Thank you for your reviews! Your comments will be useful in revising our paper. We appreciate the positive comments, e.g., that our work addresses an "under-studied"/"timely" problem "of extreme importance," the method is "nice"/"simple"/"clean"/" well validated," the paper is "(exceptionally) clear"/"well written," results are "solid"/"significant"/"clear," with "significant performance gains."

We briefly summarize the different *major* concerns of the 3 reviewers before we respond to them separately. We note that **there were no shared weaknesses pointed out by the 3 reviewers.** The main concern of @R1 (overall score=4) is about the novelty and contribution of the paper, especially in the context of a specific recent line of work by Foulds et al. The reviewer also viewed focusing on assessment of fairness instead of learning fair models as a weakness. @R2 (overall score=7) characterizes our work as being a useful approach to quantifying and reducing uncertainty in fairness metrics, with broad applicability and significant performance gains, and "does not think the weaknesses of our work is consequential enough to prevent publication". @R4 (overall score=4) identified some potential weaknesses (prior sensitivity, desirability of more theoretical results) but did not identify major issues with the paper.

@R1: "in light of Foulds et al 2019, the contribution is rather minor" Thank you for pointing out this paper. We will certainly cite and discuss it in our revised paper since it shares a common starting point with our work of using Bayesian methods to assess fairness. *However, we disagree that our contribution is minor in light of this paper. The two papers complement each other: they address different problems and take different technical approaches*. We believe our work is substantially different from [1] in objective, methods and results and we ask Reviewer 1 to reconsider our paper in this light. **Objective:**, our work specifically focuses on the problem of *leveraging unlabeled data to generate better estimates of fairness metrics* given limited labeled data; in contrast, [1] focuses on assessing *intersectional fairness* when the amount of labeled data is extremely small and unlabeled data is unavailable. We also make a point in our work of emphasizing that estimates of fairness metrics can suffer from high variance even in the presence of relatively large amounts of labeled data (see Fig 1). **Methods**: We use Bayesian methods to calibrate scores for unlabeled datapoints for improved estimation of group fairness metrics, whereas [1] uses Bayesian methods to provide parametric smoothing among groups for improved estimation of intersectional fairness metrics based on labeled data. **Results**: We evaluate our methods across a broad range of different datasets and different models (mechanisms) and demonstrate substantial improvements in accuracy of estimates of multiple group fairness metrics, whereas [1]'s results are focused mainly about the accuracy ratio between intersectional groups on two (semi-synthetic) datasets.

@R1: "focuses only on estimation and not on how to obtain fair policies". We agree that obtaining fair policies is an important problem, but we also believe that for a given model, trained fairly or not, *independent and accurate assessment of its fairness is important and under-studied* (also the motivation in [1]), particularly when users only have access to a blackbox predictor. We will make this point clearer in revision. **@R1: "seems limited to fairness settings where Bayesian calibration is applicable"**: We emphasize that our approach is applicable to any classification setting and we **do not need** any special setup to apply our method. The only requirement is that there is both labeled and unlabeled data available from the deployment environment. @R1: notation in l.105-107: We agree this notation is confusing and will remove this equation (its not needed). **@R1: "discuss how you would extend this idea to other fairness metrics."** Good point, we agree. We can directly extend our approach to handle metrics such as calibration and balance as well as ratio-based metrics and we will this discussion of such extensions to the paper. All fairness metrics which are defined as deterministic functions of model score S, label Y and sensitive attribute A (for concreteness we demonstrated with 3 popular fairness metrics in the paper) can be approximated on unlabeled data with our proposed method.

@R2: "the possibility that the CIs might be overconfident if there are many labeled examples." Good point. On page 13 of Appendix, we empirically validated that our method provides reasonably well-calibrated CIs, but we believe there is room for further improvement in this area. **@R2: "the challenge of balancing the bias-variance tradeoff for this method."** We agree that this is an interesting direction for future work. We provide some theoretical considerations in lines 165-177 and also acknowledge this issue in a brief discussion in lines 242-250: but there is certainly room for more work on this front. **@R2:Eskimo -> Inuit**: Thank you for spotting this! We will update it in the paper.

@R4: "proper sensitivity analysis": We agree that systematic sensitivity analysis is lacking and we will add it to the Supplement. The values for the priors were selected based on consideration on knowledge of ranges of miscalibration one typically sees with trained classifiers—and we also found that the same priors worked well across a large range of datasets and models without any need for tuning (see also lines 153-156). **@R4: "non-hierarchical variance**: Thanks for suggesting the ablation study, we will add it to the Supplement. **@R4: Equation 1**: The phrase "For example" was loosely worded and we will remove it in the revision to avoid potential confusion. **@R4: "more theoretical results to establish the relevance"**: We agree that more theoretical results are important going forward—please see our response to R2 on this point. **@R4: "posterior computations contain no novelties or contributions"**: We use standard MCMC for posterior computation and it works well in our experiments in terms of both accuracy and runtime.