The Role of Forgetting in the Evolution and Learning of Language

Jeffrey Barrett
University of California, Irvine

Kevin J.S. Zollman
University of California, Irvine

October 18, 2007

Abstract

Lewis signaling games illustrate how language might evolve from random behavior. The probability of evolving an optimal signaling language is, in part, a function of what learning strategy the agents use. Here we investigate three learning strategies, each of which allows agents to forget old experience. In each case, we find that forgetting increases the probability of evolving an optimal language. It does this by making it less likely that past partial success will continue to reinforce suboptimal practice. The learning strategies considered here show how forgetting past experience can promote learning in the context of games with suboptimal equilibria.

We often decry our own forgetfulness, wishing that we could remember more of the past in order to successfully guide our actions today. It is tacitly believed by many that forgetfulness is a human frailty, which should be reduced wherever possible. The present study may shed doubt on this widely held belief.

In this paper, we present a model of language evolution, where forgetfulness plays an important role. In this model, a learning rule which remembers the entire past, basic Herrnstein reinforcement learning, fairs much worse than three other learning rules that discard varying amounts of past experience. These results show how forgetting can be a virtue in the context of games with suboptimal equilibria.

We begin, in Section 1, by describing a game theoretic model for the evolution of language from random signaling – the Lewis signaling game. Unsurprisingly, the probability of evolving an optimal signaling language in such a game depends on the learning strategy used. In Section 2, we describe the successes and failures of basic Herrnstein reinforcement learning in evolving an optimal language. In the next three sections we present three different learning rules, all of which outperform basic Herrnstein reinforcement learning in developing near-optimal languages. All three of these learning rules feature a type of forgetting that helps to achieve optimality.

1 Signaling Games

David Lewis (1969) describes a type of game which can provide a model for the emergence of signaling systems. These games have since been used to investigate the evolution of language (cf. Barrett, 2006, 2007, 2008; Grim et al., 2004; Huttegger, 2007a,b; Huttegger et al., 2007; Skyrms, 1996, 2006; Zollman, 2005).

Lewis signaling games provide a general model of language that can be extended not just to the evolution of human languages but also to the evolution of simple signals in other living organisms (cf. Skyrms, 2006). These games model two individuals, the *sender* and the *receiver*, who have common interest. The sender is aware of some state of the world and has at her disposal several terms which she can send the receiver. The receiver must then take some action, which will determine if he and the sender are rewarded. The correct action depends on the state, of which the receiver is ignorant. In the simplest model there are equal numbers of states, terms, and acts and each state has one and only one appropriate act.

In this model there are a limited number of strategies that achieve the maximum payoff for the sender and receiver. In these strategies the sender uses a different term in each state and the receiver chooses the appropriate act based on the term. Lewis calls these strategies signaling systems.

That signaling systems are Nash equilibria in such games is insufficient to guarantee the evolution of signaling systems. If the sender sends the same term regardless of the state it does not much matter what the receiver does, and vice versa. Consequently, there are many Nash equilibria in such signaling games that fail to achieve the highest possible payoff. As a result, the question of how signaling emerges even in simple signaling games requires

a careful answer.

Attempts to provide this answer have focused on two strategies. One strategy is to identify features of the signaling system equilibria which would motivate intelligent players to settle on those and not the others.[1] Alternatively, one might use an evolutionary approach, asking which of the equilibria are the likely end points of evolutionary or learning dynamics.

Skyrms (1996, 2006) is one the first to have investigated these games using the tools of evolutionary game theory.[2] Skyrms (1996) investigated the two-state, two-term, two-act Lewis signaling game both using the replicator dynamics for the evolution of a population of individuals who performed as both senders and receivers. In his (2006), Skyrms investigates the same game using Herrnstein reinforcement learning for the evolution of the dispositions of a single sender-receiver pair. In both cases, he found that, when the states are equiprobable, every run of a computer simulation converges to a signaling system. More recently, an analytical proof of this result for the two-state, two-term population model with replicator dynamics has been supplied (Huttegger, 2007a).

Based on these successful results, it was conjectured that perfect signaling would also evolve when there were more states, terms, and acts, and also when the states were not equiprobable. However, it has since been discovered that this is not the case. Signaling systems are not guaranteed to evolve for population models using replicator dynamics (Huttegger et al., 2007) or for individual learning models using Herrnstein reinforcement learning (Barrett, 2006). In each case, the systems sometimes converge to suboptimal equilibria.

The failure to evolve perfect signaling when there are more than two states, terms, and acts is a result of a type of equilibria known as partial pooling equilibria. One such equilibrium is illustrated in Figure 1. Here the sender uses term 1 in states 1 and 2 and randomizes between terms 2 and 3 in state 3. The receiver randomizes between act 1 and 2 when he receives term 1 and deterministically takes act 3 when he receives terms 2 and 3. To see that this strategy set constitutes a Nash equilibrium, we must consider the possible deviations. Suppose that the states are equiprobable and x and y equal 0.5. In the state described in Figure 1 the payoff to both players is 2/3. In state 3, both players always coordinate and in states 1 and 2 they coordinate half of the time. Since each state occurs with probability 1/3, the

[1] This was the strategy suggested by Lewis (1969) and Crawford and Sobel (1982).

[2] Earlier investigations include (Wärneryd, 1993; Blume et al., 1993).

