Roth-Erev learning in Signaling and Language games

David Catteeuw [a]   Joachim De Beule [b]   Bernard Manderick [a]

[a] *Computational Modeling Lab, Vrije Universiteit Brussel, Pleinlaan 2, 1050 Brussels*
[b] *Artificial Intelligence Lab, Vrije Universiteit Brussel, Pleinlaan 2, 1050 Brussels*

Abstract

The relation between Lewis' signaling game framework and Steels' language game framework is discussed. The problem of pooling equilibria in signaling games is approached from both angles. Previous results about the requirements on learning for escaping pooling equilibria and achieving convergence are refined, and it is discussed why Roth Erev learning with forgetting should be a good learning rule in this context. This is confirmed empirically in simulation.

1 Introduction

Signaling, or the transmission of meaningful information, is an essential part of human activity [19, 17, 12, 18]. It is also believed to play a major role in the organization of biological systems and their capacity to generate complexity [4, 20, 23]. It is therefore desirable to have a good understanding of the mechanisms by which efficient signaling systems may be established, both from an analytic (scientific) point of view and from the more synthetic perspectives of artificial life and artificial intelligence.

The origin and evolution of signaling and language has been investigated previously and independently in the domains of (artificial) language evolution (see e.g. [24]) and signaling games (see e.g. [19, 23]). A first contribution of this paper is to clarify the relation between these two fields. As they have traditionally focused on different aspects of signaling, connecting them immediately allows for a transfer of knowledge. For instance, within the context of signaling games, it is known that the appearance of suboptimal equilibra, the so called "pooling equilibria", may prevent the emergence of optimal signaling [5, 16].[1] It was also already suggested that "forgetting" can help signalers to avoid or escape from such suboptimal behavior [6]. More precisely, it can reduce the stability or the size of the basin of attraction of pooling equilibria. The same effect can be achieved in population models by introducing stochasticity, see [14, 6, 1] for more details. By virtue of the relation between signaling games and language games as identified in this paper, it becomes clear that these findings are instances of a more general finding, namely that the evolution of conventions requires signalers to remain adaptive indepedent of time "ergodic" [11].

The second contribution of this paper is a further resolution of the problem of suboptimal equilibria in signaling games. It is known that the "Win-Stay/Lose-Randomize" learning rule – embodying a drastic form of forgetting – in theory always leads to an optimal equilibrium [6]. The reason for this is obvious: as long as signaling is suboptimal, signalers will randomly change their signaling behavior. This renders (partial) pooling equilibria structurally unstable and only leaves optimal signaling as valid (stable) alternatives. However, as was already shown in [11], this "absorbing state argument" is not very strong in any practical sense, precisely because it relies on trial-and-error, that is, on *random* drift. A more acceptable solution to the problem of pooling equilibria should not introduce a random but a *guided* drift towards more optimal signaling. As is shown in the remainder of this paper, this can be accomplished for instance with Roth-Erev learning with forgetting.

[1] The main body of the paper provides a more detailed account of the issues involved, including the proper definitions.

2 Signaling Games and Language games

Signaling games were introduced by philosopher David Lewis in order to provide a game theoretic approach to the problem of the emergence of conventions [19]. A convention, like a language, is a system of arbitrary rules, shared between different players, and enables them to exchange "meaningful information". Meaningful information, in turn, is "any difference that makes a difference" [7]. For instance, intuitively the observation of a signal (a difference) is meaningful if it indicates which action to take from a set of possible (different) actions.

More formally, a signaling game is a two-player extensive form game such as the one shown in Figure 1. First, Nature provides one player, the *Sender*, with one of T possible world states $t_1, \dots, t_T$. The world state is private information to the Sender and hence in game theoretic terminology it would be called his *type*. Based on the state, the Sender sends one of S signals $s_1, \dots, s_S$ to a second player, the *Receiver*. The Receiver in turn selects one of A actions $a_1, \dots, a_A$. Depending on the selected world state, the signal and the action, a payoff is determined for both Sender and Receiver. In the remainder of the paper we assume that the number of states, signals and actions are the same ($T = S = A$), and that with every world state there corresponds exactly one "correct" action (i.e. state t_1 with action a_1, t_2 with a_2 etc.). If the Sender manages to select the correct action for the selected world state, the game succeeds and both players rewarded with a payoff of $u = 1$. In any other case the game fails and the players get nothing (payoff $u = 0$).

