HYBRID LEARNING IN SIGNALING GAMES

JEFFREY A. BARRETT, CALVIN T. COCHRAN, SIMON HUTTEGGER, NAOKI FUJIWARA

Abstract. Lewis-Skyrms signaling games (Lewis 1969; Skyrms 2010) have been studied under a variety of low-rationality learning dynamics (Barrett 2006; Barrett and Zollman 2009; Huttegger, Skyrms, Smead, and Zollman 2010; Huttegger, Skyrms, Tarrès, and Wagner 2014; Huttegger, Skyrms, and Zollman 2014). Reinforcement dynamics are stable but slow and prone to evolving suboptimal signaling conventions. A low-inertia trial-and-error dynamical like win-stay/lose-randomize is fast and reliable at finding perfect signaling conventions but unstable in the context of noise or agent error. Here we consider a low-rationality hybrid of reinforcement and win-stay/lose-randomize learning that exhibits the virtues of both. This hybrid dynamics is reliable, stable, and exceptionally fast.

1. introduction

Lewis-Skyrms signaling games illustrate how it is possible for agents with limited dispositional resources to evolve successful signaling systems as they interact with the world and each other.[1] The simplest sort of signaling game consists of one sender and one receiver. In an $n \times n \times n$ signaling game, there are n states of nature, n signals the sender might use, and n actions the receiver may perform on receiving a signal. Each action is appropriate to precisely one state of nature. On a play of the game, the sender observes a randomly selected state of nature then sends a signal. The receiver, who cannot see the state of nature, observes the signal then performs an action that either matches the state of nature and is successful or does not and is unsuccessful. The agents update their dispositions to signal (conditional on the state of nature) and to act (conditional on the signal) based on the success or failure of the receiver's action.

To be successful the sender and receiver must evolve a simple signaling system by updating their dispositions on repeated plays of the game. This is a subtle learning task since they must simultaneously establish interrelated conventions and learn to use these conventions for successful action in the context of many degrees

Date: February 6, 2017.

[1]See David Lewis's (1969) characterization of signaling games. See Barrett (2007) for an example of the evolution of a simple grammar in a two-sender signaling game, Skyrms (2010) for an overview signaling games, and Barrett and Skyrms (2015) for a discussion of how signaling games may themselves evolve.

of freedom. How the agents update their dispositions on repeated plays of the game is given by their learning dynamics.

Perhaps the simplest learning dynamics is basic reinforcement. On this dynamics, an agent who experiences success on an action on a particular stimulus is slightly more likely to perform that action when presented that stimulus on a future play. Reinforcement learning does well on the $2 \times 2 \times 2$ signaling game with unbiased nature but typically leads to the establishment of suboptimal conventions for more complicated games. This is not just a problem with basic reinforcement learning. It has proven difficult to find a low-rationality learning dynamics where the agents typically establish optimal signaling conventions in an $n \times n \times n$ signaling game for n significantly greater than 2 (Barrett 2006). A learning dynamics that can accomplish this is strong.

In order to provide a compelling story regarding how linguistic conventions might evolve in nature, one would like to show how it is possible for agents to evolve successful signaling systems by means of a learning dynamics that is simple and generic. One wants a strong dynamics, but the more sophisticated or ad hoc it appears, the less compelling the explanations. Further, given that evolutionary success often depends on speed, that a dynamics is sufficiently quick might matter more to an explanation than the properties of the dynamics in the limit. Hence, we will focus here on simple, generic learning dynamics with an eye to relative speed. Significantly, such properties are also among the most salient for applications in automated decision-making and artificial intelligence.

