We thank the reviewers for their comments. All reviewers agree that our approach on exploring computational complexity as a way of comparing model interpretability would be of interest to the NeurIPS community, which is very encouraging for us. **Reviewer 1** and **Reviewer 2** also welcomed our formal approach to it (**R1**: *"Classifying the complexity of explanations is a natural step. Therefore this paper is a welcome contribution"*, **R2**: *"a thoughtful approach to an important problem"*). Both reviewers mention as a strong point the use of parameterized complexity when analyzing explanations for MLPs. **R1** even encourages us to present in more detail the classes that we use in our results as they *"could have other uses in deep learning"* (because of space limitations it was difficult to do so in the current version of the paper, but we would definitely add it in an extended version). We deal with specific comments by the reviewers below.

**Reviewer 4** appreciates that we take the risk of formally studying the elusive notion of interpretability (**R4**: *"Important goal of pinpointing an elusive concept"*), also praising our theoretical results (**R4**: *"Novel and sound complexity results"*). Nevertheless, the reviewer questions the significance of the approach, which can be summarized in the following comment: *"There does not seem to be a well-founded basis for [linking model interpretability with computational complexity], since interpretability is, to the best of my knowledge, an inherently subjective concept."* We agree with the reviewer: interpretability is largely a subjective concept and the community is far from having a definitive answer for what interpretability exactly means. But we actually consider this subjectivity as our main motivation: we wanted to explore ways into which some formal ground can be given to this subjective concept, even if in a preliminary form. This concern can thus be reduced to our choice of computational complexity as such a possible ground. **R4** considers this choice as *"not well-founded"* while the other two reviewers consider the study of the computational complexity of explanations as a plausible choice/step. We think that our results in the paper are a preliminary proof of this plausibility/adequacy, as the complexity of explanation queries actually correlates with the informal views on interpretability that can be found in the literature.

**R4** also questions our position in the following comment: *"[...] perhaps it would be more interesting to set out from a more neutral position of looking for a correlation between [complexity and interpretability]."* We disagree with this comment. Neutrality was a central concern when we were wording our submission and we were very careful in not claiming that computational complexity is the right or only way to understand interpretability. This is why we phrased our results explicitly as *"correlations"* between complexity and interpretability (exactly as the reviewer suggests). For instance, in the Abstract (line 4) we write that *"We make a step towards such a notion by studying whether folklore interpretability claims have a correlate in terms of computational complexity"*, and also in the Introduction with a similar comment (line 46). We were also careful in naming our notion as *complexity-based interpretability* when we formalize it (Section 2, line 110) to make it explicit that we are not formalizing a general notion of interpretability, but exploring a (new) narrow yet formal one. Although we disagree with the reviewer's view, this comment pushes us to make our position even more explicit in the paper. For the next version of the paper, we will include an additional discussion about this in the **Limitations** section and move this section to be right after the Introduction (it is the last section in the current submission).

We now focus on the comments by **R1** and **R2**. Both reviewers point out that some of the problems we treat, or slight variations of them, have already been discussed in the literature over different settings and under different names. We agree that our paper would improve by adding a discussion of the different names and contexts under which the literature has discussed similar problems, and explicitly mentioning their differences with our setting. We were under space constraints but in case of acceptance we will use the additional space to improve our Related Work section. **R1** and **R2** also suggest that our paper would benefit from being more explicit about the novelty and difficulty of some of the proofs by distinguishing those that involve new insights (e.g., parameterized complexity) from those that follow by already used techniques for the models that we consider. We will also implement this change in the next version. Thank you very much for the comments and the pointers. We appreciate **R1** pointing out certificate complexity and the study of boolean sensitivity as related subjects that could make for interesting lines of future research.

**R2** also comments about our choice of minimum sufficient reason (MSR) as one of the explainability queries as *"it's not clear why it would need to be minimum"*. We consider the minimality as a desirable property since an explanation (as a sufficient reason) can always be padded with superfluous information while remaining valid. Our rationale was that, among the different minimal sufficient reasons, it is arguably better to provide explanations that are as succinct as possible (observe that an input $x$ is always a sufficient reason for its own classification, but this is not really interesting). This comment also pushes us to provide a more technical discussion of the relevance of the chosen queries in the final version of our paper. Unfortunately due to space constraints we cannot address other interesting comments by **R2**, like the relationship between monotone DNFs and MSR, that suggests a further study of asymmetric models (for which MSR for positive and negative examples would have different complexities), or the need of using SHORTEST IMPLICANT CORE instead of just SHORTEST IMPLICANT. But we thank the reviewer as both comments give us the opportunity to improve our paper.