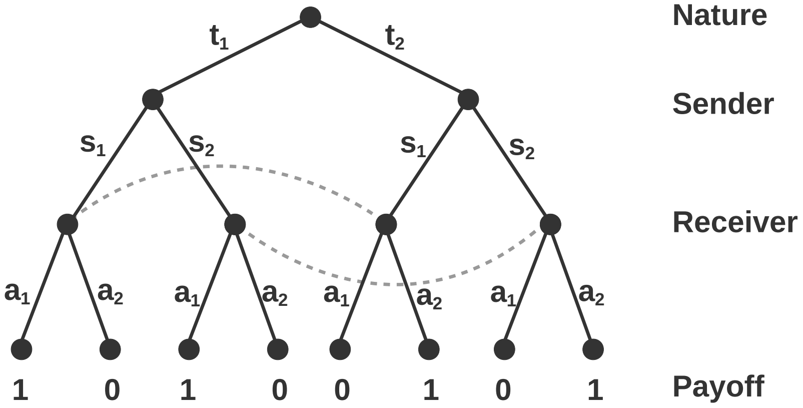
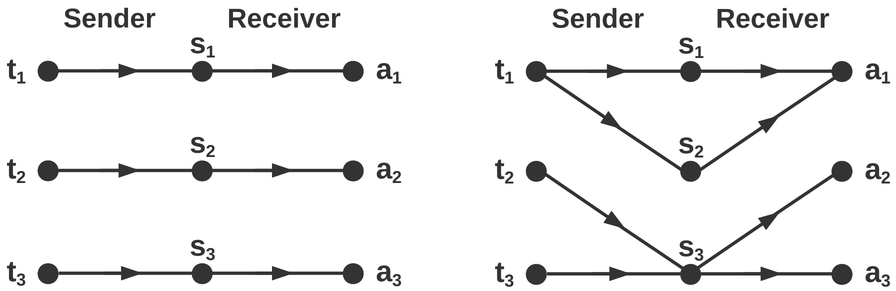


Figure 1: A signaling game is an extensive form game as explained in the text. This one has $T = S = A = 2$ states, signals and actions respectively. Since the Receiver is not directly informed of Nature's state t_i, he cannot distinguish between the nodes connected by a dashed line.

Under these conditions, the relation between signaling games and language games becomes particularly clear. Luc Steels, inspired by Wittgenstein's notion of language games [25] and seeking to apply it within the context of artificial intelligence, proposed to study the evolution of language by implementing robots "playing language games" [24]. Formally, such a game is again a two player extensive form game. In this case, both players perceive a shared "context", e.g. a number of objects. One player (the Speaker) then describes one of the objects to the second player (the Hearer). The game succeeds if the Hearer can identify the object based on the Speaker's description. It is clear that, under the stated assumptions, this scenario is formally equivalent with that of a signaling game.

Whatever the perspective taken, the essence is that players face a communication (or transfer of meaningful information) problem, which they can solve by establishing a shared language or convention, that is, by adopting compatible mappings from states to signals (in the Sender) and from signals to actions (in the Receiver). Their mappings will be compatible if they, when applied one after the other (first Sender, then Receiver), lead to the correct action for all world states. In language game parlance, this means that the speaker's object is always correctly identified by the hearer, and that neither "synonyms" –different signals encoding the same world state– nor "homonyms" –the same signal encoding different actions– are allowed. It is easily verified that, with S the number of signals again, there are $S!$ different but equally valid and optimal conventions possible, corresponding to the $S!$ unique mappings from world states to signals or from signals to actions. One such convention is shown in Figure 2a for $S = 3$.



(a) An optimal equilibrium or signaling system

(b) A partial pooling equilibrium

Figure 2: Example player configurations for a signaling game with 3 world states, signals and actions.

