as listener	as speaker: R	as speaker: H	
R	0.85	0.50	random
		0.45	direct
		0.61	belief (ours)
H*	0.5	0.77	
	0.45		
	0.57		

Table 2: Belief evaluation results for the driving game. Driving states are challenging to identify based on messages alone (as evidenced by the comparatively low scores obtained by single-language pairs) . Translation based on belief achieves the best overall performance in both directions.

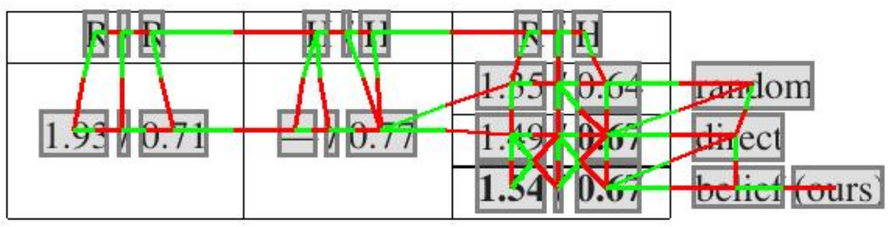

Table 3: Behavior evaluation results for the driving game. Scores are presented in the form "reward / completion rate". While less accurate than either humans or DCPs with a shared language, the models that employ a translation layer obtain higher reward and a greater overall success rate than baselines.

cases, while a model trained to communicate in natural language achieves somewhat lower performance. Regardless of whether the speaker is a DCP and the listener a model human or vice-versa, translation based on the belief-matching criterion in Section 5 achieves the best performance; indeed, when translating neuralese color names to natural language, the listener is able to achieve a slightly higher score than it is natively. This suggests that the automated agent has discovered a more effective strategy than the one demonstrated by humans in the dataset, and that the effectiveness of this strategy is preserved by translation. Example translations from the reference games are depicted in Figure 2 and Figure 7.

Driving game Behavior evaluation of the driving game is shown in Table 3, and belief evaluation is shown in Table 2. Translation of messages in the driving game is considerably more challenging than in the reference games, and scores are uniformly lower; however, a clear benefit from the belief-matching model is still visible. Belief matching leads to higher scores on the belief evaluation in both directions, and allows agents to obtain a higher reward on average (though task completion rates remain roughly the same across all agents). Some example translations of driving game messages are shown in Figure 8.

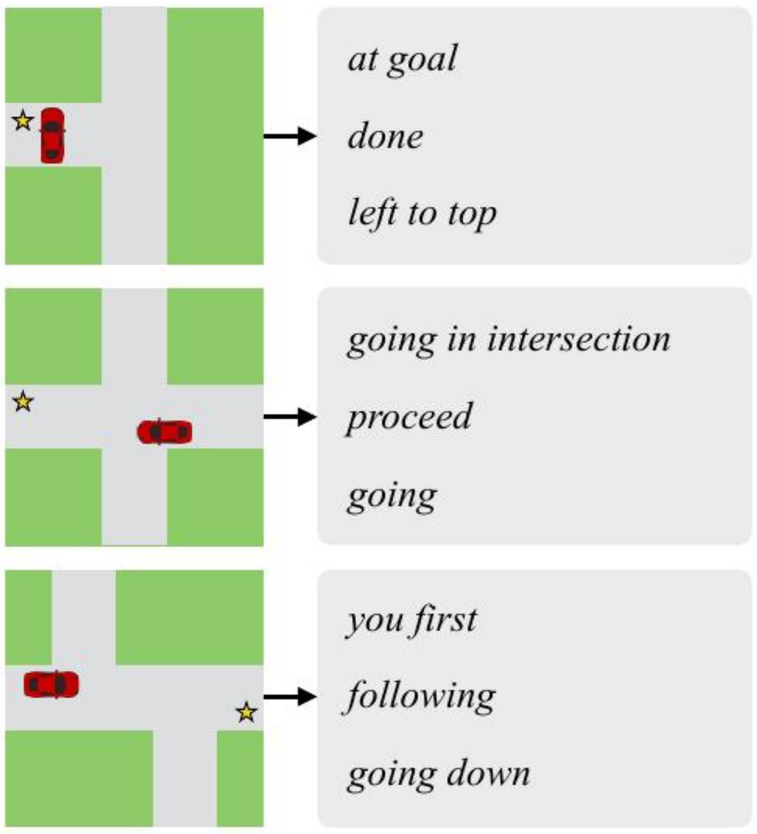

Figure 8: Best-scoring translations generated for driving task generated from the given speaker state.

9 Conclusion

We have investigated the problem of interpreting message vectors from deep networks by translating them. After introducing a translation criterion based on matching listener beliefs about speaker states, we presented both theoretical and empirical evidence that this criterion outperforms a conventional machine translation approach at recovering the content of message vectors and facilitating collaboration between humans and learned agents.

While our evaluation has focused on understanding the behavior of deep communicating policies, the framework proposed in this paper could be much more generally applied. Any encoder–decoder model (Sutskever et al., 2014) can be thought of as a kind of communication game played between the encoder and the decoder, so we can analogously imagine computing and translating "beliefs" induced by the encoding to explain what features of the input are being transmitted. The current work has focused on learning a purely categorical model of the translation process, supported by an unstructured inventory of translation candidates, and future work could explore the *compositional* structure of messages, and attempt to synthesize novel natural language or neuralese messages from scratch. More broadly, the work here shows that the denotational perspective from formal semantics provides a framework for precisely framing the demands of interpretable machine learning (Wilson et al., 2016), and particularly for ensuring that human users without prior exposure to a learned model are able to interoperate with it, predict its behavior, and diagnose its errors.