Reviewer 1 Re: suggested improvements: Section 2: Thank you for the reference and note. We will add the citation and discussion. That is *exactly* why we provide sensitivity analysis alongside the monotonicity assumption; in this context, it attends to “Moral 2” of Dawid (2000) by changing the goal to set- rather than point-identification, under varying strengths of additional assumptions, which may be plausible in practice.

Re: ignorability: we actually only need weak ignorability; thanks for catching. Section 3: Thank you for the references; we will add to the related work section on partial identification, alongside the Balke and Pearl ref. To clarify, the main point of departure from previous work is in addressing non-identifiability when **conditioning on the counterfactual potential outcome** and in **providing bounds for non-linear functionals**. Re: dependencies: We will add: our code uses numpy/sklearn/pandas, etc. We use the R Generalized Random Forests package for causal effect estimates.

Reviewer 2 1) We disagree. Our introduction cites **many** works that learn CATE (personalized causal effect) and personalized interventions from RCT (or, observational data); e.g. [17,23] for homelessness prevention and job training interventions. To clarify: these personalization approaches learn CATE and policies in “batch” rather than online fashion. The aim is still personalization; but the batch data *must* necessarily involve some randomization/overlap/exploration. When assessing the potential impact of a personalized policy, we show that this causal setting poses identification challenges for fairness metrics and provide estimators and sensitivity analyses.

2) Firstly, we do provide means of adjustment via Hardt et al. [26]. But we do highlight that direct adjustment of group-specific thresholds is controversial in practice and its relevance context-dependent, and this is not limited to our setting. We therefore defer the substantive (and less technically contributory) discussion to the appendix. In the appendix, we extensively discuss alternative approaches for minimizing disparities, including adjustment and covariate choice. Because TPR/FPR disparities could arise for a variety of reasons, it is not clear that adjustment of predictions is necessarily beneficial; we discuss reasons for caution in the appendix.

3) There is no typo there. Thanks for checking!

Reviewer 3: “I was wondering what the authors’ thoughts are on these two papers ...” Thank you, we will include these two references and discussion. There are different types of interference: 1. A universal budget/resource constraint; 2. Operational constraints (e.g. assignment under unit capacity constraints), 3. Network-type interference (violations of SUTVA) such as peer effects, and 4. general-equilibrium interference. 1&2 are related to resources. In the case of 1, under a universal budget, the optimal policy is to treat everyone above some quantile of CATE (e.g. [15]). This is an important motivation for our approach, since realistic budget constraints would lead to optimal decision policies which threshold CATE; we will highlight this further. Re: 2: Instead of taking Z to be a threshold on CATE, our approach also applies to assessing TPR/FPR of any policy Z, which may optimize assignment under more complicated resource constraints. 3&4 are types of interference that we do not address, we focus on assignment under heterogeneous effects under SUTVA.

Re: Nabi et al 2019: Their approach is complementary. While they adjust for fairness via constrained estimation (constraining pre-specified path-specific effects), they assess policy value via utility that marginalizes over the individuals’ labels (essentially utility-weighted accuracy). *If* their approach sought to *also* compare the analogous TPR and FPR (e.g. whether the disutility of fair policies falls on actually-guilty or actually-innocent), they too would have the issue of non-identifiability that we study and address in our work. Similarly with Kusner et al. 2019: their parity constraints are resource equity constraints, not classification parity, conditional on potential outcomes under assignment.

Reviewer 4 Re: choosing uncertainty sets: The magnitude of B can be directly calibrated against ATE effect size estimates from similar interventions, mechanistic knowledge, negative controls, or prior distributions on effect sizes, which practitioners typically can reason about. But instead of choosing a single B, usually sensitivity analysis is viewed as determining how big a violation is needed to overturn a conclusion. For example, it is unlikely that job training causes someone to not get a job, so if we need $B \geq 0.05$ to overturn a conclusion then it is robust if it is unrealistic 5% of the population would experience a negative causal effect. Re: estimating level of violation: Unfortunately, the level of violation is itself also unidentifiable without additional data like negative controls (see above). Re more datasets: There are not many *publicly available* datasets that were both large enough to reasonably support learning CATE as well as out-of-sample evaluation, had convincing protected group info, binary outcomes, *and* plausible monotonicity. That is why we introduced the Behaghel et al dataset, which we think is an exciting new dataset for considering fairness.

Re: Robust ROC and xROC: These are intended to provide additional information, in analogy to the use of ROC curves in assessing risk scores. Since sensitivity analysis focuses on illustrating how the extent of various claims (here, possible conditional disparities in performance) changes with the varying violations of assumptions (defier probability), we show how the bounds loosen with increasing violation. That some bounds are large should caution an analyst to draw hasty conclusions, while tight bounds imply a robust conclusion: our case study includes both examples. In the main text, the curves are overlaid: we will break out these as individual figures in the appendix and explain further how an analyst should interpret regions of overlap or non-overlap of these curves.