3 Pooling Equilibria and Win-Stay/Lose-Randomize

Besides the $S!$ optimal conventions (or signaling systems as Lewis would call them), several other mappings are possible that do not support an optimal communication between both players. In language game parlance, and under the stated assumptions in Section 2, these are the encodings that contain one or more synonyms or homonyms. In the context of signaling games, these suboptimal configurations are called *(partial) pooling equilibria*. In a total pooling equilibrium, all states map to a single signal. Figure 2b shows an example of a partial pooling equilibrium containing two synonyms (s_1 and s_2) and one homonym (s_3). Assuming that all states of Nature occur with equal probability, the expected payoff in a sequence of games in this case will be $2/3$ instead of 1 in the case of a convention. In Figure 2a, all world states always lead to a success, whereas in Figure 2b, the second and third states only half of the time.

The reason that these suboptimal configurations are called pooling *equilibria* is that neither the Sender nor the Receiver can increase their expected payoff by changing their behavior. This is in agreement with a game theoretic notion of equilibrium (e.g. Nash equilibrium). A player using the "Win-stay/Lose-Randomize" learning rule will continue to change his mapping unless it is optimal. It is easily deduced that if both players employ this rule, they will continue to change their mappings until, eventually, a fully optimal convention is found [6].[2]

Although this is an interesting conclusion from a theoretical point of view, it only has a limited practical significance. The reason for this is that the Win-Stay/Lose-Randomize rule does not induce any preference between the possible mappings, it simply randomizes between them. This means that, if the number of possible mappings becomes large, e.g. when the number of signals increases, it might take the players a very long time to actually find a fully optimal convention. By the theory of large deviations in randomly perturbed dynamical systems [13] , we can get an idea of how this time scales with S. In [11], this theory is applied to a similar setting as ours, resulting in an expected convergence time that is exponential in the problem size S. This can be considered an unacceptable characteristic of any learning rule "solving" a convention problem. Similar observations were also reported in [8, 6].

Another source of difficulty that may prevent players from reaching a convention is when the world states do not occur with equal probability. For instance, if 90% of the games are about state t_1, then both players may already achieve an expected payoff of 0.9 simply by always ignoring the signal. Barrett [5] reports that basic Roth-Erev learning is very sensitive to such inhomogeneities in state distributions. As discussed in the remainder of the paper, both difficulties can be overcome with Roth-Erev reinforcement learning with forgetting. But first, a number of previous results are briefly discussed in the next paragraph.

4 Previous work

Argiento et al. [1] proved that basic Roth-Erev always converges to a signaling system if the number of states $T = 2$ and if both states occur with the same frequency. Basic Roth-Erev learning can fail whenever the number of states $T > 2$ or when the probability distribution over the states is non-uniform [5, 15].

[2] The earliest clear accounts of the Win-Stay/Lose-Randomize learning rule that we know of are given in [21] (where it was called "Win-Stay/Lose-Switch") and [2, 3]. In particular, Ashby already noted that, in principle, this learning rule *guarantees* convergence and used this finding as the basis for the concept of ultra-stability, of which the homeostat principle is a well known example.

Barrett and Zollman [6] note that when the initial action weights are sufficiently smaller than the payoff, the probability of converging to a signaling system is much higher. We turn back to this finding in the conclusion.

Barrett and Zollman also discuss Bush-Mosteller learning [10], Win-Stay/Lose-Randomize and a more complex learning rule called ARP [9]. They prove that Win-Stay/Lose-Randomize always converges to a signaling system. As was discussed already in the introduction, and as is confirmed by our simulations (section 6), while this may be true in principle, it is of limited practical use. For both Bush-Mosteller and the ARP model, convergence was found to be sensitive to tuning of learning parameters.

Barrett studied two other variations of Roth-Erev learning [5] for signaling games with arbitrary number of states $T \geq 2$ but uniform state distributions. One variation allows for negative rewards, the other randomizes action weights. Both variations seem to improve convergence properties compared to basic Roth-Erev, but neither of them guarantee it.

5 Roth-Erev Reinforcement Learning

We tested a battery of different learning rules on signaling games. Roth-Erev learning with forgetting immediately stood out since it never failed to find an optimal convention. The reason for this becomes clear after careful analysis of this rule. We consider the Sender side, the Receiver side can be treated in a similar way. The Sender keeps track of a time-varying weight $w_{t,s}$ per state-signal pair and chooses to send a signal with a probability proportional to its weight. So, if during game k, the player observes state $t(k)$, then it sends signal s with probability $p_s(k)$:

$$p_s(k) = \frac{w_{t,s}(k)}{\sum_{i=1}^{S} w_{t,s_i}(k)}. \tag{1}$$

It is assumed that all weights are positive ($w_{t,s}(k) \geq 0$) and are initially equally distributed over all signals for each state. In the case that all weights $w_{t,s}(k)$ for a specific state t are zero, all signals are chosen with equal probability.

