We thank the reviewers for their incredibly detailed and thoughtful reviews. We really appreciate the time you put into reading and thinking about our work. The top issues we understood from the reviews were 1) lack of clarity in Section 2.1 on the usefulness of the [Higgins et al.] framework in this context and the formalization of the indirect influence definition, 2) lack of clarity about the implementation, and 3) concerns about the experimental section. Below, we include partial rewrites of Sections 2.1 and 2.2 that we will include directly in our revision and we hope clarify the first two of these issues (we describe planned improvements to the experimental section below). Note the indirect influence definition that is now more explicit about the use of the disentangled representations theory and the hopefully clearer explanation that the method uses a reductions framework.

Our primary goal is to leverage recent developments in disentangled representations to help solve the indirect influence problem for individuals, and not to supersede current work in disentangled representations. Since our goal is to compute individual influence scores, we cannot use global metrics such as demographic parity difference as reviewer 2 suggests. Given a sufficiently strong adversary, our disentanglement error metric reports high error when there is mutual information between p and x', and has low error when there is low mutual information between p and x', making it very related to the mutual information gap. We appreciate the reviewer's suggestion to report the mutual information gap as well, and will add this to our analyses.

In accordance with the suggestions of the reviewers, we intend to improve the experimental section in the following ways: 1) We will update the dSprites experiment to operate on the full 64×64 pixel images, using the standard CNN architectures in the literature. 2) We will add existing error metrics for our disentangled representations, including mutual information. 3) We will add the additional baseline of LIME, as well as more detailed exposition about how our method compares to the baselines.

We note that in the dSprites experiment our goal is to compute the influence of the latent factors on the predictions of a model trained only on the pixels, by learning how the pixels act as proxies for the latent factors. We are not trying to model relationships between the latent features themselves. It is convenient for verifying our method that these latent factors are independent, so that the ground-truth is easily interpretable. This independence does not compromise our experimental goals. Still, we agree that adding relationships between the latent factors would be an interesting extension.

We thank the reviewers for pointing out that we should emphasize the importance of indirect influence to fair machine learning and we will revise the introduction to further emphasize this application. We also appreciate the reviewers' links to other relevant papers and will incorporate these into our related work description as well as the detailed stylistic suggestions.

Section 2.1 Partial Revision:

Definitions We use the term *world state* [Higgins et al.] to represent the actual nature of the objects or people represented in the data, void of all of the errors and omissions incurred during observation. We say the world state consists of two independent factors of variation, (p, x') which correspond to the protected and unprotected aspects of the world state respectively. The unprotected features $\mathbf{x}$ are generated from the world state by an observation process $b : W \to \mathcal{X}$ so that $b(p, \mathbf{x}') = (p, \mathbf{x})$. Furthermore, let b_i be the function such that $b_i(p, \mathbf{x}')$ is the observation of only $x_i \in \mathbf{x}$. To provide generality to multiple forms of direct influence, we assume an arbitrary direct influence function $\mathcal{DI} : \mathcal{X} \times \hat{\mathcal{Y}} \to \mathbb{R}$. We formulate our implementations using SHAP as our notion of direct influence, but our framework is general and is compatible with other common local interpretability methods such as LIME and gradient based methods. We propose the following definition of indirect influence via a reduction to direct influence:

$$\mathcal{II}_p[M(p, x)] \coloneqq \mathcal{DI}_p[(M \circ b)(p, x')]$$

The above states that the indirect influence of p is the direct influence of p when considering the model as acting on world states instead of features. Whereas direct influence measures the sensitivity of a model to changes in each feature independently, indirect influence attempts to model how proxies for p change along with p. Note that indirect influence is inherently specific to a data distribution, since our goal is to understand proxy relationships between features. All indirect influence audits should then be interpreted as with respect to the dependence structures observed during training.

Section 2.2 Partial Revision:

Implementation We train a disentangled representation to estimate (p, x') for each feature of interest p. This allows us to compute representations with only two factors in a supervised manner, avoiding many of the issues in the current disentangled representations literature noted by [Locatello et al.]. A key limitation of this approach is that while easier to train, it potentially requires one to train many disentangled representations. This means the technique may be most useful in domains such as fairness where we care specifically about the impact of one or a small collection of distinguished features that may or may not be directly used as inputs to the model.