We thank all four reviewers for their helpful suggestions and positive feedback. R1 and R3 noticed that using deep generative models for Bayesian decision-making was an important and largely unaddressed problem. R3 emphasized that our three-step method outperformed more simple alternatives—an important point. R4 appreciated the thoroughness of our experiments, and our substantial improvement on biological data analysis. For each other comment in the reviews, we revised to the manuscript to address it.

Reviewer 1

Posterior collapse is an important issue, and while a thorough treatment of it is largely beyond the scope of our work, we have added to our manuscript a discussion of "Don't Blame the ELBO! A Linear VAE Perspective on Posterior Collapse". Additionally, we have added experiments comparing our method to inference procedures designed to mitigate posterior collapse: monotonic as well as cyclical KL annealing and lagging inference networks. In all experiments, these approaches outperform the VAE, but they are outperformed by the method we propose. For example, in the pPCA experiment (Table 1), the best performing annealing scheme yield a mean absolute error (MAE) of 0.0589, whereas MAE is 0.1026 for the VAE and 0.0247 for our three-step method.

We added an **algorithm box** explicitly describing our three-step method, as well as a discussion of the **computational overhead** of our method compared to a standard VAE. In short, the overhead is not large (roughly a constant factor of three) since our method simply consists of training three VAEs, each with a different loss function. In the pPCA experiment, training a single VAE takes 12 seconds while fitting step 1 and 2 of our method takes 53 seconds. Step 3 has the exact same complexity. In cases where an offline decision is made (for example in biology), this overhead is not a bottleneck.

Because all the experiments are comparisons with existing frameworks, we are confused by the feedback about the **lack of comparative results**. We have attempted to clarify the algorithms we are comparing to by changing the color scheme of Figure 2, 3, 4, to highlight what is related work. There are four or five blue squares in each of these figures, and we now cite a publication demonstrating the existing framework corresponding to each in the caption (except χ-VAE).

For **reproducibility**, we posted the code for our experiments publicly on GitHub; we excluded the link to it in our submission only to preserve our anonymity. Instead, the code used to produce the results in the paper was included in the supplement. During the author response period, we added experimental details in supplementary notes (including dataset source, size, preprocessing, split but also neural networks architecture, hyperparameters, and training / evaluation procedures). Also, we extended the **broader impact section** to note the risks of making decisions based on complex black-box models, and to highlight the importance of worst-case performance guarantees for some applications.

Reviewer 2

We added a discussion about extending the proposed method to a **broader class of losses**, which is an interesting direction. Although we expect that the optimal action will be in closed-form for most practical problems (such as the ones in the manuscript), our method may still provide substantial improvement in this extended setting. Indeed, the risk for each action is a posterior expectation. Further investigations are left as future work.

Our view is that current common practice for making decisions with VAEs, such as using the a single posterior approximation both for calculating predictive densities for and model learning, lacks **formal justification**. Our approach removes this unjustifiable restriction. Regarding **theoretical analysis**, we modified the abstract and the introduction to emphasize that this is limited to pPCA. For **computational overhead**, please see our comment to R1.

Reviewer 3

We agree that **AMCI** is interesting work, and have augmented our discussion of it and cited the extended version it in JMLR. Our method could be extended to incorporate loss-calibrated inference with alternative divergences (such as χ^2), but this is left as future work. One limitation of AMCI not shared by our approach is the runtime: for our biological application (experiment 3), AMCI requires learning a proposal for each gene; there are more than 3,000 genes in our dataset. The runtime for our method does not scale with the number of genes/decisions. Another difference that we address is fitting a model too, whereas AMCI only addresses computing an integral.

As R3 points out, our contribution is independent of **whether IWAE or WW works better** because we choose the best performing model. Nonetheless, we have re-run the experiment with 200 particles (added to the supplement) on the pPCA dataset: WW learns a better generative model than IWAE and the proposed outperforms all baselines in terms of mean average error.

Regarding **R3's questions**: yes, R3 understood the nomenclature well (more details in answer to R1).

Reviewer 4

Regarding the **reproducibility** of the results (resp. our **theoretical treatment**), please refer to our answer to R1 (resp. R2). Regarding the **particular typos**, we have fixed them.