Emergence of Honest Signaling through Learning and Evolution

David Catteeuw

December 2014

The emergence of honest (or reliable) signaling is a *multi-disciplinary* problem. Linguists and philosophers have long wondered how conventions, such as human language, can emerge without a pre-existing language. Biologists noticed that the many signals in nature can only exist because they are honest. Otherwise they would be ignored and so, not worth the trouble sending. Economists created a real breakthrough by recognizing that many interactions are characterized by private information—where one party knows more than the other and signals may, or may not, reveal that information. It explains, for example, why the free market does not work for health insurance: those willing to buy costly insurance are most likely those who expect to need it the most. I contributed to this research in three domains: common interest; costly signals; and costly, social punishment.

One reason why signals are honest is *common interest*: both the sender and the receiver of the signal benefit from conveying the correct information. Under common interest, the only question that remains is how a signal acquires its meaning. One explanation that may also explain the origins of language is that this happens by chance. My findings support this idea. In Chapter 3,

- I introduce a new behavioral rule, called 'win-stay/lose-inaction' or 'WSLI:' initially play random, repeat forever what was once successful. When two repeatedly interacting players apply WSLI they always end up signaling honestly in all Lewis signaling games (the standard game-theoretic model to study the emergence of signaling under common interest). I prove that the expected number of iterations is only polynomial in the number of signals. No such algorithm was known before.

- I show that three well-known reinforcement learning algorithms (Q-learning, Roth-Erev learning, and Learning Automata) behave exactly like WSLI in Lewis signaling games for certain parameter configurations.

- While WSLI is not robust to errors, these reinforcement learning algorithms are robust for certain parameter configurations and still reach honest signaling in a polynomial number of iterations.

Economists and biologists independently discovered that when interests conflict signals may be honest if they are costly. This is known as the *'handicap principle'* and is almost exclusively studied assuming infinite populations and by means of static equilibrium analyses—verifying if honest signaling is an equilibrium while ignoring the dynamics that may or may not lead to it. In Chapter 4,

I apply learning and evolutionary dynamics in finite populations to the Philip Sidney game:

- In many cases where honest signaling is an equilibrium, it does not emerge: equilibrium analyses wrongfully predict honest signaling.
- Dynamics reveal (partially) honest signaling in some cases where it is not an equilibrium: equilibrium analyses fail to predict (partially) honest signaling.

Costly, social *punishment* is known to promote the evolution of cooperation but its effect on the evolution of honest signaling is merely studied. In Chapter 5, I distinguish four ways of deviating from honest signaling: the sender can lie or be timid and the receiver can be greedy or worried. I extend the Philip Sidney game to explicitly allow for punishment of such behavior and study its effect on the evolution of honest signaling:

- When punishment targets lying individuals, honest signaling emerges also for cost-free signals. So, punishment provides an alternative to the handicap principle.
- When punishment targets greedy individuals, honest signaling emerges also in cases with strong conflicts, similar to the punishment of defectors to promote cooperation.
- The evolution of honest signaling does not benefit from punishment of timid or worried individuals.