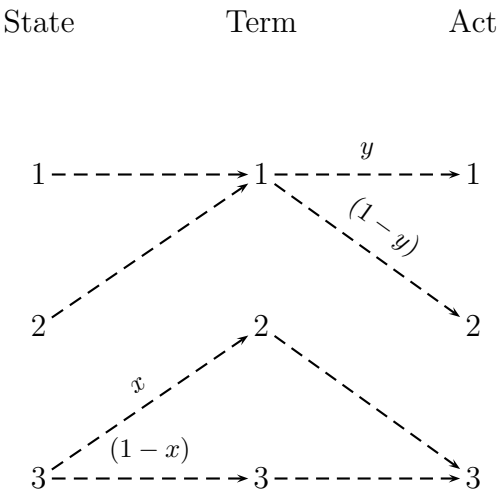

Figure 1: An illustration of a pooling equilibrium

average payoff is 2/3.

Suppose the sender were to switch to a strategy which used a different term in each state, for instance, 1-1, 2-2, 3-3. The sender and receiver continue to perfectly coordinate in state 3, coordinate half of the time in state 1, but they always fail in state 2. As a result, the average payoff to both players is 1/2 (lower than it is in the pooling equilibrium). Similar arguments can be made for alternative sender strategies and alternative receiver strategies.

2 Herrnstein Reinforcement Learning

One model of learning which has been used often in evolutionary game theory is Herrnstein reinforcement learning (cf. Roth and Erev, 1995). The underlying motivation is Richard Herrnstein's *matching law* (Herrnstein, 1970), that players will play a strategy in proportion to the accumulated payoffs for that action. Formally, this is achieved by postulating propensities that determine the probability of an agent's action on each round and that are updated according to success or failure in signaling attempts. The Herrnstein learning model is characterized by (1) the updating rule, which determines how the propensities evolve, (2) the response rule, which determines how the propensities influence behavior, and (3) the initial propensities, which determine

the starting point of the process.

1. The Updating Rule. Let $q_i(t)$ be and agent's propensity for strategy i at time t. In Herrnstein reinforcement, the propensity evolve according to the following updating rule:

$$q_i(t+1) = \begin{cases} q_i(t) + \pi(t) & \text{if action } i \text{ was taken} \\ q_i(t) & \text{otherwise} \end{cases} \tag{1}$$

Where $\pi(t)$ represents the payoff received by our agent on round t.

The propensities determine the probability of a given action on each round via a response rule.

2. The Response Rule. Let $p_i(t)$ represent the probability that our agent takes action i on round t. Herrnstein reinforcement uses the following linear response rule:

$$p_i(t) = \frac{q_i(t)}{\sum_j q_j(t)} \tag{2}$$

These two rules implement Herrnstein's suggestion that individuals will choose a behavior in proportion to the accumulated payoffs they have received by engaging in that behavior as compared to other available behaviors. Finally, we must specify the initial propensities.

3. The Initial Propensities. Each strategy is given equal initial weight, $q_i(0) = 1$ for all i.[3]

There are two general approaches to implementing a particular learning model in the context of signaling games. Alternatively, on a strategy-based implementation, players reinforce propensities for complete contingency plans for each possible state of the world or each term on each play of the game. On an act-based model, players reinforce propensities associated with particular conditional actions, either sending a particular term given a state or acting in a particular way given the reception of a term. In such an implementa-

[3]Note that the initial propensities specified in Rule 3 are equal to the magnitude of the reinforcements specified in Rule 1. Even in the context of simple Herrnstein reinforcement learning, one observes quite different behavior if this is not the case. If the initial propensities are significantly less than the magnitude of the reinforcements, then this significantly lowers the probability of converging to a suboptimal equilibrium in signaling games. This effect is apparently due to rapid initial exploration of possible strategies by the agents. If so, it represents a significant consideration in the analysis of learning strategies but is also relatively independent of the effects of forgetting.

tion, each conditional propensity to send a term or to act may be updated independently of each other conditional propensity.

We will start with an act-based implementation of Herrnstein reinforcement learning. One can think of this as a simple urn process. On each round of the game, the state of the world is randomly determined, the sender is informed of the state of the world and then consults the urn corresponding to the current state and draws a ball at random, where each ball in the urn has the same probability of being drawn. A ball represents a term, and the term represented by the drawn ball is sent to the receiver. The receiver then consults his urn corresponding to the sent term and draws a ball at random, which represents the act. The receiver then takes the selected act, and if it matches the current state of the world, both players are rewarded. If the players are rewarded both return their drawn ball to the respective urn and add another ball to the urn with the same label as the drawn ball; otherwise, the sender and receiver just return their drawn ball to the respective urn without modification. (On the basic urn learning strategy, there is no penalty to the agents for the act failing to match the state.) The game is then repeated with the updated urns.

While the basic 2-state/2-term signaling game with urn learning is relatively simple, it seems to present a difficult context for the evolution of a successful language. The space of possible states is symmetric with no special saliencies and the learning dynamic is simple reinforcement learning with no penalty for failure. The argument that is sometimes made is that if a successful term language can evolve in this context, then it is all the more plausible that a successful language might similarly evolve in contexts where there are special saliencies or more sophisticated learning strategies.[4]

Recently Argiento et al. (2007) have proven that in two state, two term, two act signaling games with equiprobable states, Herrnstein reinforcement learning will converge to a signaling system. Skyrms (2006) has also shown, with simulation, that perfect signaling evolves is a system with two senders

[4]While an even distribution of states may seem to contribute to a difficult environment for language evolution, it is harder for perfect signaling to evolve under simple reinforcement learning when the probability distribution over states of the world is not uniform. The agents might get a good enough success rate by always choosing the more likely state to reinforce the use of more than one term for this state; and since there is no punishment for failure on this learning strategy, there is no evolutionary pressure to undo these reinforced dispositions. This effect is described for the replicator dynamics in (Huttegger, 2007a).

Model	Run Failure Rate
3-state/3-term	0.096
4-state/4-term	0.219
8-state/8-term	0.594

Table 1: Run failure rates for Lewis signaling games with urn learning

and one receiver when the senders observe different, prearranged two-cell partitions of a four-state space.

