We thank the reviewers for their comments and for acknowledging that we address a relevant problem for the NeurIPS community [R1,R2,R3], that our experiments show the utility / effectiveness of the proposed method [R1,R2,R3] and that the setup and theoretical claims seem sound [R1,R2].

R1 **Comparison to [11,24] not very fair:** Please note that the claimed missing result "sensitive attribute included" ***is in the paper*** (specifically see the columns in Tabs. 1 & 2 named "Sensitive feature in the functional form of the model"); See also ll. 233-234. **They do not discuss shifts in $P(Y|X,S)$, which will not affect fairness but will affect accuracy:** The reviewer raises a very interesting point. Note that our method (cf. Eq. (6)) looks for a shared low complexity representation between the tasks. Shifts in $P(Y|X,S)$ are permitted, provided there is a common predictive representation for the task outputs. Searching for such representation is indeed one of the main ideas behind transfer learning (see e.g. Caruana 1997; Argyriou et al. 2008, etc.) and it is aided by the first term in Eq. (6), measuring the average empirical risk on the training tasks. However, differently from the above works, our method also crucially involves the second term, which encourages representations that approximately satisfies DP on average over the training tasks. Our fairness violation bound (Thm 1) involves only the marginal distributions (of each sensitive group within the task) since we care to measure this at the representation level, but see ll. 97-103 and Lemma 3 for how this affects DP at the output level. In this sense our method learn a shared and transferable representation, one based on which accurate and fair models can be learned on tasks sampled from the environment ρ. **Would we expect accuracy to degrade at all?** Yes, accuracy decreases due to the fairness constraint. This property is common to any algorithmic fairness method. **Other suggestions:** These will be addressed in the revision. In particular we'll: make our code publicly available upon acceptance, add at l. 140 that σ is applied component-wise, say at l. 159 that the function "Gap" is quantified by the r.h.s. of Ineq. (10), clarify at l. 163 that "approximately" means that the marginal distributions of the two sensitive groups within the task are closed according to some suitable measure (e.g. see that at l. 242), improve the "M [24]" notation (it stays for "method [24]" but we could just write "[24]"). Finally concerning **fairness/accuracy tradeoff in the generalization setting:** standard bounds from the learning-to-learn literature (e.g. [3]) can be readily used to bound the risk (or accuracy) of the representation h found by our method on future tasks by the minimal multitask empirical risk (over the specifications g_t). Now since our method in Eq. (6) minimizes a tradeoff between the multitask empirical risk and fairness violation, the larger γ the larger the former term, so risk on future tasks will reflect this tradeoff too.

R2 **Submission focuses entirely on dem. parity:** We agree DP is not the ultimate fairness notion. Still it is frequently studied in the literature and our study is a valuable starting point for fair representation learning within the multitask setting. **May techniques be deployed with good conscious?** Yes, our theoretical and experimental results give an indication that the method could be valuable in practice and safely deployed. Of course more experiments would be needed to assess its robustness on on real-world problems. **Fictitious multitask setting:** it is true the paper is more on theory and methodology which is not always close to practice, but imagine the following real-life scenario: each task is associated with an hospital in country X and the task is to predict whether a patient who visits emergency should be hospitalized. The environment (meta-distribution) ρ may be the uniform distribution (or weight larger hospitals more). The sensitive attribute may be race and other non-sensitive variables may measure cough frequency and body temperature. Our main result, Thm 1 (in conjunction with Cor 2) then says that if we use our method to learn a predictive representation and observe it to be fair according to DP on the random training task datasets, then it will also be fair according to DP on average on all possible hospitals at the population level (i.e. on average over random patients visiting the hospital), which is a very appealing property. **Description of ρ:** Yes, we'll mentioning also at l. 104. **How is the multitask setting realized in the empirical experiments?** Please see ll. 227-34: we test either on different data for the same tasks using during training or on a new task in leave-one-task out setting. **How is the Sinkhorn divergence in eq. (5) estimated from finite samples?** Consider $\hat{P}$ and $\hat{Q}$ the same as in ll. 116-117. Denote by $\mathsf{p} = (\mathsf{p}_1, \dots, \mathsf{p}_\mathsf{n}) := 1/n\mathbf{1}_n$ and $\mathsf{q} = (\mathsf{q}_1, \dots, \mathsf{q}_\mathsf{m}) := 1/m\mathbf{1}_m$, with $\mathbf{1}_k$ the vector with k entries equal to one (p and q denote the weights of the empirical distributions $\hat{P}$ and $\hat{Q}$). Then, $\mathsf{OT}_\varepsilon(\hat{P}, \hat{Q}) = \min_{T \in \Pi(\mathsf{p},\mathsf{q})} \langle T, C \rangle + \varepsilon \sum_{i,j=1}^{n,m} \log(T_{ij}/\mathsf{p}_i \mathsf{q}_j) T_{ij}$, where $\Pi(\mathsf{p},\mathsf{q}) := \{T \in \mathbb{R}_+^{n \times m} \,|\, T\mathbf{1}_m = \mathsf{p},\, T\mathbf{1}_n = \mathsf{q}\}$ and $C_{ij} = \|x_i - z_j\|^2$; see ref. [30] for more explanations. **Restriction to 1-hidden layer nets:** Our method and Thm 1 apply to general classes of representation functions h of suitably bounded complexity. For simplicity we illustrated them on 1-hidden layer networks, both theoretically (Thm 1 + Cor 2) and empirically. However, bounds in [Chain Rule for the Expected Suprema of Gaussian Processes, ALT 2014] could be used in place of Cor 2 for multi-layer representations. **Extending Thm 1 to Sinkhorn divergence (l. 188):** There are two main obstacles at this stage: first, the term A_h at l. 425 would not be zero, because the estimators is biased. Second, the Lipschitz behaviour needed to factor out the constant (see l. 436) is not clear in the case of Sinkhorn.

R3 **1. Main contribution:** Please see l. 46. To the best of our knowledge this is the first paper using multitask learning for fair representation. **2. Significance of Thm 1:** This result together with a bound on the Rademacher average of the representation class (e.g. Cor. 2 in the case of linear representations) gives a justification for our method – see also the reply to R2 concerning the point *fictitious multitask setting*. **3. Contribution insufficient:** We disagree: MMD and SNK have been used only recently for algorithmic fairness. The proposed method is novel, empirically competitive and theoretically grounded (see point 2 above). **4,5. Formula mistakes and missing definitions:** We'll carefully check our formulas. Thanks for Eq. (2), but note $d(\cdot)$ is already defined at l. 109 of the paper.