effect: a theoretical framework. *J. Comput. Soc. Sci.*. 2024;7(3):2221–2253. doi: 10.1007/S42001-024-00298-Z
122. Rapanos T. What makes an opinion leader: Expertise vs popularity. *Games Econ. Behav.*. 2023;138:355–372. doi: 10.1016/J.GEB.2023.01.003
123. Morgan AC, Economou DJ, Way SF, Clauset A. Prestige drives epistemic inequality in the diffusion of scientific ideas. *EPJ Data Sci.*. 2018;7(1):40. doi: 10.1140/EPJDS/S13688-018-0166-4
124. Karimi S, Alizadeh Noughabi H, Zarrinkalam F, Zihayat M. Who Gets to be an Expert? The Hidden Bias in Expert Finding. In: SIGIR-AP 2025. Association for Computing Machinery 2025; New York, NY, USA:111–115
125. Gomez CJ, Herman AC, Parigi P. Leading countries in global science increasingly receive more citations than other countries doing similar research. *Nature Human Behaviour.* 2022;6(7):919–929.
126. Pan RK, Kaski K, Fortunato S. World citation and collaboration networks: uncovering the role of geography in science. *Scientific reports.* 2012;2(1):902.
127. Perc M. The Matthew effect in empirical data. *Journal of The Royal Society Interface.* 2014;11(98):20140378. doi: 10.1098/rsif.2014.0378
128. Huang Y, Liu X, Li R, Zhang L. The science of team science (SciTS): an emerging and evolving field of interdisciplinary collaboration. *Profesional de la información.* 2023;32(2).
129. Love HB, Fosdick BK, Cross JE, et al. Towards understanding the characteristics of successful and unsuccessful collaborations: a case-based team science study. *Humanities and Social Sciences Communications.* 2022;9(1):1–11.
130. Stokols D, Hall KL, Taylor BK, Moser RP. The Science of Team Science: Overview of the Field and Introduction to the Supplement. *American Journal of Preventive Medicine.* 2008;35(2, Supplement):S77-S89. The Science of Team Sciencedoi: https://doi.org/10.1016/j.amepre.2008.05.002
131. Swaab RI, Schaerer M, Anicich EM, Ronay R, Galinsky AD. The too-much-talent effect: Team interdependence determines when more talent is too much or not enough. *Psychological Science.* 2014;25(8):1581–1591.
132. Zhu J, Zhou J, Pan J, Gu F, Guo J. Ranking Influential Non-Content Factors on Scientific Papers' Citation Impact: A Multidomain Comparative Analysis. *Big Data and Cognitive Computing.* 2025;9(2):30.
133. Alves JV, Leitão D, Jesus SM, et al. Cost-Sensitive Learning to Defer to Multiple Experts with Workload Constraints. *Trans. Mach. Learn. Res.*. 2024;2024.
134. McDonald DW, Ackerman MS. Expertise recommender: a flexible recommendation system and architecture. In: ACM 2000:231–240
135. Zou K, Sun A. A Survey of Real-World Recommender Systems: Challenges, Constraints, and Industrial Perspectives. *CoRR.* 2025;abs/2509.06002. doi: 10.48550/ARXIV.2509.06002
136. Klimashevskaia A, Jannach D, Elahi M, Trattner C. A survey on popularity bias in recommender systems. *User Model. User Adapt. Interact.*. 2024;34(5):1777–1834. doi: 10.1007/S11257-024-09406-0
137. Musto C, Lops P, Semeraro G. Fairness and Popularity Bias in Recommender Systems: an Empirical Evaluation. In: . 3078 of *CEUR Workshop Proceedings*. CEUR-WS.org 2021:77–91.
138. Färber M, Coutinho M, Yuan S. Biases in scholarly recommender systems: impact, prevalence, and mitigation. *Scientometrics.* 2023;128(5):2703–2736. doi: 10.1007/S11192-023-04636-2
139. Nikzad-Khasmakhi N, Balafar MA, Feizi-Derakhshi M. The state-of-the-art in expert recommendation systems. *Eng. Appl. Artif. Intell.*. 2019;82:126–147. doi: 10.1016/J.ENGAPPAI.2019.03.020
140. Ramanath R, Inan H, Polatkan G, et al. Towards Deep and Representation Learning for Talent Search at LinkedIn. In: ACM 2018:2253–2261
141. Fabbri F, Croci ML, Bonchi F, Castillo C. Exposure Inequality in People Recommender Systems: The Long-Term Effects. In: AAAI Press 2022:194–204.
