Experiments and Baseline Reviewer 1 is concerned that our paper is incomplete without experiments. We respectfully disagree, and would like to make several points:

1. Our paper asks and answers a basic theoretical question: is satisfying statistical fairness constraints at every round compatible *even in principle* with obtaining the optimal $\sqrt{T}$ rate-of-regret in one-sided feedback settings? Prior to our work, this question had not been asked, and there was some reason for pessimism: Joseph et al. [20] had shown that stronger individual notions of fairness were not necessarily compatible with optimal regret bounds.

2. One reason experiments are useful is to establish if a method outperforms previously studied algorithms for the same problem. But as we are the first to study the problem of satisfying statistical fairness constraints in partial-information online classification settings, there is no relevant prior work to compare to. We note that prior work on fair batch classification does suggest one basic explore-then-exploit style baseline: during an "exploration round" classify all examples as positive, and then run a fair batch classification algorithm to pick a hypothesis to play for all subsequent rounds. Its not hard to see that this baseline, when optimized, obtains only a sub-optimal $T^{2/3}$ regret bound, as compared to our optimal $\sqrt{T}$ regret bound. We are happy to elaborate on this simple baseline in the paper.

3. Another reason for experimental results is to demonstrate that the proposed algorithm is implementable, and not simply a theoretical "upper bound". But as we note in the paper, our algorithm involves a careful combination of two works: "A Reductions Approach to Fair Classification" from ICML 2018 and "Taming the Monster" from ICML 2014. We need to make modifications to the details of these algorithms to make them work in our setting, but not in any way that would affect their running time. Both algorithms are known to run efficiently in practice, with code available.

In summary, although we agree with the reviewer that experimental work could be interesting, we disagree that it would be a major contribution of this work, or that the paper is incomplete without it. The main results of the paper cannot be established experimentally, and there is no prior work to make empirical comparisons to.

Clarity and Discussion Thank you for your useful suggestions about clarity: we agree, and will try and follow your advice. In particular, we will attempt to re-work Section 3 to focus less on technical details and more on giving intuition for where and why the constraints of fairness and partial feedback require modification of existing techniques, deferring more technical details to the supplement. Briefly, in addition to converting our partial feedback setting to the more standard contextual bandits setting using the reduction from Section 2 of our paper, we face three high level obstacles: 1) We need to modify the "fair reductions" oracle to be able to handle fairness constraints arising from a *different data set* than the error objective. 2) We need to modify the "mini-monster" contextual bandit algorithm to work with an approximate oracle, rather than an exact one (because we cannot implement an exact oracle using step 1). 3) We need to modify the analysis of "mini-monster" to be able to handle our infinite hypothesis class (because our fairness constraints force our hypothesis class to consist of all convex combinations of base classifiers, even if our base class is finite). Each of these steps requires some technical work, but we agree that focusing on the basic outline will be clearer.

We will also expand discussion at your suggestion about the distinction between three fairness constraints that one could ask for, at increasing levels of strength: 1) Satisfying the equal false-positive-rate condition only asymptotically (i.e. on average, in hindsight after the completion of the algorithm). 2) Satisfying the equal false-positive-rate condition at every round (what we do in this paper). 3) Satisfying the equal false-positive-rate condition at an individual level (i.e. over only the randomness of the classifier and not the population).

Prior work had only considered condition 3 (Joseph et al. [20]), and non-trivially satisfying this condition requires realizability assumptions. Our paper asks whether we can achieve optimal regret bounds without any assumptions while still satisfying 2. In particular, this means we are also the first paper to satisfy condition 1. Since we already obtain the optimal $\sqrt{T}$ regret bound satisfying the stronger condition 2, no asymptotic improvement would be possible if we relaxed to condition 1. Condition 1 would also be meaningfully weaker, since it would offer no guarantees at any particular time step: it would allow, for example, strong discrimination at every round, so long as the direction of that discrimination varied with time.

Reviewer 2 asks if experiments would help establish if our method "ends up being fairer than some other standard method." We think the confusion here comes from the fact that we used asymptotic notation to state both the fairness and regret guarantees for our algorithm. But our algorithm satisfies a hard fairness guarantee. We will clarify this, and state our theorem using only asymptotic notation for the regret guarantee. (We also re-iterate that there are no existing standard methods for the setting we consider).

Finally, sources of unfairness: This is a good discussion to add, thank you. As with all papers that take false positive rates seriously, we implicitly assume that the labels we obtain are either correct, or have symmetric ("unbiased") errors. The sources of unfairness that we deal with relate to the differential ability of models in our class to fit both populations (which we inherit from the batch setting), and the biased data collection inherent in online partial information settings. We use EO only as a canonical example of a statistical fairness constraint, and don't take the position that it is always the right one. Our techniques also apply to other constraints like statistical parity.