At any time, action weights represent the discounted sum of all payoffs earned so far for that action(including the initial weight). More concretely, if a game is played at times $k = 0, 1, 2, ...$ involving the states $t(k) = t(1), t(2), t(3), ...$, then weights are updated according to Equations (2) below, with $u(k)$ the payoff received after the game k.

$$w_{t,s}(k+1) = \begin{cases} \lambda w_{t,s}(k) + u(k) & \text{if } t = t(k) \text{ and signal } s \text{ was chosen,} \\ \lambda w_{t,s}(k) & \text{if } t = t(k) \text{ and signal } s \text{ was not chosen,} \\ w_{t,s}(k) & \text{if } t \neq t(k). \end{cases} \tag{2}$$

The learning parameter $\lambda \in [0, 1]$ is a discount factor. It can be seen as a way of forgetting the past. When it is close to 1, the player forgets slowly, and when it is close to 0, the player forgets fast. Thus, discounting past payoffs puts more weight on recent and future payoffs, and the faster a player forgets the faster it can adapt to a changing environment. This way, Roth-Erev learning with forgetting avoids what is called the "Power Law of Practice" or too strong habit forming. In language game parlance, it is an *ergodic* learning rule, which is desired in the context of convention problems [11].

In the extreme case that the discount factor $\lambda = 1$, the update rule is the same as Roth and Erev's basic model, see [22], p. 172, which is also known as cumulative payoff matching or Herrnstein learning. In this case the probability of taking an action becomes proportional to the total payoff accumulated for that action. If the discount factor $\lambda = 0$, the learning rule reduces to Win-Stay/Lose-Randomize.

The most interesting case occurs in between these extreme cases, that is when $\lambda \in]0, 1[$. Note first that the update rules above bound the weights to a finite value. Weights of actions that are always successful will converge to $u/(1 - \lambda)$, while weights of actions that are always unsuccessful will converge to 0. As a consequence, once Sender and Receiver have reached a convention, their behavior will become fully deterministic, since only the successful mappings will converge to above-zero weights. The players will thus stick to the convention. On the other hand, as long as no full convention is established, some weights will continue to change. In sum then, the learning rule keeps the basic requirements for avoiding suboptimal configurations as already fulfilled by the Win-Stay/Lose-Randomize rule. Moreover, it introduces a preference in the player's behaviors that *drives them towards a convention*. This is not the case for the Win-Stay/Lose-Randomize rule. We are currently in the process of proving these statements using the concept of response analysis [11]. In the remainder of this paper, the statement is supported with empirical evidence.

6 Empirical Results

6.1 Setup and Measures

A large number of games was simulated involving a Sender and a Receiver employing Roth-Erev learning with forgetting. In each batch of games, all weights were initialized to 1. The discount factor $\lambda \in [0, 1]$ was varied between batches. The same value was used in the Sender and Receiver. The probability distribution over world states (determining the probability for each state to occur in a game) was also varied between batches. As mentioned, non-uniform distributions are known to cause problems for most learning rules.

The *signaling success rate* is defined as the expected payoff for the next round.[3] It can simply be calculated by summing the payoffs of all possible outcomes of the game weighed by their probability of occurrence. An outcome $t - s - a$ is defined by a state t, a signal s and an action a and corresponds to a path from the top of the tree in Figure 1 down. The probability of seeing an outcome $t - s - a$ equals the probability of Nature drawing state t, times the (conditional) probability that the Sender uses signal s given state t, times the (conditional) probability that the Receiver chooses action a after seeing signal s. The conditional probabilities can directly be inferred from the player's mapping weights and from the proportional action-selection rule (Equation 1). The signaling success rate lies between a positive chance value (depending on the problem size, i.e. S), and 1. A value of 1 indicates optimal signaling.

