We thank the reviewers for their detailed comments and helpful suggestions. We were delighted by the enthusiasm expressed by all reviewers, and the unanimous decision to place our paper above the acceptance threshold. We are grateful to Reviewer 2 for the confident & positive assessment of the paper, and humbly request that Reviewers 1, 3, & 4 consider raising their scores from 6/10 to 7/10 if they would like this work to reach the broader NeurIPS community. We will first discuss general points raised by multiple reviewers, then address reviewer-specific comments.

• **Concerns that REINFORCE is the only learning rule considered** (**R1**, **R2**): We apologize for this confusion—we were using the term "REINFORCE" loosely, but in fact our framework can easily be applied to the family of policy-gradient learning rules. In fact, two of the rules we considered, AAR and RAR, are non-REINFORCE policy-gradient rules. For example, the RAR rule is derived by optimizing a different objective function compared to that for REINFORCE (see SM Eq.1). In the paper, we branded them as "variants" of REINFORCE, intending to make it easier for readers unfamiliar with RL language. We will clarify this distinction in our revision.

• **Concerns that value function based models are not considered** (**R1**, **R2**): We agree that it would be exciting to explore the space of other learning models, and one of our future directions is to replace the fixed baseline that we currently use, with the value function (and, thus, incorporating a TD-component) in the $\mathtt{RF}_\beta$ model. We also thank the reviewers for suggesting a model comparison with variants of the Rescorla-Wagner model; this would be useful to contextualize our results, and we would be happy to add this to the final paper. Meanwhile, we would like to point out that while the computational cognitive science community tends to focus on TD-learning methods, a substantial number of papers have proposed modeling decision-making behavior with variants of REINFORCE (e.g. Dayan & Daw (2008), Kastner et al. (2019)); while Li & Daw (2011), provide support for the view that humans may use policy-gradient methods instead of value prediction.

• **Concerns about model identifiability** (**R1**, **R3**): We agree that the identifiability of our models should be shown explicitly. We will include this analysis in our revision, as well as an examination of the impact of model mismatch and a broader exploration of hyperparameter space.

Reviewer 1 : **(a)** *The descriptive approach provides limited insight into how animals learn* — We apologize for mischaracterizing our approach as purely descriptive; in fact, we view it as a platform for inferring the parameters of normative models / testing normative hypotheses about animal learning. We would agree that our finding (that negative baselines are required to account for animals' learning trajectories) is non-intuitive, and will add supplemental analyses (e.g. conditioning on incorrect choice when bias is positive) to provide more insight into how the rule affects choices. **(b)** *Correlating weights with empirical measurements like accuracy* — This is an excellent suggestion and is straightforward to show; we will certainly include supplemental figures to address this.

Reviewer 2 : **(a)** *Primary behavioral data to show the learning curves* — Great point, we will add learning curves and other analyses to show that the inferred rules *do* indeed match behavior. **(b)** *Differences in learning rates for bias and stimulus* — We should have thought of that! We will certainly discuss this in our revision, thank you for pointing this out. **(c)** *Include additional references/lessen claims of novelty* — Thank you for the additional citations, we will add these and remove the novelty claim in lines 133-4. **(d)** *Where is* $\mathtt{RF}_\beta$ *in Figure 4c/d?* — Apologies, we mistakenly labeled $\mathtt{RF}_\beta$ as R+B, and $\mathtt{RF}_K$ as R in Fig. 4c; we will fix this in the revision.

Reviewer 3 : **(a)** *Evaluation of the quality of fit* — Thank you, we will add a quantification of performance in terms of increase in log-likelihood on test set data. **(b)** *Sub-optimal behavior calls for direct validation* — Another great point, we will use primary behavioral data to quantitatively validate the predictions made by the $\mathtt{RF}_\beta$ model. **(c)** *Statistics of simulated vs. animal data* — Again, great point, we will include this in our revision. **(d)** *Outlier excluded in Fig. 4a* — Apologies, this phrase was a typo and will be removed. **(e)** *Why is* $\mathtt{RF}_\beta$ *missing is Fig. 4d* — Apologies, we mistakenly labeled $\mathtt{RF}_\beta$ as R+B, and $\mathtt{RF}_K$ as R in Fig. 4c; we will fix this in the revision.

Reviewer 4 : **(a)** *Clarifying the relationship to other work (ref. 19)* — Thank you for pointing this out, we will absolutely make the contributions of this work clear in relation to the computational framework from [19]. **(b)** *Comparing full* $\mathtt{RF}_\beta$ *model to* $\mathtt{RF}_\beta$ *without noise* — This is an interesting point, and we will include a comparison of the current $\mathtt{RF}_\beta$ fit (REINFORCE with baseline, with noise) to one without a noise term in our revision. We respectfully disagree with the comment that a noise term is unnecessary for the $\mathtt{RF}_K$ models. In fact, it was by studying the structure of the noise component that we were inspired to consider the $\mathtt{RF}_\beta$ model. As was also observed, $\mathtt{RF}_\beta$ is still not perfect for the animal we display in the second dataset, but through examining the structure of the retrieved noise component, we may be able to suggest a better normative learning rule (which may be exactly what was suggested by the reviewer – a learning rule with time-dependent baseline parameters).