We thank the reviewers for their insightful feedback and encouraging words. We are pleased that all reviewers acknowledge the relevance of learning certified individually fair representations. Below, we address the reviewers' comments and concerns, all of which we will incorporate into the next version of our work.

R1: Can you investigate the impact of robustly training the classifier on accuracy and certifiability? The impact of adversarial training on logistic regression (for both accuracy and fairness) is limited due to the smoothness of the decision boundary. In fact, over all datasets and a wide range of γ, the largest increase in certification is roughly 7%, with a simultaneous accuracy drop of 0.3%. In contrast, for a more complex classifier, such as a feedforward neural network with 2 hidden layers of 20 nodes each, adversarial training doubles the certification rate (from 34% to 70.8%) while decreasing the accuracy only by 1.6%. We will provide a more thorough investigation in the next revision.

R2: How does your work compare with counterfactual and indirect fairness? In contrast to counterfactual and indirect fairness, we do not require a causal model of the underlying data distribution. This allows for more flexibility in applying our approach but prohibits us from making causal/counterfactual claims. Nevertheless, the combination of logical constraints and causal/counterfactual fairness represents an interesting direction for future research.

R2: Can you extend your discussion of the framework from McNamara et al. [10]? Yes, a key difference to [10] is that their approach requires to know statistics of the data distribution to obtain guarantees. This allows them to compute probabilistic bounds for individual (and group) fairness: for a new data point from the same distribution, the constraint will hold with a certain probability. Our result is different in that we obtain a certificate for a fixed data point, which ensures that the fairness constraint holds (independent of the other data points). While both approaches are valid and practically relevant, they are also fundamentally different, which renders experimental comparison meaningless.

R2: You could move fair transfer learning to the main body as this is another major contribution. We agree that the compatibility of our method with existing fairness notions is an important contribution, and we would use the additional space of the camera-ready version to move the fair transfer learning section to the main body.

R2 & R3: What is the impact of the balance parameter γ on the accuracy-fairness tradeoff? We compare the accuracy and certified individual fairness for different loss balancing factors γ for the CAT + NOISE constraint on the CRIME dataset in Table 1. We observe that increasing γ up to 10 yields significant fairness gains while keeping the accuracy roughly constant. For larger values of γ, the fairness constraint dominates the loss and causes the classifier to resort to majority class prediction (which is perfectly fair). As mentioned by reviewer 3, our method increases both accuracy (albeit only by a small amount) and fairness for certain values of γ (e.g., $\gamma = 2$). Based on our observations, we conjecture that this effect is due to randomness in the training procedure and sufficient model capacity for simultaneous accuracy and fairness for $\gamma \leq 5$. As we observed the same trend on all datasets, we recommend data producers to increase γ up to the point where the downstream validation accuracy drops below their requirements.

Table 1: Accuracy and certified individual fairness for the CAT + NOISE constraint on the CRIME dataset for different loss balancing factors γ. Compared to the baseline $\gamma = 0$, our method ($\gamma \neq 0$) incurs minimal changes in accuracy while significantly increasing the percentage of certified individual fairness for a wide range of γ.

γ	0	0.1	0.2	0.5	1	2	5	10	20	50
accuracy (%)	84.36	84.62	84.87	84.36	84.10	84.62	84.36	81.79	50.77	50.77
certified (%)	6.15	9.23	12.05	18.46	33.08	52.31	61.28	62.82	100	100

R3: Can you compare with other fair representation learning methods? Yes, we will provide more in-depth comparison, even though we do not believe works, e.g., Zemel et al. [8], which either do not focus on proximity in latent space or use nonlinear methods that cannot be efficiently certified, will yield certified individually fair representations.

R3: Can you comment on the relationship with differential privacy? The close relationship between individual fairness and differential privacy (DP) has been discussed in previous work (see, e.g., [12]). However, DP crucially differs from our work in that it obtains a probabilistic guarantee, similar to [10] mentioned above, whereas we compute absolute guarantees for every data point. The DP analog of LCIFR is to limit the sensitivity of f_θ by injecting noise, and to consider h_ψ as a post-processing step. We will include an extended version of this discussion in the next revision.

R3: Is your method merely an adoption of existing work? No, recent advances in training with logical constraints and proving constraint satisfaction enable us to tackle a previously unsolved problem: provable individual fairness across *modular* components. As outlined by reviewer 2, learning certified individually fair representations and demonstrating feasibility and effectiveness are indeed essential contributions.

R3: Please clarify that ϕ and μ are logical formulas taking value either 0 or 1. We stated this in L43/L44 and L74.

R3: How does $h_\psi^{(y)}(z)$ differ from $h_\psi(z)$? We will clarify that $h_\psi^{(y)}(z)$ is the logit corresponding to label y.