It is easy to get a sense of how this works in the 2-state/2-term Lewis signaling game with simple reinforcement learning. Adding balls to the term and act urns when an act is successful changes the relative proportion of balls in each urn, which changes the conditional probabilities of the sender's terms (conditional on the state) and the receiver's acts (conditional on the term). The change in the proportion of balls of each type in each urn increases the likelihood that the sender and receiver will draw a type of ball that will lead to successful coordinated action. Here the sender and receiver are simultaneously evolving and learning a meaningful language. That they have done so is reflected in their track-record of successful action.

The situation, however, is more complicated for signaling games with more states or terms or if the distribution of states is biased (see Barrett, 2006; Huttegger, 2007a). In such modified games, partial pooling equilibria may develop and prevent convergence to perfect signaling. Table 1 shows the run failure rates for Lewis signaling games with more than two states and terms (see Barrett 2006 for more details). Here there are 10^3 runs of each model with 10^6 plays/run. A run is taken to fail if the signal success rate is less than 0.8 after 10^6 plays.

While these results illustrate failures in uniform convergence to perfect signaling, each system is always observed to do better than chance and hence to evolve a more or less effective language. In those cases where perfect signaling fails to evolve in the 3-state/3-term game, the system nevertheless approaches a signaling success rate of about 2/3.[5] Similarly, in the 4-state/4-term game, when a system does not approach perfect signaling, it approaches

[5]Systems that approach a signaling success rate of 2/3 here do not learn to signal reliably with two out of three terms; rather, such systems approach a partial pooling equilibrium like the one described above. See Barrett (2006) for more details.

Signal Success Rate Interval	Proportion of Runs
[0.0, 0.50)	0.000
[0.50, 0.625)	0.001
[0.625, 0.75)	0.045
[0.75, 0.875)	0.548
[0.825, 1.0]	0.406

Table 2: Distribution of signal success rates in the 8-state/8-term signaling game

a success rate of about 3/4.[6]

The behavior of the 8-state/8-term system is more complicated since there are several partial pooling equilibria corresponding to different signal success rates. The distribution of signal success rates in the 8-state/8-term game with 10^3 runs and 10^6 plays/run is given in Table 2.

The partial pooling equilibria that limit convergence to perfect signaling in such games is in part an artifact of simple reinforcement learning. If one allows for a slightly more sophisticated learning strategy, then one gets better convergence to perfect signaling. On the 8-state/8-term (+2, -1) signaling game, success is rewarded by adding to the relevant urns two balls of the type that led to success and failure is punished by removing from the relevant urns one ball of the type that led to failure. As illustrated in Table 3, this learning strategy more than doubles the chance of perfect signaling evolving in the 8-state/8-term game.

The overall effectiveness of learning here is improved by a punishment that lowers the agents' propensities when they fail to coordinate. The essential difference between this learning model and Herrnstein reinforcement learning, then, is that here there is a mechanism by which agents might forget past reinforcements that might otherwise have driven them toward suboptimal pooling equilibria. This provides a positive role for forgetting in learning and motivates our investigation of three other learning strategies which also allow

[6]It is a curious feature of these games that the signal success rate is always observed to be greater than 1/2. While Simon Huttegger has an argument for why the success rate should be better than chance signaling, it is unclear, at least to us, why it should always be better than even. This may be a property related to the sure-fire evolution to perfect signaling in the context of the original two-state Lewis. If so, it may also depend on the even distribution of states.

Signal Success Rate Interval	Proportion of Runs
[0.0, 0.50)	0.000
[0.50, 0.625)	0.000
[0.625, 0.75)	0.002
[0.75, 0.875)	0.110
[0.825, 1.0]	0.888

Table 3: Distribution of signal success rates in the 8-state/8-term (+2, -1) signaling game

for reductions in past reinforcement weights, but do so in another fashion.

Each of the following learning rules is a modification of the basic idea of Herrnstein reinforcement, that past success and failure determines future action, but each includes some method for the reduction of past propensities.[7] These models show how forgetting the past can aid in learning by avoiding suboptimal equilibria.

3 The ARP Model

3.1 The Model

The Adjustable Reference Point with Truncation (ARP) learning model is a generalization of reinforcement learning designed to capture that fact that perceived reward is a function of one's experience and that learning in the context of perceived loss can be faster than in the context of perceived gain. The model allows that one may become accustomed to a level of payoff in such a way that one values the same payoff less over time and begins to perceive even positive payoffs as punishments if they are below the accustomed level.

This evolving perception of rewards is seen in both animal and human data. A classic example by Tinklepaugh (1928) illustrates the effect of past payoffs on future perceptions of rewards. Tinklepaugh taught monkeys a simple discrimination task. One group was reinforced with bananas and

[7] These three rules were each chosen because they have some purchase in the experimental or modeling literature surrounding game theory. While they do not exhaust the space of possible learning rules, they represent three very different approaches to an underlying reinforcement process.

another with lettuce, and both groups learned quickly. But when a monkey that was usually paid in bananas got lettuce instead, the accuracy exhibited on the discrimination task dropped significantly suggesting that the monkey perceived the lettuce as a punishment rather than as a reward given its past experience with banana payoffs. The ARP model is designed to account for such reference point effects.

Like Herrnstein Reinforcement, the ARP model can be characterized by specifying the updating rule, the response rule, and the initial propensities.

