Reviewer #1

To clarify the origins of ASVs, we will modify lines 146-7: "In the game theory literature, this axiom was first relaxed by [43], which termed the result 'random-order values'; [30] referred to them as 'quasivalues'." We will add references to line 150: "ASVs uniquely satisfy Axioms 1–3 (q.v. Theorems 12 and 13 in [43], or Theorem 3 in [30])."

To clarify the notion of accuracy in the global Shapley sum rule, we will add: "The accuracy of randomly drawing from f's predicted probability distribution is distinct from the accuracy of predicting the max-probability class."

In response to R1's statement that the seizure (cf. Sec 4.3) could occur at any point in the time series: Each time series represents 1 sec, whereas most seizures last 30–120 sec, so a seizure is occurring (or not) for the entirety of each time series. We will add a sentence in the text to clarify this and hope this makes the application seem less odd.

Regarding R1's concern about the inefficiency of ASVs for feature selection, we propose to reframe Sec 4.4 as demonstrating a property of ASVs rather than a primary application.

Please also see lines 29–32 below in our response to R3.

Reviewer #3

R3's largest concern is that our paper does not discuss the difference between our approach and [19], which appears to reach a conclusion opposite to ours. To clarify, [19] studies the causality of the *prediction process* rather than the *data-generating process*. In particular, see Fig 2 in [19] which shows the causal process considered there: features ($\tilde{X}$'s) $\rightarrow$ model inputs (X's) $\rightarrow$ model output (Y). As [19] does not consider causal structure among the features themselves, their conclusions are not relevant for the goal of our work: to incorporate causal structure present in the data into model explainability. We will make the following addition to the end of Sec 3.2:

"The distribution $w(\pi)$ incorporates the user's knowledge of the data's causal structure into explanations of the model's predictions. Note that this is quite distinct from other work [19], which considers the model's prediction process itself to be a causal process (features $\rightarrow$ model inputs $\rightarrow$ model output) and finds ordinary Shapley values to be sufficient to explain that process. In contrast, ASVs incorporate causal structure present in the data itself."

R3 finds ASVs' incorporation of causality to be mainly based on intuition. We would distinguish between: (i) gaining causal knowledge about the data, and (ii) incorporating it into a model explainability algorithm. ASVs are solely focussed on tackling (ii); domain expertise or causal inference should generally be employed for (i). It is ASV's handling of (ii) that we claim is mathematically principled: one preserves the 3 important Shapley axioms by restricting to permutations of features consistent with causality. We will clarify this in our introduction to ASVs.

R3 is correct that the ASVs of Sec 4.2 place gender and department choice out-of-causal ordering. To measure *unresolved discrimination* with ASVs, the causal structure needs to used differently – namely, in reverse – to detect whether a protected attribute is causally mediated by a resolving variable [20]. To forecast this to the reader, we will modify line 160 (just after ASVs' definition) to read: "Alternatively, anti-causal orderings can also lead to specific insights; e.g. in Sec 4.2 we define ASVs that detect unfair model decisions."

R3 questions the definition of fairness in Sec 4.2. That definition does not allow just *any* indirect dependence on the protected attribute: only dependence on the protected attribute that is mediated by an explicitly specified *resolving variable* (like free department choice) is permitted. This is a common definition considered by [20] and others.

R3 stated that addressing the points above "could strengthen the paper tremendously". With the proposed modifications, we hope R3 will deem our paper worthy of acceptance.

Reviewer #4

R4 wonders whether ASVs explain the model or the data. The answer (cf. Sec 3.3) lies somewhere in between. As R4 states, "ASVs can be useful if one's goal is to adjust the input to get a different model prediction". However, this goal is not in opposition to "understanding the model" – it cannot be done otherwise. We will note this in the text.

R4 wonders how ASVs advance the state-of-the-art. We claim there is currently no state-of-the-art in causality-based model explainability. See e.g. lines 14–23 in our response to R3 above. For a guideline to incorporate a causal graph into ASVs, see Eq 11. Also see lines 29–32 in our response to R3 above.

References

[19] Janzing et al, "Feature relevance quantification in explainable AI: a causal problem" (2019).
[20] Kilbertus et al, "Avoiding discrimination through causal reasoning" (2017).
[30] Weber, "Probabilistic values for games" (1988).
[43] Monderer & Samet, "Variations on the Shapley value" (2002).