Synthese (2018) 195:603–621
https://doi.org/10.1007/s11229-014-0535-x

David Lewis in the lab: experimental results on the emergence of meaning

Justin Bruner · Cailin O'Connor · Hannah Rubin · Simon M. Huttegger

Received: 30 November 2013 / Accepted: 1 August 2014 / Published online: 9 September 2014

Abstract In this paper we use an experimental approach to investigate how linguistic conventions can emerge in a society without explicit agreement. As a starting point we consider the signaling game introduced by Lewis (Convention 1969). We find that in experimental settings, small groups can quickly develop conventions of signal meaning in these games. We also investigate versions of the game where the theoretical literature indicates that meaning will be less likely to arise—when there are more than two states for actors to transfer meaning about and when some states are more likely than others. In these cases, we find that actors are less likely to arrive at strategies where signals have clear conventional meaning. We conclude with a proposal for extending the use of the methodology of experimental economics in experimental philosophy.

Keywords Signaling · Experimental philosophy · Meaning · Evolution

1 Introduction

Lewis (1969) famously launched an attack on Quinean skepticism about conventions of meaning. His basic idea that such conventions could emerge in a society without explicit agreement has a number of precursors. To mention just one, Adam Smith notes that

> Two Savages, who had never been taught to speak, but had been bred up remote from the societies of men, would naturally begin to form that language by which

J. Bruner (✉)
Australian National University, Canberra, Australia

C. O'Connor · H. Rubin · S. M. Huttegger
UC Irvine, Irvine, CA, USA
e-mail: cailino@uci.edu

> they would endeavour to make their mutual wants intelligible to each other, by uttering certain sounds, whenever they meant to denote certain objects (Smith 1761).

David Lewis, Adam Smith, and many others hold that, at least on the basic level of simple signals, the emergence of meaning should be expected to arise spontaneously. Such a claim is of obvious philosophical significance, for it tells us something about how content, information, representation, and meaning enter the world. Yet how are we to assess this claim?

In this paper we use an experimental approach to study this question. Recent years have witnessed an increasing interest in experimental philosophy. Most of these investigations study intuitions about traditional philosophical issues ranging from epistemological problems to moral problems to questions in the philosophy of mind. These studies usually apply experimental methods from psychology. We shall depart from this practice by drawing on the methodology of experimental economics. We do so because methods from experimental economics are particularly germane for studying conventions in the Lewisian framework. The reason for this is that Lewis imported tools from game theory in order to explicate conventions of meaning. In particular, Lewis appealed to the signaling game, a model of information transfer between two agents, in his arguments for the conventionality of language.

In fact, some experimental economists have already begun to investigate signaling games and the emergence of meaning in the lab.[1] In this paper, we draw heavily on the experimental methods of Blume et al. (1998) who found that small populations of subjects could develop conventions of meaning in relatively short time periods in an experimental setting. We go further than Blume et al. and find that meaning can emerge in a variety of settings. We also explore scenarios where information may only be partially transferred.

The paper will proceed as follows. In Sect. 2, we describe the basic models that we explore experimentally. In Sect. 3, we describe theoretical results regarding the evolution of signaling games and outline our related experimental predictions. In Sect. 4, we briefly discuss the methodology of experimental economics. In Sect. 5 we describe our experimental set-up in detail. In Sect. 6 we outline our results. And, lastly, in Sect. 7, we discuss their significance, their limitations, and, finally, outline a proposal for extending the use of the methodology of experimental economics in experimental philosophy.

2 Signaling games and equilibria

The Lewis signaling games studied here involve two actors. Play of the game proceeds in three stages. In the first stage Nature, or some exogenous force, chooses a state of the world. In the next stage, the first actor, usually called the sender, observes this state of the world and sends a signal. Finally, the second actor, referred to as the receiver, observes this signal and guesses which state of the world has occurred. The goal of

[1] There exists a significant experimental literature on conflict of interest signaling games. Lewis signaling games are common interest and have been investigated less thoroughly.

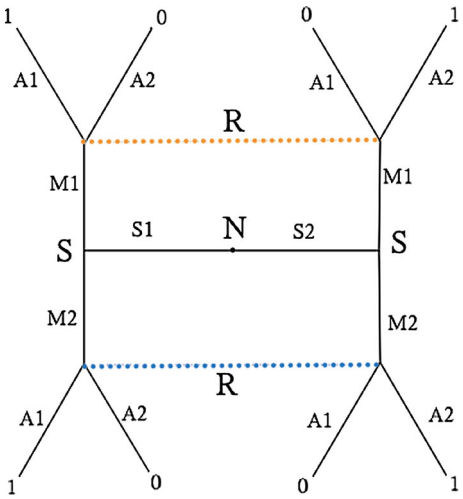

Fig. 1 A 2 × 2 signaling game. The nodes are labeled according to who makes a decision at that stage of the game. The central node representing the first stage of the game is labeled N for nature. The next two nodes are labeled S for the sender and the final set of nodes R for receiver. The *dotted lines* between R nodes represent information sets. The nodes within an information set are indistinguishable to the actor making a decision at that stage. The final nodes are payoff nodes and show what payoff the actors receive for the combination of actions leading up to that node

these actors is to use the signal to help the receiver guess correctly. If the actors do this successfully, they receive a payoff. If they fail, they do not.

In its simplest version, this game has two states of the world, s_1 and s_2, and two signals (or messages), m_1 and m_2. This version is referred to as a 2 × 2 signaling game. Figure 1 shows the extensive form of this game. This figure should be read as a decision tree. It shows the order in which the decisions are made, the options available to each actor, and the amount of information the actors have when making a decision. As mentioned, the game begins when Nature chooses one of two states. In an *unbiased* signaling game, Nature chooses each of the possible states with equal probability, .5. In the *biased* signaling game, Nature chooses one state with higher probability. Our experiment investigates behavior in both unbiased and biased 2 × 2 signaling games. We also address a slightly more complicated Lewis signaling game in which there are three states of the world and three signals. This is referred to as a 3 × 3 signaling game. For this game we assume that nature is unbiased and assigns an equal probability to all of the states.[2]

In order to inform subsequent results and discussion it will be useful to say something about the equilibrium properties of these games. A *strategy* in a game is a complete plan of action for a player. A *Nash equilibrium* for a game is a set of strategies where no player can unilaterally deviate from her strategy and improve her payoff. Lewis described Nash equilibria of signaling games that have especially nice features and called these *signaling systems*. In a signaling system, the sender and receiver use the signals in such a way that their interactions result in success regardless of the state of the world. Signaling systems are also referred to as *separating equilibria* because actors use separate signals for each state of the world.