1. The Updating Rule. The agent's propensities evolve over plays of the game by the rule

$$q_i(t+1) = max[v, (1-\phi)q_i(t) + E_k(i, R_t(\pi_i))] \tag{3}$$

Here $v > 0$ is a truncation parameter that ensures positive propensities, and ϕ is a forgetting parameter that slowly reduces the significance of past experience. The reward function

$$R_t(\pi_i) = \pi_i - \rho(t) \tag{4}$$

translates the payoff π_i into a reward given the agent's expectations from experience. The function $\rho(t)$ is the reference point against which the agent judges the current payoff. The reference point is updated by the rule

$$\rho(t+1) = \begin{cases} (1-w^+)\rho(t) + (w^+)\pi_i & \text{if } \pi_i \geq \rho(t) \\ (1-w^-)\rho(t) + (w^-)\pi_i & \text{otherwise} \end{cases} \tag{5}$$

where w^+ and w^- are the weights associated with positive and negative reinforcement respectively. The experience function

$$E_k(i, R_t(\pi_k)) = \begin{cases} R_t(\pi_k)(1-\epsilon) & \text{if } j = k \\ R_t(\pi_k)\epsilon & \text{otherwise} \end{cases} \tag{6}$$

expresses how the experience of playing k and getting the reward $R_t(\pi_i)$ affects the agent's propensity to play strategy i, and ϵ is the associated parameter.

2. The Response Rule. The probability $p_i(t)$ that i will be played at time t is again given by the linear response rule.

$$p_i(t) = \frac{q_i(t)}{\sum_j q_j(t)} \tag{7}$$

where the sum is over all pure strategies.

3. The Initial Propensities. At time $t = 1$, before the first play of the game, the agent's propensity to play pure strategy i is given by the number $q_i(1)$. In the ARP model Bereby-Meyer and Erev (1998) use the sum of initial propensities divided by the average reinforcement on a random action $S(1)$ to characterize initial propensities. We here will vary the initial propensities in order to judge the robustness of our results over this modification.

3.2 An Act-Based Implementation of the ARP Learning Model

The ARP learning dynamics may be used to update conditional propensities to signal and act in the context of a Lewis signaling game. On this implementation q_k^s represents the sender's propensity to send term k on state s, q_a^k represents the receiver's propensity to do action a on term k, and the conditional propensities are updated using the ARP dynamics; the sender's propensities q_k^s for the actual state s are updated after each play treating each k as a possible pure strategy, and the receiver's propensities q_a^k for the actual term k are updated after each play treating each a as a possible pure strategy. The probabilistic response rule only sums over propensities that corresponds to the current state for the sender and over the propensities that correspond to the current term for the receiver. The payoff for a successful signal is 1.0 and 0.0 for failure.

The APR model has seven free parameters. While the values of these parameters for human subjects would certainly depend on the particular game being played, we will start by assuming an even state distribution and with the values estimated by Erev and Roth (1998) for the first six model parameters, $\epsilon = 0.2$, $v = 0.0001$, $\phi = 0.001$, $\rho(1) = 0$, $w^+ = 0.01$, and $w^- = 0.02$, and set the initial sender and receiver conditional propensities q_k for each act to 27.0. We will then vary the experience parameter, the initial propensities, and the forgetting parameter in turn to see how each affects the evolution of an effective language in the context of the 3-state/3-term Lewis signaling game. There are 10^3 runs and 10^6 plays/run in each trial.

In the ARP model the experience parameter affects how propensities for strategies that were not played are updated – the larger the value of this parameter the greater the effect. Table 4 shows how changing the experience parameter affects the mean signal success rate and the exception rate (the

Experience Parameter	Forgetting Parameter	Initial Propensities	Mean Signal Success Rate	Exception Rate (0.8)
0.2	0.001	27	0.822	0.035
0.1	0.001	27	0.966	0.000
0.0	0.001	27	0.995	0.003

Table 4: Affect of varying the experience parameter in the ARP learning model

Experience Parameter	Forgetting Parameter	Initial Propensities	Mean Signal Success Rate	Exception Rate (0.9)
0.0	0.001	27	0.995	0.006
0.0	0.001	9	0.994	0.006
0.0	0.001	3	0.995	0.011
0.0	0.001	1	0.996	0.006

Table 5: Affect of varying the initial propensities in the APR learning model

cutoff for an efficient language here is set at a signal success rate of 0.8). The lower the experience parameter, the higher the mean signal success rate. The exception rate is also generally lower for lower experience parameters. The moral is that one does best in learning to signal in this context if one only updates the propensities corresponding to the strategy that was actually played on each play of the game and not others. We will set the experience parameter ARP model to 0.0 in order to consider the conditions under which one has the best chance of learning to signal. Note, however, that suboptimal equilibria are still observed in the ARP model with an even state distribution and with the experience parameter set to zero.

Different initial propensities do not affect the behavior of the ARP learning model much. As suggested by the data in Table 5, both the mean signal success rate and the exception rate (here set to 0.9 mean signal success rate) are roughly constant for different initial propensities. So the ARP model is relatively stable under different initial propensities for 10^6 plays/run.

Since the ARP model explains how it is possible for an effective language to evolve most of the time with an even state distribution in a 3-state/3-term signaling game, let's consider uneven state distributions. It is here where forgetting can play a significant role in helping agents avoid suboptimal

Experience Parameter	Forgetting Parameter	Initial Propensities	Mean Signal Success Rate (Max)	Exception Rate (0.93)
0.0	0.001	27	0.944 (0.997)	0.364
0.0	0.01	27	0.986 (0.987)	0.000
0.0	0.1	27	0.964 (0.965)	0.000
0.0	0.3	27	0.937 (0.938)	0.000

Table 6: Affect of varying the forgetting parameter in the APR learning model

equilibria.