Signaling games have been studied under a variety of low-rationality learning dynamics.[2] Most of these involve either some sort of reinforcement or low-inertia trial-and-error learning. On reinforcement learning, the sender and receiver gradually tune their conditional dispositions on the basis of their success and failure in action. Such reinforcements act as to evolving a memory of what has worked well in the past. This gradually evolving memory also provides the agents with stable dispositions. If an agent makes a mistake observing the state of nature, sending or receiving a signal, or performing an action or if the signal is flipped by channel noise on a play of the game, that play may not lead to successful action. But such one-shot errors typically do little to change the dispositions the agents have forged by gradual reinforcement over time. This stability, however, comes at a significant

[2]See Barrett (2006), Barrett and Zollman (2009), Huttegger, Skyrms, Smead, and Zollman (2010), Huttegger, Skyrms, Tarrès, and Wagner (2014), and Huttegger, Skyrms, and Zollman (2014) for examples. Low-rationality learning is also seen in a variety of other applications. See Khamis and Gomaa (2012) for a practical example of how reinforcement learning might be used as an enhancement to a traffic signal system. The learning dynamics proposed in the present paper might be thought of as using reinforcement learning as an enhancement to low-inertial trial-and-error learning.

cost. Reinforcement learning is slow, and its sluggishness frequently leaves agents stranded playing suboptimal signaling strategies, especially for large n.

In contrast, low-inertia trial-and-error learning often allows agents to evolve a perfect signaling system quickly. Win-stay/lose-randomize is an example of this sort of learning. On this dynamics an agent randomly selects a conditional strategy, sticks with a strategy as long as it works, but immediately shifts to a new strategy if it fails. Since agents immediately shift on failure, they do not get stuck playing suboptimal strategies.[3] But the cost of such flexibility is instability. If the sender and receiver are playing optimal signaling strategies and an agent makes a mistake or a signal is flipped by channel noise, then they will fail on that play of the game and, consequently, shift away from the optimal coordinated conventions they have evolved. They then have to find their way back to optimal play by randomly shifting on subsequent failures.

In the present paper we consider the learning dynamics win-stay/lose-randomize with reinforcement (WS/LRwR), a low-rationality hybrid of reinforcement and win-stay/lose-randomize learning that preserves the virtues of both. It is reliable, stable, and exceptionally fast. In particular, it easily outperforms both simple reinforcement learning and win-stay/lose-randomize learning in the difficult task of evolving a successful signaling system.

2. REINFORCEMENT LEARNING

On simple reinforcement learning (SR) in an $n \times n \times n$ signaling game, the sender has one urn for each possible state of nature, with each urn initially containing a single ball of each of possible signal type; and the receiver has one urn for each possible signal type, with each urn initially containing a single ball of each possible act type.[4] On a play of the game, the sender sees the state of nature, draws a ball at random from the corresponding sender urn, then sends the signal indicated on that ball. The receiver sees the signal, draws a ball at random from the corresponding receiver urn, then performs the action indicated on that ball. If the action matches the state, it is successful, and each agent returns the ball she drew to the urn from which she drew it then adds a new ball to that urn of the same type; otherwise, each agent simply returns the ball she drew to the urn from which she drew it.

[3]See Skyrms (2014) and Huttegger, Skyrms, and Zollman (2014) for discussions of convergence to optimal signaling for probe and adjust, another example of low-inertia trial-and-error learning.

[4]See Herrnstein (1970) for an early characterization of simple reinforcement learning. More sophisticated forms of reinforcement learning have also been studied. Some of these model human behavior well in some learning contexts. See Roth and Erev (1995) and (1998), Barrett (2006), and Barrett and Zollman (2009).

One of the virtues of reinforcement learning is that it is exceptionally simple. As with the other low-rationality dynamics considered here, it might be implemented in the dispositions of a simple physical system.[5]

Initially, such agents will only be successful by blind luck, but as they reinforce on success, they may evolve a signaling system that coordinates the receiver's action with the current state of nature. For the $2 \times 2 \times 2$ signaling game with unbiased nature, one can show that simple reinforcement learning will lead to a perfect signaling system with probability one.[6] But for a $n \times n \times n$ game where n is greater than 2 or if nature is biased, the agents will often (indeed for larger n or significant bias, typically) not converge to an optimal signaling system.[7] When they fail, they end up in a suboptimal partial pooling equilibria playing mixed conditional strategies.

