We thank the reviewers for their effort and very constructive feedback on our work. We did our best to address their suggestions in order to improve our paper. We provide our replies to each of the reviewers' comments. In case when the same concerns are raised by different reviewers, we repeat the same or a similar reply to each reviewer to make reading easier to them (as opposed to referring them to the same reply to another reviewer).

Dear Authors: This paper could be of interest for this journal, but the authors are requested to address and implement point-by-point all concerns of the expert reviewers. Please see below all comments:

Reviewer #1: Paper surely deals with and interesting topic but paper need to address several problems

1. State of the art is largely incomplete, several key works in the literature are not even mentioned

Reply: It would be of great help if the reviewer would refer to specific important work we have omitted. We will gratefully pay due attention to it. Now we have added three relevant recent methods (min-max based ones, like ours is) to the experimental comparison and commented on these methods in the related work section. We hope this addition contributed to the quality of the work.

1. Novelty of the proposal should be better underlined in comparison with the state of the art

Reply: Our introduction already contains a bulleted list meant to emphasize the novelty of our approach. Now we tried to better emphasize the differences to existing groups of methods in the related work. Also, we added more related methods to the related work.

1. Better intuitions behind the methods and why is should work should be inserted

Reply: We agree that this aspect of our work could have been better. We think most of the information was present in the paper, but not presented as well as it could have been. At the beginning of the section “Fair Adversarial Instance Re-weighting – FAIR” we provide some intuition about the method and why it should work. Namely, we mean the following sentences:

“Unfairness which AI models learn is introduced through data instances containing unfair decisions. Therefore, we strive to recognize if a particular instance in a dataset is unfair. The main principle of FAIR is to reweight log likelihood of each instance, according to the trade-off between fairness and prediction performance, in order to obtain a fair and useful predictor of the target variable.”

“In order to incorporate the fairness objective, FAIR weights log likelihood of instances, so that the ones that are strongly informative of the sensitive features, but not of the target variable are assigned low weights and the ones that are informative of the target variable, but not of the sensitive attributes are assigned high weights.”

However, there is more information in the theoretical analysis section which we now reorganized in order to better convey the relevant info. Namely, we have tried to improve the presentation of that content by i) emphasizing the relevant elements near the beginning of the “Theoretical analysis of model properties” section, ii) emphasizing the role of the hyperparameter alpha not as the most important variable of the analysis, but as of the variable which sets these elements in motion and allows us to “observe” their interaction, and iii) rearranging some of the the content into a separate subsection “Intuitive discussion of theoretical properties” with different ordering of the thoughts and some additional remarks which are meant to emphasize relevant points.

Also, the proposal of probabilistic variants is motivated by the guaranteed existence of mixed Nash equilibrium in the probabilistic context (while the existence of the pure Nash equilibrium is not guaranteed in the scalar context) as we state at the beginning of the subsection 3.2. Just in case that this issue is also related to the reviewer’s comment, we emphasize that we added a short discussion on the differences of the variants of FAIR method in the subsection “Discussion and qualitative study”.

1. Experimental results is missing a strong baseline

Reply: So far we used 6 baselines for comparison in our paper and we consider several of them to be results of rather serious research. Of course, we were happy to make our experimental comparison even stronger, so we added 3 methods to the comparison and one of these methods comes in two variants. Hence, we added 4 baselines to our experimental comparison, resulting in 10 baselines in total. For more relevant comparison, as additional methods we choose min-max based ones, like ours is. Our conclusions remained the same as before – that FAIR is strongly competitive (roughly equal or better) to other methods. However, we would like to underline that we consider somewhat better performance of one method over the other an important, but not crucial trait. Therefore, instead of promoting our work as significantly better than the others, we are happy to say it is at least as good as others, but that it offers an additional feature. Namely, the interpretable model structure as reflected through the interpretable instance weighting, which we managed to incorporate into an end-to-end optimisation problem, unlike earlier methods which rely on instance weighting. We already stress that in the paper.

Reviewer #2: The overall goal of the paper is interesting: to combine reweighing and adversarial approaches to fair classification. The proposed solution seems novel but also very surprising in its simplicity. The authors suggest training three independent neural networks (one to predict the label, one to predict the sensitive attribute, and one to pick training instance weights) simultaneously. The common loss function will prompt the instance weighting network to put large weights on samples whose labels can be correctly predicted but whose sensitive attributes cannot, and small weights on other samples. Training the label network on these reweighted instances is supposedly enough to achieve a fair and accurate classifier (both of the other two networks are discarded after training). While it seems intuitively plausible that doing this might lead to some notion of "fairness" it is very difficult to see how it will enforce the formal definitions of fairness used in the field. After all, none of them are directly optimized for.