Suppose that the random state distribution is (0.8, 0.1, 0.1) over the three states in the 3-state/3-term signaling game, and consider varying the forgetting parameter. First, note that with the uneven state distribution and a low forgetting parameter, the exception rate is extremely high for the ARP model with more than 36% of runs failing to evolve an efficient language. Raising the forgetting parameter discounts the effect of past experience on current propensities, and it thus allows agents to evolve an efficient language even in the context of a very uneven state distribution. As suggested by the data of Table 6, there is a trade off: the more forgetful the agents, the less likely they are to get stuck in suboptimal equilibria, but also the lower the maximum signal success rate on a run. Forgetful agents forget the evidence that might send them to a suboptimal equilibrium, but in this case they also forget the evidence that would allow them to converge to perfect signaling, and hence evolve an imperfect language where the terms have only approximate meanings.

Agents may, however, do quite well here. With a forgetting parameter of 0.01, a very efficient language (with a mean signal success rate of better than 98%) is always observed to evolve in the 3-state/3-term Lewis signaling game with an uneven state distribution. So the evolved meanings of the terms of the agents' language are sharply approximate, and, in this sense, similar to the terms of human natural languages.

The ARP learning model was designed to capture the psychology of how human agents learn. Here we see how the very human trait of forgetfulness can facilitate the successful evolution and learning of a term language.

4 Smoothed Reinforcement Learning

Forgetting also provides benefits in the context of other learning models. The smoothed reinforcement learning model results from a modification of the updating and response rules of Herrnstein reinforcement learning.[8]

1. The Updating Rule. The weights are updated according to this rule:

$$q_i(t+1) = \begin{cases} (1-\delta)q_i(t) + \delta\pi_i(t) & \text{if action } i \text{ was taken} \\ q_i(t) & \text{otherwise} \end{cases} \tag{8}$$

Instead of summing the current payoff with the previous payoff, in this learning rule the current payoff is averaged with the prior weights using a parameter δ. This results in past payoffs becoming less and less relevant to the current play, effectively being discounted.

2. The Response Rule. Rather than using a simple averaging, the probability of an action being chosen uses a logistic response rule:

$$p_i(t) = \frac{e^{\lambda q_i(t)}}{\sum_j e^{\lambda q_j(t)}} \tag{9}$$

3. The Initial Propensities. Like the previous two models we will set $q_i(1) = 1$ for all i.

We thus have a two parameter model. δ represents the degree of averaging. A high δ (close to 1.0) represents a learner who puts the most stock in recent events at the cost to previous ones; a low δ represents the opposite. λ represents the degree of "smoothness" to the function. The higher λ the more small difference affect the probability.

Consider an act-based implementation of this learning model. Supposing that there are two strategies, 1 and 2, and that $q_2(t) = 1.5$, Figure 2 shows $p_1(t)$ for varying values of $q_1(t)$ and λ. This shows that as λ becomes larger small difference in past payoffs correspond to greater differences in response probabilities. Since $q_2(t) = 1.5$, this represents a situation where strategy 2 has been reinforced already. Suppose instead an early situation where no action has yet been rewarded. In this case, $q_2(t) = 1.0$. The varying values of $q_1(t)$ and λ are represented in Figure 3. Here we see that as λ grows, the

[8]This model was suggested in conversation (with Brian Skyrms) by Ed Hopkins and is similar to a version of stochastic fictitious play analyzed in Benaïm et al. (2006).