A few examples of simple reinforcement learning in $n \times n \times n$ signaling games where $n > 2$ will be useful for the purpose of comparison. On simulations of the n=3 game, the system fails to do better than a 0.8 cumulative success rate on just 0.096 of 10^3 runs with 10^6 plays per run. For the n=4 game, the failure rate is higher at 0.219. And for the n=8 game, the failure rate is much higher at 0.594. In those runs where the game fails to evolve perfect signaling, the agents end up playing a suboptimal combination of mixed conditional strategies.[8]

While simple reinforcement learning is relatively slow and often fails to evolve perfect signaling for larger n, whatever degree of success the composite system does attain is stable since it is these historically successful dispositions that are most strongly reinforced.

3. WIN-STAY/LOSE-RANDOMIZE

Win-stay/lose-randomize learning (WS/LR) is a form of low-inertia trial-and-error learning. It can be applied to complete strategies or to individual actions.[9] Here we will consider WS/LR applied to individual conditional actions.

In an $n \times n \times n$ signaling game on WS/LR the sender starts by randomly assigning a signal to each of the n states of nature, and the receiver starts by randomly

[5] See Barrett and Skyrms (2015) for a description of how such systems then might evolve a signaling game by ritualization.

[6] See Argiento, Pemantle, Skyrms, and Volkov (2009).

[7] See Barrett (2006) and (2007) and Huttegger (2007) for discussions.

[8] See Barrett (2006), Skyrms (2010), Hofbauer and Huttegger (2008), and Huttegger, Skyrms, Smead, and Zollman (2010) for further details regarding the behavior of simple reinforcement learning.

[9] See Huttegger, Skyrms, and Zollman (2014) for the former sort of application. Win-stay/lose-randomize is closely related to win-stay/lose-shift and probe-and-adjust learning. Win-stay/lose-shift requires the agent to move to a new strategy if the current strategy fails. Among the virtues of win-stay/lose-randomize, is that the agent need not track what strategy led to failure when a new strategy is selected.

assigning an act to each of the n signals.[10] On each play of the game, the sender observes the state of nature, then sends the signal currently assigned to that state. The receiver observes the signal, then performs the action currently assigned to that signal. If the act is successful, the agents keep their assignments of signals to states and acts to signals (win-stay); otherwise, the sender randomly selects a signal with uniform probabilities and assigns it to the current state and the receiver randomly selects an act with uniform probabilities and assigns it to the current signal (lose-randomize).[11]

The first two lines of table 1 show the mean and median number of plays for convergence to perfect signaling for an $n \times n \times n$ signaling game with WS/LR learning for each n from 2 to 8.[12] While the differences between the two learning dynamics prevent a simple comparison, WS/LR is roughly three orders of magnitude faster than SR for the games considered here. Further, unlike SR, WS/LR is not susceptible to suboptimal pooling equilibria. On the simulations we discuss here WS/LR was always observed to converge to perfect signaling in finite times.[13]

	2x2x2	3x3x3	4x4x4	5x5x5	6x6x6	7x7x7	8x8x8
WS/LR mean	11.2	90	674	6,590	83,912	1,362,729	25,605,008
WS/LR median	7	62	463	4,583	58,001	940,789	17,539,708
WS/LRwR mean	9.77 [0.9985]	50 [0.9968]	165 [0.9949]	545 [0.9953]	3,322 [0.9964]	40,021 [0.9958]	565,594 [0.9950]
WS/LRwR median	6	32	88	188	353	598	973

TABLE 1. Speeds for WS/LR and WS/LRwR (noise-free)

It is the forgetfulness of WS/LR that prevents the agents from getting stuck playing suboptimal strategies, but it is the same forgetfulness that makes the dynamics extremely unstable. If perfect signaling has evolved under WS/LR, then a

[10]Note that the initial random assignment of signals to states may not use all of the signals and that the initial assignment of acts to signals may not use all of the acts.

[11]Note that the currently used signal or act may serve as the new signal or act since there is no restriction on the assignment when it is randomized on failure. Note also that this dynamics may assign the same signal to different states or the same act to different signals. But if a perfect signaling system evolves, and it will on this dynamics, such assignments will be temporary.