[2] Many variations exist on the signaling game. We do not, for example, consider games where the interests of the actors conflict, or where approximately correct guesses of the state of nature are rewarded.

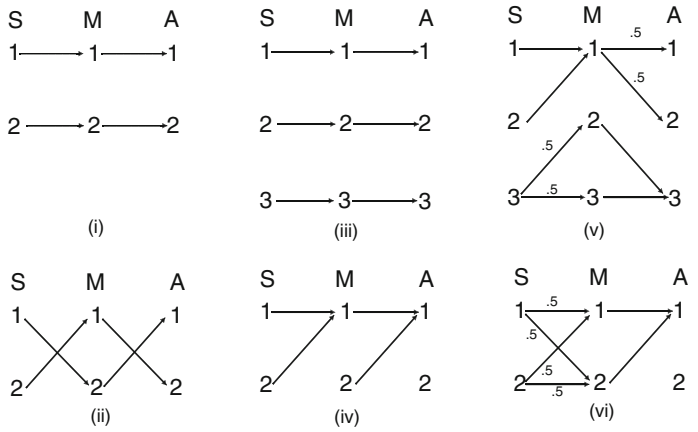

Fig. 2 In this figure, S stands for state, M for signal (or message), and A for act (which specifies the receiver's guess as to which state obtains). The numbers in the column below S represent the states of the world, those below M represent the signals, and those below A the receiver's guess. The *arrows* between the *left columns* specify the sender's strategy, the *right* the receiver's strategy

In the 2×2 game there are two possible signaling systems. In Fig. 2, diagrams (i) and (ii) show representations of the combinations of sender and receiver strategies in these signaling systems. In (i) the sender sends m_1 in s_1 and m_2 in s_2. The receiver guesses s_1 in response to m_1 and s_2 in response to m_2. In (ii), the signals, and the guesses, are reversed. In the 3×3 signaling game, there are six signaling systems, one corresponding to each possible strategy where the actors use a separate signal to denote each state. Diagram (iii) shows one such signaling system.

Signaling systems are what are called *strict Nash equilibria*. This means that unilateral deviation results in a strictly worse payoff. They are the only strict Nash equilibria of the signaling games addressed here. These games do have other Nash equilibria, though. Some of these are referred to as *pooling equilibria*. With a *pooling strategy* the sender sends the same signal in each state or the receiver makes the same guess no matter the signal. Diagram (iv) in Fig. 2 shows an example of a pooling equilibrium for the 2×2 game in which both the sender and receiver use pooling strategies. This outcome is obviously a poor one, but since neither player can unilaterally deviate and improve her payoff it is a Nash equilibrium.

To this point, the strategies described have all been *pure strategies*, but signaling games also have Nash equilibria in *mixed strategies*. A mixed strategy, unlike a pure strategy, is one where the actor behaves probabilistically. For example, a sender might send m_1 with probability .5 and m_2 with probability .5. Figure 2 diagram (vi) shows a pooling equilibrium that exists in a 2×2 game. The decimals in this figure represent the probability with which each signal is sent in each state. In this example, both m_1 and m_2 are sent with probability .5 in each state of the world and the receiver always guesses s_1.

The important feature of this Nash equilibrium is that the receiver pools by always guessing s_1. In a game where this state has a very high probability, choosing the corresponding action while ignoring the messages will lead to a success most of the time. The reason these mixed pooling equilibria are of interest here is that theoretical

work indicates that they are expected to arise with greater frequency in games of higher bias (Hofbauer and Huttegger 2008).

Unlike the 2×2 signaling game, in the 3×3 game there exist a number of equilibria that result in more success than the pooling equilibria but are not maximally informative like signaling systems. In these *partial pooling equilibria* information about one state of the world is transferred perfectly, while information about the other two states is pooled. Figure 2 diagram (v) shows an example of such an equilibrium.

3 Theory and predictions

Lewis (1969) described signaling systems, and thus conventional meaning, in terms of rational choice, common knowledge, and salience. In line with how Adam Smith described the process of arriving at simple languages in the quote above, there is an obvious alternative to this high-rationality approach. The signaling game can be played repeatedly with the sender and receiver adjusting their strategy choices according to past experience. The original problem of the emergence of meaning now turns into a clean, albeit more specific question: Can learning from experience lead players to arrive at a signaling system of a Lewis signaling game?

The theoretical side of this problem is quite well understood. As a first approximation one can use the *evolutionary replicator dynamics* to investigate whether populations of simple agents will evolve to signal successfully. The replicator dynamics are the most commonly used model of evolutionary change in evolutionary game theory. Previous work has indicated that the outcomes of these models are closely related to the outcomes of various learning models (Hopkins 2002; Börgers and Sarin 1997). For this reason, outcomes of models using the replicator dynamics can inform our expectations for how agents will learn in the lab. Models of learning are also investigated by evolutionary game theorists and our experimental predictions will be informed by the results of a number of investigations that apply learning dynamics to signaling games.

Huttegger (2007) and Pawlowitsch (2008) provide a fairly complete analysis of signaling games for the replicator dynamics. An extension that adds mutation to the replicator dynamics is studied in Hofbauer and Huttegger (2008). Argiento et al. (2009) analyze the unbiased signaling game introduced above within the context of a reinforcement learning rule.[3] Huttegger and Zollman (2011) provide a number of results related to the replicator dynamics in 2×2 and 3×3 signaling games. Huttegger et al. (2010, 2014) provide an overview of much of this literature. Skyms (2010) gives a more general overview of signaling games.

This theoretical literature leads to fairly robust qualitative predictions of the effect of learning in repeated Lewis signaling games when players are randomly matched. We use these results to generate the following predictions for experimental subjects playing Lewis signaling games.

> Prediction 1. In the unbiased 2×2 Lewis signaling game, observed play will converge to one of the signaling systems.

[3] Reinforcement learning is one type of learning model commonly used in evolutionary game theory.

This prediction is supported by at least two theoretical findings. Huttegger (2007) proves that in the unbiased 2×2 signaling game, the evolutionary replicator dynamics converges to one of the two signaling systems. Argiento et al. (2009) prove essentially the same result for reinforcement learning. This result is particularly germane as Blume et al. (2002) find that reinforcement learning tracks human subject performance in signaling games. Lastly, Blume et al. (1998) provide experimental results that support this prediction.

