We thank the reviewers for their thoughtful feedback and helpful suggestions. We address specific points below.

R1: missing assumptions. The standard assumptions made in the causality literature are required when we observe only one outcome per unit and cannot observe the counterfactual outcomes. These assumptions are needed for identification of causal effects from observed outcomes. However, in our case, we perform two interventions on every unit (and on the mediators) and observe all counterfactual outcomes. Therefore, these assumptions are not needed for our calculations of mediation effects, and no statistical bias is expected from the analysis.

R1: broader applicability. While we focused in this work on binary interventions and outcomes, the existing literature on causal mediation analysis enables the study of more general scenarios, including a different combination of variable types (binary, categorical or continuous) for interventions, mediators and the outcomes.

R1: defining bias. Dwork (2012) defines an algorithm to be fair if it gives similar predictions to similar individuals. The formalization of this definition was extended into Counterfactual Fairness (Kusner, 2017). We will explicitly define bias as the extent to which an algorithm is not counterfactually fair and draw the connection to our outcome variable y.

R1: dataset differences. We believe the difference in NIE between the professions dataset (NIE concentrated in initial layers) and the Winograd-style datasets (NIE concentrated in middle layers) reflects the fact that the former relates to bias in word embeddings (lexical semantics), while the latter relates to bias in coreference, a higher-level phenomenon.

R1: harms of gender binarization We acknowledge that our current discussion of the unintended harms of treating grammatical gender as binary variable is insufficient. Experimental results on he/they show very similar total effects to he/she ($\pm 15\%$), although with a lower variance. We will add these results and a discussion of the measuring difficulties of this effect under the singular/plural "they" confounder, as well as suggestions for mitigation, to the main body, and will extend the impact statement.

R1/R3: other model variants. We now have additional results for Tranformer-XL, BERT, DistilBERT, RoBERTa, and XLNet, which are consistent with the results from GPT-2.

R2: insights and takeaways. Debiasing is an important research direction. Although we feel it is beyond the scope of this paper, we believe our insights point to promising applications in evaluating and developing debiasing techniques. One could envision manipulating mediators found through our method to reduce gender bias, e.g., setting them to a null/neutral value. Further study is needed to evaluate how this approach impacts model bias and general performance.

R2: heads targeting anti-stereotypical candidates. Attention may capture negative as well as positive relationships, depending on the head-specific value vectors to which the attention weights are applied. We hypothesize that attention towards an anti-stereotypical candidate may decrease the probability of it being treated as the antecedent.

R2: concentration of attention in specific heads. Past work has shown that attention in middle layers correlates with coreference (as you allude to), which is tightly related to our analysis of gender bias. Specialization of attention heads has also been observed more generally, e.g., for various types of dependency relations (Clark et al., 2019). We will expand the discussion of this point in the camera-ready version.

R2: other types of biases. For this novel adaptation of mediation analysis, we perform an extensive analysis of a specific case study rather than a broader study of multiple phenomena, which would be a great area for future work.

R2: correctness. Would R2 point out the methodological problems that warrant a "no" answer to the correctness question, such that we may address them?

R2: missing related work. Thank you for pointing this out. We will add this to the related work.

R2/R3: limited scale and setup. We are constrained by available resources. However, we find consistent results across multiple models/datasets. In addition, the Winograd-style datasets are fairly nuanced in the linguistic phenomena.

R3: only the reporting clause is reported. In the professions dataset, we deliberately use verbs that are as neutral as possible to focus on the profession word and the bias it leads to. In the Winograd-style datasets, the examples are much more nuanced, in fact containing similar examples to the second one suggested by the reviewer. In this case, the bias depends on the entire context in the prompt. We agree that our method doesn't take into account potential bias in the continuation itself.

R3: only examines nouns. This is true for the professions dataset, though in the Winograd-style datasets the verbs play a role as well. We feel that a focused analysis of bias in verbs, while valuable, would warrant a separate study.

R3: inter-sentential context. Indeed, we have only looked at intra-sentential context. We note that some contexts are rather nuanced in these Winograd-style datasets. Our methodology may be applied to inter-sentential contexts as well.

R3: direct effects. Figure 5 is representative of the direct effects we observed in all models: direct effects approximate the difference between the total effects and the indirect effects. We will include additional results on direct effects.

R3: difference from other studies on gender bias. The main difference is that our research questions and methodology focus specifically on mediators in bias by performing interventions. We will clarify this in the related work.

R4: computational cost. We recognize that that computational cost is non-trivial. We discuss computational complexity in Appendix D with respect to the subset selection algorithm, but we will also discuss more generally in the main body.