We would like to thank the reviewers for their thoughtful comments. Our paper makes the empirical point that state of the art deep learning techniques don't work in collaborative settings and we need human data to fix this. It also introduces the Overcooked environment as one in which coordination is key to achieve high reward. We are glad that both R1 and R2 see the significance of these contributions, and R1 is correct: we have gotten a lot of interest in the environment, and so we are planning to release code soon. However, R2 and R3 had some concerns.

R2 ***would like to see further experimentation***, and points out that one question of particular interest is when it is useful to model the human, and how that relates to a more diverse set of environments. We agree that these are two important research directions, as are many others, some of which we listed in our future work section. We see the paper as seeding many lines of inquiry, many more than we can investigate ourselves – we packed quite a bit in the paper and the appendix already. Based on initial interest, it seems that many groups will investigate these questions.

R3 ***is concerned that the HRI community has already made our empirical points***. We ask R3 to consider that *there is a difference between previous results that compared planning-based approaches and our work which emphasizes their deep learning counterparts*. We think there's value in making these points in the context of these newer techniques. There is also more nuance to our results that R3 is not giving us credit for. First, PBT is supposed to do better than pure self-play because it's designed to be robust. We find that it isn't. Second, as R2 noted, the qualitative analysis of what exactly happens when you run RL and self-play here and try it with humans is illustrative.

In addition, note that *it was not obvious a priori that self-play is insufficient*. OpenAI Five has played Dota cooperatively with humans, while For The Win agents have played Quake Capture the Flag with a human teammate (see citations in main paper). This has inspired many people to think that collaboration as a whole is addressable this way – talks about alternative methods for HRI inevitably lead to the question "what about self-play, wouldn't that solve it?"

We do think our results are of interest to the HRI community as well. Most HRI papers, including the ones R3 cites, replace an optimal human with a noisily optimal human (where data is used to learn the reward function or goal parameters). Obviously noisy rationality is a better model than perfect rationality, but we show something stronger: in our setting, *even behavior cloning does better than rationality*. This is because noisy optimality attributes mistakes to randomness (which can't be predicted), while behavior cloning can learn *how* the human makes mistakes, and so can correct for them. Our qualitative analysis shows that there are *systematic* ways in which humans deviate from optimality, and so we expect the black-box approach would work significantly better than an approach based on noisy optimality. Unfortunately, this discussion didn't make it into our paper because our target audience was different.

We did run an experiment to test this for this rebuttal. We created a simple hierarchical agent A, which chooses a subgoal (e.g. "get an onion"), and then chooses an action to pursue the subgoal. Both choices are modeled using noisy optimality, with β chosen so that the behavior looks "human-like". We trained PPO$_A$ (analogous to PPO$_{BC}$) and evaluated against H_{proxy}. As expected, PPO$_A$ performed better than SP but worse than PPO$_{BC}$. We could likely improve A with more time, but we expect it would not change the qualitative results.

We think these experiments suggest that the HRI community consider how to obtain human models that can predict *systematic* deviations, unlike noisy rationality. If nothing else, the experiments are valuable because they demonstrate that the Overcooked environment is an excellent testbed for HRI techniques.

R2: Thanks for the reference on the importance sampling line of work! We agree that these techniques could help. One difficulty is that human-AI gameplay visits previously unseen states as the agent learns to deal with human failures. While behavioral cloning suffers from distributional shift, importance sampling would fail entirely.

In Figure 5a, we used layouts as a covariate, but did not test interaction effects between layout and the agent type. So, the effect we got was in aggregate – it does not mean that BC helped in every layout individually. Running the test on Figure 10a produces similar but less pronounced results. There is still a significant main effect for agent type on reward ($F(2, 251) = 3.56$, $p = .03$), and the post-hoc analysis with Tukey HSD corrections shows that PPO$_{BC}$ significantly outperforms PBT ($p = .03$) and non-significantly outperforms SP ($p = .12$). We will add this to the appendix.

R3: The imitation learning condition is an agent that acts like a human. We included it because we have also heard (less often) the suggestion that human-AI collaboration could happen just by having the AI imitate the human.

Yes, the layout order was fixed, which could lead to learning effects. If we randomized the order, different humans would have different learning effects for the same layout, leading to higher variance results and requiring both more data to train models and more evaluation in order to detect a statistically significant effect.

With short horizons and few deliveries, policies of different skill could get the same reward. However, long horizons make DRL training harder, and might bore human players. We chose a horizon of 400 to balance these tradeoffs.

One potential explanation for our deep RL results is we didn't properly tune our self-play algorithms. We included the planning experiments to demonstrate that the problem lies squarely with the assumption of optimality, not DRL training.