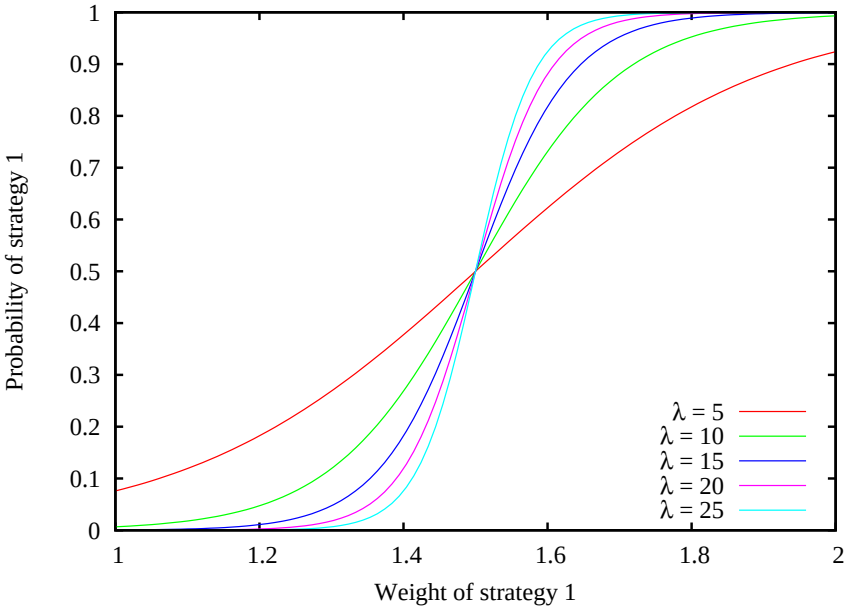

Figure 2: $p_1(t)$ for several λ's, where $q_2(t) = 1.5$

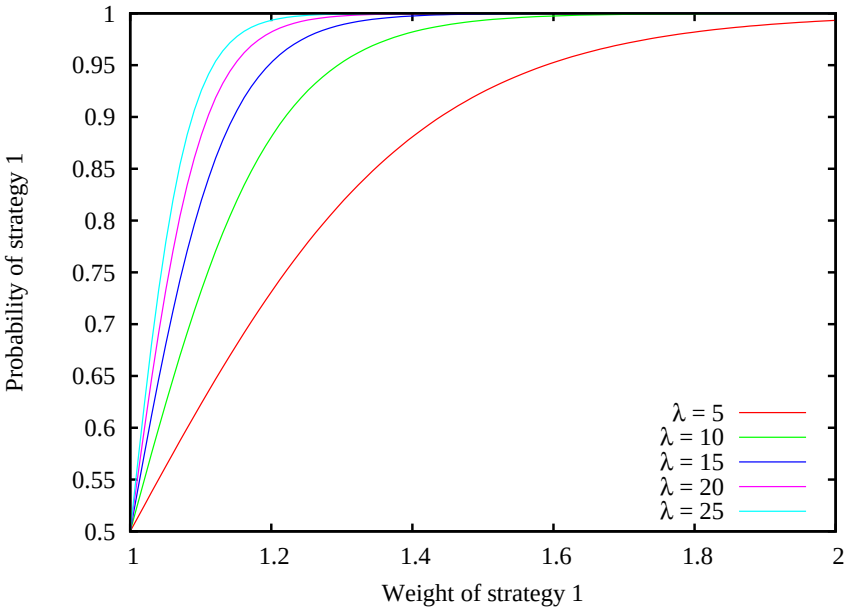

Figure 3: $p_1(t)$ for several λ's, where $q_2(t) = 1.0$

	λ				
δ	5.0	10.0	15.0	20.0	25.0
0.001	0.380	0.922	0.973	0.984	0.988
0.01	0.832	0.989	0.996	0.998	0.999
0.1	0.774	0.992	0.999	0.999	0.999
0.99	0.793	0.999	0.999	1.000	1.000

Table 7: Average efficiency

	λ				
δ	5.0	10.0	15.0	20.0	25.0
0.001	0.000	0.511	0.898	0.943	0.962
0.01	0.000	0.967	0.989	0.993	0.996
0.1	0.000	0.976	0.997	0.999	0.999
0.99	0.000	0.999	1.000	1.000	1.000

Table 8: Converged to Signaling System ($> 0.99\%$ efficiency)

response function tends to take any initial success more seriously, responding by almost deterministically choosing the reinforced action.[9]

In order to determine the convergence properties of this learning rule, we will study simulations for several parameters. An instance of successful signaling results in a payoff of 2.0.[10] As in previous models we will treat each state and term as a distinct learning situation. Tables 7 and 8 show the results for varying δ's and λ's. These results are from 1,000 runs each of 10,000 generations of a 3-state/3-term signaling game with equiprobable states. Table 8 represent the proportion of these runs that converge to almost perfect signaling ($p_i(t)$ was greater than 0.99 for both the sender and receiver for a given term-act).

These simulation result show that for low λ's none of the tested values for δ are sufficient to result in convergence to optimal signaling. This occurs because players continue to randomize sufficiently long that all available

[9]This response mimics the "win-stay" response rule discussed in more detail in the next section.

[10]This is necessary because the initial propensities are equal to 1 and averaging requires that the payoff be greater than 1.

strategies are equally reinforced. As a result, no convergence is ever achieved. For larger λ's, initial success increases the probability of choosing one strategy over another, which results in disproportionate use of those strategies in the future. This begins a process that leads to successful convergence.

With these larger values of λ, it appears that larger δ's are better for the evolution of efficient languages. The larger δ is the more weight individuals place on recent payoffs. This again represents a greater degree of forgetfulness.

In this model we find two features appear to assist in the evolution of optimal signaling. First, a large λ helps substantially. A large λ intuitively corresponds to one taking small differences in payoffs more seriously. Second, a large δ, which corresponds to taking the recent past more seriously than the distant past also helps. Again, forgetting proves efficacious in the evolution and learning of a successful language. And here, unlike in the ARP learning model, perfect signaling may evolve even with very high forgetting ($\delta = 0.99$).

5 Win-Stay/Lose-Randomize

We will consider a final learning rule, win-stay/lose-randomize, which is in a sense maximally forgetful. Here we imagine that individuals only remember their most recent successes for each state/term. If their last action in a given state/term was successful they keep that strategy, otherwise they choose a new strategy for the current state or term at random.[11] This represents a sort of extreme version of smoothed reinforcement, where $\delta = 1.0$ and λ goes to infinity.

In-keeping with our convention, we will restrict individuals to learning only in the context of a given state or term. I.e., individuals who fail to

[11]From the description presented in (Wilcox and Jackson, 2002) it appears that the *Portia* jumping spider employs the strategy when attempting to fool prey. This strategy is also similar to another learning rule, win-stay/lose-switch, which was first introduced in the context of learning in bandit Problems (Robbins, 1952). Bandit problems are a class of learning problems where one is intent on maximizing a payoff in an uncertain environment. Win stay, lose switch was first applied to game theory by Nowak and Sigmund (1993). Interestingly, win-stay/lose-switch is a terrible learning rule in signaling games. While the only fixed points in this learning rule are signaling systems, these states are not accessible from others in a 2-state/2-term/2-act signaling game. So, only players that begin in a signaling system will ever reach one. All other initial states follow closed loops, repeating inefficient strategies forever.

coordinate in a given state/term pair will only randomize their action for that state or that term and not their contingency plans for other states or terms. The same result would hold (with a far simpler proof) if allowed individuals to randomize over the set of all contingency plans.

In the context of signaling games, win-stay/lose-randomize only has fixed points where both individuals play complementary signaling systems.[12] In fact, not only are these the only fixed points, but one can also prove that as the number of plays goes to infinity, the probability that the players achieve optimal signaling systems approaches 1. This is true regardless of the number of states, terms, and acts (so long as they are the same) and regardless of the probability distribution over the states (as long as each state receives non-zero probability). Informally this occurs because, the players randomly try things out succeeding and failing until the states and actions proceed in the right order to result in their perfect coordination. Once there, they never leave.

