Author Response

Following the reviews we received for NeurIPS we have taken substantial measures to address reviewer concerns both during the initial rebuttal period and beyond.

1 Additional Datasets

A common concern across reviewers was limited evaluation on real-world datasets. While the scope of our initial experiments was already well in line with the existing related literature, we agreed with the reviewers and have added two additional tabular datasets from the social sciences domain as well as one additional dataset from the vision domain. As mentioned in the paper, we have resorted to datasets commonly used in the literature.

A note on image datasets

Related work on plausibility of counterfactuals has largely relied on small image datasets like *MNIST* (Dhurandhar et al. 2018; Schut et al. 2021). This may be due to the fact that generating counterfactuals for high-dimensional input data is computationally very challenging. An exception to this rule is the work on *REVISE* (Joshi et al. 2019), which uses a larger image dataset. *REVISE* is suitable for this task, because it maps counterfactuals to a lower-dimensional latent space. Similarly, our proposed *ECCCo+* should also be applicable to high-dimensional input data. In our benchmarks, however, we include other generators that search directly in the input space. Since our benchmarks required us to generate a very large number of counterfactuals, it was not at this time feasible to include larger image datasets. That is despite our best efforts to optimize the code and parallelize the computations through multi-threading and multi-processing on a high-performance computing cluster.

2 Constraining Energy Directly

In our initial work we used our unfaithfulness metric directly as a penalty term in *ECCCo*'s counterfactual search objective. This generally achieves the highest levels of faithfulness but it has several disadvantages, some of which were pointed out by the reviewers. Our new approach constrains the energy directly, which is more theoretically grounded and leads to better results across the board. Since it does not depend on generating samples through SGLD, our new approach is much more computationally efficient as well. Additionally, it also addresses the following reviewer concerns:

Results were biased with respect to unfaithfulness metric

One reviewer raised concern about the fact the using the unfaithfulness metric as a penalty biases the results. This is a valid concern which we have addressed now.

Counterfactuals looked homogeneous

Another reviewer pointed out that the counterfactuals generated by *ECCCo* looked homogeneous, which is also a valid concern. The observed homogeneity most likely stemmed form the fact that the samples generated through SGLD for the underlying models were fairly homogenous. With our new approach we no longer rely on SGLD samples and the homogeneity issue is no longer present.

Closeness criterium was violated

A related concern was that large perturbations induced by *ECCCo* seemed to violate the closeness criterium. As we discuss in the paper, our findings do not suggest that *ECCCo* yields unnecessarily costly counterfactuals. Indeed, with reference to the vision data, *ECCCo* seems to keep useful parts of the factual largely in tact, which reduces costs. As we already argued during the rebuttal and in the paper, additional costs cannot be avoided entirely when faithfulness and plausibility are prioritized. This applies to *ECCCo* as much as to other generators like *REVISE*.

Generalizability

This was not an explicit concern but some reviewers wondered if *ECCCo* could also be applied to non-differentiable models. While our initial approach that relied on SGLD samples was not suitable for non-differentiable models, our new approach is. This is because none of its

penalties rely on differentiability. Of course, we still framed *ECCCo* in terms of gradient-based optimization, but the proposed penalties could be applied to other, non-gradient-based counterfactual generators as well such as *FeatureTweak*, for example (Tolomei et al. 2017).

3 Beyond JEMs

Another common concern was that *ECCCo* primiarly achieved good results for JEMs. This has been addressed by introducing *ECCCo+* for situations when plausibility is crucial. We find that *ECCCo+* achieves good plausibility and faithfulness across the board. We have added convolutional neural networks to our analysis and find that *ECCCo+* achieves results for these models that are at least on par with the results for JEMs.

4 Mathematical notation and concepts

One reviewer took issue with our mathematical notation, a concern that was not shared by any of the other reviewers. Nonetheless, we have revisited the notation and hope that it is now more clear. That same reviewer also raised concern about our definitions of plausibility and faithfulness that rely on distributional properties. In particular, the reviewer argued that "[…] the class-condition distribution $p(\mathbf{x}|\mathbf{y}^+)$ is existed but unknown and learning this distribution is very challenging especially for structural data". We have extensively argued our case during the rebuttal and pointed to a potential reviewer misunderstanding in this context. In particular, we argued:

> We do not see this as a weakness of our paper. While we agree that learning this distribution is not always trivial, we note that this task is at the very core of Generative Modelling and AI—a field that has recently enjoyed success, especially in the context of large unstructured data like images and language. Learning the generative task is also at the core of related approaches mentioned in the paper like REVISE: as we mention in line 89, the authors of REVISE "propose using a generative model such as a Variational Autoencoder (VAE)" to learn $p(\mathbf{x})$. We also point to other related approaches towards plausibility that all centre around learning the data-generating process of the inputs X (lines 85 to 104). Learning $p(\mathbf{x}|\mathbf{y}^+)$ should generally be easier than learning the unconditional distribution $p(\mathbf{x})$, because the information contained in labels can be leveraged in the latter case.

None of the other reviewers found any issue with our definitions and we have made no changes in this regard. We did, however, make a minor change with respect to the related evaluation metrics. We are now more careful about our choice of the distance function. In particular, we investigated various distance metrics for image data and decided to rely on structural

dissimilarity. For all other data we use the L2 Norm, where we previously used the L1 Norm. This has no impact on the results, but there was no obvious reason to use the L1 Norm in the first place other than the fact that it is typically used to assess closeness.

5 Conformal prediction was introduced too suddenly

One reviewer pointed out that conformal prediction was introduced too suddenly. We have moved the introduction of conformal prediction forward and added more detail in line with reviewer feedback.

6 Limitations section

We have extended the limitations section to address reviewer concerns.

7 Other improvements

As discussed above, counterfactual explanations do not scale very well to high-dimensional input data. The NeurIPS feedback has motivated us to work on this issue by enabling intuitive support for multi-threading and multi-processing to our code. This has not only allowed us to include additional datasets but also to run extensive experiments to fine-tune hyperparameter choices. All of our code will be open-sourced as a package and we hope that it will be as useful to the community as was to us during our research.

Dhurandhar, Amit, Pin-Yu Chen, Ronny Luss, Chun-Chen Tu, Paishun Ting, Karthikeyan Shanmugam, and Payel Das. 2018. "Explanations Based on the Missing: Towards Contrastive Explanations with Pertinent Negatives." *Advances in Neural Information Processing Systems* 31.

Joshi, Shalmali, Oluwasanmi Koyejo, Warut Vijitbenjaronk, Been Kim, and Joydeep Ghosh. 2019. "Towards Realistic Individual Recourse and Actionable Explanations in Black-Box Decision Making Systems." *arXiv Preprint arXiv:1907.09615*.

Schut, Lisa, Oscar Key, Rory Mc Grath, Luca Costabello, Bogdan Sacaleanu, Yarin Gal, et al. 2021. "Generating Interpretable Counterfactual Explanations By Implicit Minimisation of Epistemic and Aleatoric Uncertainties." In *International Conference on Artificial Intelligence and Statistics*, 1756–64. PMLR.

Tolomei, Gabriele, Fabrizio Silvestri, Andrew Haines, and Mounia Lalmas. 2017. "Interpretable Predictions of Tree-Based Ensembles via Actionable Feature Tweaking." In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 465–74. https://doi.org/10.1145/3097983.3098039.