These theoretical results do not carry over to the cases where Nature is biased, and neither are there existing experimental results for these cases. Huttegger (2007) proves that if one of the states in the 2×2 game has probability strictly greater than .5, the evolutionary replicator dynamics will carry some of the populations to one of the signaling systems, but others will be carried to other outcomes such as pooling equilibria. The results in Hofbauer and Huttegger (2008) are qualitatively similar. This leads to a prediction for the biased 2×2 case.

> Prediction 2. In the biased 2×2 Lewis signaling game, observed play will converge to one of the signaling systems or to a pooling equilibrium.

The results in Hofbauer and Huttegger (2008) suggest that the strength of the bias of Nature influences how often one should expect players to converge to pooling outcomes. The greater the bias, the more often players do not converge to one of the signaling systems. This yields a third prediction.

> Prediction 3. In two treatments that differ in their probability for the more likely state in the biased 2×2 Lewis signaling game, the observed play will converge to a pooling outcome more often in the treatment with the higher probability for the more likely state.

Our last and final prediction concerns the unbiased 3×3 game. As mentioned, these games have partial pooling equilibria. It is shown in Huttegger (2007) and, more thoroughly, in Pawlowitsch (2008) that partial pooling equilibria are evolutionarily significant. The replicator dynamics, as well as many qualitatively similar evolutionary and learning dynamics, will sometimes converge to a partial pooling equilibrium, though usually not as frequently as to a signaling system. In particular, Huttegger et al. (2010) find that 4.7 % of population starting points converge to partial pooling equilibria under the discrete time replicator dynamics. Blume et al. (2001) investigate these games experimentally. They find that populations quickly converge to signaling systems, but they also provide experimental subjects with a complete history of play and with psychologically salient signals that facilitate coordination. These works lead to the final prediction.

> Prediction 4: In the unbiased 3×3 Lewis signaling game, observed play will sometimes converge to a partial pooling equilibrium, but will more often converge to a signaling system.

In the remainder of the paper, we explore these predictions using experimental methods that will be described in the next section.

4 The methodology of experimental economics

Subject-based laboratory experiments, while once limited primarily to psychology, have in recent decades played a large role in economics, sociology, anthropology, and philosophy. The burgeoning use of laboratory experiments across these disciplines has led to a number of distinct methodological schools. While recent work in experimental philosophy tends to utilize methods familiar to psychologists, our experiment will adhere to the tenets of experimental economics. This section will briefly discuss the norms of experimental design and why it makes sense for our experiment to conform to these standards.

Experiments in economics go as far back as Allais (1953) and are now a staple of the discipline. The laboratory setting provides the experimenter with a high level of control that is often impossible to attain in the field, making the laboratory an ideal venue to test the predictions of economic theory. Experimental work has investigated a number of topics in economics, from price theory (Smith 1962), to bargaining (Guth et al. 1982; Fehr and Gächter 2000), backwards induction (Binmore et al. 2002), and social or other-regarding preferences (Charness and Rabin 2002).

Unlike psychology experiments, where subjects often are instructed to imagine how they would behave in hypothetical situations, economics experiments require individuals to actively participate in a game or strategic situation in which real money is at stake. Smith (1976) refers to this payment scheme as one of induced valuation. In other words, a subject's payment is not just a flat fee, but importantly hinges on the subject's decisions in the laboratory in a way that is consistent with the underlying theory being tested. The reasons for this payment structure should be obvious—economic theory only provides concrete predictions as to how individuals will *actually* behave (Croson 2005). It is silent on the topic of how individuals *think* they would behave in particular strategic situations.

A second important and distinctive feature of economics experiments is that they are by and large context-free. Participants are presented with the most abstract formulation of the experiment possible, even if the purpose of the experiment is to investigate some real-world phenomenon. This is done in part because explicit reference to the real-world case could introduce a systematic bias to the data. Consider, for example, a study on pollution and taxation. An individual wishing to appear environmentally conscious might decide to levy high taxes on companies that pollute. Presenting this subject with the same strategic situation in the abstract could lead to the adoption of a drastically different taxation scheme.

With respect to our experiment, both of the above two points about incentives and context apply. Evolutionary game theory and models of individual learning provide concrete predictions on whether or not individuals can converge to signaling conventions. Thus for the experiment to be a test of the theory we need to properly incentivize individuals. Secondly, ensuring that our experiment is context-free is important. Since signaling and communication is obviously something our subjects do quite regularly outside of the lab, presenting the game as having to do with communication may result in significantly higher levels of coordination than would occur if the presentation of the game was context-free. This is especially significant as our study attempts to inves-

tigate how meaning can arise when no prior communication is present. The details of these two points will be spelled out further in the next section.

5 Experimental setup

In keeping with the decision to use design protocols from experimental economics, and our goal of extending the results of Blume et al. (1998) to signaling games besides the unbiased 2 × 2 game, our experimental set-up was largely drawn from the one developed by these authors.

During the experiment, subjects were asked to play 2×2 and 3×3 games of the types described in the first section. The experiment was run in 20 sessions, each of which involved 12 experimental subjects. The subject pool consisted of undergraduate and graduate students from the University of California, Irvine. Subjects were recruited by e-mails announcing the experiment to students registered in the Experimental Social Science Laboratory subject pool. At the start of each session, experimental subjects were asked to sit at a randomly assigned computer terminal where they were presented with a set of instructions. These instructions, like the rest of the experiment, were programmed and conducted with the software z-Tree (Fischbacher 2007). The set of instructions provided subjects with knowledge of the game and the payment structure employed.[4] Player knowledge of the game was complete with one exception, which will be elaborated shortly.

Within each session, six participants were randomly assigned to act as senders and six as receivers throughout the experiment. Sessions of the 2 × 2 cases were divided into two treatments of 60 rounds each. For each of the total 120 rounds, every sender was randomly paired with a receiver in his or her group. Every round consisted of two stages. During the first stage, each sender was randomly presented with one of two possible symbols by the computer. The symbols observed by each of the senders were chosen independently using a predetermined probability. In some treatments the computer's choice was unbiased, i.e., both symbols were chosen with equal probability (.5/.5). In other treatments, the choice of the computer was biased, i.e., one symbol was chosen with a greater chance than the other (.7/.3 or .9/.1). After observing the symbol chosen by the computer, each sender was then asked to choose one of two signals to send to the receiver he or she had been paired with. In the second stage of the round, each receiver observed the signal which had been sent by his or her partner and was asked to choose one of the two symbols initially shown to the senders. At the end of the round, all players were informed as to which symbol the computer displayed, which signal the sender chose, and which symbol the receiver chose. If the receiver's choice in stage two matched the symbol displayed by the computer in stage one, the two players were informed that they had 'succeeded'. If not, they were informed that they had 'failed' that round of the experiment.

