We would like to thank the reviewers for their comments and feedback. We are aware that in a largely conceptual paper like ours there are many subtleties that do take time and effort to digest.

Reviewer #1: Causal Shapley values (SVs) are defined in Section 2. These do *not* coincide with the interventional (and thus in our terms, marginal) SVs of [9] and others. Janzing et al. [9] write down the same equation, but then choose to ignore any dependencies between the features in the real world (e.g., that in summer it tends to be warmer than in winter). We make the different choice to incorporate these dependencies and hence do not simplify to $P(\mathbf{X}_{\bar{S}}|do(\mathbf{X}_S = \mathbf{x}_S)) = P(\mathbf{X}_{\bar{S}})$, but keep $P(\mathbf{X}_{\bar{S}}|do(\mathbf{X}_S = \mathbf{x}_S))$ in our definition of the causal SVs. We will follow the reviewer's suggestion to make this more explicit in Section 2.

This distinction then hopefully also resolves the reviewers' issue about the indirect effect: it obviously vanishes for marginal SVs, but need not vanish for causal and conditional SVs. This should also be clear from the examples in Section 4 (Figure 2). The decomposition for conditional SVs follows by replacing "conditioning by intervention" with "conditioning by observation", i.e., by replacing $do(\mathbf{X} = \mathbf{x})$ with $\mathbf{x}$. The decomposition is introduced in Section 3 to assist our illustration of how the different SVs attribute a model's prediction to the features involved in this prediction in Section 4 for different causal models. Here we also discuss in which cases (most notably the fork and the confounder) conditional SVs fail to provide an intuitive causal attribution.

The causal chain graphs are introduced as a means to compute causal SVs (whether symmetric or asymmetric) when users are willing/able to specify a (partial) causal ordering, but not a full-fletched causal model. This is indeed the same information that the asymmetric SVs of [6] rely on. On top of [6] we offer (1) a formalization in terms of causal chain graphs, (2) the notion that, with "conditioning by intervention" instead of "by observation" as in [6] there is no strict need for introducing asymmetry in the SVs, and (3) that this then also leads to the intuitively correct way of computing SVs in the case of common confounding between the features.

Reviewer #2: Section 4 aims to illustrate the behavior of the various SVs in simple cases that can be analyzed analytically and then to argue which is the most intuitive, indeed also linking to psychological literature when appropriate. Here one prominent theory, dating back to [15], states that humans sample over different possible scenarios to judge causation. Translating this to a situation in which there are two possible causes, X_1 and X_2, where it is unknown which one is intervened upon first, may suggest that the natural interpretation is to consider both options and average over them.

We fully agree that quantifying causal influence is a difficult topic and any method has its weaknesses, but causal SVs appear to fare better than the reviewer suggests. Discontinuity w.r.t. arrows with zero strength is an issue for the asymmetric SVs, but not for the symmetric SVs that consider all orderings, not just those consistent with the causal DAG. After averaging over all these orderings, the indirect effect already does incorporate all possible paths (so we do not see how or why it needs to be generalized), but of course in the game-specific way inherent to the Shapley value approach. We will add comments and disclaimers to clarify this and adapt our description of Janzing et al. and related work as suggested by the reviewer. Our statement 'not every causal query need be identifiable (see e.g., [24])' did not presume DAGs with all variables observed, but more general causal structures possibly including latent variables.

Reviewer #3: As indicated in the introduction and in the discussion, improving counterfactual explanations is indeed an interesting topic, but beyond the scope of the current paper. For the record, counterfactual explanations as in e.g. [33] are quite different from the counterfactual question posed by the reviewer, which can be answered simply by reading off the output of the model.

As also indicated at the end of the introduction, we ourselves do not consider Theorem 1 the main contribution of the paper. See also the third paragraph of our answer to Reviewer #1 and the beginning of Section 5: causal SVs are generally applicable when a user is willing/able to specify a causal model among the features that are used as input to the model and when all causal queries are indeed identifiable. Specifying when this is the case is a topic on its own: we will add more references (see also the supplement). We introduce causal chain graphs as a practical approach to handle partial causal knowledge. In causal chain graphs, all causal queries are guaranteed to be identifiable. They do include the possibility to handle cycles, confounders, etc, as should be clear from Figure 2. In fact, all the examples in Figure 1 are easily translated to causal chain graphs. An illustrative example for the fork could be predicting hotel occupation (Y), based on season (X_2) and temperature (X_1).

Causal relationships are indeed asymmetric, but that does not prevent the causal SVs from being symmetric according to the standard symmetry axiom for SVs (see the definition in Section 2 and the elaborate discussion in [9], Section 3 in response to Sundarajan and Najmi, 2019). We chose not to repeat this argumentation, but will add a reference.

Figure 4 is meant to illustrate the difference between the various SVs (asymmetric SVs focus on the root cause, marginal SVs on the direct effect, symmetric causal SVs consider both), not necessarily to claim that one is always better than the other. We will extend the supplement with additional empirical analyses, e.g., on (deep) neural networks.

(7) indeed should have been (6). We will fix the other minor issues, also those rightfully indicated by **Reviewer #4**.