We thank the reviewers for their constructive feedback and suggestions to strengthen this work. All points raised will be addressed in the revised version of the manuscript.

R1: As suggested we will include motivating examples in the introduction (see new fig under **R2**). We will emphasise our explanations are local (what is expected to happen given the current state and action), and that our target audience is RL practitioners trying to understand agents. We will ensure that acronyms and mathematical notations are specified. This includes changing $\pi^*(a|s)$ to $\pi^*(s)$ and explaining that θ^- is the target network parameters. We will also clarify the instance of our chosen contrastive explanation.

HCI Comparison with [10]. Our approach shows what the agent expects to be the consequences of the action while [10] highlights the components of the environment that lead an agent to select an action. These are complementary explanations, and we feel that a user study comparison on any particular task would simply show if information about the environment or future state is more relevant to that task.

Large state-action spaces and potential pitfalls. This is a suggestion for potential future research which was mentioned in the conclusion. Unfortunately, combining concepts with RL explanations will require significant further work.

Code is missing. Apologies for any confusion. We meant to say we will provide the full code on publication. We could not make our code repository public without breaking the anonymity of the review process.

R2: *In blackjack why doesn't the dealer's hand change?* The dealer shows one card, and then the agent plays in response. The agent learns that once played, the revealed card never changes, as shown by the belief maps.

Interpreting figure 2c and 2d. These are four terminal states created to make the reward function deterministic: *lose*, *win*, *bust*, and *draw* and one transition state *hit, no bust*. The figure will be updated so the states are clearly labelled.

Compare your explanations to forward simulation. This is a great idea. N.B. the unpublished work [van der Waa et al., 2018] performs these forwards simulations. We have added Fig 1 to the paper to illustrate these differences. It shows the belief of a blackjack agent mistrained on a small and biased replay buffer. This leads the agent to believe that hitting on a 15 always leads to a 20. Meanwhile the forward simulation on the right shows the "real" behaviour of the system and therefore provides no insight into problems at training time.

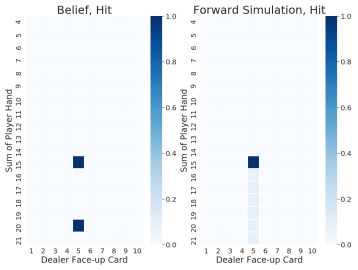

Figure 1: Belief map (left) vs forward simulation (right).

We thank the reviewer for directing us to Dayan [1993], which is closely related, albeit outside the context of XAI. We will discuss this carefully in the final version.

How theorem 1 implies it is never possible to produce a post-hoc interpretation. Theorem 1 shows that a general explainer can not always generate correct post-hoc interpretations, by showing an ambiguous example where it fails. The reviewer is correct, it is sometimes possible to give correct post-hoc interpretations and there is always a trivial explainer that exactly describes the behaviour of any given agent and no others. In contrast theorem 2 guarantees our explainer is always correct.

Fuzzy DQN belief estimates in cartpole We suspect the fuzzy estimates may be due to: (1) low learning rate; the sparse belief map updates could adversely affect the magnitude of gradients (also present in tabular belief map trained at a very low learning rate). (2) distribution of the experience replay buffer; poor experiences can cause a shift in the agent which is exacerbated by batch learning. This is another example of the insights provided by our method.

Bias not present in Figures 4 and 5. This text was accidentally left in from an earlier draft with different results. You are correct that the current figures do not show any such bias.

Why isn't col 4 of fig 6(b) blank since col 1 and 2 match. Although every cell visited in subfig 1 is also visited in subfig 2, the strengths of the belief in those states differs due to the discount factor of future states ($\gamma = 0.9$). Also note the caption for column 3 should read $\min(0, \mathbf{H}(s, a_1) - \mathbf{H}(s, a_0))$.

R3: *Evidence that the belief maps indeed demonstrate the agent's 'intent'.* The new Fig. 1 (see R2.3) shows the intention of a mistrained agent. Moreover, theorem 2 is a proof that the explanations really are consistent with what the agent expects to happen, and the result has also been validated numerically.

Why the formulation is desired beyond other kinds? We don't claim that our formulation is always more desired, however it provides a new type of information no other method offers. These explanations are more like those psychologists say people offer (see intro) and can be used to identify a gap between replay buffers and the true environment (see fig. 1).

Why is $p(s_t + n|s_t, a_t, \pi)$ useful to understand the agent intent compared to $p(a_t + n|s_t, a_t, \pi)$? It is simple for our method to compute the latter. However, we feel it provides less insight: In taxi it would show how often the taxi moves in each direction over an entire episode, while in blackjack it would show the average number of sticks and hits. Meanwhile our approach returns the route of the taxi, and the sequence of card values.

It is incorrect to say there are just two groups of interpretation methods. Yes this is an oversimplification. We will clarify.

R4: The reviewer did not ask for any clarifications in the rebuttal. Nevertheless we greatly appreciate the kind words and the affirmation of the importance of this research.