In each session, treatment I was identical to treatment II with the exception that the predetermined probability with which the computer selected symbols for observation by the senders was varied. Furthermore, the symbols displayed by the computer were

[4] Instructions are available upon request.

altered. The treatment order was varied so as to minimize ordering effects. In total, there were four runs of the unbiased treatment, and eight runs each of the .7/.3 and .9/.1 treatments.[5]

The 3×3 experimental set-up was the same as the unbiased 2×2 case with two exceptions. First, obviously, there were three possible states of the world and three available signals for the senders. Each possible state appeared with equal probability. Second, since we were not considering a biased variation there was only need for one treatment consisting of 60 rounds. There were a total of ten runs of the 3×3 treatment.

The only exception to players' complete knowledge of the game regarded the levels of bias with which the computer selected symbols to present to the senders. Players in 2×2 cases were informed that this bias might vary from treatment to treatment, but were not given the actual probabilities used, or even a list of those that might be used during the experiment. Subjects only gained information about this bias through observation over the course of the treatment. This was done to prevent receivers from using their knowledge of the game to pre-emptively choose a strategy such as 'always pick the more common state' rather than developing this strategy through experience.

Subjects received a \$7 payment for showing up to the experiment. In addition, subjects were able to earn more based on their success throughout the experiment. For each treatment, the payment received by each subject was determined by randomly selecting two rounds of play and observing whether that subject had succeeded or failed in those rounds. Each round of success was rewarded with \$4, and each round of failure with \$1, so in addition to the show up fee subjects could earn from \$4 to \$16 in the 2×2 cases and \$2 to \$8 in the 3×3 case. The rounds determining the payoff were not selected from the first ten rounds of play in order to give subjects a period in which to familiarize themselves with the game.

Risk aversion is higher when a greater sum of money is at stake (Holt and Laury 2002). Therefore an incentive structure where performance in each round was potentially worth a significant amount was used to prevent subjects from engaging in non-optimal (risky) behavior for fun. The alternative would be an incentive structure where, say, subjects received a small amount of money for each successful round. It has also been argued that as players accumulate payoff over the course of an experiment, their perceived level of wealth may change, influencing play (Davis and Holt 1993). The payoff structure employed here mitigates these effects because subjects remain unsure of payoff gained until the end of the experiment. Subjects were paid in cash immediately following each session.

Schelling (1960) introduced the idea that there often exist 'focal points' of coordination games that allow players to coordinate behavior in these games without communication. Experimental results have since confirmed Schelling's theory (Mehta et al. 1994). In order to prevent such saliency effects from influencing the results of our experiment, certain steps were taken. First, the symbols presented to the senders and chosen by the receivers were @/! during treatment I and \$/$\sim$ during treatment II. The signals chosen by the senders and viewed by the receivers were #/% for treatment I and */ $\wedge$ for treatment II. These symbols were chosen because they are not naturally

[5] We completed only four unbiased 2×2 runs because this experimental set-up had already been considered by Blume et al. (1998).

ordered or ranked by saliency. (Numbers and letters, in contrast, are easily ordered.) Neither is there any obvious way to associate these symbols with each other and thus solve the coordination problem. Furthermore, the ordering of choices presented to senders and receivers was picked randomly by the computer. In this way, the subjects were prevented from using the physical ordering of the symbols on the screen to improve coordination. Lastly, subjects made their selections by clicking buttons on the screen with a mouse, rather than pressing keys to prevent any keyboard ordering effects from assisting coordination.[6]

The subjects engaged in the experiment were, obviously, language users. As a result they were likely predisposed towards certain assumptions about communication and information transfer. In describing the experiment to subjects, we primarily chose language that conveyed information about the game without explicitly describing the situation as one of information transfer or communication. For instance, players were informed that they would be divided into 'role 1 participants' and 'role 2 participants' in the experiment rather than 'senders' and 'receivers'. There was one exception to this rule, which was that the sender's choice was described as a 'signal' to his or her partner.

As mentioned, our experimental design draws heavily on that of Blume et al. (1998). The primary significant differences between the design employed here and that of Blume et al. regard the information presented to the players. Blume et al. primarily performed experimental trials of signaling games in which at the end of every round, subjects were presented with information about the choices made by every other subject in his or her group both for the previous round and all past rounds of play. Our experiment, however, did not provide subjects with this information. Instead, as noted, at the end of each round subjects were informed only of their choices and the choices of their partners in the previous round. Blume et al. also presented subjects with a pre-experiment trial in which states and signals were labeled with the letters a and b rather than symbols in order to, "[ensure] that players understand the structure of a sender-receiver game, message space, and population history" (1328). We did not provide such a trial. We chose to differ from the experimental set-up of Blume et al. in these ways because (1) we are interested in the possibility of low-rationality strategies leading to the emergence of meaning and (2) we wanted to minimize contextual effects of the set-up on experimental participants. We also, of course, depart from Blume et al. in considering treatments of biased and 3×3 signaling games (though Blume et al. (2001) consider 3×3 games).

6 Results

We will proceed by addressing each of the four predictions made in Sect. 1. In each case we will present the results relevant to the prediction in question.

[6] It might be argued that players could still use keyboard ordering to assist coordination. By checking individual trials, we verified that different groups associated different signals with states, indicating that these associations were not formed using keyboard ordering.

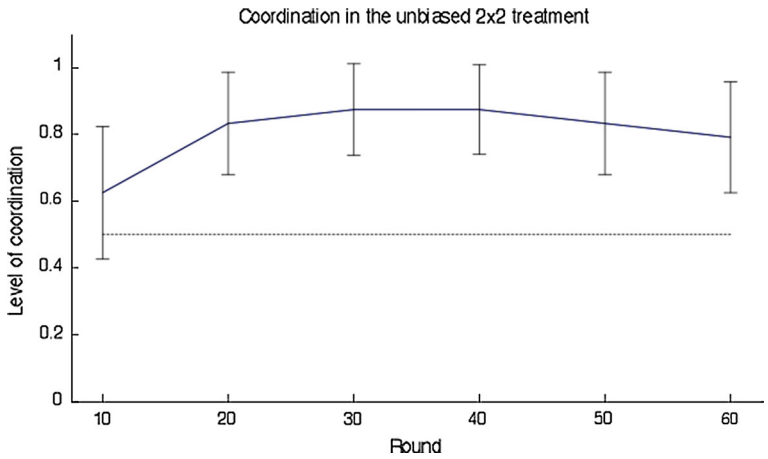

Fig. 3 Level of coordination in the unbiased 2×2 treatment. Results are averaged over all four runs of this treatment, with average level of coordination calculated for every ten rounds. The *dashed line* represents the average level of coordination expected if the receiver ignores the sender's message and simply acts randomly

Prediction 1. In the unbiased 2×2 Lewis signaling game, observed play will converge to one of the signaling systems.

