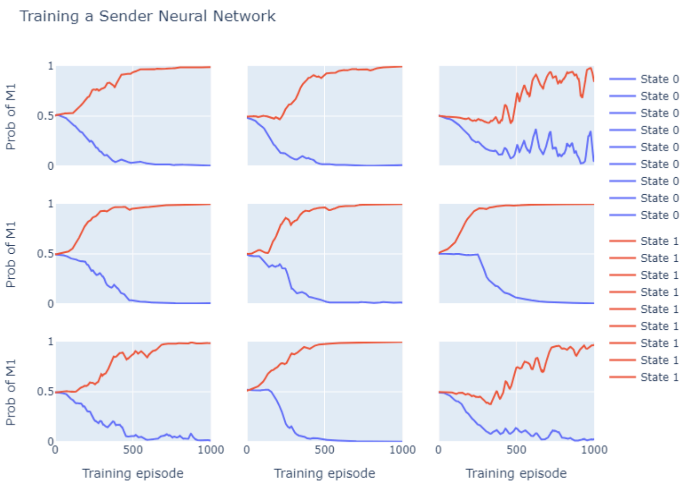
Context is Everything: Evidence Against the Lewis Signalling Game as a Model of the Minimal Conditions for the Evolution of Meaning

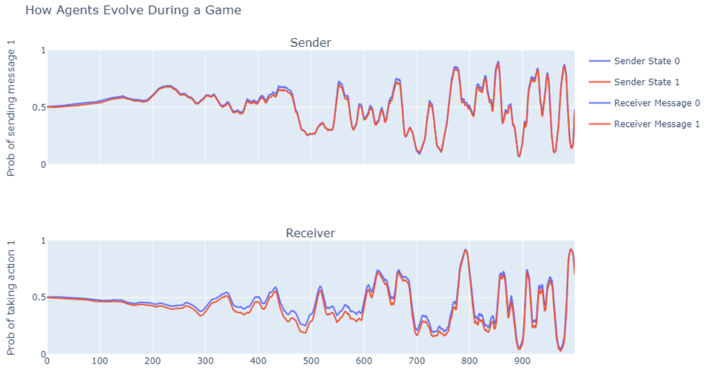
By Elijah Sandler



This project uses a neural network built using Python's Keras library, which allows users to build configurable neural networks.[1] The neural network will be applied to several problems within the scope of the Lewis signalling game, focusing mainly on the most basic form of the game, which has two states, messages and actions.

First, it is important to show that the model can learn to play in a game theoretic manner. Shown left is an image of nine runs training a sender from scratch. The x-axis represents the number of episodes the model had been training for, and the y-axis represents how likely the model is to play message 1. The sender was partnered with a receiver who would always play the action directly corresponding to the message received, and would not change over time. In this game, the model was consistently able to quickly learn the optimal strategy of sending message 0 when seeing state 0, and message 1 when seeing state 1. This is shown in the graph by the divergence of the two lines of each plot, which represent the preferences of the model given each of the two environmental states.

The next test is to determine if it is possible to train a sender and a receiver simultaneously. This is the test that is most analogous to the traditional evolutionary game theory notion of the signalling game, where both players enter with zero information and attempt to create meaning by coordinating their responses. In this test, the neural networks were unable to derive meaning. Shown below are two stacked plots, with axes identical to the graph above, but this time only representing one iteration of the learning phase. When run in conjunction, both models fail to extract meaning from the game. This is illustrated by the fact that the orange and blue lines for each model are basically identical, meaning that the state or message the model received did not alter its decision making. In fact, over 1000 episodes of training, no information was passed or received in the game at all.



[1] The code for this project is available in this GitHub repository.

What went wrong? Essentially, the two players only get rewarded if they successfully coordinate. If neither player has any idea of what the other player is doing, and what the other player is doing is constantly changing, there is no correlation between the states, messages, actions, and rewards that the neural networks can interpret. Because the neural networks learn recurrently, a model that has trained for 1000 turns but has never received information from its partner will produce the same output as a model that was just created, and hence has no information. Because both models are in this position, there is no way for them to interpret what their partner is doing. The Roth-Erev reinforcement learning model, which is the only model that can consistently solve a 2x2 Lewis signalling game from scratch, ignores this issue by brute forcing the problem. This may lead to some questions about whether pattern recognition is necessary for meaning, but this paper won't address those concerns.

It has been noted by Bruner *et al*. that humans are very good at solving abstract versions of the Lewis signalling game, and are consistently able to solve versions that are more complex than the simple 2x2 game. I suggest that the discrepancies between the human and machine results are not indicative of a failing of the neural network, but rather that these results provide evidence that the Lewis signalling game does not fully capture the minimal requirements for the evolution of meaning, and that the reason for this is that it is missing one key attribute.

Bruner *et al.* note that "[an] important and distinctive feature of economics experiments is that they are by and large context-free," and that they will account for this in the setup for their experiment. However, I do not believe that this is the case, as the instructions "provided subjects with knowledge of the game and the payment structure employed." Additionally, the players were told after each round the full results of the game, including which state, message, and action was picked.

The neural networks don't know if they are the sender or the receiver, but humans do. In fact, in the code for these models, both the sender and receiver are modelled by an instance of the same class, because that's how the players in the original Lewis signalling game are portrayed. They don't have a concept of the rest of the game; the receiver doesn't just not know what the state is, *they don't know that the state exists*. When Bruner *et al.* provide their players with "complete knowledge of the game," they are creating a disanalogy sufficient enough that their empirical experiments no longer support the Lewis signalling game as a representation of the minimal conditions for the emergence of meaning. This is an oversight that has been made frequently when discussing the results of empirical Lewis signalling game experiments.

Bibliography

Bhatia, Sudeep, and Russell Golman. *A Recurrent Neural Network for Game Theoretic Decision Making*, www.cmu.edu/dietrich/sds//docs/golman/a-recurrent-neural-network-for-game-theoretic-decision-making-bhatia_golman.pdf.

Bruner, J., O'Connor, C., Rubin, H. *et al.* "David Lewis in the lab: experimental results on the emergence of meaning." *Synthese,* vol. 195, 2018, pp. 603–21. https://doi.org/10.1007/s11229-014-0535-x.

Erev, Ido, and Alvin E. Roth. "Predicting How People Play Games: Reinforcement Learning in Experimental Games with Unique, Mixed Strategy Equilibria." *The American Economic Review*, vol. 88, no. 4, 1998, pp. 848–81.

Korbak, Tomek. "Introduction to Lewis Signaling Games with Python." *Tomek Korbak*, tomekkorbak.com/2019/10/08/lewis-signaling-games/.

Lewis, David. *Convention. A Philosophical Study*, Harvard University Press, 1969.

Skyrms, Brian. *Signals Evolution, Learning, & Information*. Oxford University Press, 2013.