We thank the reviewers for their insightful feedback. We will incorporate all the suggestions/clarifications in the final version. Our detailed comments are provided below. All references are to the citations in the submission.

Novelty and Contributions: While there has been a lot of prior work on generating individual recourses via local counterfactual explanations, there is little to no work on *global* counterfactual explanations which can provide a high-level summary of recourses associated with a given (black box) model. This work makes the first attempt at addressing this critical gap. Our main contributions are: 1) We introduce the notion of **global counterfactual explanations** and propose the first framework, AReS, to generate them. Our explanations provide interpretable, customizable, and accurate summaries of *actionable* recourses for the *entire population* with emphasis on specific subgroups (which are either input by end users or learned automatically). 2) Our work also outlines one of the first solutions for **learning feature costs** from user inputs on pairwise feature comparisons. While we demonstrate (Theorem 2.2) that our optimization problem reduces to the generalized constrained optimization formulation for local counterfactuals [31,33] and can thereby generate individual recourses as well, the main goal of this result is to establish connections with prior research and not to suggest that generating individual recourses is one of the main contributions of our work.

Feature costs and feature changes: We consider two kinds of recourse costs: `featurecost` which captures the notion that some features can be intrinsically harder to change than others; and `featurechange` which captures the notion that changing feature values gets harder as the magnitude of the change increases. Depending on the specific application setting, one of these notions might be more important than the other. So, instead of combining these two notions into a single cost function, we provide end users with the flexibility to choose their relative importance (by setting λ_3 and λ_4). Our optimization framework is also generic enough to incorporate multiple definitions of the aforementioned costs (e.g., `featurechange` can be defined using the percentile shifts in feature values as done in [31]).

R1: (i) **User studies**: In addition to the biased two-level model discussed in Section 5, we also experimented with introducing racial biases into a 3-layer neural network (3-NN) and a logistic regression (LR) model via trial and error. We then carried out similar user studies (as in Section 5) with 36 participants to evaluate how our explanations compared with aggregates of individual recourses. In case of 3-NN, AReS clearly outperformed AR-LIME (88.9% vs. 44.4% on bias detection; 55.6% vs. 11.1% on bias description). In case of LR, AReS and AR-LIME performed comparably (88.9% in both cases on bias detection; 66.7% vs. 44.4% on bias description). This was omitted due to space constraints, but will be included in the final version. Also, see *R2: Bias detection* below. (ii) **Two-level decision sets** carry semantic meaning – with outer level rules describing *subgroups* and inner level rules representing recourses for the corresponding subgroups. As shown by prior work [14], this interpretation makes it very easy for end users to understand explanations. Furthermore, our preliminary studies have also shown that users can easily distinguish between subgroups and their corresponding recourses with two-level rule sets, but experience difficulties in doing so with 1 or > 2 levels. (iii) We account for **uncertainty in actionability** of features by using the *probabilistic* Bradley-Terry model (See defn p_{ij} in line 195) to learn feature costs. We will make this connection clearer in the final writeup. (iv) **Table 2**: We concur with the reviewer that our main contribution is recourse summaries. The goal of Table 2 is not to claim that we outperform individual recourse techniques but to assure the reader that we are not sacrificing recourse accuracy or costs in our attempt to construct interpretable summaries (as also pointed out by R3). The *mean fcost* metric shows lower values for AReS compared to AR not due to parameter errors but because the log-percentile shift optimized for by [31] is different from what is captured by this metric (Lines 308-310). We also compared AReS and AR using the cost function from [31] and found that AR achieves about 8 to 10% lower costs than AReS, as expected. We will include these clarifications in the final writeup. (v) **Unifying prior work**: As correctly pointed out, we are just writing down the Lagrangian form of a general constrained optimization formulation so that it can later be used for proving Theorem 2.2.

R2: **Bias detection**: We would like to emphasize that AReS, at its core, is an explainability technique and is not explicitly optimized for detecting model biases or fairness violations. That said, explainability techniques are commonly used to detect "potential" model biases or discriminatory behavior [24,33]. The bias detection study (Section 5) is meant to be a proof-of-concept to demonstrate that AReS can also be used to highlight potential biases, similar to other explainability techniques.

R3: (i) **Feature cost calculations:** We had conducted a user study to obtain pairwise feature comparisons for the Credit dataset, and had leveraged these inputs to learn feature costs and generate recourse summaries using AReS. We did not find significant drops/differences in recourse accuracies or mean fcosts (in comparison with the setting of uniform feature costs). We also experimented with non-uniform feature costs and observed similar results. (ii) **Recourse interactions:** It is theoretically possible for rules to contradict each other, but in practice we observed this occurs very rarely ($< 0.2\%$). Our objective already optimizes for *coverage* and *interpretability*, thus providing little incentive to choose multiple rules that apply to same sets of data points (thereby reducing the chance of contradictions).

R4: (i) **Text and image data:** It is easy to extend AReS to domains beyond tabular data, as long as the input features are interpretable. For example, *bag-of-words* features in case of text, and *super-pixels* of images can be used as inputs to AReS. Explainability techniques commonly use these kinds of interpretable representations as features [24].