142. Abdollahpouri H, Mansoury M, Burke R, Mobasher B, Malthouse EC. User-centered Evaluation of Popularity Bias in Recommender Systems. In: Masthoff J, Herder E, Tintarev N, Tkalcic M., eds. *Proceedings of the 29th ACM Conference on User Modeling, Adaptation and Personalization, UMAP 2021, Utrecht, The Netherlands, June, 21-25, 2021*ACM 2021:119–129
143. Elahi M, Kholgh DK, Kiarostami MS, Saghari S, Rad SP, Tkalcic M. Investigating the impact of recommender systems on user-based and item-based popularity bias. *Inf. Process. Manag.*. 2021;58(5):102655. doi: 10.1016/J.IPM.2021.102655
144. Genderize . Genderize API. https://genderize.io/; 2023. Accessed: 16-June-2023.
145. Alkhathlan M, Cachel K, Shrestha H, Harrison L, Rundensteiner EA. Balancing Act: Evaluating People's Perceptions of Fair Ranking Metrics. In: ACM 2024:1940–1970
146. Zhu Z, Wang J, Caverlee J. Measuring and Mitigating Item Under-Recommendation Bias in Personalized Ranking Systems. In: Huang JX, Chang Y, Cheng X, et al., eds. *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval, SIGIR 2020, Virtual Event, China, July 25-30, 2020*ACM 2020:449–458
147. Cherumanal SP, Spina D, Scholer F, Croft WB. Evaluating Fairness in Argument Retrieval. In: Demartini G, Zuccon G, Culpepper JS, Huang Z, Tong H., eds. *CIKM '21: The 30th ACM International Conference on Information and Knowledge Management, Virtual Event, Queensland, Australia, November 1 - 5, 2021*ACM 2021:3363–3367

148. Ghosh A, Dutt R, Wilson C. When Fair Ranking Meets Uncertain Inference. In: Diaz F, Shah C, Suel T, Castells P, Jones R, Sakai T., eds. *SIGIR '21: The 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, Virtual Event, Canada, July 11-15, 2021*ACM 2021:1033–1043
149. Seth A, Hemani M, Agarwal C. DeAR: Debiasing Vision-Language Models with Additive Residuals. In: IEEE 2023:6820–6829
150. Morik M, Singh A, Hong J, Joachims T. Controlling Fairness and Bias in Dynamic Learning-to-Rank. In: Huang JX, Chang Y, Cheng X, et al., eds. *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval, SIGIR 2020, Virtual Event, China, July 25-30, 2020*ACM 2020:429–438
151. Diaz F, Mitra B, Ekstrand MD, Biega AJ, Carterette B. Evaluating Stochastic Rankings with Expected Exposure. In: d'Aquin M, Dietze S, Hauff C, Curry E, Cudré-Mauroux P., eds. *CIKM '20: The 29th ACM International Conference on Information and Knowledge Management, Virtual Event, Ireland, October 19-23, 2020*ACM 2020:275–284
152. Patro GK, Biswas A, Ganguly N, Gummadi KP, Chakraborty A. FairRec: Two-Sided Fairness for Personalized Recommendations in Two-Sided Platforms. In: Huang Y, King I, Liu T, Steen vM., eds. *WWW '20: The Web Conference 2020, Taipei, Taiwan, April 20-24, 2020*ACM / IW3C2 2020:1194–1204
153. Qi T, Wu F, Wu C, et al. ProFairRec: Provider Fairness-aware News Recommendation. In: Amigó E, Castells P, Gonzalo J, Carterette B, Culpepper JS, Kazai G., eds. *SIGIR '22: The 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, Madrid, Spain, July 11 - 15, 2022*ACM 2022:1164–1173
154. Mansoury M, Abdollahpouri H, Pechenizkiy M, Mobasher B, Burke R. A Graph-Based Approach for Mitigating Multi-Sided Exposure Bias in Recommender Systems. *ACM Trans. Inf. Syst.*. 2022;40(2):32:1–32:31. doi: 10.1145/3470948
155. Balagopalan A, Jacobs AZ, Biega AJ. The Role of Relevance in Fair Ranking. In: Chen H, Duh WE, Huang H, Kato MP, Mothe J, Poblete B., eds. *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2023, Taipei, Taiwan, July 23-27, 2023*ACM 2023:2650–2660
156. Heuss M, Sarvi F, Rijke dM. Fairness of Exposure in Light of Incomplete Exposure Estimation. In: Amigó E, Castells P, Gonzalo J, Carterette B, Culpepper JS, Kazai G., eds. *SIGIR '22: The 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, Madrid, Spain, July 11 - 15, 2022*ACM 2022:759–769
157. Wu H, Mitra B, Ma C, Diaz F, Liu X. Joint Multisided Exposure Fairness for Recommendation. In: Amigó E, Castells P, Gonzalo J, Carterette B, Culpepper JS, Kazai G., eds. *SIGIR '22: The 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, Madrid, Spain, July 11 - 15, 2022*ACM 2022:703–714