Assume an N-state/N-term signaling game, where the number of states of the world, number of terms, and number of actions are all the same. We will represent a sender's strategy as a function, $s : N \to N$ and the receiver's strategy as another function, $r : N \to N$. Let S and R represents the set of all sender and receiver strategies respectively.[13] The state of the system at any given time can be represented as an element in $S \times R$.

Since there is a randomizing component we can represent this system as a Markov chain. Certain states in this Markov chain are *absorbing*, that is once the system enters these states it never leaves. It is straight-forward to see that signaling systems are the only absorbing states, since there is a potential loss (and thus a potential switch) in any other state. This is however, not sufficient to prove that the system will converge to signaling systems in the long run, it is also necessary to prove that the signaling systems are accessible from every other state.

Definition 1 *Suppose two states,* $< s_a, r_a >, < s_b, r_b > \in S \times R$. $< s_b, r_b >$ *is* directly accessible *from* $< s_a, r_a >$ *if and only if:*

1. *(Sender failure) If* $s_a(x) \neq s_b(x)$ *then* $r_a(s_a(x)) \neq x$

[12] A fixed point of the learning rule is a strategy set such that if both players play it on one round they will play it forever thereafter.

[13] $r(\cdot)$ and $s(\cdot)$ represent full contingency plans for every state and every signal. Thus our learning rule does not allow for an agent to choose an arbitrary r or s when they fail.

2. *(Receiver failure) If* $r_a(x) \neq r_b(x)$ *then* $r_a(s_a(x)) \neq x$

3. *(One change) There is at most one* x *such that,* $s_a(x) \neq s_b(x)$ *and there is at most one* y *such that* $r_a(y) \neq r_b(y)$ and

4. *(Coordinated change) If there is an* x *such that* $s_a(x) \neq s_b(x)$ *and a* y *such that* $r_a(y) \neq r_b(y)$*, then* $y = s_a(x)$*.*

This definition coincides with there being a non-zero probability of reaching one state from another using win-stay/lose-randomize. By using only a definition of accessibility, we are remaining neutral with respect to both the distribution over the states and the distribution over the strategies used by an agent when he switches strategies. It suffices then to show that (1) a signaling system is accessible from any state and (2) signaling systems are the only absorbing states.

Definition 2 *Two states* $a, b \in S \times R$ *are accessible if and only if there is a sequence* $< a, c_1, c_2, \ldots, b >$ *such that each is directly accessible from the previous state in the sequence.*

In order to prove our main result we will divide all states into two classes. Let $P(s) = |\{n : s(n) = s(m) \text{ for some } m \neq n\}|$, this represents the number of states that map onto the same term. We will first show the following result:

Lemma 1 *For every state* $a =< s_a, r_a >$ *where* $P(s_a) > 0$ *there is some state* $b =< s_b, r_a >$ *such that* b *is directly accessible from* a *and* $P(s_b) < P(s_a)$*.*

Proof. Since $P(s_a) > 0$ there is at least one n and $m \neq n$ such that $s_a(n) = s_a(m)$. Since r_a is a function, at least one of the following must be true: (1) $r_a(s_a(n)) \neq n$ or (2) $r_a(s_a(m)) \neq m$. WLOG assume (1). Because n and m both map to the same term, there must be one $t \in \{1, 2, ...N\}$ which is not in the range of s_a (an unused term). Let $s_b(x) = s_a(x)$ for all $x \neq n$. Let $s_b(n) = t$ (the unused term). It should be clear that $P(s_b) < P(s_a)$, so it is sufficient to prove that $b =< s_b, r_a >$ is accessible from $a =< s_a, r_a >$.

By assumption $r_a(s_a(n)) \neq n$, and by definition $s_a(x) = s_b(x)$ for all $x \neq n$, so b satisfies *sender failure*. Since r_a is constant across both states, b trivially satisfies *receiver failure* and *coordinated change*. By definition b satisfies *one change*. □

As a result of this lemma, we can show that a state $< s_b, r_a >$, where $P(s_b) = 0$ is accessible from any initial state $< s_a, r_a >$.

We will now define a function Q which measures the average success of the sender/receiver pair. It will count the number of states where the two fail to coordinate, $Q(< s, r >) = |\{n : r(s(n)) \neq n\}|$

Lemma 2 *For any state $a =< s_a, r_a >$ where $P(s_a) = 0$ and $Q(a) > 0$, there is a state $b =< s_a, r_b >$ such that b is directly accessible from a and $Q(b) < Q(a)$.*

Proof. Since $Q(a) > 0$, there is at least one n such that $r_a(s_a(n)) \neq n$. Choose such an n. Let $r_b(x) = r_a(x)$ for all $x \neq n$. Let $r_b(n) = s_a^{-1}(n)$ (since $P(s_a) = 0$, this is unique). It should be clear that $Q(b) < Q(a)$, so it sufficient to prove that b is accessible from a.

Since s_a is constant between a and b, *sender failure* and *coordinated change* are trivially satisfied. By assumption $r_a(s_a(n)) \neq n$ and $r_b(x) = r_a(x)$ for all $x \neq n$, satisfying *receiver failure* and *one change*. □

Lemmas 1 and 2 together entail that the agents will always approach perfect signaling. Lemma 1 shows that from any state we can access a state where $P(s) = 0$, and the receiver strategy is unchanged. Lemma 2 shows that from that state we can access a state where $Q(a) = 0$, i.e. a signaling system.[14] Since a signaling system is accessible from any state, and signaling systems are the only absorbing states, the probability that a random state converges to a signaling system approaches 1 as the number of runs goes to infinity.[15]

The upshot is that while win-stay/lose-randomize is in a sense maximally forgetful, it is also perfectly successful in the evolution and learning of a language. Here we see how an extreme form of forgetting might altogether avoid the threat of suboptimal equilibria.