[12]The simulations of WS/LR and WS/LRwR were run in JAVA using the Eclipse integrated development environment. The game was run 10^4 times for each n.

[13]Indeed, one can prove that WS/LR will converge to an optimal signaling system with probability one. See Huttegger, Skyrms, and Zollman (2014) and Huttegger, Skyrms, and Zollman (2014) for proofs of convergence in finite times for closely associated learning dynamics.

single mistake by either player or a single signal flipped by channel noise may kick the agents out of equilibrium. And, as we will see later, returning to equilibrium after being kicked out is a difficult task for WS/LR.

4. WIN-STAY/LOSE-RANDOMIZE WITH REINFORCEMENT

Ideally, one would like to have a low-rationality dynamics with the stability of SR and the speed of WS/LR. Win-stay/lose-randomize with reinforcement (WS/L-RwR) is a hybrid dynamics that provides both virtues.

In an $n \times n \times n$ signaling game on WS/LRwR learning, the sender (and receiver) are again equipped with one urn for every state of nature (signal), each containing one ball of every possible signal (act) type. As in WSLR, the sender starts by randomly assigning a signal to each of the n states of nature, and the receiver starts by randomly assigning an act to each of the n possible signals. On a play of the game, the sender observes the state of nature, then sends the signal currently assigned to that state. The receiver observes the signal, then performs the action currently assigned to that signal. If the act is successful, the agents keep their assignments of signals to states and acts to signals. Finally, the agents tally the success for the current signal and act. They do so by adding a ball to the current state (signal) urn of the type of the signal (act) just performed [win-stay and reinforce].[14] Otherwise, the sender draws a ball from the urn corresponding to the state of nature just observed and assigns that state of nature to the signal indicated by the ball. Similarly, the receiver draws a ball from the urn corresponding to the signal just observed and assigns that signal to the act indicated by the ball [lose-randomize using probabilities determined by past reinforcements]. These assignments are used for future signals (conditional on states) and acts (conditional on signals) until updated if they fail.[15]

The third and fourth lines of table 1 show the mean and median number of plays to convergence for an $n \times n \times n$ signaling game with WS/LRwR learning for each n. While the thought was to capture the stability of SR with the hybrid dynamics, an issue we will consider in the next section, WS/LRwR is also much faster than WS/LR, especially for larger n. The mean time to convergence is two orders of magnitude faster, and the median is five orders of magnitude faster for $n = 8$. The reinforcements allow WS/LRwR to hold steady those parts of the signaling system that are working while finding conventions for the parts where successful conventions have not yet been formed.

[14]This is the reinforcement mechanism borrowed from SR. Rather than determine the probability of a particular action, however, reinforcement on this dynamics determines the probability of each conditional action being selected as a new strategy if the currently strategy fails.

[15]As with WS/LR, the initial random assignment may not use all of the signals or acts, and the current signal or act may serve as the new randomly selected signal or act on WS/LRwR.

Before considering the relative stability of the two dynamics, it is important to note that there is one sense in which WS/LR is better behaved than WS/LRwR. While every run of WS/LR was successful, WS/LRwR occasionally failed to find a signaling system even when run for twice as long as the longest time it took for WS/LR to find a signaling system.[16] The numbers in square brackets on the second row of table 1 indicate the proportion of runs where the agents converged to perfect signaling on WS/LRwR for each n. The upshot is that WS/LRwR gains a burst in speed on most runs, but at the cost of occasionally taking longer than WS/LR in finding an optimal signaling system.[17]

5. NOISE

While the increase in speed of WS/LRwR over WS/LR is welcome, the motivation behind the hybrid dynamics was the instability of WS/LR when there is possibility of agent error or noise. To compare the two dynamics in this regard, we considered their stability in the context of random channel noise where, with probability 0.01, the receiver observes a random signal rather than the signal sent by the sender.[18]

The mean and median number of plays to reach optimal signaling for both dynamics is given in table 2. Again, the numbers in square brackets indicate the proportion of runs where the agents found optimal signaling on WS/LRwR for each n. Tables 1 and 2 are comparable.

