We thank all the reviewers for their exceptionally careful reading of the paper, and their very helpful comments.

Two reviews commented that comparisons to more other algorithms would be helpful. The algorithm of clearest relevance here is that introduced in Jaques et al. (2018). This algorithm can be seen as introducing a bias encouraging the speaker to increase positive listening. We already note in the paper that this bias did not help with solving the MNIST task, but will emphasize this point. Further, we elaborate the discussion of this in three ways. Firstly, we will display the results in full, including the analogue of Figure 2, rather than simply state that the algorithm from Jaques et al. does not outperform the no-communication baselines. Secondly, we will run a comparison using the bias from Jaques et al. on Treasure Hunt. Finally, we will explain why we believe the bias from Jaques et al. is unhelpful in this setting. In short, this is because for a fixed listener, the speaker policy which optimizes positive listening has no relation to speaker's input. Thus this bias does not force the speaker to produce different message for different inputs, and so does not increase the learning signal for the listener. This is true when the observations of the speaker and listener are independent (such as in the MNIST task). In that case, the expected positive listening for the speaker is a function only of its message, not its observation. And so the best policy to increase this expectation doesn't depend on the observation. In Treasure Hunt, it's not quite true that the speaker should ignore its state to increase positive signaling - it should only use the state to deduce the listeners state, and therefore how it might best be influenced.

The other suggested baselines are RIAL from Foerster et al. 2016 and Sukhbaatar et al. (2016). RIAL is very similar to the baseline used in the paper; the difference is in the underlying RL algorithms (DQN in RIAL, and A3C here). Sukhbaatar et al. (2016), and other approaches we are aware of, use a differentiable model of communication. While also interesting, this is not the domain we are examining here, so we have not run these baselines.

There were two suggestions for improvements to the related work section; other emergent communication work (e.g. COMA from Foerster et al. (2017), Sukhbaatar et al. (2016)), and the extensive literature on auxiliary losses, particularly in MARL. We agree with both, and we updated the paper accordingly.

Two of the reviews expressed a desire to see these algorithms tried on more complex tasks, which we agree is an important direction. However, we believe that it is a novel contribution to achieve emergent communication in any many-step RL environment with non-differentiable communication channels. We therefore think that these methods represent an important advance, and are likely to contribute to solutions of harder communication problems, even if more advances are needed. We updated the discussion in the paper to consider the possible scaling of the algorithm in more detail - adding discussion of $n > 2$ agents, and discussing the implications of larger communication channels.

Reviewer 1 asks two related questions; why (2) is formulated with trajectories, and why line 12 of algorithm 2 includes the speaker's hidden state. This is because the message policy is recurrent (which is necessary for a good protocol in Treasure Hunt, due its partial observability); so the calculation of the message policy needs the hidden state, and the natural formulation of the mutual information in positive speaking is with the trajectory rather than the state.

Reviewer 1 is correct to point out that the baselines in the MNIST task are not heavily optimized; with a curriculum approach as suggested, we would not be surprised if this task could be solved (at least sometimes) without the biases we use. However, the main purpose of this task is to examine how these biases affect learning, we don't think optimising the baselines is as important as in other contexts.

Reviewer 1 comments that runnable code or message information would be helpful. In an appendix, we will provide a wider variety of message protocols, summarizing them in a similar way to those examined in the paper.

Reviewer 2 comments that it would be good to provide a plot or table with the final reward across all runs. We will add this to Table 2. As the reviewer notes, this will have a significant difference between our methods and the baseline, due to the increased rate with which the agents learn to communicate with the biases.

Reviewer 2 asks why the agents fail to achieve optimal performance, and notes that the performance is little better than the baseline conditional on communication happening. The answer to this is in Section 4.2; these methods are primarily aimed at getting communication to emerge in the first place, rather than at reaching the global optimum protocol. There is still much work to be done in joint exploration in emergent communication.

Reviewer 2 asks about the total loss used. The loss is the usual loss for the RL algorithm being used (REINFORCE for the MNIST problem, and A3C for Treasure Hunt), plus the losses we outline. We will clarify this in the paper.

Reviewer 2 asks about why we use multi-step, rather than single-step, CIC. Intuitively, we expect messages from several steps previously to be relevant to the listener's decision. Empirically, we found the multi-step version to be superior; we state this in the text and provide an ablation study for this (perhaps in an appendix, for want of space).

All minor corrections to notation and spelling are gratefully received, and we will make them for the next version. We will also improve the legibility of Figure 2, and add a more detailed explanation of Equation 6 in the supplementary material.