In the unbiased 2×2 treatment our results conform to the previous findings of Blume et al. (1998). Specifically, we find our experimental subjects are able to consistently converge to signaling systems. Figure 3 shows results for the average coordination achieved in the unbiased 2×2 runs as compared to the expected coordination if the experimental subjects were choosing actions randomly. Data points were calculated by determining the proportion of successful sender-receiver interactions that occurred in the span of ten rounds. As is clear from Fig. 3, signaling systems emerge rapidly, and by the 30th round of play, the majority of subjects behave in a manner consistent with this signaling system. This figure includes 95 % confidence intervals—meaning that we are 95 % confident that the intervals encompass the the true value of the parameter measuring coordination in these systems.[7]

It should be noted that most of our runs did not converge to a *perfect* signaling system. This was not just due to the subjects making occasional errors or briefly experimenting with different strategies, which occurred with some frequency. In some runs there existed subjects who stubbornly sent the wrong signal or performed the incorrect act given the predominant signaling system. For example, one sender sent the same signal in both states, even though all receivers were employing the same separating strategy.

Prediction 2. In the biased 2×2 Lewis signaling game, observed play will converge to one of the signaling systems or to a pooling equilibrium.

[7] Using individual behavior from the last round of our four unbiased sessions as independent observations, we also employ a one-sided t test to reject the null hypothesis of independence (that states are independent of signals sent and signals are independent of actions taken) with $p << .01$.

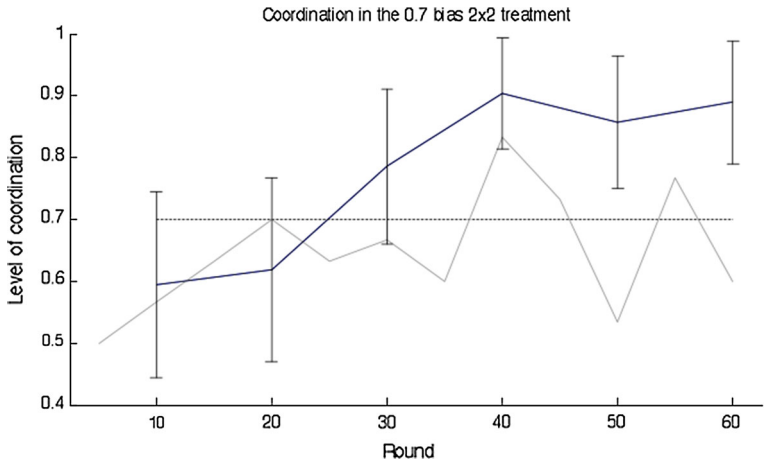

Fig. 4 Level of coordination in the low bias (0.7) 2×2 treatment. Results for the seven treatments that reached separating equilibria are averaged, with average level of coordination calculated for every ten rounds. Average levels of coordination for the eighth treatment are represented with the *gray line*. The *dashed line* represents the average level of coordination expected if the receiver ignores the senders message and simply acts randomly

As described in Sect. 3, we ran two types of biased treatments of the 2×2 game. In the low bias treatment the more likely state of the world was chosen by nature with probability .7, while in the high bias treatment this probability was raised to .9. We conducted eight runs each for low and high bias cases. In general, the experimental data from both bias treatments support prediction 2—observed play tends to converge to either a signaling system or a pooling equilibrium. Unlike the unbiased case, non-optimal pooling behavior was observed in both treatments.

In assessing prediction 2, we will first discuss the .7 bias treatment. In the eight runs conducted, only one resulted in a pooling equilibrium. In this case the receivers picked the state of the world more likely to be chosen by nature, while the senders mixed over the available signals in both states of the world. Thus little to no information was transferred and, not surprisingly, coordination was achieved approximately 70 % of the time. The other seven runs of the .7 bias treatment all converged to a signaling system (or something close to a signaling system). By the last ten rounds of the experiment, receivers in the separating runs were able to determine the correct state of the world on average 90 % of the time, significantly above the expected pooling rate.

Figure 4 shows the average coordination rate over the runs of the .7 treatment that resulted in signaling systems, again with confidence intervals. We also show the lone pooling outcome which is centered at the chance line but vacillates quite wildly. This is due to the fact that when the less likely state of the world occurs, the receivers fail to coordinate. Thus their success level is stochastic as is the selection of states by nature.[8]

[8] We once again employ a one-sided t test with a null hypothesis that a sender and receiver successfully coordinate 70 % of the time. This is the highest possible success rate if no information is tranferred. Using data taken from the last round of play in our eight .7 bias runs, we reject the null hypothesis (p value $<<0.01$).

	Sender			Receiver		
Group 2		m_1	m_2		s_1	s_2
	s_1	.8182	.1818	m_1	1	0
	s_2	0	1	m_2	.6	.4
Group 4		m_1	m_2		s_1	s_2
	s_1	0	1	m_1	0	1
	s_2	.25	.75	m_2	1	0
Group 5		m_1	m_2		s_1	s_2
	s_1	.8246	.1754	m_1	.9792	.0208
	s_2	.3333	.6667	m_2	.75	.25
Group 6		m_1	m_2		s_1	s_2
	s_1	.3333	.6667	m_1	.9	.1
	s_2	.3333	.6667	m_2	.9	.1

Fig. 5 Selected experimental runs for the 2 × 2 game with a 0.9 bias. In each session we record the proportion of senders in the last ten rounds of the experiment utilizing m_1 or m_2 in s_1 and s_2. Likewise, we document the proportion of receivers that pick s_1 and s_2 upon receiving m_1 and m_2 from the sender

The data from the .9 bias treatment also support prediction 2 as will be discussed below.

> Prediction 3. In two treatments that differ in their probability for the more likely state in the biased 2 × 2 Lewis signaling game, the observed play will converge to a pooling outcome more often in the treatment with the higher probability for the more likely state.