	2x2x2	3x3x3	4x4x4	5x5x5	6x6x6	7x7x7	8x8x8
WS/LR mean	11.4	92	693	6,957	92,014	1,512,384	29,400,236
WS/LR median	8	65	482	4,852	64,006	1,037,499	19,941,227
WS/LRwR mean	10.0 [0.9985]	50 [0.9956]	175 [0.9938]	511 [0.9934]	2,101 [0.9895]	32,328 [0.9874]	409,742 [0.9823]
WS/LRwR median	7	33	91	192	369	620	971

TABLE 2. Speeds for WS/LR and WS/LRwR (0.01 noise)

[16]We used a cutoff of twice the longest time it took for WS/LR to find an optimal signaling system for each n when we ran WS/LRwR. The means and medians are calculated on the runs that converged before hitting the cut-off.

[17]Note that whether WS/LRwR is sure to eventually find an optimal signaling system is currently an open question.

[18]The probability of each random signal is $1/n$ for each game. Note that it is possible that the randomly selected signal matches the signal sent.

In the noise-free case, both WS/LR and WS/LRwR are stable since the agents are always successful after finding an optimal signaling equilibrium. But channel noise makes optimal play unstable for both learning dynamics. Once they find an optimal signaling system, if any signal is randomly flipped by channel noise, the agents will fail on that play. Since WS/LR has no memory, the agents will immediately randomize their conditional strategies with uniform probabilities for each. Agents on WS/LRwR, however, typically continue to play their current conditional strategies insofar as these strategies have been historically successful and hence have been significantly reinforced.

	2x2x2	3x3x3	4x4x4	5x5x5	6x6x6	7x7x7	8x8x8
WS/LR mean	265	169	144	132	122	120	116
WS/LR median	183	117	101	90	85	83	80
WS/LRwR mean	90 [0.08]	243 [0.2967]	450 [0.5491]	640 [0.7526]	799 [0.8764]	879 [0.9391]	936 [0.9747]
WS/LRwR median	37	114	199	255	301	307	335

TABLE 3. Mean plays to break optimal signaling (0.01 noise)

Table 3 indicates the mean and median number of plays that it takes for the 0.01 channel noise to break the optimal signaling system for those runs which are observed to leave the first evolved signaling system.[19] Note that WS/LRwR is an order of magnitude more stable than WS/LR for larger n. Indeed, if a single signal is randomly flipped, WS/LR nearly always breaks and WS/LRwR nearly always preserves optimal play. Further, when WS/LR breaks, with no memory of what has worked well in the past, it typically unravels the entire system of evolved conventions. The first two lines of table 4 give the mean and median number of plays it takes for WS/LR to find an optimal signaling system when it is kicked out by channel error—in short, it typically takes as long as it took to find optimal signaling in the first place. Finally, when it finds an optimal signaling system again, the new system is typically different from the one that initially evolved.[20] In short, WS/LR is extremely unstable in the context of error or noise.

[19]The values in brackets are the proportion of runs that reach first convergence and eventually leave. Many runs under WS/LRwR hit the run-length cap never leaving the signaling system they initially evolve. For $n = 2$, for example, 0.92 of the runs are never observed to leave the initially evolved signaling system. For $n = 4$, the proportion of runs that never leave is somewhat lower at 0.45.

[20]For the $n = 6$ game, for example, the agents almost always evolve a new signaling system.