[14]The path detailed in the proof of Lemma 1 and 2 is typically not the most efficient or the most probable path to a signaling system. However, since we are only here concerned with limiting behavior, demonstrating that one such path exists is sufficient.

[15]We could have got this result directly with a learning rule that always allowed for a positive probability of switching to an absorbing state. Implementing win-stay/lose-randomize on entire contingency plans is such a learning rule.

6 Conclusion

The APR learning model, smoothed reinforcement learning, and win-stay/lose randomize all outperform traditional Herrnstein reinforcement learning in the evolution of optimal languages in signaling games. Each of these learning models provide mechanisms whereby agents may forget past evidence that would otherwise have driven them toward suboptimal equilibria.

In each case, forgetting allows for a persistent randomness which can help move a sender-receiver pair away from the suboptimal equilibria and toward optimal ones. In the case of ARP learning, this persistent randomness prevented the model from ever achieving optimality. The modified response rule used in smoothed reinforcement learning, however, helped to overcome the persistent randomness introduced by discarding the past, and to settle on the optimal equilibrium. This settling effect is taken to an extreme in the last learning rule, win-stay/lose-randomize. Here individuals are persistently random until they are optimal in which case, they stick to optimality because it is an absorbing state. This illustrates a sense in which being maximally forgetful can be maximally beneficial in achieving perfect signaling.[16]

The moral is that forgetful learning rules outperform their retentive counterparts in the evolution and learning of language in signaling games. More generally, some form of forgetfulness may prove to be a virtue whenever there is the threat of suboptimal equilibria. In this, something that might have seemed unquestionably detrimental may in fact be beneficial.

[16]There are other ways to introduce persistent randomness into learning. In the context of reinforcement learning, perhaps the most direct way is to randomly perturb the memories of each agent on each play of the signaling game to a degree proportional to the current level of reinforcement. Such models do in fact outperform Herrnstein reinforcement learning in signaling games (Barrett, 2006). They also illustrate a form of forgetting that provides the direct benefit of persistent randomness.

References

Argiento, R., R. Pemantle, B. Skyrms, and S. Volkov (2007). Learning to signal: Analysis of a micro-level reinforcement model. *Manuscript*.

Barrett, J. (2008). Dynamic paritioning and the conventionality of kinds. *Philosophy of Science*.

Barrett, J. A. (2006). Numerical simulations of the lewis signaling game: Learning strategies, pooling equilibria, and the evolution of grammar. Technical Report MBS 06-09, University of California, Irvine: Institute for Mathematical Behavioral Sciences.

Barrett, J. A. (2007). The evolution of coding in signaling games. *Theory and Decision*.

Benaïm, M., J. Hofbauer, and E. Hopkins (2006). Learning in games with unstable equilibria. Technical report.

Bereby-Meyer, Y. and I. Erev (1998). On learning to become a successful loser: A comparison of alternative abstractions of learning processes in the loss domain. *Journal of Mathematical Psychology 42*, 266–286.

Blume, A., Y.-G. Kim, and J. Sobel (1993). Evolutionary stability in games of communication. *Games and Economic Behavior 5*, 547–575.

Crawford, V. and J. Sobel (1982). Strategic information transmission. *Econometrica 50*(6), 1431–1451.

Erev, I. and A. E. Roth (1998, September). Predicting how people play games: Reinforcement learning in experimental games with unique, mixed strategy equilibria. *The American Economic Review 88*(4), 848–881.

Grim, P., T. Kokalis, A. Alai-Tafti, N. Kilb, and P. St Denis (2004). Making meaning happen. *Journal of Experimental and Theoretical Artificial Intelligence 16*(4), 209–243.

Herrnstein, R. J. (1970). On the law of effect. *Journal of the Experimental Analysis of Behavior 15*, 245–266.

Huttegger, S. (2007a, January). Evolution and explanation of meaning. *Philosophy of Science 74*(1), 1–27.

Huttegger, S. (2007b). Evolutionary explanations of indicatives and imperatives. *Erkenntnis 66*, 409–436.

Huttegger, S., B. Skyrms, R. Smead, and K. Zollman (2007). Evolutionary dynamics of lewis signaling games: Signaling systems vs. partial pooling. Technical Report MBS 07-01, University of California, Irvine: Institute for Mathematical Behavioral Sciences.

Lewis, D. (1969). *Convention: A Philosophical Study*. Cambridge: Harvard University Press.

Nowak, M. and K. Sigmund (1993, July 1). A strategy of win-stay, lose-shift that outperforms tit-for-tat in the Prisoner's Dilemma game. *Nature 364*, 56–58.

Robbins, H. (1952). Some aspects of the sequential design of experiments. *Bulletin of the American Mathematical Society 58*, 527–535.

Roth, A. E. and I. Erev (1995). Learning in extensive-form games: Experimental data and simple dynamics models in the intermediate term. *Games and Economic Behavior 8*, 164–212.

Skyrms, B. (1996). *Evolution of the Social Contract*. Cambridge: Cambridge University Press.

Skyrms, B. (2006). Signals. In *Presidential Address, Philosophy of Science Associate Meeting*, Vancouver, B.C.

Tinklepaugh, O. L. (1928). An experimental study of representative factors in monkeys. *Journal of Comparative Psychology 8*(3), 197–236.

Wärneryd, K. (1993). Cheap talk, coordination, and evolutionary stability. *Games and Economic Behavior 5*, 532–546.

Wilcox, S. and R. Jackson (2002). Jumping spider tricksters: Deceit, predation, and cognition. In M. Bekoff, C. Allen, and G. M. Burghardt (Eds.), *The Cognitive Animal*. Cambridge: MIT Press.

Zollman, K. J. (2005). Talking to neighbors: The evolution of regional meaning. *Philosophy of Science 72*, 69–85.