As stated above, prediction 3 stipulates that the stronger the bias, the less likely the population is to converge to a signaling system. To assess this claim we compare the .7 runs to the .9 runs. Specifically, we turn our attention to the observed play of the last ten rounds of these runs. Data for selected experimental runs for the .9 bias treatment is presented in Fig. 5. We examine the last ten rounds of each .9 run and count the number of times each sender uses a particular signal in each of the states. Likewise, we track the number of times each receiver picks a particular state upon receiving a signal from the sender.

We find that in four of the eight .9 runs the population failed to arrive at a signaling system.[9] Group 6 (see Fig. 5) arrived at a pooling equilibrium. In this case the senders signal with the same probabilities, regardless of state of the world. The receivers in turn ignore the signal and perform the act that is appropriate for the more likely state of the world. In groups 2 and 5, pooling-like behavior was observed. In these cases, the receivers were more likely to choose the high-probability state of the world, regardless of signal. In group 4 we observed out of equilibrium behavior, meaning that the participants' behavior was not consistent with any Nash equilibrium. In this case, the receivers separated even though the senders were pooling. They would have

[9] We conducted a one-sided *t* test and in these four runs, we cannot reject the null hypothesis that sender-receiver pairs are successful 90 % of the time. This may seem like an odd null hypothesis because in the unbiased and .7 bias cases, 90 % coordination levels counted as signaling systems. Inspection of the strategies of the actors in these cases (see Fig. 5) though, makes clear that these populations truly did not reach separating strategies.

done better to unilaterally switch to a pooling strategy.[10] The fact that our laboratory subjects can get locked into such sub-optimal behavior for 20 or 30 rounds may indicate that more time in the laboratory is required to accurately assess the feasibility of the evolution of signaling in high bias cases.

To sum up, the results in the .9 bias case were noisier and harder to interpret than those in the .7 bias and unbiased cases. Generally, though, they support prediction 3. In the treatments with greater levels of bias, more pooling behavior was seen. Note that they also support prediction 2.[11]

> Prediction 4: In the unbiased 3 × 3 Lewis signaling game, observed play will sometimes converge to a partial pooling equilibrium, but will more often converge to a signaling system.

We conducted ten runs of the 3 × 3 treatment to assess whether signaling systems would emerge in this slightly more complicated signaling game. We find that signaling systems sometimes emerge but are just as likely to fail to emerge. Even though nature is unbiased, signaling systems are difficult to coordinate on in this more complicated game because there are many more strategies available to both senders and receivers.[12] Three runs led to robust signaling systems with high levels of coordination, the remaining seven runs did not. Figure 6 tracks the average level of coordination over 60 rounds averaged over the runs that reached signaling systems and those that did not.

Even though relatively few of the 3 × 3 treatments resulted in clear-cut signaling systems, the vast majority of these experiments outperformed chance, meaning that sender-receiver pairs were successful more than 33 % of the time.[13] Additionally, while separating systems were not ubiquitous, none of the sessions resulted in a pooling or a partial pooling equilibrium.

It should be noted that there was a robust tendency for the levels of coordination to increase over the course of 60 rounds. This occurred in all 3 × 3 runs, with there being no run in which the average level of coordination in the first ten rounds was greater than the average level of coordination in the last ten rounds. Furthermore, as is evident from Fig. 6, at the end of the experiment the subjects seemed to still be improving their coordination. Thus there is good reason to believe that if our experiment had been carried out beyond 60 rounds, we would observe more coordination and perhaps more of the experimental runs would have fixed on a signaling system or even a partial pooling equilibrium.

To summarize, as in the 2 × 2 high bias case, the results in the 3 × 3 unbiased case were somewhat messy and difficult to interpret. Better than chance coordination,

[10] Because this case was unusual, we looked at data from the last 25 rounds of play as well. In this larger sample, the senders sent m_1 in s_2 with higher probability (.66 %). This partial separation on the part of the senders may help explain why non-equilibrium separating was seen by the receivers.

[11] One caveat should be noted which is that in the biased cases actors encountered the unlikely state of the world less often than in the unbiased cases. This may mean that they simply has less time to learn a signaling system.

[12] The number of strategies goes from four to twenty seven when we move from the 2 × 2 game to the 3 × 3 game.

[13] A one-sided t test confirms this. As in the unbiased 2 × 2 case, we once again reject the null hypothesis of independence (with a p value$<<0.01$).

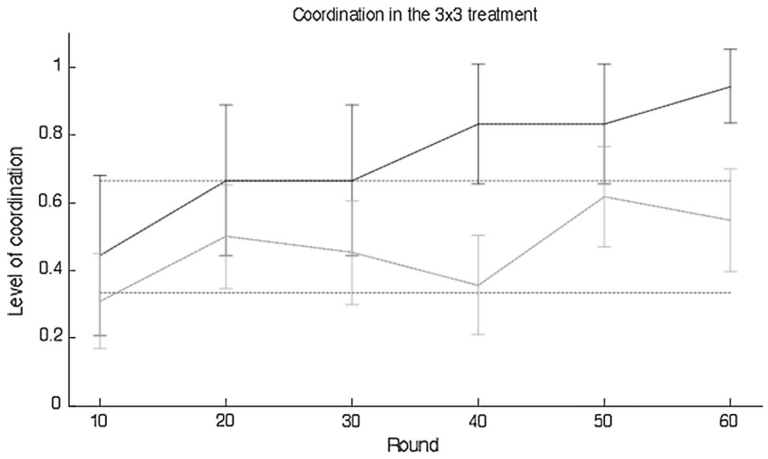

Fig. 6 Level of coordination in the 3×3 treatment. Average results for the three runs that reached separating equilibria are represented by the *black line*, average results for the seven remaining runs are represented by the *gray one*. The *lower dashed line* shows the average level of coordination possible if the receiver ignores the senders message and simply acts randomly. The *upper dashed line* shows the average level of coordination possible at the partial pooling equilibrium

and the observation of some signaling systems, support the first part of prediction 4. There was no obvious support, however, for the prediction of partial pooling equilibria. However, for simple reinforcement learning partial pooling equilibria do not have a large basin of attraction in 3×3 games.[14] The basin of attraction is larger in signaling games with more states and signals.[15] Further experiments with these signaling games might lead to more information concerning partial pooling equilibria.