The third and fourth lines of table 4 give the mean and median number of plays for WS/LRwR to return to an optimal signaling system per channel error. Measured by the mean, WS/LRwR is three orders of magnitude more stable than WS/LR for $n = 8$. Measured by the median, it is seven orders of magnitude more stable. Further, as the median indicates, WS/LRwR typically either does not leave optimal signaling at all or just bounces back to its initial optimal signaling system when it gets kicked out by error. Indeed, on simulation, the system is usually observed to find the same equilibrium it had initially evolved. Optimal signaling conventions do sometimes unravel under WS/LRwR. But, while this is typical for WS/LR, it is rare for WS/LRwR. The reinforcements under WS/LRwR provide a memory that typically guides the system back to optimal play by biasing the random shifts on failure to favor conditional actions that have worked well in the past.[21]

	2x2x2	3x3x3	4x4x4	5x5x5	6x6x6	7x7x7	8x8x8
WS/LR mean	7.46	64	556	5,836	82,227	1,370,470	27,000,015
WS/LR median	4	38	341	3,837	53,915	901,404	24,584,681
WS/LRwR mean	0.10	1.35	6.84	20.50	169	6,702	73,516
WS/LRwR median	0.00	0.00	0.14	0.60	1.00	1.50	2.00

Table 4. Plays to return to optimal signaling (0.01 noise)

Significantly, the longer the system spends at optimal play, the stronger the reinforcements that guide it back when it leaves. One can, consequently, expect agent error and system noise to become less effective at ever kicking the system out of optimality the longer the game is played. One can also expect a yet surer path back to optimality if error does kick the system out of optimal play. And finally, since the system typically does not stray too far from the initial optimal signaling system it evolved, one can expect a high degree of success while it is returning. In other words, given the memory provided by past reinforcements on success, one can expect WS/LRwR typically to produce ever more stable successful dispositions.

[21] Note that there is nothing special about signal noise 0.01. When there is less noise, both WS/LR and WS/LRwR spend more time on average in the first signaling system attained. This additional time reinforces the signaling system for WS/LRwR players, meaning that it takes less time on average for them to return to a signaling system when it does break. WS/LR players do not get this advantage, so it still takes a long time to find their way back to equilibrium. When there is more noise, WS/LRwR takes longer to return to equilibrium after the first break since it spends less time at equilibrium when it first gets there.

6. DISCUSSION

Since the agents must establish and learn conventions in the context of many degrees of freedom, signaling games pose a difficult learning problem. Simple reinforcement learning (SR) is successful in finding optimal conventions only in the context of the simplest games. The agents almost always find an optimal equilibrium for the n=2 game with unbiased nature, but they are highly unlikely to do so if n is significantly greater than 2 or if nature exhibits a strong bias in states. And even when the agents are successful in evolving a perfect signaling system under SR, they do so very slowly. On the other hand, the dispositions the agents evolve by reinforcement are stable since occasional agent error or system noise do little to disturb their past reinforcements.

Win-stay/lose-randomize (WS/LR) learning is much faster than SR, it always eventually finds an optimal signaling strategy, and it sticks with the optimal strategy it finds in the noise-free case. But WS/LR is wildly unstable. A single mistake or noisy signal typically knocks the system out of a coordinated system of optimal signaling conventions. When it does, it takes as long to find new signaling conventions as it took to find the first. And, when they are found, they are typically different.

The problem is that WS/LR is perfectly forgetful. All that matters is what happened in the last play of the game. This prevents past reinforcements on marginally successful strategies from standing in the way of the agents finding a perfect signaling system, but it also means that the agents have no record of what worked well in the past to guide them if an error occurs on a particular play. Hence, their dispositions under the dynamics are maximally unstable.

Win-stay/lose-randomize with reinforcement (WS/LRwR) combines the virtues of SR and WS/LR. While suboptimal partial pooling equilibria are possible, they are rare for both noise-free and noisy systems. It is typically many orders of magnitude faster than WS/LR in finding a perfect signaling system, and the stability it inherits from SR helps it keep the system it evolves. The trial-and-error feature of the dynamics allows for the free exploration of alternative strategies and the reinforcements act as a memory of what has worked well in the past. The memory typically keeps the system close to optimal play and provides increasing stability over time for successful dispositions.