Reply: The reviewer correctly notices that we do not directly optimize some specific fairness metric. However, this is not an unusual approach to fairness. We even argue that it is a common one. Namely, several adversarial methods we refer to in the related work (e.g., FAD) also work in a similar way – they learn a representation which is uninformative of a sensitive feature, without explicitly relying on some fairness metric. Even methods which explicitly refer to some fairness metric in their construction perform some smooth relaxation of the metric, thus diverging from the declared objective. Therefore, we consider this a commonplace in construction of fairness related methods.

The promised "theoretical analysis of properties of adversarial re-weighting" offers no light on this point. Instead, it is simply a (mathematical) explanation of how varying the hyperparameter will change the threshold at which the weighting network will be indifferent between assigning a weight of 0 or 1 to a datapoint. While this may be a useful clarification, it gives no information as to the model's performance in terms of accuracy or fairness and probably shouldn't be included as a headline contribution.

Reply: We agree that our presentation of our theoretical results deserves the reviewer’s criticism. Indeed, we have put it in a way which suggests that the absolute focus of the analysis is on the hyperparameter alpha. However, we believe that more useful information is contained in the analysis. Namely, it is not only about the hyperparameter alpha, but also about the interaction of several relevant elements of the model which directly relate to accuracy and fairness. Therefore, we have tried to improve the presentation of that content by i) emphasizing the relevant elements near the beginning of the “Theoretical analysis of model properties” section, ii) emphasizing the role of the hyperparameter alpha not as the most important variable of the analysis, but as of the variable which sets these elements in motion and allows us to “observe” their interaction, and iii) rearranging some of the the content into a separate subsection “Intuitive discussion of theoretical properties” with different ordering of the thoughts and some additional remarks which are meant to emphasize relevant points.

Similarly, the exploration of "several variants of instance weight estimation" is disappointing. The authors provide only a quick definition of the four suggested variants and make no efforts to discuss their respective tradeoffs or how to pick which to use.

Reply: We agree that this aspect of our work merits more attention. Main motivation for introducing probabilistic variants is the fact that pure Nash equilibrium does not always exist and that therefore the convergence is not guaranteed, while the probabilistic approach guarantees the existence of the so-called mixed Nash equilibrium. We stated this at the beginning of subsection 3.2 and now we added a reference to this issue in that part of the text. Since we had no clear prior preference for the form of the probabilistic model we tried two - Bernoulli and beta which are already developed in literature in other fields. Main advantage of the FAIR-scalar approach is its simplicity. While the plots in our results section suggest that FAIR-beta approach is best performing, FAIR-scalar also provides good performance. Also, in our experiments we did not observe relevant empirical differences in convergence properties between FAIR-beta and FAIR-scalar, so given its simplicity FAIR-scalar and that the differences in performance are not great, FAIR-scalar could be considered a preferred variant. To clarify this in the paper, we have added this kind of discussion in subsection “Discussion and qualitative study”.

This puts a lot of importance on the results section since, otherwise, all the authors have done is propose a very simple optimization and given trivial theoretical analysis of the hyperparameter. Taken at face value, the results seem excellent, often beating other state of the art methods. However, there are multiple oddities and unclarities that cast doubt on these results:

1. The main results are pareto curves of AUC\_y (as a measure of performance) vs various measures of fairness. This doesn't make any sense. AUC is calculated before a threshold (and therefore a label) is picked, but the various (un)fairness measures depend on the label and therefore on the specific threshold picked. Why aren't the authors comparing simple accuracy to the (un)fairness measure (with a simple 0.5 threshold)?

Reply: We accept the reviewer’s criticism. We reworked our experimental evaluation in accordance with this suggestion and now we use accuracy instead of AUC. However, the overall picture and the conclusions remained the same.

2. The different models/benchmarks use wildly different architectures. Looking at Table B.7 makes it seem that FAD's prediction network is consistently given drastically fewer layers than the FAIR equivalents (18 vs 62 in Adult, 58 vs 464 in Readmission, etc). There may well be a good reason for this but it isn't really discussed. Perhaps equalizing some common "cost" (like overall training cpu time) would make the comparaison more fair?