7 Discussion

We find that signaling systems can emerge under a variety of circumstances. That said, there are certain conditions which make it substantially more likely for signaling systems to emerge in the lab. Our empirical results generally confirm theoretical predictions of signaling games. When nature is biased in the simple 2×2 signaling game, sub-optimal pooling equilibria can emerge. Such inefficient arrangements become more commonplace as nature is increasingly biased.

The results were clearer in some cases than others. In particular, in the 3×3 case, the final strategies developed by the experimental subjects were sometimes hard to interpret compared to the unbiased 2×2 case. As we have indicated, this messiness may result from the difficulty of reaching coordination in this slightly more complicated

[14] Huttegger et al. (2010) found convergence to partial pooling equilibria to be rare (4.7 % of initial populations) in unbiased 3×3 signaling games under the discrete time replicator dynamics.

[15] Barrett (2006) found that under Herrnstein reinforcement learning 8×8 Lewis signaling games converged to partial pooling equilibria with greater frequency than 4×4 games, which in turn converged to these more often than 3×3 games.

game given the short time period. In order to assess this possibility, we constructed a simulation tailored to mimic the experimental set-up. Using a simple learning model we compared the time to reach coordination in each treatment.[16] We found that there is an order of magnitude difference in time to coordination between the unbiased 2×2 case and the 3×3 case.[17] This could give insight into why the subjects in the 3×3 case appear to still be improving coordination after 60 rounds.[18] It may be that in this case, even if signaling is a possibility, learning to signal may take a substantial amount of time.

Before continuing, we will briefly discuss how the results of our unbiased 2×2 runs relate to those of Blume et al. (1998). First it should be noted that our results are more readily applied to non-human animals and non-rational agents. This is the case because, as mentioned, our experimental set-up was designed to more significantly limit the amount of information available to our subjects, thus preventing the use of certain high-rationality strategies such as 'best-respond to the population in the last round' or 'best-respond to the population history'. Our subjects could only respond to their own history and that of their partners in each round.

Furthermore, our results were obtained in an environment which may have been less favorable to the emergence of meaning. Because subjects were not given access to population history, they had to instead rely on memory to determine which action to perform. As mentioned, we also did not have subjects perform practice rounds using symbols that had pre-established meaning (such as 'a' and 'b'), as did Blume et al. Such practice rounds may have helped Blume et al.'s subjects coordinate earlier and better because signaling systems became salient. The fact that our subjects still managed to reach coordinating strategies adds weight to the claim that meaning can emerge naturally through repeated interaction.

As is the case with any experiment, there are limitations in our experimental design that should be mentioned. For one, in the 2×2 cases subjects completed two treatments in each session, meaning that there was the possibility of ordering effects influencing the results. The subjects may have learned how to reach a signaling system in one treatment and learned a new signaling system more readily in the second. To control for ordering effects, we alternated which treatment the subjects completed first. In an attempt to assess whether there existed an ordering effect, we calculated the average level of success in the last ten rounds of each experimental run and compared those that occurred in the first half of a session with those in the second half. No significant

[16] Herrnstein reinforcement learning dynamics were used for this model. For a more detailed description of this dynamics see Skyrms (2010). Blume et al. (2002) find that Herrnstein reinforcement learning provides a good approximation of human learning in signaling games. The model otherwise conformed to the features of the experimental setup, i.e., there were 12 agents, etc.

[17] In 200 runs of simulation, we found that actors playing the unbiased 2×2 game reached a success rate of .95 (as defined by expected payoff given learned strategy divided by highest possible expected payoff) in 6,789 rounds on average. Actors in a 3×3 game, on the other hand, took 43,993 rounds to reach this success rate.

[18] Perhaps surprisingly, in these simulations, actors in biased 2×2 games had even more difficulty reaching separating strategies than those in the 3×3 game. This is contrary to our experimental results. One potential explanation for this discrepancy may be that as our subjects are language users they are predisposed to use signaling strategies. In the 2×2 biased cases they were able to do so because the number of available strategies was still quite small. In the 3×3 cases, they were stymied by the greater number of strategies.

differences were noted. Additionally, due to the repetitive nature of the experiment subjects may have lost interest and become less attentive in the later rounds. This could slow learning if they had not already reached a signaling system. It could also disrupt an existing signaling system, deteriorating progress already made. This possibility was partially ameliorated by our payment structure, but still may have influenced results.

One possible complaint about this work relates to the number of states in the game. If the population has difficulty reaching separating equilibria when there are only three states of the world, how can these results inform real-world signaling where there are, obviously, many more states. It has recently been argued by O'Connor (2013) that actors can learn to transfer information about many states of the world when there is more structure in the state space of the signaling game. One possible extension of the work here would involve investigating whether or not these results transfer to an experimental setting.

Another possible extension of these results regards recent work on signaling games with partial conflict of interest. Interestingly, Wagner (2013) as well as Huttegger and Zollman (2010) have demonstrated that there can be partial transfer of information at equilibrium even when the interests of sender and receiver are not completely aligned. The results of this paper have only explored the transfer of information in cases of common interest. Investigating how meaning can emerge under less favorable conditions is a natural extension.[19]

To date, there have only been a few philosophers who have employed methods from experimental economics to address questions in philosophy. Notable examples include Bicchieri and Chavez (2013) and Bicchieri and Lev-On (2007). These authors have mostly focused on questions regarding norms and ethics. Game theory and evolutionary game theory, however, have been applied more broadly in philosophy, from the evolution of meaning (Skyrms 2010), to the origin of logic (Skyms 2010), to political and social philosophy (Vanderschraaf 2007, 2006; Muldoon et al. 2011), and to issues in decision and rational choice theory (Alexander 2010). When game-theoretic concepts are applied in philosophy, it becomes natural to use experimental economics to gain further traction on these philosophical topics.

In this paper, we present results for one such exploration. We would also like to suggest that there is a great deal of important work still to be done using methods of experimental economics to address philosophical questions. Experimental philosophers are natural candidates to perform this work. Perhaps unsurprisingly, philosophically minded economists have already made inroads in a number of different subfields of philosophy, such as social contract theory (Powell and Wilson 2008; Smith et al. 2012), political philosophy (Frolich and Oppenheimer 1992), and epistemic game theory (Binmore et al. 2002). Ernst (2007) has suggested that results from experimental economics have philosophical import. This is clearly right, and adds bite to the idea that experimental philosophy stands to benefit by adopting methods from experimental economics.

[19] There is an existing literature on information transfer with costly signals in experimental economics. The proposed work would explore signaling when signal costs are less than the amount needed to sustain full information transfer.