We say that Sender and Receiver have failed to reach a convention if, after many iterations, the signaling success rate is still below a *threshold* θ. This threshold is chosen halfway between the optimal value, which is 1, and the signaling success rate of the best suboptimal equilibrium. The *best suboptimal equilibrium* has a signaling success rate of 1 minus the probability of the least occurring state: $1 - \min_t(Pr(t))$, where $Pr(t)$ is the probability that Nature draws state t. The best suboptimal equilibrium is actually a partial pooling equilibrium such as the one in Figure 2b, where the Sender uses the same signal for different states. When seeing this signal, the Receiver cannot distinguish between the states, and hence can do no better than assuming the most frequent state. so whenever *the other* state occurs the game will fail. In all other cases, the game succeeds. Thus, the signaling success rate in the *best* partial pooling equilibrium is determined by the frequency of the *least* frequent state.

6.2 Results

Simulation results are summarized in Table 1. The table shows the failure rates for different values of the discount factor λ for different signaling games. It corresponds to the fraction of runs for which the signaling success rate did exceed the threshold θ by the end of the batch (10^6 iterations). None of the batches failed in which $\lambda \in]0, 1[$. For basic Roth-Erev learning ($\lambda = 1$), the failure rate gradually increases with the problem size S and with the non-uniformity of the world-state distribution. When $\lambda = 0$ (Win-Stay/Lose-Randomize), eventhough convergence is expected in principle, in practice this clearly is not the case, especially when the problem size is increased (e.g. $S = 8$).

Figure 3 shows four batches in more detail. The figures show the evolution of the signaling success rates over time (games). Ignoring the results for $\lambda = 0$ (Win-Stay/Lose-Randomize), it can be seen that decreasing the discount factor λ speeds up convergence. This corresponds to a larger degree of forgetting or more adaptive behavior. Although each line is an average over 100 runs, it is obvious that the Win-Stay/Lose-Randomize rule produces very erratic behavior. The signaling success rate jumps irregularly and may even decrease, illustrating that with this rule it is quite possible for the players to unlearn successful behavior (see also the discussion in section 3). In general higher discount factors seem to induce more smooth dynamics. Obviously, convergence times also depend on problem size.

7 Conclusion and Future Work

By identifying the relation between Lewis' framework of signaling games and Steels' framework of language games, a direct transfer of knowledge between both frameworks becomes possible. For instance, the problem of pooling equilibria in signaling games can be approached from a broader angle. In particular, and although some of the details still need to be worked out, our preliminary analysis and all of our simulations confirmed that there is a difference between previously theoretically established convergence conditions and their practical value. In particular, we showed that the Win-Stay/Lose-Randomize learning rule, although

[3] Note that this is different from [5], where it is defined as the average payoff over all past rounds.

Table 1: Failure rates over 1000 runs of at most 10^6 iterations, for signaling games with different number of signals S and world state distributions. The results are reported for different discount factors λ. For Roth-Erev learning with forgetting no run ever failed. A run is considered a failure if the signaling success rate at the end of the run is still less than the threshold θ. The threshold is chosen as explained in the text.

S	distribution	θ	$\lambda = 1$	$\lambda = 0$	$\lambda \in]0, 1[$
2	uniform	0.75	0.000	0.000	0.000
3	uniform	0.8333...	0.100	0.000	0.000
3	$(0.3, 0.3, 0.4)$	0.85	0.125	0.000	0.000
3	$(0.25, 0.25, 0.5)$	0.875	0.206	0.000	0.000
3	$(0.2, 0.2, 0.6)$	0.9	0.316	0.000	0.000
3	$(0.1, 0.1, 0.8)$	0.95	0.637	0.000	0.000
4	uniform	0.875	0.208	0.000	0.000
8	uniform	0.9375	0.635	0.961	0.000
8	$(1, 2, \ldots, 8)/36$	0.986111...	0.883	0.974	0.000
16	uniform	0.96875	0.911	1.000	0.000
32	uniform	0.984375	0.996	1.000	0.000