Reply: We understand and acknowledge the reviewer’s concern. Still, we would like to point to the fact that in the case of FAD architecture the component networks are stacked, leading to greater depth of the complete network. In our case, the component networks are not stacked (they communicate through the loss only), so the depth is not accumulated. Therefore the impact of the difference in the number of layers need not be as big as it initially might seem. Also, in the first version of the paper we already tuned the architectures in order to obtain best results for each method (and we mention that in the Models paragraph of the Experimental setup section). However, in this revision we paid more attention to this issue and we tuned all architectures again using a wider range of hyperparameters, paying special attention to the number of layers (and thus investing a few weeks in repeating all the experiments). Indeed, there were considerable changes in the best architectures and the gap in numbers of layers has decreased. However, the overall picture of the results did not change substantially. Anyway, we updated the tables with new architecture information.

3. The current presentation is hard to read (6 pages of sparse graphs and dense tables) and not very clear. Comparing pareto curves for different model families is a good idea, but constructing these by picking a few arbitrary hyperparameter values for each model family doesn't seem like a good solution. Perhaps instead one could pick a few unfairness constraints, train models from each family satisfying those constraints (by picking the correct hyperparameter) and compare the accuracy of these results.

Reply: The reviewer’s idea makes perfect sense, and we already considered it when we planned our experimental evaluation. The good side is that if implemented correctly it would lead to succinct comparison of model accuracies at the same fairness level. However, tuning all models to the same fairness level is a very nontrivial task since the hyperparameters of different methods are not measured at the same scale and do not have the same meaning. Hence, it was not clear to us that in the end such an attempt would be successful, but it might require considerable computation. Therefore we opted for the Pareto front based comparison. While it is true that the models which constitute different families will not have the same fairness or accuracy values, if the hyperparameter range is reasonable, an imperfect but useful approximation of the Pareto curve is obtained. We believe that the choice we made is practically more feasible and that, although it might be imperfect, it allows for a meaningful comparison. We believe this make good sense in the context in which we do not aim to demonstrate that our method is definitely superior to all alternatives in terms of performance (the feasibility of such a task is doubtful to us regardless of the method considered), and instead we are happy to demonstrate that it is roughly at least as good as others, but that it offers an additional feature. Namely, the interpretable model structure as reflected through the interpretable instance weighting, which we managed to incorporate into the end-to-end optimisation problem, unlike earlier methods which rely on instance weighting.

Reviewer #4: This paper proposes Fair Adversarial Instance Re-weighting. In general, this paper is interesting. The method is new and valid based on the results. However, the following issues should be addressed.

1. The literature review should be enhanced with more related works to introduce the background.

Reply: We have added three relevant recent methods (min-max based ones, like ours is) to the experimental comparison and commented on these methods in the related work section. We hope this addition contributed to the quality of the work. If the reviewer has specific papers he/she thinks that our work is lacking, we will gratefully pay due attention to them.

2. A flow chart can be helpful for readers to better understand the method.

Reply: We refer the reviewer to two flow charts in Figure 1. If these are insufficient, we kindly ask the reviewer to suggest which aspects are lacking. We will gladly expand or improve our flow charts based on such advice.

3. The experiments seem a little bit weak. I suggest to introduce a more challenging case to fully examine the proposed method, as well as ablation studies to examine the effects of the key parameters within the reweighting mechanism.

Reply: So far we used 6 baselines for comparison in our paper and we consider several of them to be results of rather serious research. Of course, we were happy to make our experimental comparison even stronger, so we added 3 methods to the comparison and one of these methods comes in two variants. Hence, we added 4 baselines to our experimental comparison, resulting in 10 baselines in total. For more relevant comparison, as additional methods we choose min-max based ones, like ours is. Our conclusions remained the same as before – that FAIR is strongly competitive (roughly equal or better) to other methods. However, we would like to underline that we consider somewhat better performance of one method over the other an important, but not crucial trait. Therefore, instead of promoting our work as significantly better than the others, we are happy to say it is at least as good as others, but that it offers an additional feature. Namely, the interpretable model structure as reflected through the interpretable instance weighting, which we managed to incorporate into an end-to-end optimisation problem, unlike earlier methods which rely on instance weighting. We already stress that in the paper.

Regarding ablation studies, each part of the model (feature network, classifier network, and sensitive network) is essential to the method and cannot be removed in order to assess its individual contribution to the system. Also, since we do not combine any additional tricks over the basic architecture, we think there is no opportunity for an ablation study. We think that the closest analysis to what the reviewer requests is actually provided in section 5.2 of the paper in which we observe the behaviour of the system when we vary the hyperparameter alpha and provide a qualitative study.