There are a number of natural extensions of the present research. To begin, each of the signaling games considered above has separating equilibria—one would like to study what would happen if that were not the case. A natural place to begin is with signaling games where there are fewer signals than states. Consider the $3 \times 2 \times 3$ game with unbiased nature. Here the agents do not have enough signals to represent nature, so they will never evolve a perfectly reliable signaling system. This does not cause additional problems for reinforcement learning. Simulated agents are nearly always found to evolve a stable signaling system that does as well as possible on SR in the unbiased $3 \times 2 \times 3$ game (Barrett 2006). But it does cause problems for WS/LR. Since there are, by necessity, frequent errors, the agents keep reseting their dispositions with the result that they never evolve stable conventions.[22] While we have not systematically studied such games for WS/LRwR, the evidence we do have is promising. In the $3 \times 2 \times 3$ game with unbiased nature we found WS/LRwR to exhibit a cumulative success rate of 0.6660 on 10^4 runs with 10^6 plays per run. This is close to the theoretically optimal success rate of 2/3. Further, we again found WS/LRwR to be faster than SR.[23]

One might also consider the extent to which WS/LRwR models the learning of human agents. Both simple reinforcement learning Herrnstein (1970) and some types of hybrid reinforcement learning (Roth and Erev 1995 and 1998) have been compared with model human learning in the context of formal games.[24] Experimenters have also investigated the learning behavior of agents in the context of a number of different signaling games (Blume, DeJong, Kim, and Sprinkle 1998; Moreno and Baggio (2015); and Bruner, OConnor, Huttegger, and Rubin 2014 and 2016). While the experimental evidence is subtle, human agents tend to establish signaling conventions faster than they would by simple reinforcement learning in the unbiased $2 \times 2 \times 2$ game.[25] A natural extension of the present paper would be to compare human learning in such games against WS/LRwR. While human learning is clearly more complex in general, WS/LRwR may provide a better model than simple reinforcement learning for the speed and type of exploration that one finds with human agents in the context of signaling games. While a systematic analysis is beyond the scope of the present paper, there is good evidence that the mean behavior of both WS/LR and WS/LRwR is an order of magnitude faster than human learning in the unbiased $2 \times 2 \times 2$ signaling game.[26]

---

[22]On the unbiased $3 \times 2 \times 3$ game, WS/LR exhibits a mean cumulative success rate of 0.4440.

[23]WS/LRwR exhibits a mean cumulative success rate of 0.645 compared to 0.576 for SR after 1000 plays for this particular game. We would like to thank one of the reviewers for suggesting that we consider a game without separating equilibria.

[24]The type of hybrid dynamics studied are rather different than what we have considered here. Roth and Erev (1995) and (1998) introduced such features as forgetting and an adjustable reference point to reinforcement learning. With forgetting past reinforcements are discounted over time. This makes it easier to escape from suboptimal pooling equilibria. With adjustable-reference-point learning a fixed level of reward changes the agents' dispositions less over time as they get used to the reward. The result is that agents are less likely to remain content with suboptimal payoffs.

[25]The data that Bruner, OConnor, Huttegger, and Rubin (2014) present for the $2 \times 2 \times 2$ signaling game with unbiased nature puts their experimental subjects better than an order of magnitude faster than simple reinforcement learning as they typically have at least a rough signaling system within 30 plays.

[26]Compare the simulation results in table 1 to the Bruner, OConnor, Huttegger, and Rubin (2014) result of about 30 plays for human agents. In their experiments, human learning was as fast as

Our aim in the present paper is to characterize WS/LRwR and its properties. WS/LRwR is just trial-and-error learning where the likelihood of a conditional action being tried on failure is determined by how well that action has worked in the past. As such, it is a generic, low-rationality dynamics that is easily implemented in an artificial system and that one might also expect to find exhibited in the behavior of natural agents.

WS/LR and WS/LRwR for some trials of the unbiased $3 \times 3 \times 3$ game, but it was not nearly as consistent as these low-rationality dynamics.

References

[1] Argiento, R., R. Pemantle, B. Skyrms and S. Volkov (2009) Learning to signal: analysis of a micro-level reinforcement model. Stochastic Processes and Their Applications vol 119, n. 2, pp. 373–390.