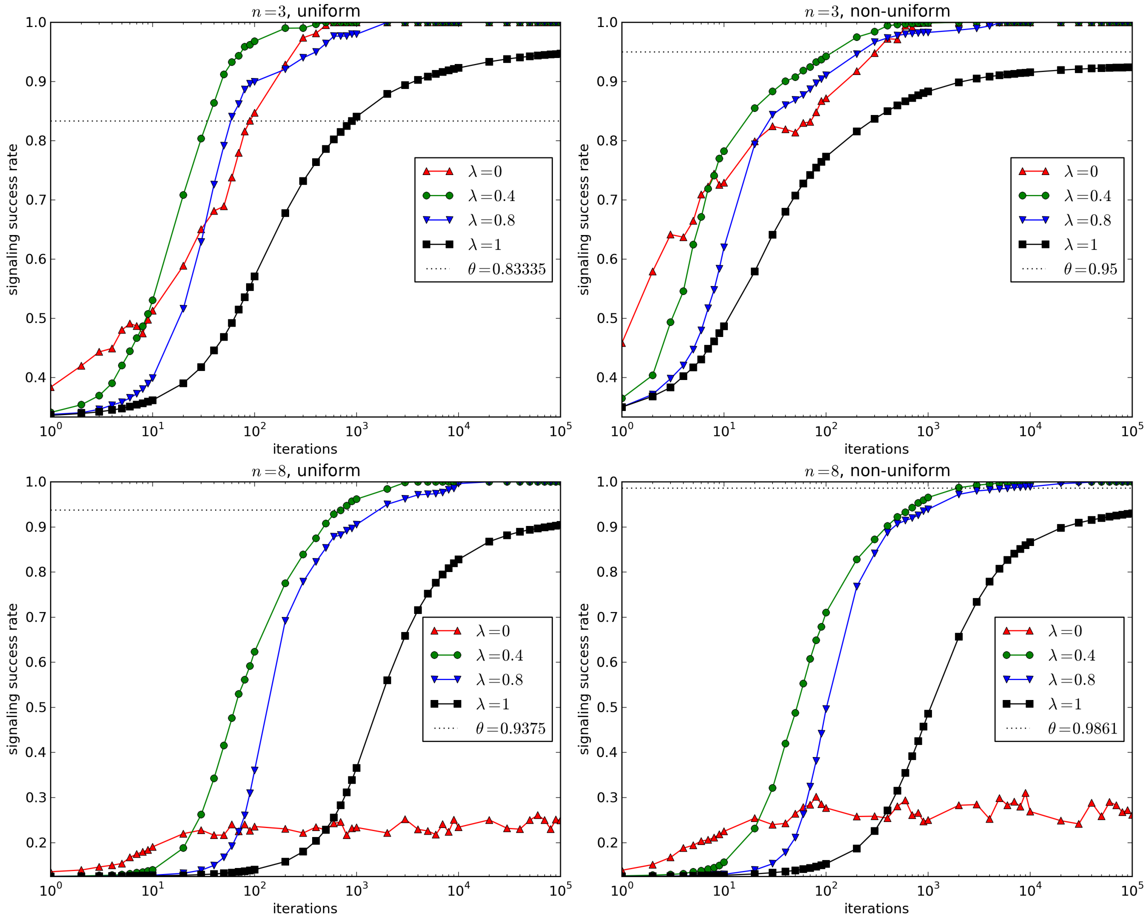


Figure 3: Evolution of signaling success rates during the first 10^5 iterations. Each line represents the average over 100 runs for one specific value of the discount factor λ. The four figures each correspond to another signaling game. *Top-left*: $S = 3$ states with equal probability. *Top-right*: $S = 3$ states with probability distribution $(0.1, 0.1, 0.8)$. *Bottom-left*: $S = 8$ states with equal probability. *Bottom-right*: $S = 8$ states with probability distribution $(1, 2, \ldots, 8)/36$.

theoretically sound, is of negligible practical use, especially for large problem sizes. In particular, it leads to convergence times that increase exponentially with the problem size. Furthermore, our preliminary analysis suggested, and our simulations confirmed, that Roth-Erev learning with forgetting, while also guaranteeing convergence, does not suffer from this drawback. That is, it guarantees much more "reasonable" convergence times, for arbitrary problem sizes, and even for arbitrary world state distributions.

There is a potentially interesting link between our findings and the ones discussed by Barrett and Zollman (see Section 4). As mentioned, they note that when the initial mapping weights are sufficiently smaller than the payoff, the probability of converging to a signaling system is much higher. Roth-Erev learning with forgetting continues to decrease the weights of unsuccessful mappings until they hit a value of 0. This means that these weights will eventually become smaller than the payoff, which automatically puts the players in the desired regime as identified by Barrett and Zollman.

