We thank the reviewers for their consideration of our paper and for their feedback. The consensus appears to be that this is a generally well written paper, exploring a *"potentially impressive"* (R2), *"useful"* (R1, R3) and *"novel"* (R1, R2) connection between fairness and disentanglement with claims backed up by substantial empirical evidence (R1, R2, R3). There appears to be one major concern and 4 minor questions/suggestions, which we kindly address below.

Main concern of R2: use of ground-truth factors to compute disentanglement scores

The main concern of R2 seems to be that the paper relies on *"disentanglement scores, which are computed based on ground-truth factors of variations"* but that these factors *"are generally not available in practice"*. In our view, there appears to be a misunderstanding with regards to the exact goals and contributions of this paper:

- **Motivation: usefulness of disentanglement notions vs new method.** The main goal of this paper is to *evaluate the usefulness* of disentangled representations (in terms of leading to fair predictions), rather than proposing new methods for *learning* disentangled representations. This distinction is critical: While we do train various methods for the purpose of this study, we do not make any assumptions about the *feasibility* of (and *methods* for) learning disentangled representations in the absence of ground-truth factors, and our results have implications for the supervised, semi-supervised, and unsupervised settings. This paper provides evidence that, if one were to find a reliable method for disentanglement (in terms of current metrics), predictions on representations of such an approach would be more fair.
- **Relevance: validate the motivation of >15 recent ML papers.** Recently, numerous papers have been concerned with learning disentangled representations [7, 8, 12, 17, 18, 20, 31, 32, 33, 34, 43, 50, 57, 61, 66, 74, 76, 82, 84]. The key motivation (but also assumption) of these works is that current notions of disentanglement (MIG, DCI, *etc.*) are desirable. Until now there has been little empirical evidence verifying this. In particular, the study of Locatello et al., ICML 2019 was inconclusive in this regard, see their Section 5.5. This paper fills this gap by investigating whether disentanglement is useful for improving fairness of downstream predictions. Our results are novel and provide motivation for further research into (a) disentanglement methods and (b) their application for ML fairness.
- **We present a heuristic to select fair representations.** As described in Section 4.2, as a by-product of our investigation, we noticed that downstream predictive performance may be used as a way to select fair representations among the representations we trained using SOTA unsupervised disentanglement methods. We argue that this might be an interesting heuristic as it only requires labels for the single downstream prediction task but not for the representation learning.

Other concerns:

R1-R2: *Motivation for adjusted metrics in Section 4.2.*

We compute the adjusted metrics to answer the question **"Given two representations with the same downstream performance, is the more disentangled one also more fair?"**. The key difficulty is that for a given representation there may not be other ones with exactly the same downstream performance. Hence, we compute these adjusted metrics which intuitively measure how much fairer (more disentangled) a representation is compared to an average representation with the same downstream performance. To compute the average fairness (disentanglement) for a given downstream performance, we use a nearest neighbor regressor as a robust non-parametric 1D regression model.

R1: *Chain of arguments in "How do we identify fair models?" not fully clear. + Is the argument by R1 correct?*

The understanding of the reviewer is correct. We will add the suggested graph "fairness <- disentangling -> accuracy" and revise the manuscript such that this argument is explained more concisely.

R2: *Theorem 1.*

Entangled representations affect the fairness of downstream classifiers. The theorem proves a point: the Bayes optimal classifier does not imply fairness on an entangled representation. The proof is by counterexample. We further support this claim with Figure 2, showing that this is a practical issue that arises on trained classifiers and not simply a theoretical corner case. If we can disentangle target and sensitive variables then the classifier will be fair, see lines 145-155. In the particular counterexample of the theorem, there is a trade-off between fairness and accuracy as is common in the fairness literature. This is discussed in lines 135-138 and in the footnote.

R3: *Demographic parity.*

We will add more motivation and examples, in particular relating to previous work (e.g. [9, 92]) and to relevant settings, for example to settings where sensitive variables may not be recorded due to privacy reasons.

Clarity, title, and other comments regarding the presentations of our results.

We thank the reviewers for their suggestions to improve the presentation. We will use their feedback to revise the manuscript. In particular, as suggested, we will change the title to "On the Fairness of Disentangled Representations", improve the discussion on section 4.2 and add a discussion on non-VAE based methods such as ALI. By gradient boosted trees (GBT10000) we mean gradient boosting of decision trees [Friedman, 1999] trained on 10000 examples.