Acknowledgements The authors would like to thank Andreas Blume and Elliott Wagner for comments on the paper. We would like to thank Michael McBride for advice on experimental economics and Sabine Kunrath for help with the statistical analysis of our data. Thanks to helpful audiences at GIRL 2013 and the ESSL workshop at UC Irvine 2012. This material is based upon work supported by the National Science Foundation under Grant No. EF 1038456. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

References

Alexander, J. (2010). Local interactions and dynamics of rational deliberation. *Philosophical Studies*, *147*(1), 103–121.

Allais, P. M. (1953). Le comportement de l'homme rationnel devant le risque: Critique des postulats et axiomes de l'ecole americane. *Econometrica*, *21*, 503–546.

Argiento, A., Pemantle, R., Skyrms, B., & Volkov, S. (2009). Learning to signal: Analysis of a micro-level reinforcement model. *Stochastic Processes and their Applications*, *119*, 373–390.

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.

Bicchieri, C., & Chavez, A. (2013). Norm manipulation, norm evasion: Experimental evidence. *Economics and Philosophy*, *29*(2), 175–198.

Bicchieri, C., & Lev-On, A. (2007). Computer-mediated communication and cooperation in social dilemmas: An experimental analysis. *Politics, Philosophy, and Economics*, *6*(2), 139–168.

Binmore, K., McCarthy, J., Ponti, G., Samuelson, L., & Shaked, A. (2002). A backward induction experiment. *Journal of Economic Theory*, *104*, 48–88.

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

Blume, A., DeJong, D. V., Kim, Y. G., & Sprinkle, G. B. (2001). Evolution and communication with partial common interest. *Games and Economic Behavior*, *37*(1), 79–120.

Blume, A., DeJong, D. V., Neumann, G. R., & Savin, N. E. (2002). Learning and communication in sender-receiver games: An econometric investigation. *Journal of Applied Economics*, *17*, 225–247.

Börgers, T., & Sarin, R. (1997). Learning through reinforcement and replicator dynamics. *Journal of Economic Theory*, *77*, 1–14.

Charness, G., & Rabin, M. (2002). Understanding social preferences with simple tests. *The Quarterly Journal of Economics*, *117*(3), 817–869.

Croson, R. (2005). The method of experimental economics. *International Negotiation*, *10*, 131–148.

Davis, D., & Holt, C. A. (1993). Experimental economics: Methods, problems and promise. *Estudios Económicos*, *8*(2), 179–212.

Ernst, Z. (2007). Philosophical issues arising from experimental economics. *Philosophy Compass*, *2*(3), 497–507.

Fehr, E., & Gächter, S. (2000). Cooperation and punishment in public goods experiments. *The American Economic Review*, *90*(4), 980–994.

Fischbacher, U. (2007). z-Tree: Zurich toolbox for ready-made economics experiments. *Experimental Economics*, *10*(2), 171–178.

Frolich, N., & Oppenheimer, J. (1992). *Choosing justice: An experimental approach to ethical theory*. Berkeley: California University Press.

Guth, W., Schmittberger, R., & Schwarze, B. (1982). An experimental analysis of ultimatum bargaining. *Journal of Economic Behavior and Organization*, *3*, 367–388.

Hofbauer, J., & Huttegger, S. (2008). Feasibility of communication in binary signaling games. *Journal of Theoretical Biology*, *254*, 843–849.

Holt, C., & Laury, S. (2002). Risk aversion and incentive effects. *The American Economic Review*, *92*(5), 1644–1655.

Hopkins, E. (2002). Two competing models of how people learn in games. *Econometrica*, *70*(6), 2141–2166.

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

Huttegger, S. M., & Zollman, K. (2010). Dynamic stability and basins of attraction in the sir philip sidney game. *Proceedings of the Royal Society B*, *94*, 1–8.

Huttegger, S. M., & Zollman, K. J. S. (2011). Signaling games: Dynamics of evolution and learning. *Language, games, and evolution* (pp. 160–176). Berlin: Springer.
Huttegger, S. M., Skyrms, B., Smead, R., & Zollman, K. J. S. (2010). Evolutionary dynamics of Lewis signaling games: Signaling systems versus partial pooling. *Synthese*, *172*(1), 177–191.
Huttegger, S. M., Skyrms, B., Tarrès, P., & Wagner, E. O. (2010). Some dynamics of signaling games. *Proceedings of the National Academy of Sciences USA*, *111*, 10873–10880.
Lewis, D. K. (1969). *Convention*. Cambridge, MA: Harvard University Press.
Mehta, J., Starmer, C., & Sugden, R. (1994). The nature of salience: An experimental investigation of pure coordination games. *The American Economic Review*, *84*(3), 658–673.
Muldoon, R., Borgida, M., & Cuffaro, M. (2011). The conditions of tolerance. *Politics, Philosophy, and Economics*, *11*(3), 322–344.
O'Connor, C. (2013). The evolution of vagueness. *Erkenntnis*, *79*, 704–727.
Pawlowitsch, C. (2008). Why evolution does not always lead to an optimal signaling system. *Games and Economic Behavior*, *63*, 203–226.
Powell, B., & Wilson, B. (2008). An experimental investigation of hobbesian jungles. *Journal of Economic Behavior and Organization*, *66*, 669–686.
Schelling, T. C. (1960). *The strategy of conflict*. Cambridge, MA: Harvard University Press.
Skyms, B. (2010). The flow of information in signaling games. *Philosophical Studies*, *147*, 155–165.
Skyrms, B. (2010). *Signals: Evolution, learning, and information*. Oxford: Oxford University Press.
Smith, A. (1761). Considerations concerning the first formation of languages. *Appended to the second edition of The theory of moral sentiments*.
Smith, A., Skarbek, D., & Wilson, B. (2012). Anarchy, groups, and conflict: an experiment on the emergence of protective associations. *Social Choice and Welfare*, *39*(2), 325–353.
Smith, V. (1962). An experimental study of competitive market behavior. *The Journal of Political Economy*, *70*(2), 111–137.
Smith, V. (1976). Experimental economics: induced value theory. *The American Economic Review*, *66*(2), 274–279.
Vanderschraaf, P. (2006). War on peace? a dynamical analysis of anarchy. *Economics and Philosophy*, *22*(2), 243–279.
Vanderschraaf, P. (2007). Covenants and reputations. *Synthese*, *157*, 167–195.
Wagner, E. O. (2013). The dynamics of costly signaling. *Games*, *4*, 161–183.