More work is needed in order to confirm some of the findings reported in this paper, especially from a theoretical stance. It will also be interesting to investigate *multi*-player setups (i.e. involving more than two players). We intend to work out these and other things further in the future.

References

[1] Raffaele Argiento, Robin Pemantle, Brian Skyrms, and Stanislav Volkov. Learning to signal: Analysis of a micro-level reinforcement model. *Stochastic Processes and their Applications*, 119(2):373–390, February 2009.

[2] W. Ross Ashby. *Design for a brain*. New York: Wiley, 1954.

[3] W. Ross Ashby. *An Introduction to Cybernetics*. Chapman & Hall, London, 1956.

[4] Marcello Barbieri. Biosemiotics: a new understanding of life. *Naturwissenschaften 95(7)*, pages 577–599, July 2008.

[5] Jeffrey A Barrett. Numerical Simulations of the Lewis Signaling Game : Learning Strategies , Pooling Equilibria , and the Evolution of Grammar. Technical Report September, University of California, Irvine: Institute for Mathematical Behavioral Science, 2006.

[6] Jeffrey A Barrett and Kevin J. S. Zollman. The role of forgetting in the evolution and learning of language. *Journal of Experimental & Theoretical Artificial Intelligence*, 21(4):293–309, December 2009.

[7] Gregory Bateson. Form and pathology in relationship. In *Steps to an Ecology of Mind*. Chandler Pub. Co., 1972.

[8] Alan W. Beggs. On the convergence of reinforcement learning. *Journal of Economic Theory*, 122(1):1–36, May 2005.

[9] Y Bereby-Meyer and I Erev. On Learning To Become a Successful Loser: A Comparison of Alternative Abstractions of Learning Processes in the Loss Domain. *Journal of mathematical psychology*, 42(2/3):266–86, June 1998.

[10] Robert R. Bush and Frederick Mosteller. A mathematical model for simple learning. *Psychological review*, 58(5):313–23, September 1951.

[11] Bart De Vylder. *The Evolution of Conventions in Multi-Agent Systems*. PhD thesis, Vrije Universiteit Brussel, Artificial Intelligence Lab, 2007.

[12] Terrence W. Deacon. *The symbolic species: the co-evolution of language and the brain*. W.W. Norton, New York, 1997.

[13] M. I. Freidlin and A. D. Wentzell. *Random Perturbations of Dynamical Systems*. Springer-Verlag, second edition, 1984.

[14] Josef Hofbauer and Simon M. Huttegger. Feasibility of communication in binary signaling games. *Journal of theoretical biology*, 254(4):843–849, October 2008.

[15] Simon M. Huttegger. Evolution and the Explanation of Meaning. *Philosophy of Science*, 74(1):1–27, January 2007.

[16] Simon M. Huttegger, Brian Skyrms, Rory Smead, and Kevin J. S. Zollman. Evolutionary dynamics of Lewis signaling games: signaling systems vs. partial pooling. *Synthese*, 172(1):177–191, February 2009.

[17] Ray Jackendoff. *Patterns in the mind: Language and human nature*. Harvester Wheatsheaf (Paramount Publishing), 1993.

[18] Simon Kirby. Syntax without natural selection: How compositionality emerges from vocabulary in a population of learners. In C. Knight, J. Hurford, and M. Studdert-Kennedy, editors, *The Evolutionary Emergence of Language: Social function and the origins of linguistic form*. Cambridge University Press, 2000.

[19] David Lewis. *Convention: A Philosophical Study*. Harvard University Press, Cambridge, 1969.

[20] J. Maynard-Smith and E. Szathmáry. *The major transitions in evolution*. Oxford University Press, Oxford, 1995.

[21] H. Robbins. Some aspects of the sequential design of experiments. *Bulletin of the American Mathematical Society*, 58:527535, 1952.

[22] Alvin E. Roth and Ido Erev. Learning in extensive-form games: Experimental data and simple dynamic models in the intermediate term. *Games and Economic Behavior*, 8(1):164–212, 1995.

[23] Brian Skyrms. *Signaling: Evolution, Learning and Information*. Oxford University Press, New York, 2010.

[24] Luc Steels. Language games for autonomous robots. *IEEE Intelligent Systems*, September-October 2001:17–22, 2001.

[25] Ludwig Wittgenstein. *Philosophical Investigations*. Macmillan, New York, 1953.