Thank you all for your positive feedback!

Reviewer 1: While we agree that this work in part combines previous ideas where it was not previously clear how to do so, we also believe that this work contains several key contributions that are both original and highly non-trivial. A full list of original contributions is presented in Appendix A.

For example, past uses of PBSs in RL have been limited to special cases[1] because it was not clear how to determine another player's belief distribution. We present a counter-intuitive but elegant solution to this problem in the last part of Section 5.2 and which is described in more detail in Appendix G. The description in Section 5.2 was brief due to space constraints, but we intend to expand this description with the extra space granted in the camera ready version.

We are confident this work will have a wide impact on the fields of RL and game theory. These two research communities have previously developed very different techniques for solving perfect-information versus imperfect-information games. This paper for the first time presents a common link between those two separate research threads.

Regarding multiplayer games: we agree this is an important research direction and mention it briefly in the conclusion as a direction for future work. There are major theoretical challenges that make such an extension highly non-trivial. In particular, PBSs no longer have unique values. While it is possible that the techniques described in this paper may do empirically well, extending them in a way that is theoretically sound is a major open question.

Reviewer 2: In addition to results in poker, the paper also present results in Liar's Dice that show our algorithm converges to a Nash equilibrium. We chose Liar's Dice over Leduc hold'em for several reasons: we wanted to show ReBeL works in two-player zero-sum games in general (not just poker), Liar's Dice is easily modified to be larger and more complex, and Liar's Dice has relatively simple rules.

Furthermore, we will open source our ReBeL implementation for Liar's Dice to ensure the research community can reproduce, understand, and build upon our results.

While we only played against one top human in HUNL, we believe the number of hands played is the relevant metric, particularly because we played against one of the strongest top humans in the field (Dong Kim was considered the strongest of the four humans that played against Libratus and did the best out of the four). ReBeL played enough hands to clearly show a statistically significant victory over Dong Kim. Playing against more top humans rather than doing more hands against a single top human in HUNL is problematic for a number of reasons:

- In order to incentivize strong play, substantial monetary compensation is essential.
- Due to the prevalence of poker assistance software and due to the amount of money on the line, a match against a top human must either be carefully monitored to ensure no cheating, or must involve a great deal of trust.
- Playing against more humans would necessarily mean a drop in the quality of the players. Dong Kim is considered one of the strongest players in the world, and it is likely that any other player that could be recruited would be weaker.

We tested Algorithm 1 using a warm-started policy in TEH (the purple line in Figure 2).

Please see our response to Reviewer 1 regarding multiplayer games.

We used the same LBR parameters (listed in the caption of Table 1) that were used to evaluate DeepStack.

We will improve the clarity in Section 3 and Section 4. The notation is for simultaneous-move games, but can be extended to games like poker by having a player choose a null action in turns where they cannot act. We will clarify this in the camera ready.

On line 24 we use one-ply to mean we look one move ahead.

Regarding reproducibility, we will open source our implementation of ReBeL on Liar's Dice.

Reviewer 3: We will use part of the extra page in the camera ready to expand our discussion of the experimental results.

Regarding experiments against a top human, please see our response to reviewer 2.

We will describe the Δ notation and explain exploitability in more detail. Thank you for pointing out the typo!

[1] Specifically, fully cooperative games, where players have no incentive to lie about their policy, and a limited class of adversarial games where players' belief distributions are always common knowledge. While PBS value functions have been used in the past in adversarial games too, it was unknown how to train them using RL.