[2] Barrett, J. A. (2007) Dynamic partitioning and the conventionality of kinds. Philosophy of Science, vol 74, pp. 527–546.

[3] Barrett, J. A. (2006) Numerical simulations of the Lewis signaling game: learning strategies, pooling equilibria, and the evolution of grammar. Institute for Mathematical Behavioral Sciences, paper 54, http://repositories.cdlib.org/imbs/54.

[4] Barrett, J. A. and B. Skyrms (2015) Self-assembling games. The British Journal for the Philosophy of Science. Published online 13 September 2015. doi: 10.1093/bjps/axv043

[5] Barrett, J. A. and K. Zollman (2009) The role of forgetting in the evolution and learning of language. Journal of Experimental and Theoretical Artificial Intelligence, vol 21, n. 4, pp. 293–309.

[6] Blume, A., D. V. DeJong, Y. G. Kim, G. B. Sprinkle (1998) Experimental evidence on the evolution of meaning of messages in sender-receiver games. The American Economic Review, vol 88, n. 5, pp. 1323–1340.

[7] Bruner, J., C. OConnor, S. Huttegger, and H. Rubin (2014) David Lewis in the lab: an experimental study of signaling conventions. Synthese, First Online: 9 September 2014. doi: 10.1007/s11229-014-0535-x.

[8] Bruner, J., C. OConnor, S. Huttegger, and H. Rubin (2016) Communication when interests conflict: a signaling experiment. Manuscript.

[9] Erev, I. and A. E. Roth (1998) Predicting how people play games: reinforcement learning in experimental games with unique, mixed strategy equilibria. American Economic Review, vol 88, pp. 848–81.

[10] Herrnstein, R. J. (1970) On the law of effect. Journal of the Experimental Analysis of Behavior, vol 13, pp. 243–266.

[11] Huttegger, S. (2007) Evolutionary explanations of indicatives and imperatives. Erkenntnis, vol 66, pp. 409–436.

[12] Hofbauer, J. and S. Huttegger (2008) Feasibility of communication in binary signaling games. Journal of Theoretical Biology, vol 254, n. 4, pp. 843–849.

[13] Khamis, M. A., W. Gomaa (2012) Enhanced multi-agent multi-objective reinforcement learning for urban traffic light control. 11th International Conference on Machine Learning and Applications (ICMLA), doi: 10.1109/ICMLA.2012.108.

[14] Moreno,M. and G. Baggio (2015)Role asymmetry and code transmission in signaling games: an experimental and computational investigation.Cognitive Science, vol 39, n. 5, pp. 918–43.

[15] Huttegger, S., B. Skyrms, R. Smead, K. J. S. Zollman (2010) Evolutionary dynamics of Lewis signaling games: signaling systems vs. partial pooling. Synthese, vol 172, n. 1, pp. 177–191.

[16] Huttegger, S., B. Skyrms, P. Tarrès, and E. Wagner (2014) Some dynamics of signaling games. Proceedings of the National Academy of Sciences, vol 111, n. S3, pp. 10873–10880.

[17] Huttegger, S., B. Skyrms, and K. J. S. Zollman (2014) Probe and adjust in information transfer games. Erkenntnis, vol 79, n. S4, pp. 1–19.

[18] Lewis, David (1969) *Convention.* Cambridge, MA: Harvard University Press.

[19] Roth, A. E. and I. Erev (1995) Learning in extensive form games: experimental data and simple dynamical models in the intermediate term. Games and Economic Behavior vol 8, pp. 164–212.

[20] Skyrms, B. (2014) Learning to signal with two kinds of trial and error" forthcoming in *Foundations and Methods from Mathematics to Neuroscience: Essays Inspired by Patrick Suppes* C. E. Crangle, A. Garcia de la Sienra, and H. E. Longino (eds): CSLI Publications.

[21] Skyrms, Brian (2010) *Signals Evolution, Learning, & Information.* New York: Oxford University Press.

Logic and Philosophy of Science, UC Irvine; Irvine, CA 92697, USA
E-mail address: j.barrett@uci.edu